Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2)

S. D. Bentley*, K. F. Chater†, A.-M. Cerdeño-Tárraga*, G. L. Challis†‡, N. R. Thomson*, K. D. James*, D. E. Harris*, M. A. Quail*, H. Kieser†, D. Harper*, A. Bateman*, S. Brown*, G. Chandra†, C. W. Chen§, M. Collins*, A. Cronin*, A. Fraser*, A. Goble*, J. Hidalgo*, T. Hornsby*, S. Howarth*, C.-H. Huang§, T. Kieser†, L. Larke*, L. Murphy*, K. Oliver*, S. O'Neil*, E. Rabbinowitsch*, M.-A. Rajandream*, K. Rutherford*, S. Rutter*, K. Seeger*, D. Saunders*, S. Sharp*, R. Squares*, S. Squares*, K. Taylor*, T. Warren*, A. Wietzorrek†, J. Woodward*, B. G. Barrell*, J. Parkhill* & D. A. Hopwood†

* *The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK*
† *John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK*
‡ *Department of Chemistry, University of Warwick, Coventry CV4 7AL, UK*
§ *Institute of Genetics, National Yang-Ming University, Shih-Pai, Taipei 112, Taiwan*

Streptomyces coelicolor **is a representative of the group of soil-dwelling, filamentous bacteria responsible for producing most natural antibiotics used in human and veterinary medicine. Here we report the 8,667,507 base pair linear chromosome of this organism, containing the largest number of genes so far discovered in a bacterium. The 7,825 predicted genes include more than 20 clusters coding for known or predicted secondary metabolites. The genome contains an unprecedented proportion of regulatory genes, predominantly those likely to be involved in responses to external stimuli and stresses, and many duplicated gene sets that may represent 'tissue-specific' isoforms operating in different phases of colonial development, a unique situation for a bacterium. An ancient synteny was revealed between the central 'core' of the chromosome and the whole chromosome of pathogens *Mycobacterium tuberculosis* and *Corynebacterium diphtheriae*. The genome sequence will greatly increase our understanding of microbial life in the soil as well as aiding the generation of new drug candidates by genetic engineering.**

Nutritionally, physically and biologically, soil is a particularly complex and variable environment. Streptomycetes are among the most numerous and ubiquitous soil bacteria[1]. They are crucial in this environment because of their broad range of metabolic processes and biotransformations. These include degradation of the insoluble remains of other organisms, such as lignocellulose and chitin (among the world's most abundant biopolymers), making streptomycetes central organisms in carbon recycling. Unusually for bacteria, streptomycetes exhibit complex multicellular development, with differentiation of the organism into distinct 'tissues': a branching, filamentous vegetative growth gives rise to aerial hyphae bearing long chains of reproductive spores. The importance of streptomycetes to medicine results from their production of over two-thirds of naturally derived antibiotics in current use (and many other pharmaceuticals such as anti-tumour agents and immunosuppressants), by means of complex 'secondary metabolic' pathways. Furthermore, streptomycetes are members of the same taxonomic order (Actinomycetales) as the causative agents of tuberculosis and leprosy (*Mycobacterium tuberculosis* and *M. leprae*), the genomes of which have been sequenced[2,3]. Much should be learned about these pathogens from genome-level comparisons with harmless saprophytic relatives such as streptomycetes.

Streptomyces coelicolor A3(2) is genetically the best known representative of the genus[4]. The single chromosome is linear with a centrally located origin of replication (*oriC*) and terminal inverted repeats (TIRs) carrying covalently bound protein molecules on the free 5′ ends. Replication proceeds bidirectionally from *oriC*, leaving a terminal single-stranded gap on the discontinuous strand after removal of the last RNA primer. An unusual process of 'end-patching' by DNA synthesis primed from the terminal protein fills the gap[5]. Studies of many streptomycetes, including most notably a close relative of the A3(2) strain, *Streptomyces lividans* 66, established further novelties. More than a million base pairs (bp) of DNA at either end of the chromosomes can undergo extensive deletions and amplifications without compromising viability under laboratory conditions[6], and early comparisons of linkage maps established that most streptomycetes show conservation of gene order (synteny) in the core region[7]. Here, we report the use of an ordered cosmid library[8] to sequence the *S. coelicolor* genome. The strain used, M145, is a prototrophic derivative of strain A3(2) lacking its two plasmids (SCP1, linear, 365 kb, AL590463, AL590464; and SCP2, circular, 31 kb, AL645771, which have been sequenced separately).

Genome structure

General features of the chromosome sequence are shown in Table 1 and Fig. 1. At 8,667,507 bp it is the largest completely sequenced bacterial genome. The *oriC* and *dnaA* gene are about 61 kb left of the centre, at 4,269,853–4,272,747 bp. Like many other microbial genomes, there is a slight bias (55.5%) towards coding sequences on the leading strand. Although less pronounced than for most other eubacterial chromosomes, there is a discernible decrease in the GC bias around *oriC*, thought to be related to DNA replication[9]. In contrast to all other bacterial genomes studied to date, however, the

Table 1 General features of the chromosome

Component of chromosome	Property
Total size	8,667,507 bp
Terminal inverted repeat	21,653 bp
G + C content	72.12%
Coding sequences	7,825
. . .of which pseudogenes	55
Coding density	88.9%
Average gene length	991 bp
Ribosomal RNAs	6 × (16S–23S–5S)
Transfer RNAs	63
Other stable RNAs	3

© 2002 Macmillan Magazines Ltd

S. coelicolor chromosome displays a downward rather than an upward shift, indicating a small bias towards C on the leading strand.

Coding density is largely uniform across the chromosome, with only a slight decrease in the distal regions. The distribution of different types of genes reveals, however, a central core comprising approximately half the chromosome and a pair of chromosome arms (Fig. 1). Nearly all genes likely to be unconditionally essential—such as those for cell division, DNA replication, transcription, translation and amino-acid biosynthesis—are located in the core (exceptions tend to be duplicate genes). In contrast, 'contingency' loci coding for probable non-essential functions, such as secondary metabolites, hydrolytic exoenzymes, the conservons (conserved operons) and 'gas vesicle' proteins (see below), lie in the arms. Curiously, this biphasic structure of the chromosome does not align with the position of *oriC*. The core appears to extend from around 1.5 Mb to 6.4 Mb, giving uneven arm lengths of approximately 1.5 Mb (left arm) and 2.3 Mb (right arm). The difference in arm lengths may reflect some gross rearrangement or different rates of DNA accumulation in each arm. The fact that *oriC* is roughly central suggests some selective pressure for such positioning.

