BMC
Genomics

RESEARCH ARTICLE

Open Access

The complete genome sequence of the acarbose producer *Actinoplanes* sp. SE50/110

Patrick Schwientek[1,2], Rafael Szczepanowski[3], Christian Rückert[3], Jörn Kalinowski[3], Andreas Klein[4], Klaus Selber[5], Udo F Wehmeier[6], Jens Stoye[2] and Alfred Pühler[1,7*]

Abstract

Background: *Actinoplanes* sp. SE50/110 is known as the wild type producer of the alpha-glucosidase inhibitor acarbose, a potent drug used worldwide in the treatment of type-2 diabetes mellitus. As the incidence of diabetes is rapidly rising worldwide, an ever increasing demand for diabetes drugs, such as acarbose, needs to be anticipated. Consequently, derived *Actinoplanes* strains with increased acarbose yields are being used in large scale industrial batch fermentation since 1990 and were continuously optimized by conventional mutagenesis and screening experiments. This strategy reached its limits and is generally superseded by modern genetic engineering approaches. As a prerequisite for targeted genetic modifications, the complete genome sequence of the organism has to be known.

Results: Here, we present the complete genome sequence of *Actinoplanes* sp. SE50/110 [GenBank:CP003170], the first publicly available genome of the genus *Actinoplanes*, comprising various producers of pharmaceutically and economically important secondary metabolites. The genome features a high mean G + C content of 71.32% and consists of one circular chromosome with a size of 9,239,851 bp hosting 8,270 predicted protein coding sequences. Phylogenetic analysis of the core genome revealed a rather distant relation to other sequenced species of the family Micromonosporaceae whereas *Actinoplanes utahensis* was found to be the closest species based on 16S rRNA gene sequence comparison. Besides the already published acarbose biosynthetic gene cluster sequence, several new non-ribosomal peptide synthetase-, polyketide synthase- and hybrid-clusters were identified on the *Actinoplanes* genome. Another key feature of the genome represents the discovery of a functional actinomycete integrative and conjugative element.

Conclusions: The complete genome sequence of *Actinoplanes* sp. SE50/110 marks an important step towards the rational genetic optimization of the acarbose production. In this regard, the identified actinomycete integrative and conjugative element could play a central role by providing the basis for the development of a genetic transformation system for *Actinoplanes* sp. SE50/110 and other *Actinoplanes* spp. Furthermore, the identified non-ribosomal peptide synthetase- and polyketide synthase-clusters potentially encode new antibiotics and/or other bioactive compounds, which might be of pharmacologic interest.

Keywords: Genomics, Actinomycetes, Actinoplanes, Complete genome sequence, Acarbose, AICE

Background

Actinoplanes spp. are Gram-positive aerobic bacteria growing in thin hyphae very similar to fungal mycelium [1]. Genus-specific are the formation of characteristic sporangia bearing motile spores as well as the rare cell wall components meso-2,6-diaminopimelic acid, L,L-2,6-

diaminopimelic acid and/or hydroxy-diaminopimelic acid and glycine [1-4]. Phylogenetically, the genus *Actinoplanes* is a member of the family Micromonosporaceae, order Actinomycetales belonging to the broad class of Actinobacteria, which feature G + C-rich genomes that are difficult to sequence [5,6].

Actinoplanes spp. are known for producing a variety of pharmaceutically relevant substances such as antibacterial [7-9], antifungal [10] and antineoplastic agents [11]. Other secondary metabolites were found to possess

* Correspondence: Puehler@CeBiTec.Uni-Bielefeld.de
[1]Senior research group in Genome Research of Industrial Microorganisms, Center for Biotechnology, Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany
Full list of author information is available at the end of the article

inhibitory effects on mammalian intestinal glycosidases, making them especially suitable for pharmaceutical applications [12-15]. In particular, the pseudotetrasaccharide acarbose, a potent α-glucosidase inhibitor, is used worldwide in the treatment of type-2 diabetes mellitus (non-insulin-dependent). As the prevalence of type-2 diabetes is rapidly rising worldwide [16] an ever increasing demand for acarbose and other diabetes drugs has to be anticipated.

