nature biotechnology

Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338

Markiyan Oliynyk[1–3], Markiyan Samborskyy[1,3], John B Lester[1], Tatiana Mironenko[1], Nataliya Scott[1], Shilo Dickens[1], Stephen F Haydock[1] & Peter F Leadlay[1]

Saccharopolyspora erythraea **is used for the industrial-scale production of the antibiotic erythromycin A, derivatives of which play a vital role in medicine. The sequenced chromosome of this soil bacterium comprises 8,212,805 base pairs, predicted to encode 7,264 genes. It is circular, like those of the pathogenic actinomycetes *Mycobacterium tuberculosis* and *Corynebacterium diphtheriae*, but unlike the linear chromosomes of the model actinomycete *Streptomyces coelicolor A3(2)* and the closely related *Streptomyces avermitilis*. The *S. erythraea* genome contains at least 25 gene clusters for production of known or predicted secondary metabolites, at least 72 genes predicted to confer resistance to a range of common antibiotic classes and many sets of duplicated genes to support its saprophytic lifestyle. The availability of the genome sequence of *S. erythraea* will improve insight into its biology and facilitate rational development of strains to generate high-titer producers of clinically important antibiotics.**

S. erythraea (Waksman 1923) Labeda 1987, comb. nov.[1] is a Gram-positive filamentous bacterium originally identified as *Streptomyces erythraeus* but later assigned to the genus *Saccharopolyspora*. Despite this reclassification, it has been considered as very close in its biology to genuine streptomycetes such as *S. avermitilis* and the model organism *S. coelicolor* A3(2), the complete linear genomes of which have been published[2,3]. *S. erythraea* produces erythromycin A, an important broad-spectrum antibiotic against pathogenic Gram-positive bacteria[4]. The commercial importance of erythromycin has fostered intensive research into its biosynthesis, and genetic engineering of the pathways involved promises to enhance production of potentially valuable analogs of polyketide secondary metabolites[5]. This has revived efforts to increase strain productivity. Historically, wild-type actinomycete strains have been subjected to multiple rounds of random mutagenesis and selection to obtain overproducing mutants for industrial production of a desired secondary metabolite. However, genome-scale information might allow such actinomycete strains to be more quickly optimized for production. We present the complete sequence of the *S. erythraea* genome and compare it to other actinobacteria whose genomes have been sequenced. The strain used, NRRL23338, is the original form[6] of the type strain of *S. erythraca* NRRL2338, which is now listed as NRRL23338 white.

RESULTS

Sequencing and gene annotation of the *S. erythraea* genome

The main features of the chromosome sequence are shown in **Table 1** and **Figure 1**. At 8,212,805 bp, it is comparable in size to the linear genomes of *S. coelicolor* M145 (8.7 Mbp) and *S. avermitilis* MA-4680 (9.0 Mbp). However, the *S. erythraea* genome is apparently circular rather than linear, a topology it shares with other actinobacteria such as *M. tuberculosis*[7], *C. diphtheriae*[8], *Nocardia farcinica*[9] and *Frankia spp*[10]. The *S. erythraea* chromosome contains 7,198 predicted protein-coding sequences (CDSs), whose overall features are given in **Table 1**. For 4,777 (66.4%) of these, a putative function could be ascribed to the encoded proteins (**Supplementary Table 1** online). Of the rest, 829 (11.5%) showed similarity to hypothetical proteins in other genomes, and 1,592 (22.1%) had no substantial similarity to predicted proteins in public databases. The initiation codon of the *dnaA* gene, adjacent to the origin of replication *oriC*, was chosen as the starting point for numbering the CDSs[11]. The average GC content of the *S. erythraea* chromosome is 71.1%, the GC bias being noticeably lower near *oriC*. There is a definite coding bias in favor of the leading strand (59.1%), and a pronounced GC skew inversion can be seen at *oriC* and also on the opposite side of the chromosome to *oriC*, where replication presumably terminates (**Fig. 1**). A region of the chromosome (total of 4.4 Mbp) extending either side of *oriC* (**Fig. 1**) appears to contain the majority (85%) of the genes predicted to be essential. The ends of this region are signaled by extensive regions of markedly lower GC content (**Fig. 1**). In this core region (an important feature also of the linear genomes of both *S. avermitilis*[2] and *S. coelicolor*[3]), the gene order shows substantial residual conservation when other actinobacteria are used for comparison, although numerous inversions have occurred around *oriC* (**Fig. 2**). In the regions outside the core, where the chromosome has apparently undergone major expansion,

[1]Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK. [2]Present address: EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [3]These authors contributed equally to the work. Correspondence should be addressed to P.F.L. (leadlay@mole.bio.cam.ac.uk).

Received 21 December 2006; accepted 21 February 2007; published online 18 March 2007; doi:10.1038/nbt1297

Table 1 Features of the *S. erythraea* genome

Component of the genome	Property
Topology	Circular
Length	8,212,805 bp
G+C content	71.1%
Coding density	84.9%
Coding sequences	7,264
rRNA	16 genes in four sets
tRNA	50
CDSs	7,198
conserved with assigned function	4,777 (66.4%)
conserved with unknown function	829 (11.5%)
nonconserved	1,592 (22.1%)
Average CDS length	967.9 bp

CDS, protein-coding sequence.