Streptomyces coelicolor and *M. tuberculosis* are both actinomycetes

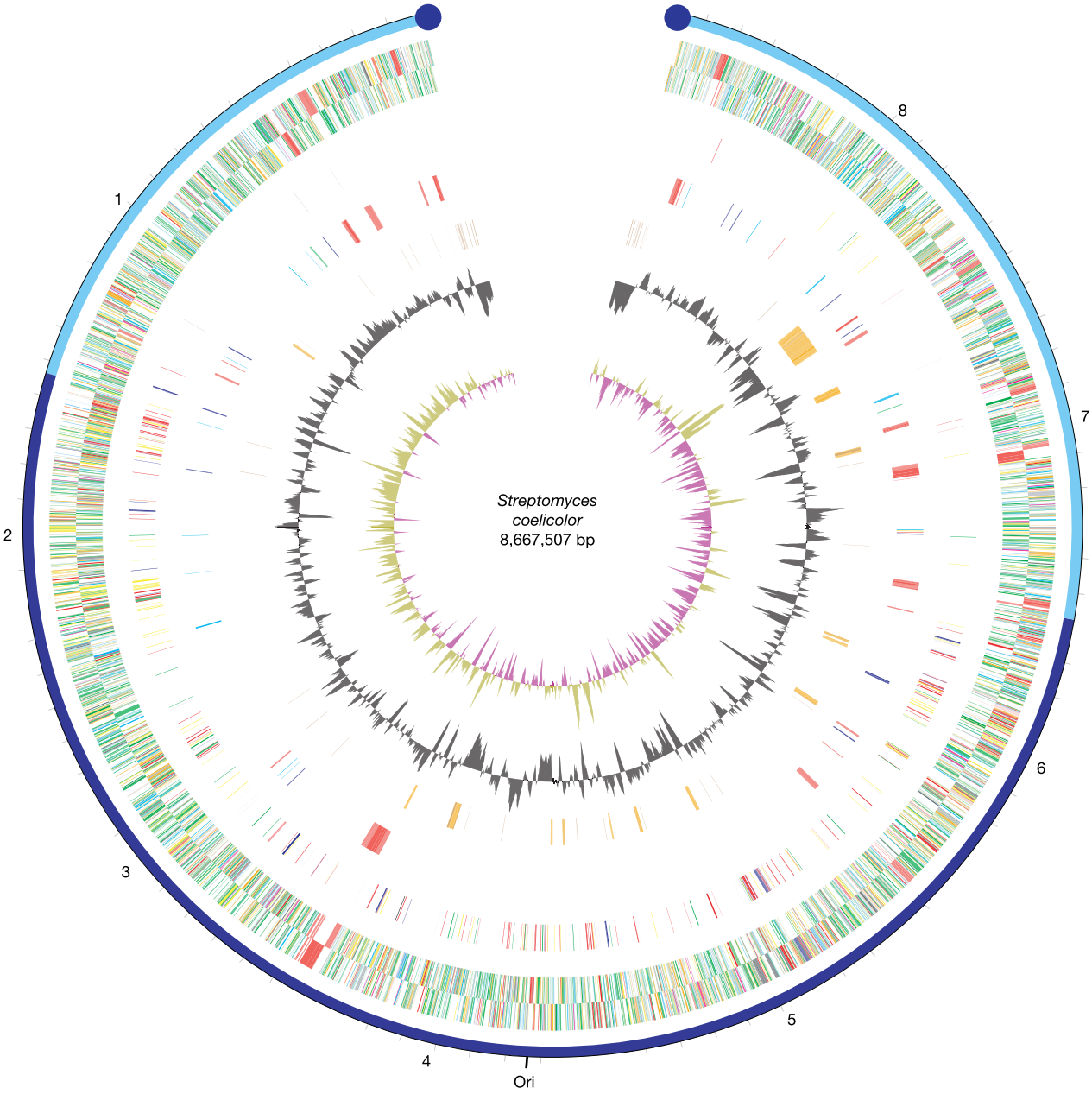

Figure 1 Circular representation of the *Streptomyces coelicolor* chromosome. The outer scale is numbered anticlockwise (to correspond with the previously published map[8]) in megabases and indicates the core (dark blue) and arm (light blue) regions of the chromosome. Circles 1 and 2 (from the outside in), all genes (reverse and forward strand, respectively) colour-coded by function (black, energy metabolism; red, information transfer and secondary metabolism; dark green, surface associated; cyan, degradation of large molecules; magenta, degradation of small molecules; yellow, central or intermediary metabolism; pale blue, regulators; orange, conserved hypothetical; brown, pseudogenes; pale green, unknown; grey, miscellaneous); circle 3, selected 'essential' genes (for cell division, DNA replication, transcription, translation and amino-acid biosynthesis, colour coding as for circles 1 and 2); circle 4, selected 'contingency' genes (red, secondary metabolism; pale blue, exoenzymes; dark blue, conservon; green, gas vesicle proteins); circle 5, mobile elements (brown, transposases; orange, putative laterally acquired genes); circle 6, G + C content; circle 7, GC bias ((G − C/G + C), khaki indicates values >1, purple <1). The origin of replication (Ori) and terminal protein (blue circles) are also indicated.

© 2002 Macmillan Magazines Ltd

but have very different lifestyles. Their genomes reveal much similarity at the level of individual gene sequences, and many similar gene clusters. Global comparison showed perceptible higher-order synteny as well, shown as a dot plot in Fig. 2a. A prominent feature is the central broken diagonal cross pattern formed by the regions of synteny. This broken-X pattern is commonly seen in comparisons of related bacteria and the breaks are attributed to multiple inversions centred on *oriC*[10]. Normally, synteny extends over the whole of the compared chromosomes; however, for the comparison between *S. coelicolor* and *M. tuberculosis*, the broken-X pattern correlates only with the core of the *S. coelicolor* chromosome. Therefore this region and the whole *M. tuberculosis* chromosome must have had a common ancestor, with the chromosome arms of *S. coelicolor* consisting of subsequently acquired DNA. The syntenic regions mainly comprise genes concerned with primary cellular functions. The most strongly conserved is the gene cluster coding for the subunits of respiratory chain NADH dehydrogenase (systematic gene numbers SCO4562–4575). Functions/proteins coded for by other regions of synteny include the origin of replication (SCO3873–3892), urease activity (SCO1231–1236), pyrimidine biosynthesis (SCO1472–1488), arginine biosynthesis (SCO1570–1580), pentose phosphate pathway/tricarboxylic acid cycle (SCO1921–1953), histidine and tryptophan biosynthesis (SCO2034–2054), cell division (SCO2077–2092) and ribosomal proteins (SCO4701–4724).