Starting in 1990, the industrial production of acarbose is performed using improved derivatives of the wild-type strain *Actinoplanes* sp. SE50 (ATCC 31042; CBS 961.70) in a large-scale fermentation process [12,17]. Since that time, laborious conventional mutagenesis and screening experiments were conducted by the producing company Bayer AG in order to develop strains with increased acarbose yield. However, the conventional strategy, although very successful [18], seems to have reached its limits and is generally superseded by modern genetic engineering approaches [19]. As a prerequisite for targeted genetic modifications, the preferably complete genome sequence of the organism has to be known. Here, a natural variant representing a first overproducer of acarbose, *Actinoplanes* sp. SE50/110 (ATCC 31044; CBS 674.73), was selected for whole genome shotgun sequencing because of its publicity in the scientific literature [17,20-22] and its elevated, well measureable acarbose production of up to 1 g/l [23]. Most notably, the acarbose biosynthesis gene cluster has already been identified [17,21,24-26] and sequenced [GenBank: Y18523.4] in this strain. For most of the identified genes in the cluster, functional protein assignment has been accomplished [17,21,22,24-27], presenting a fairly complete picture of the acarbose biosynthesis pathway in *Actinoplanes* sp. SE50/110, reviewed in [17,28,29]. However, scarcely anything is known about the remaining genome sequence and its influence on acarbose production efficiency through e.g. nutrient uptake mechanisms or competitive secondary metabolite gene clusters.

We have previously reported on the obstacles of high-throughput next generation sequencing for the *Actinoplanes* sp. SE50/110 genome sequence [30], which was carried out at the Center for Biotechnology, Bielefeld University, Germany. Having extensive experience in sequencing microbial genomes with classical Sanger methods [31-33], 454 next generation pyrosequencing technology [34,35] and especially with high G + C content genomes [36-39], we were able to identify the causes for the initial *Actinoplanes* sp. SE50/110 sequencing run to result in an unusually high number of contiguous sequences (contigs). It was found that a large number of stable secondary structures containing high G + C contents were responsible for the inability of the emulsion PCR step of the 454 library preparation