S. erythraea has many more orthologs with *N. farcinica* or *S. coelicolor* than with *M. tuberculosis*, as expected, but the orthologs (reciprocal best-hit pairs in pairwise BLASTP searches $E < 10e^{-10}$) are in each case randomly scattered on the chromosome (**Fig. 2**). Compared to the linear streptomycete genomes, this 'noncore' region of the *S. erythraea* genome contains a very high number of insertion sequences (93 in 13 separate families, 2.3% of the genome) almost all of which are associated with transposases. Recombination between these repetitive elements could well have promoted the observed randomization of orthologous gene locations compared to the streptomycete genomes. Half of the insertion sequence elements are found in two major clusters at 2.5–3.1 Mbp and 5.4–5.75 Mbp, respectively (**Fig. 1**). These regions also have a substantially lower GC content (**Fig. 1**), which hints at prior horizontal gene transfer. A third region of lower GC content (6.06–6.24 Mbp) contains several giant CDSs (SACE_5463, SACE_5483 and SACE_5523) of unknown function, which may also have been acquired by horizontal gene transfer. Further work will be required[12] to obtain an accurate identification of genes potentially acquired by horizontal gene transfer, and to identify, if possible, the likely origins of such genes[12]. There are four sets of rRNA genes, each containing, unusually, a duplicated 5S rRNA gene. *S. erythraea* also differs from sequenced streptomycete strains in having a *sel* operon (*selA-D*) for the production of selenocysteinyl-tRNA, including the selenocysteinyl-tRNA itself, which recognizes specific UGA stop codons (*selC*, SACE_3551). The previously described and partially sequenced plasmids pSE101 (ref. 13) and pSE211 (ref. 14) are both present as integrated elements of 10.9 kbp and 17.3 kbp, respectively.

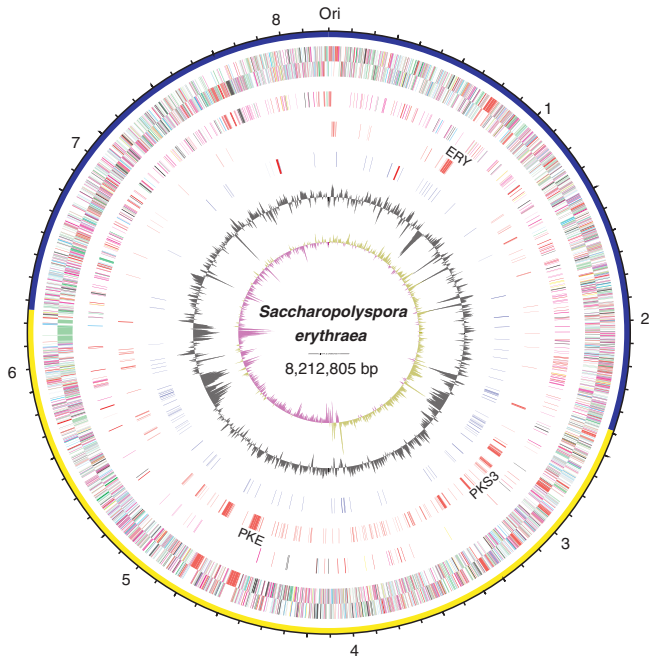
Figure 1 Schematic representation of the *S. erythraea* chromosome. The outer scale is numbered in megabases from the origin of replication (*ori*) and indicates the core (blue) and noncore (yellow) chromosomal regions. Circles 1 and 2 (from the outside in), all genes (reverse and forward strand, respectively) color-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; gray, miscellaneous); circle 3, selected essential genes (for cell division, DNA replication, transcription, translation and amino-acid biosynthesis, color coding as for circles 1 and 2); circle 4, selected secondary metabolic genes (three largest PKS clusters labeled ERY, PKE and PKS3); circle 5, mobile genetic elements (blue, transposases; red, prophages/integrated plasmids); circle 6, GC content; circle 7, GC bias ((G – C/G + C), khaki indicates values >1, purple values <1).

The topology of the chromosome

The genome of *S. erythraea* NRRL2338 has previously been reported to be linear, as judged by restriction mapping[15], but those data are also consistent with a circular genome if very small (and easily overlooked) predicted fragments from the putative termini are taken into account (**Supplementary Fig. 1** online). No evidence was found, from our sequence analysis, for the presence of terminal inverted repeat sequences, genes encoding termini-associated proteins or any other features of previously described linear streptomycete genomes[16–18]. Unlike *S. coelicolor* and *S. avermitilis*, *S. erythraea* also has genes resembling both *xerC* (SACE_2322, SACE_6041) and *xerD* (SACE_2295, SACE_3643, SACE_5095, SACE_5242), which together comprise a site-specific recombination system for resolving dimeric circular chromosomes. We avoided extensive handling or passage of the *S. erythraea* NRRL23338 strain from the NRRL Culture Collection. This is the 'white', less-pigmented form previously described[6], which produced substantially more erythromycin than the 'red' form listed as NRRL2338 red. The circularity of the genome therefore seems to be a feature of the type strain. We have recently conducted extensive shotgun sequencing of DNA from a different isolate of NRRL2338, and again found no evidence for linearity (data not shown). It remains possible that other lineages of *S. erythraea* may be found to be linear, and that the circular genome we have sequenced has arisen recently. An alternative possibility is that the difference in topology is ancient, and reflects the taxonomic separation of *S. erythraea* from the streptomycetes. It has been suggested that the unusual linear genomes of streptomycetes may have arisen through integration of a linear plasmid into a circular chromosome[19,20].