The genome of the pathogenic actinomycete *Corynebacterium diphtheriae* has been sequenced recently (http://www.sanger.ac.uk/Projects/C_diphtheriae/). Comparison with the *S. coelicolor* chromosome gives a similar pattern to that for *M. tuberculosis*, with the regions of synteny covering the entire *C. diphtheriae* chromosome and just the *S. coelicolor* core region (Fig. 2b). The syntenic regions again correspond to genes coding for primary cellular functions and several of these regions are common to all three chromosomes. *Mycobacterium tuberculosis* and *C. diphtheriae* have more extensive synteny than either has with *S. coelicolor* (Fig. 2c), reflecting taxonomic groupings: *C. diphtheriae* and *M. tuberculosis* are in the suborder Corynebacterineae of the actinomycetes, whereas *S. coelicolor* is in the Streptomycineae.

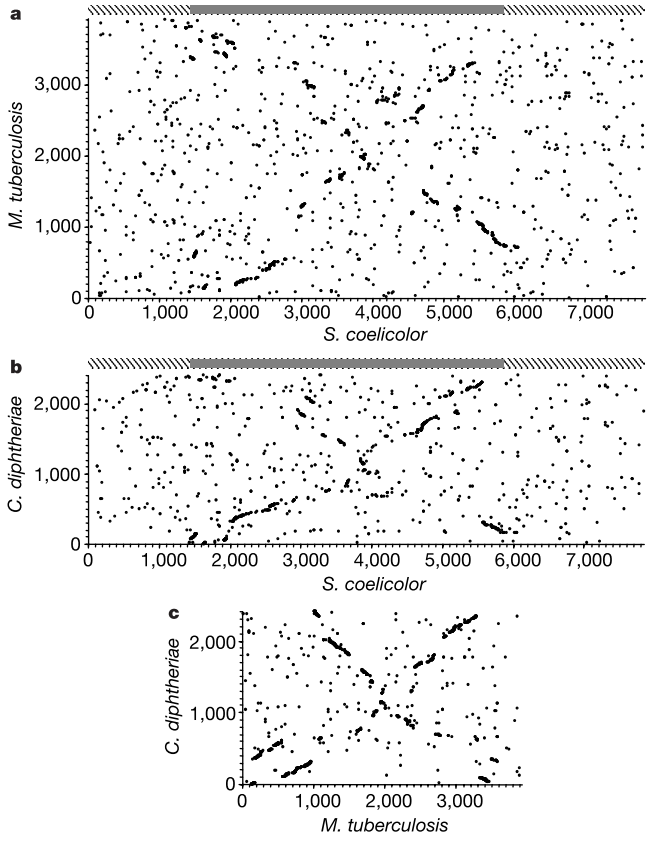

Figure 2 Comparison of chromosome structure for *S. coelicolor* versus *M. tuberculosis* (**a**), *S. coelicolor* versus *C. diphtheriae* (**b**) and *M. tuberculosis* versus *C. diphtheriae* (**c**). Axes represent the proteins coded for in the order in which they occur on the chromosomes. For each genome, DnaA is centrally located. Dots represent a reciprocal best match (by FASTA comparison[50]) between protein sets. The bars above plots **a** and **b** indicate the core (solid, SCO1440–5869) and arm (hatched) regions of the *S. coelicolor* chromosome.

By investigating regions of unusual DNA content and/or genes with sequence similarity to those from known mobile genetic elements, we designated 14 regions as potentially recently laterally acquired insertions (See Supplementary Information). By far the largest insertion contains 148 genes and is located at a transfer RNA gene[11]: as well as many hypothetical genes, it includes genes for heavy metal resistance (SCO6835–6837) and secondary metabolite production (SCO6827). Six other inserted regions have plasmid function genes in common with the integrative plasmid pSAM2 of *Streptomyces ambofaciens*[12]. Four of these pSAM2-like integrants appear to have inserted within a tRNA gene, including two that are adjacent to secondary metabolic clusters (calcium-dependent antibiotic (CDA), SCO3250–3262; *whiE*, SCO5327–5350). Notably, 11 of the 14 insertions lie to the right of *oriC*, correlating with the greater variation in DNA composition in the right half of the chromosome (Fig. 1).

Putative transposase genes are found throughout the chromosome in intact, truncated and frame-shifted forms. Many are associated with the multi-gene integrations described above. For the remainder, there is a particular concentration at the sub-TIR regions, 35–95 kb from the ends (Fig. 1). This indicates a tolerance to insertion events in these regions and thus offers a possible route for chromosome expansion. Of the 78 predicted transposase coding sequences, five are within transposons (one of which codes for a possible antibiotic resistance protein (SCO0107)), 31 form simple insertion elements and the remainder are not bounded by inverted repeats. Most fall into five families, suggesting a degree of intra-chromosomal transposition. Such events offer a route for gene duplication. Two of the insertion elements mark the inner boundaries of the TIRs, suggesting a possible role in their maintenance.

A plethora of proteins

With 7,825 predicted genes, the *S. coelicolor* chromosome has an enormous coding potential. This figure compares with 4,289 genes in the Gram-negative bacterium *Escherichia coli*; 4,099 in the Gram-positive spore-former *Bacillus subtilis*; 6,203 in the lower eukaryote *Saccharomyces cerevisiae*; and a predicted 31,780 in humans (http://www.ebi.ac.uk/genomes/). The genome contains almost twice as many genes as that of *M. tuberculosis*. This large number of genes reflects both a multiplicity of new protein families and an expansion within known families when compared with other bacteria (further information is available at http://www.sanger.ac.uk/Projects/S_coelicolor/). Many protein families that are significantly expanded in *S. coelicolor* are involved in regulation, transport and degradation of extracellular nutrients (Table 2).

The genome shows a strong emphasis on regulation, with 965 proteins (12.3%) predicted to have regulatory function. Discovery of so many regulators extends the observation that the proportion of regulatory genes increases with bacterial genome size[13]. There is a clear preference for certain regulator groups. Sigma factors act by binding to and affecting the promoter specificity of the RNA polymerase core enzyme, thus directing the selective transcription of gene sets. *Streptomyces coelicolor* codes for a remarkable 65 sigma factors (the next highest number so far found is 23 in *Mesorhizobium loti*, with a genome size of 7.6 Mb[14]), of which 45 are 'ECF' (extra-cytoplasmic function) sigma factors (41 from family 13

© 2002 Macmillan Magazines Ltd

alone; Table 2). Previously described ECF sigma factors (in *S. coelicolor*) respond to external stimuli and activate genes involved in disulphide stress, cell-wall homeostasis and aerial mycelium development[15]. Most of the other sigma factors fall into a single group (family 54, Table 2). Within this is a sub-group peculiar to Gram-positive bacteria, most of which have a single member; however, *B. subtilis* has three, controlling forespore development and the general stress response, and *S. coelicolor* has at least eight, many of them involved in responses to various stresses[16]. The numerous potentially stress-responsive sigma factors may account for the independent regulation of diverse stress response regulons in *S. coelicolor*[17]. Although widely distributed among bacteria, the atypical, enhancer-dependent sigma-54 and its cognate activators[18] are absent.