protocol to amplify these regions. This led to their absence in the Genome Sequencer FLX output and ultimately to the missing sequences between the contigs established by the Newbler assembly software. Fortunately, these problems could be solved in a second (whole-genome shotgun) run by adding a trehalose-containing emulsion PCR additive, and by increasing the read length [30]. Based on this draft genome sequence, we now present the scaffolding strategy for the remaining contigs and report on the successful gap closure procedure that led to the complete finishing of the *Actinoplanes* sp. SE50/110 genome sequence. Furthermore, results from gene finding and genome annotation are presented, revealing compelling insights into the metabolic potential of the acarbose producer.
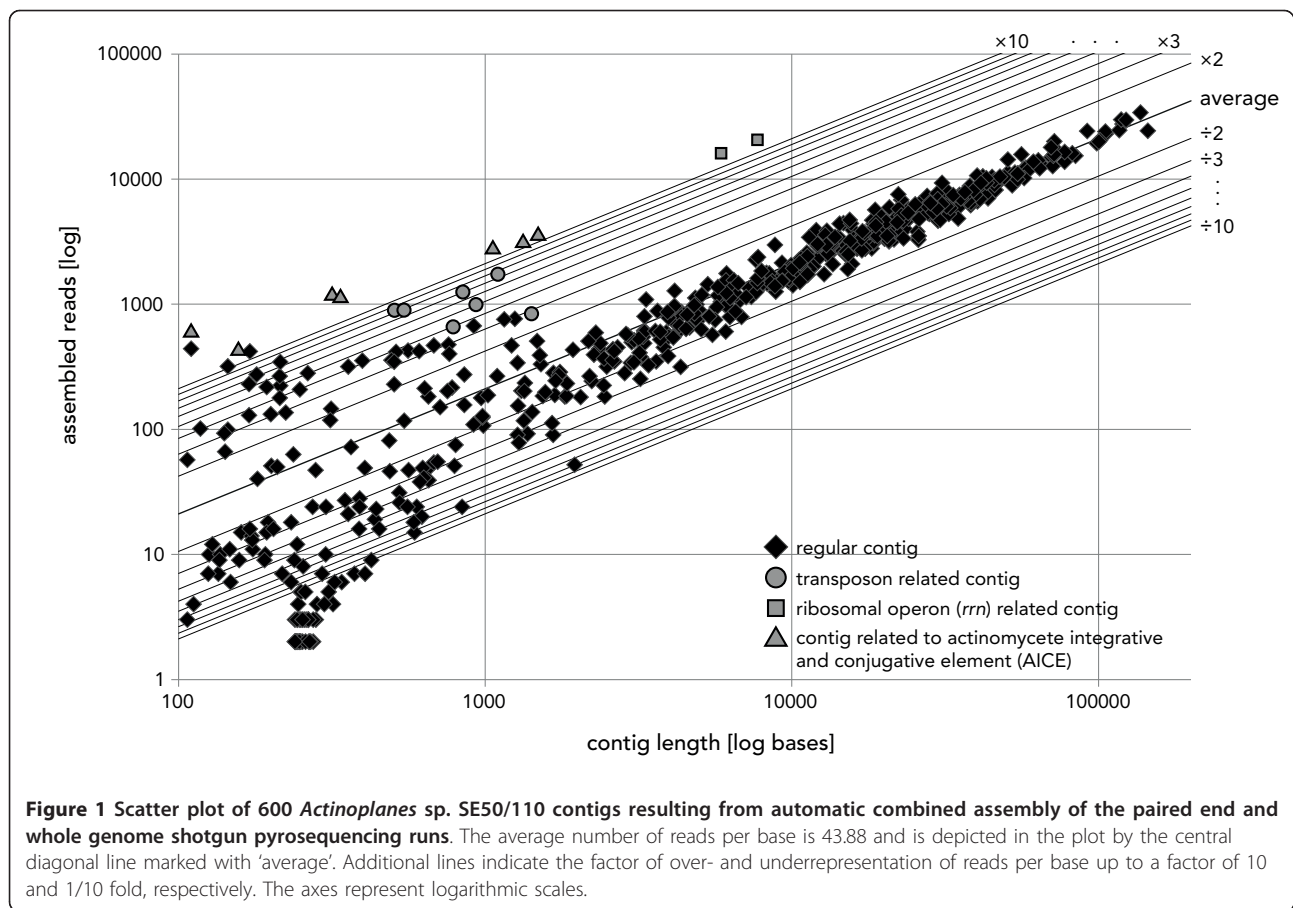
Results and discussion

High-throughput pyrosequencing and annotation of the *Actinoplanes* sp. SE50/110 genome

The complete genome determination of the arcabose producing wild-type strain *Actinoplanes* sp. SE50/110 was accomplished by combining the sequencing data generated by paired end (PE) and whole-genome shotgun pyrosequencing strategies [30]. Utilizing the Newbler software (454 Life Sciences), the combined assembly of both runs resulted in a draft genome comprising 600 contigs (476 contigs ≥ 500 bases) and 9,153,529 bases assembled from 1,968,468 reads.

The contigs of the draft genome were analyzed for over- or underrepresentation in read coverage by means of a scatter plot to identify repeats, putative plasmids or contaminations (Figure 1). While most of the large contigs show an average coverage with reads, several contigs were found to be clearly overrepresented and are of special interest as discussed later. However, the majority of the unusual high and low covered contigs are of very short length, representing short repetitive elements (overrepresented) and contigs containing only few reads of low quality (underrepresented). These findings indicate clean sequencing runs without contaminations.

Based on PE information, 8 scaffolds were constructed using 421 contigs with an estimated total length of 9,189,316 bases (Figure 2A). These PE scaffolds were used to successfully map terminal insert sequences of 609 fosmid clones randomly selected from a previously constructed fosmid library (insert size of ~37 kb). The mapping results validated the PE scaffold assemblies and allowed the further assembly of the original 8 paired end scaffolds into 3 PE/fosmid (PE/FO) scaffolds due to bridging fosmid reads (Figure 2B).