Gene comparisons with *S. coelicolor* and *S. avermitilis*

The Basic Local Alignment Search Tool protein clustering program (BLASTCLUST; minimum 70% length coverage, minimum 30% identity) was used to show that 3,589 (50%) of the predicted CDSs cluster into multigene families, which are likely to have arisen by gene duplication during evolution. The distribution and numbers of genes in these families presumably contribute to the survival of *S. erythraea*



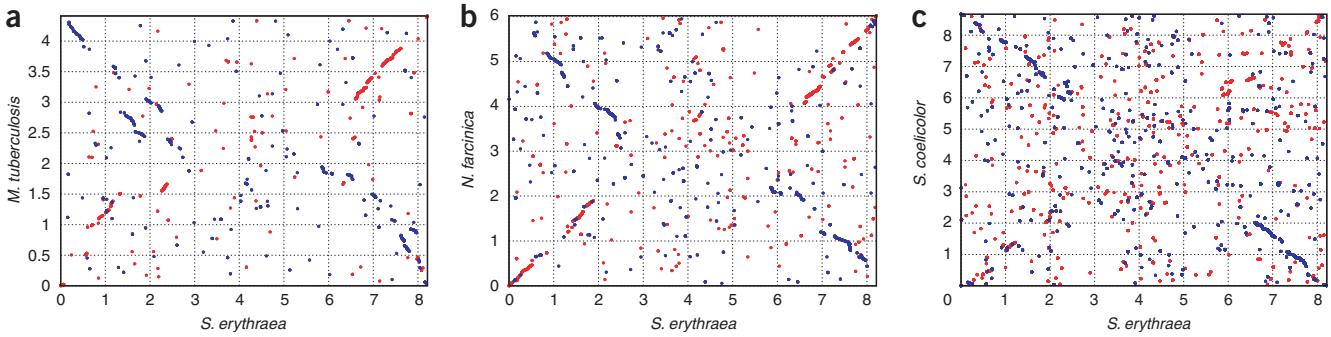


Figure 2 Whole genome comparisons of *S. erythraea*. (**a**) *M. tuberculosis*. (**b**) *N. farcinica*. (**c**) *S. coelicolor*. For each genome, DnaA is located at position 0. Dots represent a reciprocal best match (by BLAST comparison[49]) between orthologs (matches on the same strand in red and on the opposite strand in blue).

in the highly competitive, highly changeable soil environment (**Supplementary Table 1**). There is a rich array of genes potentially involved in defense or stress responses of various kinds. Intriguingly, in view of the evidence for potential acquisition of genes by horizontal gene transfer, *S. erythraea* has 20 restriction endonucleases and eight site-specific methyltransferases. Although it is not possible to specify their recognition sequences, together these restriction enzymes might have been expected to provide a formidable barrier to incoming DNA. Also present are the linked genes *pglY* (SACE_5129) and *pglZ* (SACE_5128), which mediate immunity to infection by bacteriophage ϕC31. In addition to the recombinational repair genes *recA* (SACE_1736) and *recB* (SACE_0087, SACE_1056, SACE_2242, SACE_6256), there is also *recF* (SACE_0005) and a GT-mismatch repair endonuclease (SACE_1556) with no counterpart in *S. avermitilis* or *S. coelicolor*. A total of 30 genes have products likely to be involved in ensuring or preserving correct protein folding, including two copies each of the three chaperone proteins *groEL* (SACE_0527, SACE_0543), *groES* (SACE_0927, SACE_7319) and *dnaJ* (SACE_1480, SACE_7208). Also likely to be involved in the stress response of *S. erythraea* are 22 genes related to *uspA* ('universal stress protein' of *Escherichia coli*) and genes encoding numerous cold shock proteins. A total of 1,118 genes (15.5%) are involved in regulation. As in *S. avermitilis* and *S. coelicolor*, an unusually large number of these (38) encode alternative sigma factors for the RNA polymerase, allowing for programmed transcription of particular sets of genes. A host of genes encoding other transcription factors are present, including 101 TetR-like, 34 GntR-like and 48 LysR-like regulatory proteins (**Supplementary Table 2** online). The response to changing environmental conditions and availability of nutrients is mediated by at least 42 sensor kinases, and 113 two-component response regulators. There are 40 genes encoding serine/threonine protein kinases and numerous and diverse eukaryotic-like protein phosphatases. A total of 658 genes (8.9%) appear to be involved in transport into or out of the cell, encoding large numbers of proteins acting as permeases, ion- or sugar-binding transporters, or ATP-driven transmembrane pumps. A wide range of degradative enzymes, including seven chitinases and multiple proteinases and glucanases, is predicted to be secreted from the cell, and presumably these play a key role in breaking down the heterogenous alternative food sources in soil. There are cobalamin-dependent versions of methionine synthase (SACE_3898) and ribonucleotide reductase (SACE_1764) as well as cobalamin-independent enzymes catalyzing the same reactions (SACE_4744 and SACE_1282, SACE_1283, respectively). Although *S. erythraea* is considered an obligate aerobe, we found two complete clusters of genes for nitrate reductase, indicating that alternative electron acceptors might be available under conditions where oxygen levels are low. Detoxification is a key function for soil bacteria, and *S. erythraea* like the previously sequenced streptomycetes has numerous genes for transporters mediating resistance to various heavy metals, as well as a substantial cohort (36) of cytochrome P450 enzymes. Some of these have known roles in specific hydroxylation steps during biosynthesis of erythromycin and other secondary metabolites[21–23], but others are predicted to play a role in selective oxidation and detoxification of organic materials.

Genes for antibiotic resistance

The genome of *S. erythraea* helps to explain the bacterium's intrinsic resistance to a wide range of antibiotics, because it encodes numerous enzymes predicted to inactivate common antibiotic classes. There are 17 β-lactamase genes present and two macrolide esterases. One of these esterases (SACE_0712) lies within the previously sequenced biosynthetic gene cluster for erythromycin and is inactivated by transposon insertion, but the second (SACE_1765) aligns well with authentic erythromycin esterases from erythromycin-resistant bacteria. It may have a hitherto-overlooked role in the regulation of erythromycin biosynthesis. Genes are present for efflux proteins for chloramphenicol (SACE_0228), daunorubicin (SACE_0206, SACE_0207), lincomycin, camphor (SACE_7237, SACE_7239), fosmidmycin (SACE_3971), bicyclomycin (SACE_2577, SACE_3939, SACE_6077, SACE_7323), tetracycline (SACE_1156, SACE_4211), vancomycin and related glycopeptides (SACE_7320, SACE_2593, SACE_2926), as well as genes for streptomycin phosphotransferase (SACE_5997), spectinomycin phosphotransferase (SACE_4273), aminoglycoside N-3′-acetyltransferase (SACE_3603, SACE_3604), an aminonucleoside phosphotransferase (SACE_1856) and two putative macrolide glycosyltransferases (SACE_1884, SACE_3599). At least 21 CDSs appear to encode dioxygenases related to the bleomycin-resistance protein, and, in addition to the known *ermE* rRNA methyltransferase gene (SACE_0733), 11 further ribosome-modifying rRNA methyltransferases appear to be present. In *S. avermitilis*, a second version of tryptophanyl-tRNA synthetase[2] may be resistant to the antibiotic indolmycin, which targets the usual version of this enzyme. Experiments are needed to determine whether in *S. erythraea* the instances of duplicated aminoacyl-tRNA synthetase genes (three for cysteine, two each for lysine, threonine and tryptophan) have any such significance.