Streptomyces coelicolor also has abundant two-component regulatory systems where typically, in response to an extracellular stimulus, an integral membrane sensor protein phosphorylates a response regulator, causing it to bind to specific promoter regions and thus activate or repress transcription. We identified 85 sensor kinases and 79 response regulators, including 53 sensor–regulator pairs. The genome also codes for many members of previously described regulator families such as LysR, LacI, ROK, GntR, TetR, IclR, AraC, AsnC and MerR. The TetR family regulators in *S. coelicolor* form several subfamilies, often containing few or no members from the other genomes analysed. Furthermore, there is a group (family 86, Table 2) of 25 putative DNA-binding proteins that has no members from outside *S. coelicolor* and may constitute a new *Streptomyces*-specific family of regulators. Also notable is the presence of 44 putative serine/threonine protein kinases (family 6.1, Table 2). Examples of these typically eukaryotic regulators are now known to occur in many bacteria, but in much smaller numbers.

Reflecting its many interactions with the complex soil environment, *S. coelicolor* has 614 proteins (7.8%) with predicted transport function. A large proportion of these are of the ABC transporter type, including 81 typical ABC permeases and 141 ATP-binding proteins (24 of which are fused to membrane-spanning domains). Transporters for which the substrate is predictable include those for sugars, amino acids, peptides, metals and other ions. There are also several possible drug efflux proteins. Import of specific substrates would in part be facilitated by the 75 putative surface-anchored substrate-binding proteins of *S. coelicolor*.

The ability of *S. coelicolor* to exploit nutrients in the soil is abundantly demonstrated by our prediction of 819 potentially secreted proteins (10.5%). Secreted hydrolases are particularly numerous (for example, family 7 (Table 2), which is over-represented in *S. coelicolor*). They include 60 proteases/peptidases, 13 chitinases/chitosanases, eight cellulases/endoglucanases, three amylases and two pectate lyases. As well as the complete Sec protein translocation system, *S. coelicolor* seems to contain the machinery and cognate signal sequences for the recently discovered TAT (twin arginine transport) system for exporting pre-folded proteins[19] (T. Palmer, personal communication).

A marked example of multiple paralogues in *S. coelicolor* is a four-gene cluster that we named the conservon (for conserved operon). In the 13 such clusters (*cvnA*, *B*, *C*, *D*, *1-13*) there is unidirectional transcription and often overlap of translational start and stop codons, suggesting an operon structure. The only other known *cvn* cluster is present in *M. tuberculosis*. The protein products form distinct and exclusive families (Table 2; families 178, 177, 214, 180; CvnA, B, C and D, respectively). The first gene codes for a probable membrane protein weakly resembling sensor kinases, the fourth codes for a possible ATP/GTP-binding protein, and the other two are of unknown function. In four of the clusters the immediate downstream gene codes for a predicted cytochrome P-450.

Paralogous enzymes may sometimes represent isozymes active at different stages in the developmental cycle. One such example is the differential activities of duplicate gene clusters for glycogen synthesis in the vegetative and aerial mycelium[20]. Here we highlight a further five examples of paralogues for metabolic enzymes in *S. coelicolor*. (1) Two gene clusters code for enzymes of the pentose phosphate pathway (SCO1935–1939 and SCO6657–6663). (2) Four loci for tryptophan biosynthesis (SCO2036–2043, SCO2117,

Table 2 Occurrence of a selection of protein families in six related genomes

Majority description*	Sco	Mtu	Cdi	Bsu	Mlo	Eco
ECF sigma factor (13)	41 (0.52)	10 (0.25)	7 (0.29)	7 (0.17)	16 (0.23)	1 (0.02)
Sigma factor (54)	14 (0.17)	3 (0.07)	2 (0.08)	8 (0.19)	3 (0.04)	4 (0.09)
Two-component sensor kinase (1.3)	27 (0.34)	8 (0.20)	5 (0.20)	12 (0.29)	41 (0.60)	20 (0.46)
Two-component sensor kinase (1.15)	6 (0.07)	0	0	0	0	0
Two-component regulator (1.6)	50 (0.63)	2 (0.05)	5 (0.20)	9 (0.21)	5 (0.07)	7 (0.16)
Two-component regulator (1.5)	24 (0.30)	11 (0.28)	6 (0.24)	13 (0.31)	21 (0.31)	14 (0.32)
Serine/threonine protein kinase (6.1)	44 (0.56)	13 (0.33)	5 (0.20)	8 (0.19)	14 (0.20)	8 (0.18)
Regulator (LacI) (2.4)	33 (0.42)	1 (0.02)	2 (0.08)	12 (0.29)	15 (0.22)	13 (0.30)
Regulator (ROK) (36)	23 (0.29)	3 (0.07)	3 (0.12)	3 (0.07)	6 (0.08)	7 (0.16)
Regulator (TetR) (112)	18 (0.22)	1 (0.02)	0	0	0	0
Regulator (KorSA/GntR) (2.9)	10 (0.12)	0	0	0	0	0
Regulator (WhiB-like)	8 (0.10)	4 (0.10)	3 (0.12)	0	0	0
DNA-binding (86)	25 (0.31)	0	0	0	0	0
ABC transport (ATP-binding) (2.1.3)	27 (0.34)	4 (0.10)	8 (0.33)	6 (0.14)	3 (0.04)	3 (0.06)
Transport (permease) (2.3)	36 (0.45)	4 (0.10)	1 (0.04)	8 (0.19)	25 (0.37)	3 (0.06)
Transport (sugar) (2.2)	36 (0.45)	4 (0.10)	1 (0.04)	8 (0.19)	26 (0.38)	3 (0.06)
Transport (substrate-binding) (1.8)	35 (0.44)	4 (0.10)	1 (0.04)	6 (0.14)	24 (0.35)	3 (0.06)
Integral membrane (59)	14 (0.17)	14 (0.35)	3 (0.12)	2 (0.04)	0	0
Membrane ATPase (42)	13 (0.16)	12 (0.30)	6 (0.24)	4 (0.09)	3 (0.04)	1 (0.02)
Secreted hydrolase (7)	100 (1.27)	19 (0.48)	8 (0.33)	21 (0.51)	8 (0.11)	9 (0.20)
Secreted hydrolase (7.3)	17 (0.21)	0	0	0	1 (0.01)	0
Secreted chitinase (7.8)	5 (0.06)	0	0	0	0	0
Secreted protease (191)	10 (0.12)	3 (0.07)	0	0	0	0
Secreted protease (7.6)	8 (0.10)	0	0	0	0	0
Secreted cellulase (7.4)	7 (0.08)	1 (0.02)	0	1 (0.02)	0	0
Secreted hypothetical (17)	25 (0.31)	11 (0.28)	8 (0.33)	13 (0.31)	8 (0.11)	9 (0.20)
Hypothetical (63)	25 (0.31)	0	0	6 (0.14)	0	0
Conservon (Cvn1–4; 178, 177, 214, 180)	13 (0.16)	1 (0.02)	0	0	0	0
Hypothetical (204)	13 (0.16)	0	0	0	0	0
Hypothetical (19)	12 (0.15)	0	0	0	0	0

Numbers indicate absolute number of proteins from each genome in each family and the percentage of the total proteins in that genome in parentheses. Family number is indicated in parentheses in the majority description column. The hierarchical numbering system reflects use of higher BlastP thresholds to break large complex families into discrete subfamilies. Complete data are available from http://www.sanger.ac.uk/Projects/S_coelicolor/. Sco, *S.coelicolor*; Mtu, *M. tuberculosis*; Cdi, *C. diphtheriae*; Bsu, *B. subtilis*; Mlo, *M. loti*; Eco, *E. coli*.
*Groupings are based on sequence similarity, so individual families do not necessarily include all representatives of each type of protein in each genome (see Methods).