Gap closure between the remaining contigs was carried out by fosmid walking (746 reads) and genomic PCR technology (236 reads) in cases were no fosmid was spanning the target region. Genomic PCR technology was also used

Figure 1 Scatter plot of 600 *Actinoplanes* sp. SE50/110 contigs resulting from automatic combined assembly of the paired end and whole genome shotgun pyrosequencing runs. The average number of reads per base is 43.88 and is depicted in the plot by the central diagonal line marked with 'average'. Additional lines indicate the factor of over- and underrepresentation of reads per base up to a factor of 10 and 1/10 fold, respectively. The axes represent logarithmic scales.

to determine the order and orientation of the remaining 3 PE/FO scaffolds. The finishing procedure was manually performed using the Consed software [40] and resulted in the final assembly of a complete single circular chromosome of 9,239,851 bp with an average G + C content of 71.36% (Figure 3). According to genome project standards [41], the finished *Actinoplanes* sp. SE50/110 genome meets the gold standard criteria for high quality next generation sequencing projects. The general properties of the finished genome are summarized in Table 1.
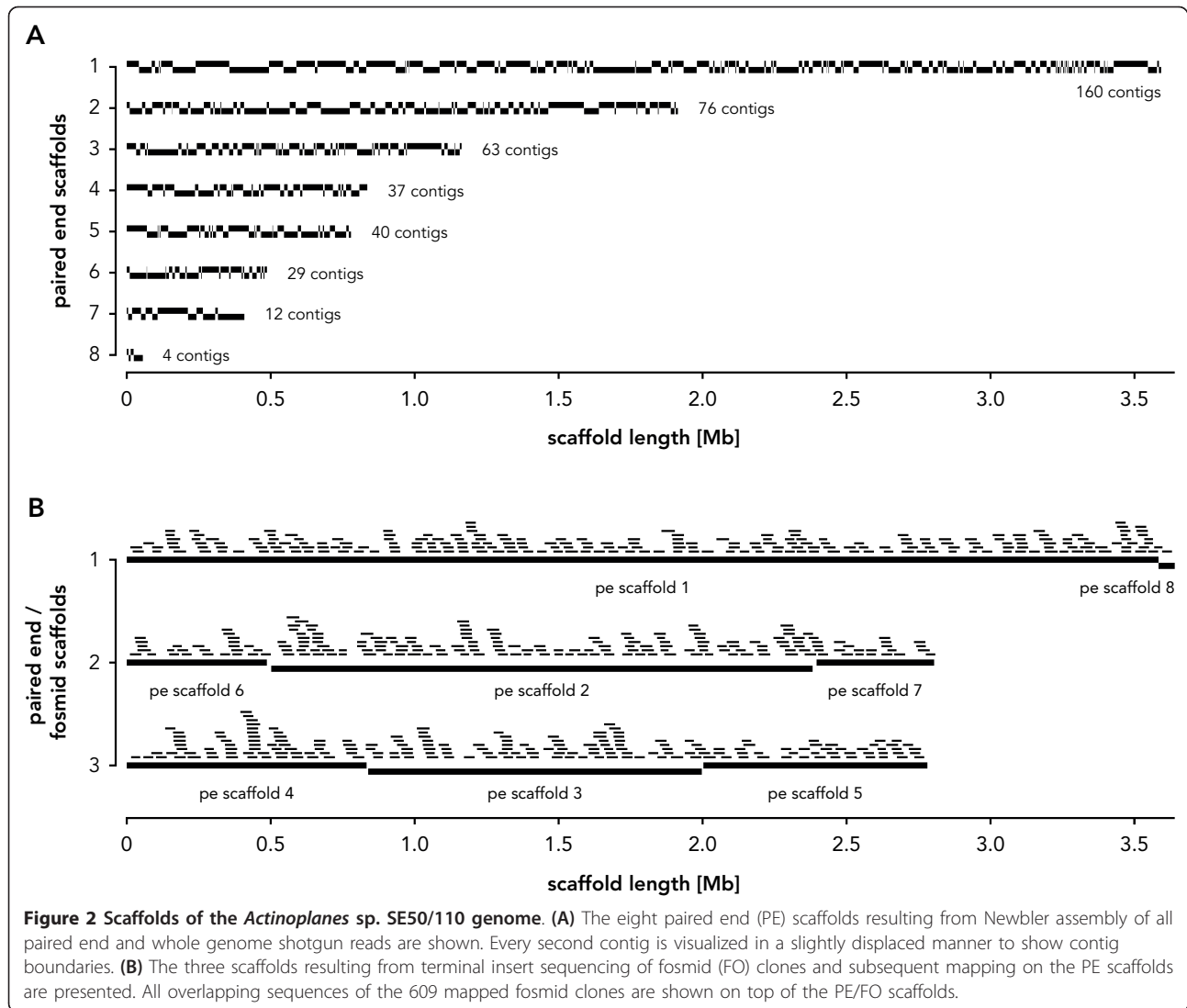
Utilizing the prokaryotic gene finders Prodigal [42] and Gismo [43] in conjunction with the GenDB annotation pipeline [44], a total of 8,270 protein-coding sequences (CDS) were determined on the *Actinoplanes* sp. SE50/110 genome (Figure 3). These include 4,999 genes (60.5%) with an associated functional COG category [45], 2,202 genes (26.6%) with a fully qualified EC-number [46] and 973 orphan genes (11.8%) with neither annotation nor any similar sequence in public databases using BLASTP search with an e-value cutoff of 0.1. In total, the amount of protein coding genes (coding density) covers 90.11% of the genome sequence with a significant difference of 4% in G + C content between non-coding (67.74%) and coding (71.78%) regions.