Potential for production of secondary metabolites

S. erythraea is best known as the organism used for industrial-scale production of the macrolide polyketide erythromycin A. The gene

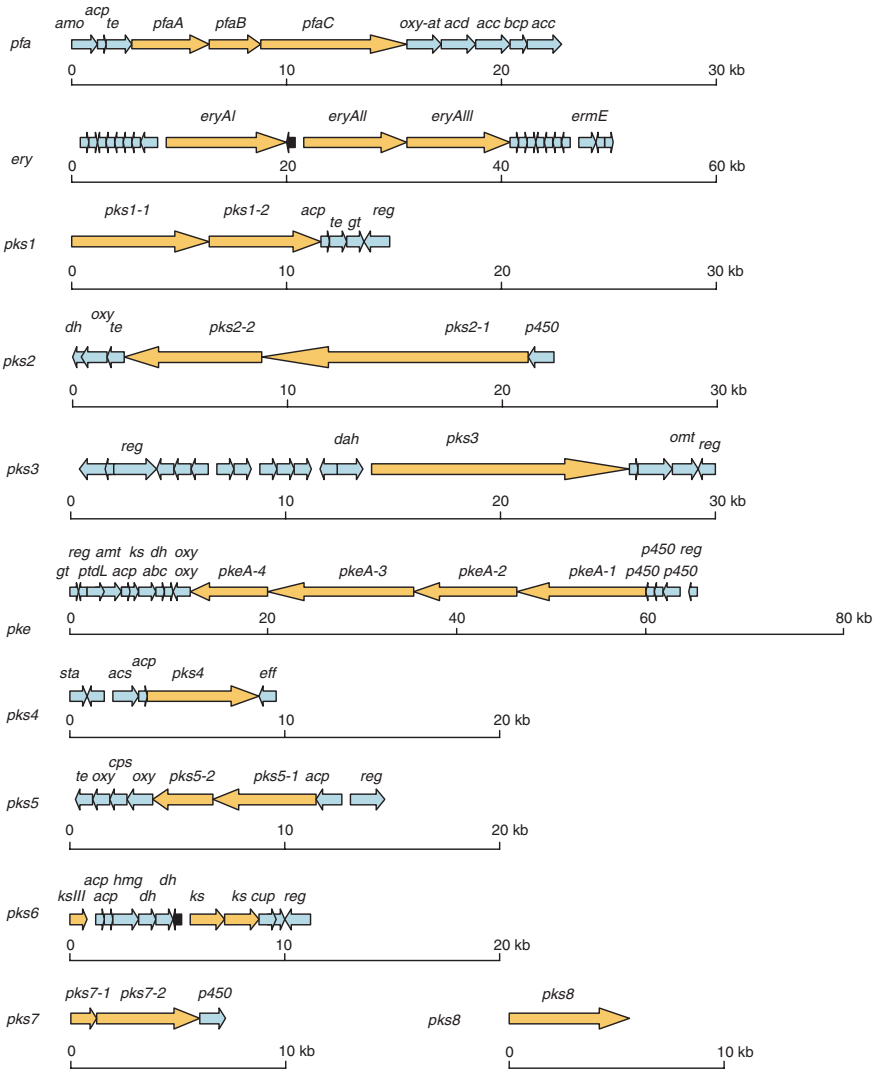


Figure 3 Gene clusters for polyketide biosynthesis. *abc*, ABC transporter; *acc*, acyl-CoA carboxylase; *acd*, acyl-CoA dehydrogenase; *acp*, acyl carrier protein; *acs*, acyl-CoA synthetase-like gene; *amo*, amine oxidase, *amt*, aminotransferase; *at*, acyltransferase; *bcp*, biotin carboxy carrier protein; *cbs*, carbamoyltransferase; *cup*, cupin-like protein; *dah*, DAHP synthase; *dh*, dehydratase; *eff*, efflux gene; *ermE*, erythromycin-resistance gene rRNA methyltransferase; *ery*, erythromycin; *gt*, glycosyltransferase; hmg, HMG-CoA synthase; *ks*, ketosynthase; *ksIII*, FabH-like protein; *omt*, *O*-methyltransferase; *oxy*, oxidoreductase; *p450*, cytochrome P450; *pfa*, polyunsaturated fatty acid synthase; *pke*, octaketide synthase; *sta*, StaD-like protein *te*, thioesterase.