© 2002 Macmillan Magazines Ltd

SCO2147, SCO3211–3214) include two *trpC*, two *trpD* and three *trpE* genes. A *trpCDGE* locus is within the gene cluster for production of CDA[21], a peptide antibiotic that contains tryptophan (in the 'unnatural' D form). The local cluster may ensure adequate tryptophan for CDA biosynthesis at the appropriate stage in the life cycle, independently of the needs of protein synthesis. (3) Five homologues of *fabH* code for a dedicated ketosynthase for the first step in fatty acid biosynthesis (condensation of acetyl-coenzyme A (CoA) with malonyl-CoA to yield acetoacetyl-CoA). One of the five (SCO2388) is in the main fatty acid biosynthetic operon and is essential[22]. Three of the other four *fabH* homologues (SCO5888, 3246, 1271) are in gene clusters for secondary metabolism: the *red* and *cda* clusters, and a cluster of unknown product (see below). At least the first two clusters determine molecules with fatty acid components, and the presence of *fabH* paralogues makes it highly probable that some of the steps in their biosynthesis use dedicated enzymes, rather than sharing enzymes functioning in primary metabolism[23]. (4) Three clusters code for a typical four-subunit respiratory nitrate reductase (SCO0216–0219, SCO4947–4950, SCO6532–6535), indicating the importance of a capacity for microaerobic growth in what was classically regarded as an obligate aerobe. (5) Flexibility in respiration is further indicated by a second (partial) copy of the operon coding for subunits of the respiratory chain NADH dehydrogenase (SCO4599–4608).

Unexpectedly, there are two gene clusters (SCO0649–0658, SCO6499–6508) similar in sequence and gene order to an operon from *Halobacterium* sp. that is involved in the production of gas vesicle proteins, including the eight genes essential for this phenotype[24]. Many overtly water-living bacteria use gas vesicles as flotation devices, but the only previous occurrence of gas vesicle genes (but not so far of the vesicles themselves) in a soil organism is in *Bacillus megaterium*[25]. The benefit of gas vesicles to *Streptomyces* is unknown, but perhaps such buoyancy devices would allow spores to remain at the oxygen-rich surface during dispersal and germination in waterlogged soil.

Many genes for secondary metabolism

Chromosomal gene clusters specifying the biosynthesis of the aromatic polyketide antibiotic actinorhodin, the so-called RED complex of red oligopyrrole prodiginine antibiotics, and the non-ribosomal peptide CDA had previously been analysed[26,27], as had the *whiE* cluster of genes coding for a type II polyketide synthase for a grey spore pigment[28]. The genome sequence reveals a further 18 clusters that would code for enzymes characteristic of secondary metabolism (Fig. 3). These include type I modular and both type I and type II iterative polyketide synthases (PKSs), chalcone synthases, non-ribosomal peptide synthetases (NRPSs), terpene cyclases, and others. The distribution of the clusters on the chromosome seems non-random, with some preponderance in the arms, but more especially in a region near the right-hand core–arm boundary (Fig. 1). Comparison with similar clusters from other organisms and the application of recently developed sequence analysis tools have, in some cases, provided insight into the probable structure of the end products determined by these genes. For example, using predictive models for substrate amino-acid recognition[29,30], the two NRPSs coded for by SCO0492 and SCO7681–