The complete annotated genome sequence was deposited at the National Center for Biotechnology Information (NCBI) [GenBank:CP003170].

General features of the *Actinoplanes* sp. SE50/110 genome

The origin of replication (*oriC*) was identified as a 1266 nt intergenic region between the two genes *dnaA* and *dnaN*, coding for the bacterial chromosome replication initiator protein and the β-sliding clamp of the DNA polymerase III, respectively. The *oriC* harbors 24 occurrences of the conserved DnaA box [TT(G/A)TCCACa], showing remarkable similarity to the *oriC* of *Streptomyces coelicolor* [47]. Almost directly opposite of the *oriC*, a putative *dif* site was found. Its 28 nt sequence 5'-CAGGTCGA-TAATGTATATTATGTCAACT-3' is in good accordance with actinobacterial *dif* sites and shows highest similarity (only 4 mismatches) to that of *Frankia alni* [48]. In addition to the identified *oriC* and *dif* sites, the calculated G/C skew [(G-C)/(G + C)] suggests two replichores composing the circular *Actinoplanes* sp. SE50/110 genome (Figure 3).

In accordance with previous findings [49], six ribosomal RNA (*rrn*) operons were identified on the genome in the typical 16S-23S-5S order along with 99 tRNA

Figure 2 Scaffolds of the *Actinoplanes* sp. SE50/110 genome. **(A)** The eight paired end (PE) scaffolds resulting from Newbler assembly of all paired end and whole genome shotgun reads are shown. Every second contig is visualized in a slightly displaced manner to show contig boundaries. **(B)** The three scaffolds resulting from terminal insert sequencing of fosmid (FO) clones and subsequent mapping on the PE scaffolds are presented. All overlapping sequences of the 609 mapped fosmid clones are shown on top of the PE/FO scaffolds.