clusters for erythromycin (*ery*)[121], and for a second modular polyketide synthase (PKS) of unknown function (*pke*)[22], have been previously analyzed, as has the gene (SACE_1243, *rppA*) for a type III PKS, which generates the reddish pigment typical of *S. erythraea*[23]. The genomic sequencing has revealed a further 22 clusters for the biosynthesis of polyketides, terpenes and nonribosomally synthesized peptides. The distribution of these clusters is not uniform around the chromosome: only four (including that for erythromycin) are in the 'core' region that contains most of the essential genes (**Fig. 1**), and one of these four clusters (*nrps1*) is inactivated by frameshift mutations. Twenty-one of the clusters are outside this region. Of the uncharacterized PKS gene clusters (**Fig. 3** and **Supplementary Table 3** online), one cluster (*pfa*) appears to govern the biosynthesis of polyunsaturated fatty acids such as eicosapentaenoic acid. Most of the others are modular, and between them are expected to generate specific polyketides in the range of 2–9 polyketide units long (for details of predicted domains, see **Supplementary Table 4** online). Two (SACE_5308, *pks7*; and SACE_5532, *pks8*) would encode multifunctional single-module PKS enzymes apparently related to the iterative polyketide synthases involved in enediyne or methylsalicylic acid synthesis. None of the hypothetical products of any of these PKSs, or of the *pke* PKS, have previously been detected, although extensive searches were made using 50 different solid and liquid media[22]. Strikingly, there are no type II PKS genes encoding the biosynthesis of aromatic polyketides, which are such a characteristic component of natural product biosynthesis in typical streptomycetes such as *S. coelicolor* and *S. avermitilis*, where they routinely govern the synthesis of antibiotics and spore pigments. Their absence in *S. erythraea* is therefore surprising. The *S. erythraea* genome also houses a number of gene clusters encoding nonribosomal peptide synthetases (NRPSs) (**Fig. 4** and **Supplementary Table 3**). These modular multienzyme systems are widespread in both bacteria and fungi and their products include (often cyclic) peptide antibiotics, immunosuppressants like cyclosporin, penicillins and iron-scavenging compounds (siderophores). Although iron is an abundant element in soil, owing to its poor bioavailability, a particularly large number of genes in soil bacteria appear to encode proteins involved in the acquisition and uptake of iron. Many streptomycetes synthesize nonpeptidic hydroxamate siderophores such as desferrioxamine E, synthesized from lysine and ornithine, but the genes required for this pathway[24] are not present in *S. erythraea*. Comparison of the NRPS sequences with those of authentic NRPSs allows the probable enzyme complement of each multienzyme to be deduced and tentative deductions to be drawn as to the likely structure of the peptide product (**Supplementary Table 5** online). Our predictions are based on a structure-based "specificity code"[25], which, together with later refinements[26,27], provides useful clues to the nature of the amino acid introduced at each stage. This will certainly be useful in guiding genome mining for the natural products of the *nrps* genes of *S. erythraea*, but it is important to stress that there are now multiple examples where the compounds isolated are different from those predicted by such methods (see, for example, the recent substantial correction of the structure of the siderophore coelichelin from *S. coelicolor*[28]). Of the *S. erythraea* NRPS-containing gene clusters, *nrps3* (SACE_2691-2703) and *nrps5* (SACE_3028-3039) may govern siderophore production, as both contain several genes whose predicted protein products are similar to proteins essential for iron-siderophore recognition and transport. *S. erythraea* produces at least one hydroxamate siderophore[29]. Intriguingly, *S. erythraea* has a complete set

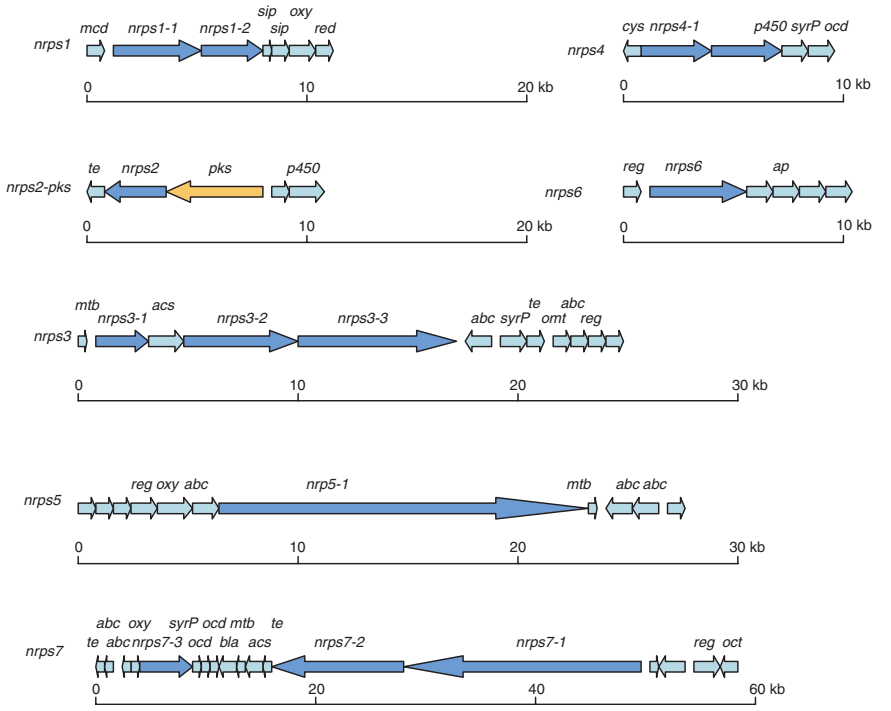


Figure 4 *S. erythraea* gene clusters for nonribosomal peptide synthetases. *abc*, ABC transporter; *acs*, acyl-CoA synthetase-like protein; *bla*, β-lactamase; *cys*, cysteine-synthase-like protein; *mcd*, malonyl-CoA decarboxylase; *mtb*, mtbH-like protein; *oct*, ornithine carbamoyltransferase; *ocd*, ornithine/lysine cyclodeaminase; *oxy*, oxidoreductase, *red*, reductase; *sip*, siderophore interacting protein; *syrP*, SyrP-like protein; *reg*, regulatory protein; *te*, thioesterase.