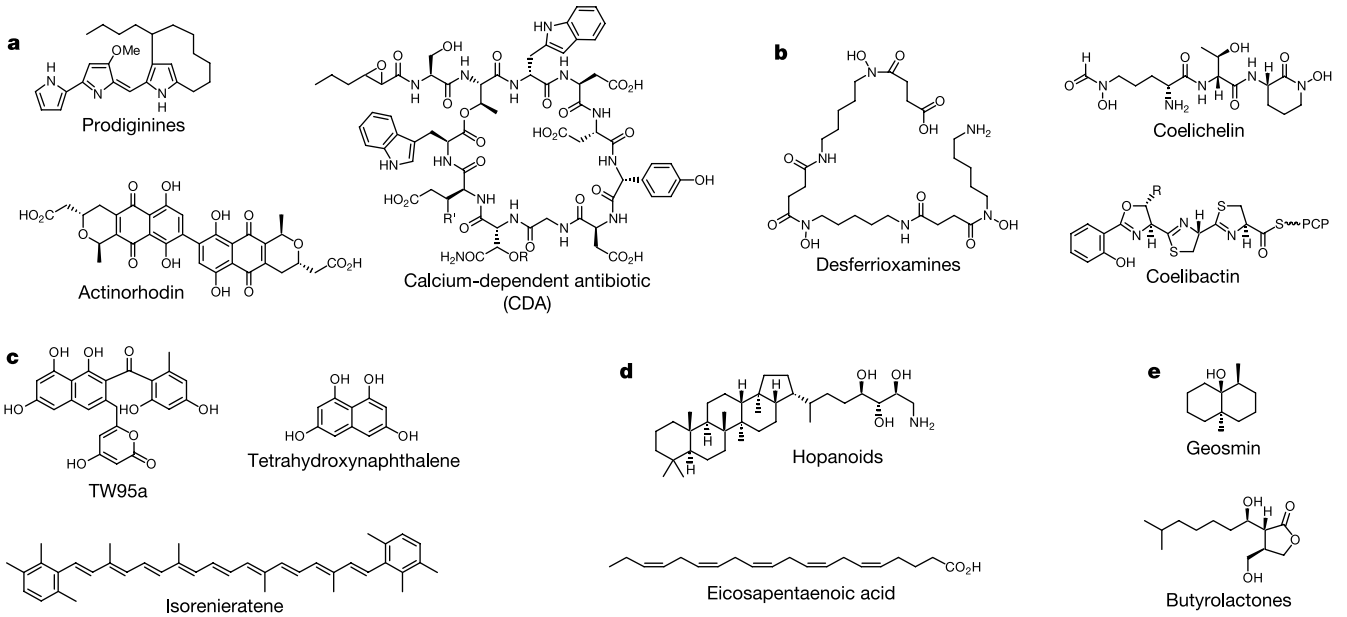

Figure 3 Secondary metabolites known or predicted to be made by *S. coelicolor* A3(2), grouped according to their putative function. These are: antibiotics (**a**), siderophores (**b**), pigments (**c**), lipids (**d**) and other molecules (**e**). The chromosomal locations of the gene clusters are: actinorhodin, SCO5071–5092; prodiginines (mixture of butyl-*meta*-cycloheptylprodiginine (shown) and undecylprodiginine), SCO5877–5898; CDA complex (CDA1, R = OPO_3H_2, R′ = H; CDA2, R = OPO_3H_2, R′ = Me; CDA3b, R = OH, R′ = H; CDA4b, R = OH, R′ = Me), SCO3210–3249; desferrioxamines (mixture of desferrioxamine G_1 (shown) and desferrioxamine E), SCO2782–2785; coelichelin, SCO0489–0499; coelibactin (structure is that predicted for a late intermediate attached to the PCP domain in the last module of the coelibactin NRPS; R = H/Me, the complete structure cannot be predicted as the regiospecificity of several methyltransferases, a cytochrome P-450 and an oxidoreductase coded for by genes in the cluster cannot be deduced), SCO7681–7691; TW95a (structure is the product obtained from heterologous expression of the *whiE* minimal PKS and the *whiE-ORFIV* genes; the structure of the grey spore pigment has not been elucidated), SCO5314–5320; tetrahydroxynaphthalene (predicted product of the chalcone synthase, which may be further modified by enzymes coded for by other genes in the cluster), SCO1206–1208; isorenieratene, SCO0185–0191; hopanoids (mixture of aminotrihydroxybacteriohopane (shown) and hopene), SCO6759–6771; eicosapentaenoic acid, SCO0124–0129; geosmin, SCO6073; butyrolactones (believed to be assembled by the *scbA* gene product), SCO6266. The structures of the remaining putative secondary metabolites are unknown. The chromosomal location of these clusters and the type of secondary metabolic enzyme(s) coded for are: SCO6429–6438, NRPS; SCO6273–6288 and SCO6826-6827, type I polyketide synthases; SCO7669–7671 and SCO7222, chalcone synthases; SCO5222–5223, sesquiterpene cyclase; SCO5799–5801, siderophore synthetase; SCO1265–1273, type II fatty acid synthase; SCO0381–0401, deoxysugar synthases/glycosyl transferases.

© 2002 Macmillan Magazines Ltd

7683 were deduced to catalyse the biosynthesis of novel siderophores named 'coelichelin'[31] and 'coelibactin' (G.L.C., unpublished data), respectively. A third cluster, SCO2782–2785, probably directs the biosynthesis of two further siderophores, desferrioxamines G1 and E[32]. Two large open reading frames (ORFs) (SCO0126 and 0127) code for multi-enzymes with a domain organization very similar to a type I iterative PKS/FAS from a Gram-negative bacterium, *Shewanella* sp., that catalyses biosynthesis of eicosapentaenoic acid[33]. We therefore predict a role for this cluster in polyunsaturated fatty acid biosynthesis. Similarly, the cluster SCO6759–6771 has been implicated in hopanoid biosynthesis[34], and SCO1206–1208 in tetrahydroxynaphthalene biosynthesis[35]. The sesquiterpene cyclase coded for by SCO6073 is probably involved in geosmin biosynthesis (B. Gust, K. Fowler, T.K., G.L.C. and K.F.C., personal communication) and SCO0185–0191 probably directs biosynthesis of the carotenoid isoreneriatine[36].

Although three of the *S. coelicolor* clusters specify antibiotics, most of the others are probably responsible for products with different functions. For example, hopanoids may protect against water loss through the plasma membrane in the aerial mycelium[34], and eicosapentaenoic acid may help to maintain membrane fluidity at low temperature. It is notable that at least three clusters probably code for siderophore biosynthesis, implying that *S. coelicolor* is under strong selective pressure to scavenge iron in situations of low iron availability. Thus, products of some of these clusters might accurately be labelled 'stress metabolites', predicted to combat stresses of a physical (desiccation, low temperature), chemical (low iron) or biological (competition) nature.

Cell and developmental biology

Escherichia coli and *B. subtilis* multiply by binary fission, whereas *S. coelicolor* grows as a non-dividing, many-branched mycelium, mainly by tip growth, with multiple copies of the genome in each hyphal compartment. Unigenomic dispersive exospores are borne as chains on specialized, little-branched aerial hyphae that probably extend by intercalary growth[37]. The genome sequence provides some new insights into this complex life cycle.

Initiation of DNA replication in *S. coelicolor* involves an *oriC*-linked *dnaA* gene, the product of which interacts with an unusually large number (17) of DnaA boxes at the replication origin[38]. In addition to its initiator function, DnaA is a transcription factor in a diverse range of bacteria[39]. It is therefore conspicuous that 42 (82%) out of the 51 'strong' DnaA boxes of *S. coelicolor* (TT(G/A)TCCACA[38]) lie in non-coding DNA upstream of genes. DnaA may conceivably coordinate the replication of multiple genomes in each hyphal compartment with cell-cycle-dependent gene expression.

Our limited understanding of bacterial chromosome partitioning is based largely on studies of low copy number plasmids of Gram-negative bacteria[40]. The *parAB* gene pair on many such plasmids codes for ParA, an ATPase of unclear function, and ParB, which binds one or more *parS* sites near *parA* and *parB*. Many bacteria (including *S. coelicolor*, but not *E. coli*) contain *parAB* genes near *oriC*, and in some cases *parS* target sites have been identified[41]. In *S. coelicolor*, there is a high concentration of putative *parS* sites surrounding *oriC*[42], with 18 'perfect' sites (GTTTCACGTGAAAC) in a 515-kb segment (4,174,551–4,689,985). Unlike DnaA boxes, nearly all of the *parS* sites are immediately downstream of genes, perhaps indicating selection for avoidance of effects on gene expression resulting from ParB–*parS* binding.

Streptomycetes have at least three different kinds of septa[43]. It is therefore surprising that genes clearly homologous to conserved 'divisome' (cell division) genes of other bacteria are generally present only once or (in the case of *ftsA*) not at all. Presumably the different kinds of cell division involve dedicated accessory proteins. This contrasts with genes coding for enzymes for peptidoglycan synthesis and metabolism: there are eight *ftsI*/*mrdA* (class 2/3 transpeptidase) and five *mrcA*/*mrcB* (peptidoglycan synthetase) homologues.