of genes resembling *mxc* genes of the myxobacterium *Stigmatella aurantiaca*[30] required for the synthesis from chorismic acid of 2,3-dihydroxybenzoic acyl-O-AMP, a key precursor of the catechol-type siderophore myxochelin (SACE_3854, mxcD; SACE_3852, MxcF; SACE_3855, mxcC; SACE_3853, mxcE). In *S. erythraea*, the C-5 precursors for terpenoid biosynthesis are apparently generated by the methylerythritol phosphate pathway, for which all the genes are present. Terpenoid metabolites play various roles in bacteria, providing for example quinone components of the electron transport chain, modified tRNA species, and carotenoid pigments for UV protection. Of the six terpene synthase genes present (**Supplementary Table 3** online), three (*tpc1*, *tpc3* and *tpc5*) show substantial similarity to terpene cyclases in other organisms that are known to produce geosmin, a sesquiterpene that provides soil with its characteristic smell. Similarly, the *hop* cluster is very similar to clusters in *S. avermitilis* (SAV1650-1654) and *S. coelicolor* (SCO6760-6764) that are thought to direct the production of hopanoids[31]. These compounds are proposed to reduce desiccation stress in aerial mycelium. Scattered elsewhere in the genome there are additional genes potentially involved in secondary metabolite–producing pathways. For example, there is a partial set of carotenoid biosynthetic genes (SACE_3271, SACE_3272, SACE_1713, SACE_3269, SACE_3539). The presence of a tryptophan halogenase (SACE_4919) and of a second halogenase (SACE_4927) may signal the production of halogen-containing metabolites in this strain. The CDSs SACE_4230-4233 closely resemble the *ramCSAB* genes of *S. coelicolor,* which are involved in production of the lantibiotic-like peptide SapB that acts as a morphogen in aiding the production of aerial hyphae[32]. Two other possible lantibiotic synthetase genes are encoded by SACE_4389 and SACE_4025.

Genes contributing to erythromycin production

Although most polyketide antibiotic gene clusters contain one or more regulatory genes, their absence from the *ery* biosynthetic gene cluster has hampered efforts to enhance erythromycin production other than by medium manipulation, random mutagenesis and selection. The availability of the genome sequence will allow global approaches to defining the mechanism by which erythromycin production is controlled in *S. erythraea*, in understanding both classical repression by certain sources of carbon, nitrogen and phosphorus and the role of pathway-specific regulators. Meanwhile, there is already evidence, for the nonfilamentous *Aeromicrobium erythreum*, that increasing the flux through feeder metabolic pathways (**Supplementary Fig. 2** online) strongly influences the erythromycin yield[33]. Production of erythromycin requires propionyl-CoA to provide a starter unit, and (2*S*)-methylmalonyl-CoA to provide extender units, for the polyketide chain of the antibiotic. Previous attempts to define the proximal pathways that furnish these building blocks have given inconclusive results[34], and analysis of the genome sequence now suggests some reasons for this. For example, biotin-dependent carboxylation of propionyl-CoA is an established route to (2*S*)-methylmalonyl-CoA (**Fig. 3**), and in *S. erythraea* it appears that there are at least five genetic loci, displaying remarkably diverse protein architectures, that might code for an enzyme catalyzing this reaction (**Supplementary Fig. 3** online)[35,36]. Further work will be required to deconvolute the contributions made by these gene sets to erythromycin biosynthesis. A second proposed route to (2*S*)-methylmalonyl-CoA proceeds by the rearrangement of succinyl-CoA catalyzed by adenosylcobalamin-dependent methylmalonyl-CoA mutase, but this yields the (2*R*)-isomer of methylmalonyl-CoA, not the (2S)-isomer. Our analysis reveals the presence of a gene encoding an authentic methylmalonyl-CoA epimerase (SACE_6238) that would interconvert (2*R*)- and (2S)-isomers[37]. There is a counterpart of this gene in each of *S. avermitilis* (SAV2857) and *S. coelicolor* (SCO5398). There is a single cluster of the adenosylcobalamin-dependent methylmalonyl-CoA mutase genes (SACE_5638-5640). *S. erythraea*, unlike many streptomycetes, has no homolog of crotonyl-CoA reductase or of adenosylcobalamin-dependent isobutyryl-CoA mutase, explaining its inability to furnish butyrate units for polyketide biosynthesis[38,39]; it also has no homolog of *meaA* in *S. coelicolor*, which has been implicated in provision of methylmalonyl-CoA from acetoacetyl-CoA[40]. The availability of the complete genome sequence now provides the basis for systematic approaches to identify and manipulate such feeder pathways with the aim of increasing polyketide production. It also provides the starting point for an integrated genome-scale analysis of *S. erythraea* metabolism, as provided recently for *S. coelicolor*[41].

DISCUSSION

As with other soil bacteria that have been studied, *S. erythraea* has a remarkable potential for the production of secondary metabolites of various kinds, many of them with antibiotic activity. Although they

are not essential for growth under laboratory conditions, it is clear that the ability to produce such compounds is a trait that confers a substantial advantage, which outweighs the considerable energetic cost of maintaining this arsenal. It has been pointed out, for example, that the presence of numerous modular polyketide synthase gene sets in the same cell might during evolution enhance the likelihood of recombination between them, leading to new and potentially fruitful biosynthetic pathways[42]. Further work will be required to establish the chemical structures and biological activities of the products of many of the pathways that have been uncovered, but our results emphasize that even in well-studied bacteria there is considerable untapped potential for producing diverse chemical compounds. Most of the predicted products appear to be unique to this strain. The sequencing of the *S. erythraea* genome has provided evidence of considerable divergence from the streptomycetes in gene organization and function, confirming previous taxonomic and biochemical insights. It will now be possible to analyze directly the genetic differences between the wild-type strain and the strains derived from it by mutation and selection that are used for industrial production of erythromycin A at high titer. Although such comparisons are unlikely to be immediately informative, because of the presence of numerous neutral or even deleterious mutations, they should give fresh impetus to the search for the causes of antibiotic overproduction in such strains.

METHODS

Sequencing and assembly. Whole-genome shotgun sequencing of the *S. erythraea* NRRL2338 genome was done using frequently cutting restriction enzymes and 2- to 10-kbp fragments were cloned into plasmid vectors. Cosmids (32–46 kbp inserts) were also generated from genomic DNA and end-sequenced to provide additional read-pair information; to provide increased coverage of selected regions; and to fill gaps. Remaining gaps and ambiguities were closed using PCR products from specifically designed oligonucleotide primers. Sequence assembly was done using the Phrap assembler[43] and editing was done using consed version 14 (ref. 44). Repeats were resolved by doing a mini-assembly for the individual sections of the genome, and the resulting consensus was integrated into the main genome assembly. The data overall were in good agreement (with one exception—see the text and **Supplementary Fig. 1**) with published *Ase*I and *Dra*I restriction maps[15]. The final assembly contained 72,537 sequence reads, including 51,626 reads from the whole genome shotgun, 5,069 from the cosmid clone shotgun, and 6,470 from cosmid ends, and 8,808 from cosmid primer walking, 262 from specific PCR products and 302 from scaffolding reads to resolve highly repetitive IS regions. Together this provided 7.1-fold coverage with an estimated error rate of <1 per 100,000 bases of the consensus sequence.

Genome analysis and annotation. CDSs were predicted and annotated using the program fgenesB[45] (http://www.softberry.com/), trained *ab initio*, and manually curated using Artemis (version 8)[46] and a set of in-house PERL scripts. CDS annotation was based on hits to KEGG and Uniprot databases, and sorted according to the COG[47] functional database. tRNA genes were predicted with tRNAscan[48]. The BLAST[49] program (NCBI version 2.2.15) was used for database searches and BLASTclust (part of the NCBI BLAST distribution) was used to generate clusters of protein families. Interproscan[50] was used to confirm domain assignments.

Accession codes. EMBL/GenBank: The genome sequence has been deposited in the database under accession number AM420293.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank Torsten Schwecke, Mohammed Sebaihia, Neil Whitehead and Melanie Wall for help with initial sequence analysis. We gratefully acknowledge support of part of this work by a project grant from the Exploiting Genomics Initiative of the UK Biotechnology and Biological Sciences Research Council.

AUTHOR CONTRIBUTIONS

P.F.L., J.B.L. and M.O. generated ideas and coordinated the project; M.O. and T.M. prepared genomic DNA, and constructed shotgun and cosmid libraries; T.M. and N.S. did shotgun template sequencing; S.D. maintained strains; M.O. and M.S. set up data collection, assembled shotgun reads, generated finishing reads and did programming for project automation and bioinformatics; M.S. did the assembly finishing and resolved repetitive element sequences; M.O. and M.S. completed gene detection; M.O., M.S. and P.F.L. annotated genes; S.F.H. gave input on annotation of secondary metabolic pathways; P.F.L., M.O. and M.S. cowrote the paper.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/naturebiotechnology.

Published online at http://www.nature.com/naturebiotechnology
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Labeda, D.P. Transfer of the type strain of *Streptomyces erythreus* (Waksman 1923) Waksman and Henrici 1948 to the genus *Saccharopolyspora* Lacey and Goodfellow 1975 as *Saccharopolyspora erythraea* sp. nov. and designation of a neotype strain for *Streptomyces erythraeus*. *Int. J. Syst. Bacteriol.* **37**, 19–22 (1987).
2. Ikeda, H. *et al.* Complete genome analysis and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat. Biotechnol.* **21**, 526–531 (2003).
3. Bentley, S.D. *et al.* Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147 (2002).
4. Staunton, J. & Weissman, K.J. Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **18**, 380–416 (2001).
5. Weissman, K.J. & Leadlay, P.F. Combinatorial biosynthesis of reduced polyketides. *Nat. Rev. Microbiol.* **3**, 925–936 (2005).
6. Hessler, P.E., Larsen, P.E., Constantinou, A.I., Schram, K.H. & Weber, J.M. Isolation of isoflavones from soy-based fermentations of the erythromycin-producing bacterium *Saccharopolyspora erythraea*. *Appl. Microbiol. Biotechnol.* **47**, 398–404 (1997).
7. Cole, S.T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
8. Cerdeno-Tarraga, A.M. *et al.* The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. *Nucleic Acids Res.* **31**, 6516–6523 (2003).
9. Ishikawa, J. *et al.* The complete genomic sequence of *Nocardia farcinica* IFM 10152. *Proc. Natl. Acad. Sci. USA* **101**, 14925–14930 (2004).
10. Normand, P. *et al.* Genome characterization of facultatively symbiotic Frankia sp. strains reflect host range and host plant biogeography. *Genome Res.* **17**, 7–15 (2007).
11. Jakimowicz, D. *et al.* Structural elements of the *Streptomyces oriC* region and their interactions with the DnaA protein. *Microbiology* **144**, 1281–1290 (1998).
12. Merkl, R. SIGI: score-based identification of genomic islands. *BMC Bioinformatics* **5**, 22 (2004).
13. Brown, D.P., Idler, K.B., Backer, D.M., Donadio, S. & Katz, L. Characterization of the genes and attachment sites for site-specific integration of plasmid pSE101 in *Saccharopolyspora erythraea* and *Streptomyces lividans*. *Mol. Gen. Genet.* **242**, 185–193 (1994).
14. Brown, D.P., Idler, K.B. & Katz, L. Characterization of the genetic elements required for site-specific integration of plasmid pSE211 in *Saccharopolyspora erythraea*. *J. Bacteriol.* **172**, 1877–1888 (1990).
15. Reeves, A.R., Post, D.A. & Van den Boom, T.J. Physical-genetic map of the erythromycin-producing organism *Saccharopolyspora erythraea*. *Microbiology* **144**, 2151–2159 (1998).
16. Lin, Y.-S., Kieser, H.M., Hopwood, D.A. & Chen, C. The chromosomal DNA of *Streptomyces lividans* is linear. *Mol. Microbiol.* **10**, 923–933 (1993).
17. Leblond, P., Redenbach, M. & Cullum, J. Physical map of the *Streptomyces lividans* 66 genome and comparison with that of the related strain *Streptomyces coelicolor* A3(2). *J. Bacteriol.* **175**, 3422–3429 (1993).
18. Bao, K. & Cohen, S.N. Recruitment of terminal protein to the ends of *Streptomyces* linear plasmids and chromosomes by a novel telomere-binding protein essential for linear DNA replication. *Genes Dev.* **17**, 774–785 (2003).
19. Volff, J.-N. & Altenbuchner, J. A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol. Lett.* **186**, 143–150 (2000).
20. Chen, C.W., Huang, C.-H., Tsai, H.-H. & Kirby, R. Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. *Trends Genet.* **18**, 522–529 (2002).
21. Staunton, J. & Wilkinson, B. Biosynthesis of erythromycin and rapamycin. *Chem. Rev.* **97**, 2611–2630 (1997).
22. Boakes, S. *et al.* A new modular polyketide synthase in the erythromycin producer *Saccharopolyspora erythraea*. *J. Mol. Microbiol. Technol.* **8**, 73–80 (2004).
23. Cortés, J. *et al.* Identification and cloning of a type III polyketide synthase required for diffusible pigment biosynthesis in *Saccharopolyspora erythraea*. *Mol. Microbiol.* **44**, 1213–1224 (2002).
24. Barona-Gomez, F., Wong, U., Giannakopulos, A.E., Derrick, P.J. & Challis, G.L. Identification of a cluster of genes that directs desferrioxamine biosynthesis in *Streptomyces coelicolor* M145. *J. Am. Chem. Soc.* **126**, 16282–16283 (2004).