A principal difference between *S. coelicolor* and unicellular rods concerns septum placement. In rods, division involves a centrally located septum, with alternative division sites close to the cell poles usually being silent. This involves the *minC*, *D* and *E* genes in *E. coli*, and the *minC*, *D* and *divIVA* genes of *B. subtilis*[44]. In hyphae of *Streptomyces*, there is no centre point, and division events are usually far from hyphal 'poles'. Consistent with this, there are no *minC*, *minE* or *divIVA*-like genes in *S. coelicolor*. On the other hand, there is a large family of perceptibly *minD*-like genes (which, notably, reveal distant similarity to *parA*). These may control the use of potential division sites at various positions (for example, polar, sub-polar, between pre-existing septa, or at branch points).

Discussion

The genome sequence of *S. coelicolor* has revealed much about the many adaptations of this model actinomycete to life in the highly competitive soil environment. Derived from an ancestor common to other actinomycetes, the chromosome has acquired the ability to replicate in a linear form and appears to have expanded by lateral acquisition and internal duplication of DNA. Chromosome expansion has provided a wealth of genes, allowing the organism a more complex life cycle, adapting to a wider range of environmental conditions and exploiting a greater variety of nutrient sources. This has coincided with an increase in regulatory systems, with a particular emphasis on detection of, and response to, extracellular stimuli. The preferential incorporation (and subsequent maintenance) of occasionally beneficial sequences outside the ancestral core has created chromosome arms comprised mostly of 'non-essential' functions. The abundance of previously uncharacterized metabolic enzymes, particularly those likely to be involved in the production of natural products, is a resource of enormous potential value. Understanding of such enzymes will facilitate the genetic engineering of pathways to produce new compounds with potential therapeutic activity, including much needed antimicrobials[45]. The incomplete genome sequence of an industrial species, *S. avermitilis*[46], appears to contain a different set of gene clusters for secondary metabolism from *S. coelicolor*. It may be that the arm regions of different streptomycete chromosomes have been accumulated separately, and therefore contain a largely different complement of contingency genes representing a huge pool of metabolic diversity. □

Methods

Genome sequencing

We sequenced the genome of *S. coelicolor* A2(3) from 325 overlapping clones. Of these, 305 were cosmids[8], one was the terminal plasmid pLUS221 and 19 were selected from a set of 3,456 bacterial artificial chromosomes mapped to the sequences of finished cosmid contigs by end sequencing. The methods for clone growth and isolation, sonication to produce 1.4–2-kb fragments, library preparation in either M13 or pUC18 vectors, and sequencing were as described previously[47]. Most of the clones were digested with *Dra*I, and insert purified, before the fragmentation step in order to remove cloning vector. This was not done for those clones known to contain *Dra*I sites, and in these cases DNA from the cloning vector was greatly over-represented in the subclone libraries. The finished 325 clones formed a contiguous sequence extending from within the left TIR to the right end of the genome. The genome sequence was completed by extending the incomplete left TIR with a 7-kb consensus sequence copied from the right TIR. The sequence was assembled, finished and annotated as described previously[2], using Artemis (http://www.sanger.ac.uk/Software/Artemis) to collate data and facilitate annotation. Protein families were constructed, independently of annotation, by performing an 'all-against-all' Blast (NCBI Blast version 2) comparison[48] of proteins within a database containing all predicted protein products from six genomes (Table 2), then single-linkage clustering using a Blast threshold of 70 bits. We checked composition of families using ClustalW[49]. Complex families were resolved by raising the Blast threshold to 100, 150 or 200 bits, as reflected in the hierarchical family numbering system (for example, family 2.1.3 was created using a Blast threshold of 150 on family 2.1).

Received 3 December 2001; accepted 27 March 2002.

1. Hodgson, D. A. Primary metabolism and its control in streptomycetes: a most unusual group of bacteria. *Adv. Microb. Physiol.* **42,** 47–238 (2000).

© 2002 Macmillan Magazines Ltd

2. Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393,** 537–544 (1998).
3. Cole, S. T. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409,** 1007–1011 (2001).
4. Hopwood, D. A. Forty years of genetics with *Streptomyces*: from *in vivo* through *in vitro* to *in silico*. *Microbiology* **145,** 2183–2202 (1999).
5. Bao, K. & Cohen, S. N. Terminal proteins essential for the replication of linear plasmids and chromosomes in *Streptomyces*. *Genes Dev.* **15,** 1518–1527 (2001).
6. Volff, J. N. & Altenbuchner, J. Genetic instability of the *Streptomyces* chromosome. *Mol. Microbiol.* **27,** 239–246 (1998).
7. Friend, E. J. & Hopwood, D. A. The linkage map of *Streptomyces rimosus*. *J. Gen. Microbiol.* **68,** 187–197 (1971).
8. Redenbach, M. *et al.* A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb *Streptomyces coelicolor* A3(2) chromosome. *Mol. Microbiol.* **21,** 77–96 (1996).
9. Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13,** 660–665 (1996).
10. Eisen, J. A., Heidelberg, J. F., White, O. & Salzberg, S. L. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* **1,** (Research) 0011 (2000).
11. Mazodier, P., Thompson, C. & Boccard, F. The chromosomal integration site of the *Streptomyces* element pSAM2 overlaps a putative tRNA gene conserved among actinomycetes. *Mol. Gen. Genet.* **222,** 431–434 (1990).
12. Sezonov, G., Duchène, A. M., Friedmann, A., Guérineau, M. & Pernodet, J. L. Replicase, excisionase, and integrase genes of the *Streptomyces* element pSAM2 constitute an operon positively regulated by the *pra* gene. *J. Bacteriol.* **180,** 3056–3061 (1998).
13. Stover, C. K. *et al.* Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406,** 959–964 (2000).
14. Kaneko, T. *et al.* Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA Res.* **7,** 331–338 (2000).
15. Paget, M. S. B., Hong, H.-J., Bibb, M. J. & Buttner, M. J. in *Control of Bacterial Gene Expression* (eds Hodgson, D. A. & Thomas, C. M.) 105–125 (Cambridge Univ. Press, Cambridge, 2002).
16. Kelemen, G. H. *et al.* A connection between stress and development in the multicellular prokaryote *Streptomyces coelicolor* A3(2). *Mol. Microbiol.* **40,** 804–814 (2001).
17. Vohradsky, J. *et al.* Developmental control of stress stimulons in *Streptomyces coelicolor* revealed by statistical analyses of global gene expression patterns. *J. Bacteriol.* **182,** 4979–4986 (2000).
18. Buck, M., Gallegos, M. T., Studholme, D. J., Guo, Y. & Gralla, J. D. The bacterial enhancer-dependent sigma(54) (sigma(N)) transcription factor. *J. Bacteriol.* **182,** 4129–4136 (2000).
19. Berks, B. C., Sargent, F. & Palmer, T. The Tat protein export pathway. *Mol. Microbiol.* **35,** 260–274 (2000).
20. Schneider, D., Bruton, C. J. & Chater, K. F. Duplicated gene clusters suggest an interplay of glycogen and trehalose metabolism during sequential stages of aerial mycelium development in *Streptomyces coelicolor* A3(2). *Mol. Gen. Genet.* **263,** 543–553 (2000).
21. Huang, J., Lih, C. J., Pan, K. H. & Cohen, S. N. Global analysis of growth phase responsive gene expression and regulation of antibiotic biosynthetic pathways in *Streptomyces coelicolor* using DNA microarrays. *Genes Dev.* **15,** 3183–3192 (2001).
22. Revill, W. P., Bibb, M. J., Scheu, A. K., Kieser, H. J. & Hopwood, D. A. Beta-ketoacyl acyl carrier protein synthase III (FabH) is essential for fatty acid biosynthesis in *Streptomyces coelicolor* A3(2). *J. Bacteriol.* **183,** 3526–3530 (2001).
23. Hopwood, D. A. Genetic contributions to understanding polyketide synthases. *Chem. Rev.* **97,** 2465–2497 (1997).
24. Offner, S., Hofacker, A., Wanner, G. & Pfeifer, F. Eight of fourteen *gvp* genes are sufficient for formation of gas vesicles in halophilic archaea. *J. Bacteriol.* **182,** 4328–4336 (2000).
25. Li, N. & Cannon, M. C. Gas vesicle genes identified in *Bacillus megaterium* and functional expression in *Escherichia coli*. *J. Bacteriol.* **180,** 2450–2458 (1998).
26. Hopwood, D. A., Chater, K. F. & Bibb, M. J. in *Genetics and Biochemistry of Antibiotic Production* (eds Vining, L. C. & Stuttard, C.) 65–102 (Butterworth-Heinemann, Newton, Massachusetts, 1995).
27. Chong, P. P. *et al.* Physical identification of a chromosomal locus encoding biosynthetic genes for the lipopeptide calcium-dependent antibiotic (CDA) of *Streptomyces coelicolor* A3(2). *Microbiology* **144,** 193–199 (1998).
28. Davis, N. K. & Chater, K. F. Spore colour in *Streptomyces coelicolor* A3(2) involves the developmentally regulated synthesis of a compound biosynthetically related to polyketide antibiotics. *Mol. Microbiol.* **4,** 1679–1691 (1990).
29. Challis, G. L., Ravel, J. & Townsend, C. A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* **7,** 211–224 (2000).
30. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6,** 493–505 (1999).
31. Challis, G. L. & Ravel, J. Coelichelin, a new peptide siderophore encoded by the *Streptomyces coelicolor* genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiol. Lett.* **187,** 111–114 (2000).
32. Imbert, M., Bechet, M. & Blondeau, R. Comparison of the main siderophores produced by some species of *Streptomyces*. *Curr. Microbiol.* **31,** 129–133 (1995).
33. Takeyama, H., Takeda, D., Yazawa, K., Yamada, A. & Matsunaga, T. Expression of the eicosapentaenoic acid synthesis gene cluster from *Shewanella sp*. in a transgenic marine cyanobacterium, *Synechococcus* sp. *Microbiology* **143,** 2725–2731 (1997).
34. Poralla, K., Muth, G. & Hartner, T. Hopanoids are formed during transition from substrate to aerial hyphae in *Streptomyces coelicolor* A3(2). *FEMS Microbiol. Lett.* **189,** 93–95 (2000).
35. Funa, N. *et al.* A new pathway for polyketide synthesis in microorganisms. *Nature* **400,** 897–899 (1999).
36. Krügel, H., Krubasik, P., Weber, K., Saluz, H. P. & Sandmann, G. Functional analysis of genes from *Streptomyces griseus* involved in the synthesis of isorenieratene, a carotenoid with aromatic end groups, revealed a novel type of carotenoid desaturase. *Biochim. Biophys. Acta* **1439,** 57–64 (1999).
37. Chater, K. F. & Losick, R. in *Bacteria as Multicellular Organisms* (eds Shapiro, J. A. & Dworkin, M.) 149–182 (Oxford Univ. Press, New York, 1997).
38. Zakrzewska-Czerwinska, J., Jakimowicz, D., Majka, J., Messer, W. & Schrempf, H. Initiation of the *Streptomyces* chromosome replication. *Antonie Van Leeuwenhoek* **78,** 211–221 (2000).
39. Messer, W. & Weigel, C. DnaA initiator—also a transcription factor. *Mol. Microbiol.* **24,** 1–6 (1997).
40. Bignell, C. & Thomas, C. M. The bacterial ParA-ParB partitioning proteins. *J. Biotechnol.* **91,** 1–34 (2001).
41. Lin, D. C. & Grossman, A. D. Identification and characterization of a bacterial chromosome partitioning site. *Cell* **92,** 675–685 (1998).
42. Kim, H. J., Calcutt, M. J., Schmidt, F. J. & Chater, K. F. Partitioning of the linear chromosome during sporulation of *Streptomyces coelicolor* A3(2) involves an *oriC*-linked *parAB* locus. *J. Bacteriol.* **182,** 1313–1320 (2000).
43. Kwak, J., Dharmatilake, A. J., Jiang, H. & Kendrick, K. E. Differential regulation of *ftsZ* transcription during septation of *Streptomyces griseus*. *J. Bacteriol.* **183,** 5092–5101 (2001).
44. Jacobs, C. & Shapiro, L. Bacterial cell division: a moveable feast. *Proc. Natl Acad. Sci. USA* **96,** 5891–5893 (1999).
45. Rodriguez, E. & McDaniel, R. Combinatorial biosynthesis of antimicrobials and other natural products. *Curr. Opin. Microbiol.* **4,** 526–534 (2001).
46. Ōmura, S. *et al.* Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. *Proc. Natl Acad. Sci. USA* **98,** 12215–12220 (2001).
47. Harris, D. & Murphy, L. in *Methods in Molecular Biology* (eds Starkey, M. P. & Elaswarapu, R.) Vol. 175, 217–234 (Humana, Totowa, New Jersey, 2001).
48. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
49. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
50. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85,** 2444–2448 (1988).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com).

Acknowledgements

We would like to acknowledge the support of the Wellcome Trust Sanger Institute core sequencing and informatics groups. This work was funded by the Biotechnology and Biological Sciences Research Council and by the Wellcome Trust through its Beowulf Genomics Initiative. C.W.C. and C.-H.H. were supported by the R.O.C. National Science Council and Ministry of Education.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.D.B. (e-mail: sdb@sanger.ac.uk) or D.A.H. (e-mail: david.hopwood@bbsrc.ac.uk). The complete sequence is deposited in GenBank/EMBL under accession number AL645882.

© 2002 Macmillan Magazines Ltd