25. Stachelhaus, T., Mootz, H. & Marahiel, M. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493–505 (1999).
26. 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).
27. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D.H. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.* **33**, 5799–5808 (2005).
28. Lautru, S., Deeth, R.J., Bailey, L.M. & Challis, G.L. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nat. Chem. Biol.* **1**, 265–269 (2005).
29. Oliveira, P.H., Batagov, A., Ward, J., Baganz, F. & Krabben, P. Identification of erythrobactin, a hydroxamate-type siderophore produced by *Saccharopolyspora erythraea*. *Lett. Appl. Microbiol.* **42**, 375–380 (2006).
30. Gaitatzis, N., Kunze, B. & Mueller, R. Novel insights into siderophore formation in myxobacteria. *ChemBioChem* **6**, 365–374 (2005).
31. 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).
32. Kodani, S. *et al.* The SapB morphogen is a lantibiotic-like peptide derived from the product of the developmental gene *ramS* in *Streptomyces coelicolor*. *Proc. Natl. Acad. Sci. USA* **101**, 11448–11453 (2004).
33. Reeves, A.R., Cernota, W.H., Brikun, I.A., Wesley, R.A. & Weber, J.M. Engineering precursor flow for increased erythromycin production in *Aeromicrobium erythreum*. *Metab. Eng.* **6**, 300–312 (2004).
34. Donadio, S., Staver, M.J. & Katz, L. Erythromycin production in *Saccharopolyspora erythraea* does not require a functional propionyl-CoA carboxylase. *Mol. Microbiol.* **19**, 977–984 (1996).
35. Oh, T.-J., Daniel, J., Kim, H.-J., Sirakova, T.D. & Kolattukudy, P.E. Identification and characterisation of Rv3281 as a novel subunit of a biotin-dependent acyl-CoA carboxylase in *Mycobacterium tuberculosis* H37Rv. *J. Biol. Chem.* **281**, 3899–3908 (2006).
36. Diacovich, L. *et al.* Kinetic and structural analysis of a new group of acyl-CoA carboxylases found in *Streptomyces coelicolor* A3(2). *J. Biol. Chem.* **277**, 31228–31236 (2002).
37. Leadlay, P.F. Purification and characterisation of methylmalonyl-CoA epimerase from *Propionibacterium shermanii*. *Biochem. J.* **197**, 413–419 (1981).
38. Liu, H. & Reynolds, K.A. Precursor supply for polyketide biosynthesis: the role of crotonyl-CoA reductase. *Metab. Eng.* **3**, 40–48 (2001).
39. Stassi, D.L. *et al.* Ethyl-substituted erythromycin derivatives produced by directed metabolic engineering. *Proc. Natl. Acad. Sci. USA* **95**, 7305–7309 (1998).
40. Zhang, W. & Reynolds, K.A. MeaA, a putative coenzyme B12-dependent mutase, provides methylmalonyl-CoA for monensin biosynthesis in *Streptomyces cinnamonensis*. *J. Bacteriol.* **183**, 2071–2080 (2001).
41. Borodina, I., Krabben, P. & Nielsen, P. Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Res.* **15**, 820–829 (2005).
42. Jenke-Kodama, H., Borner, T. & Dittmann, E. Natural biocombinatorics in the polyketide synthase genes of the actinobacterium *Streptomyces avermitilis*. *PLoS Comput. Biol.[online]* **2**, e132 (2006)(10.1371/journal.pcbi.0020132).
43. Ewing, B., Hillier, L., Wendl, M. & Green, P. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
44. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
45. Lukashin, A.V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).
46. Rutherford, K. *et al.* Artemis: sequence visualisation and annotation. *Bioinformatics* **16**, 944–945 (2000).
47. Tatusov, R.L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22–28 (2001).
48. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
49. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
50. Zdobnov, E.M. & Apweiler, R. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).