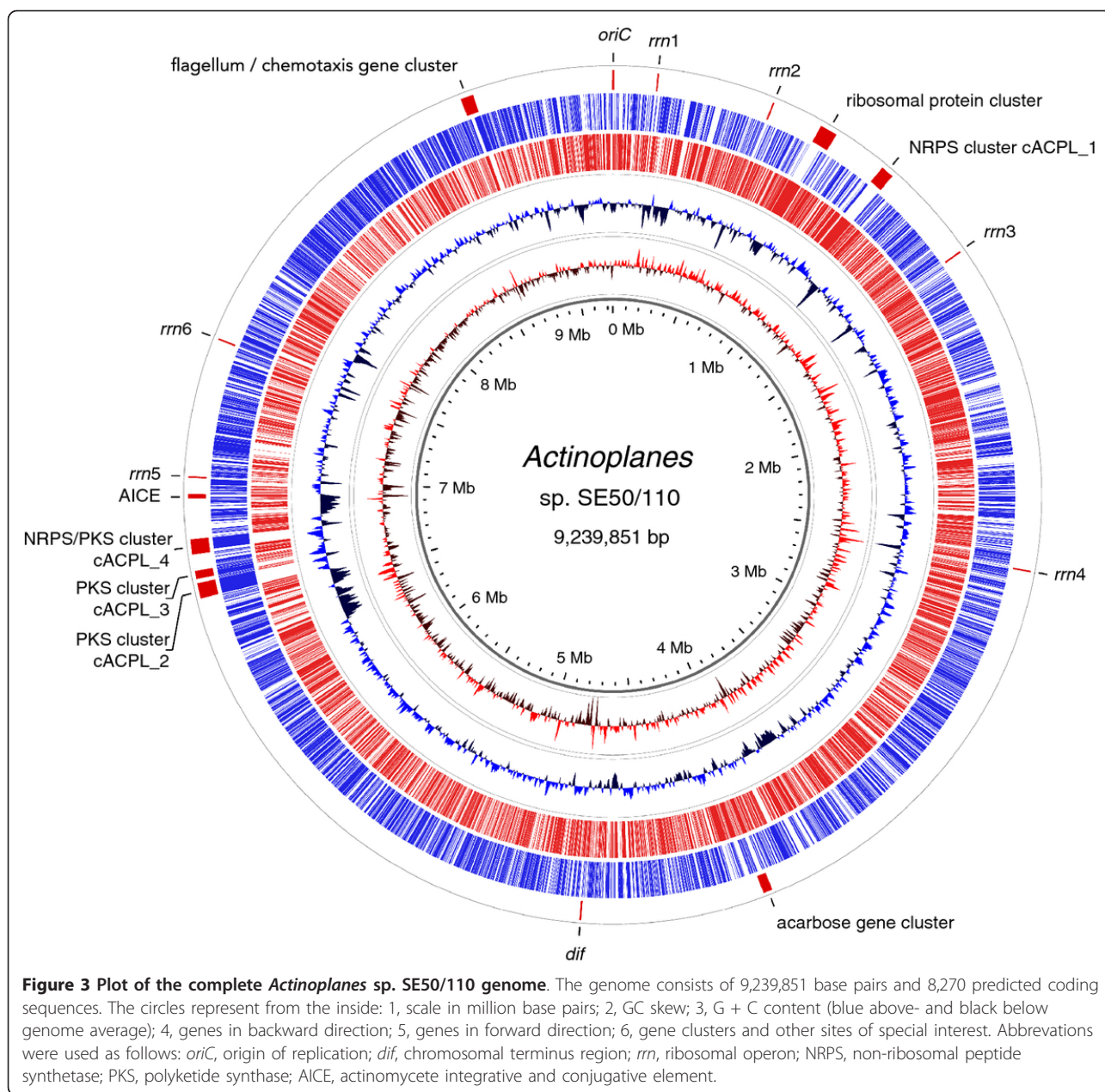
genes determined by the tRNAscan-SE software [50]. The six individual *rrn* operons were previously assembled into one operon ranging across seven contigs with a more than ten-fold overrepresentation (Figure 1). This overrepresentation might be explained by the operon's remarkably low G + C content of 57.20% in comparison to the genome average of 71.36%, which is typical for actinomycetes [49]. The low G + C content in this area may have introduced an amplification bias in favor of the *rrn* operon during the library preparation and thus, result in an overrepresentation of reads for this genomic region. To account for single nucleotide polymorphisms (SNPs) and variable regions between ribosomal genes, all six *rrn* operons were individually re-sequenced by fosmid walking. The *rrn* operons are located on the leading strands, four on the right and two on the left replichore. Interestingly, they reside in the upper half of the genome, together with a ~40 kb

gene cluster hosting more than 30 ribosomal proteins (Figure 3). Other overrepresented large contigs were identified as transposase genes or transposon related elements (Figure 1).

Approximately 500 kb upstream of the *oriC* site, a flagellum gene cluster was found. Its expression in spores is one of the characteristics discriminating the genus *Actinoplanes* from other related genera [1,4]. The cluster consists of ~50 genes spanning 45 kb. Besides flagellum associated proteins, the cluster also contains genes coding for chemotaxis related proteins.
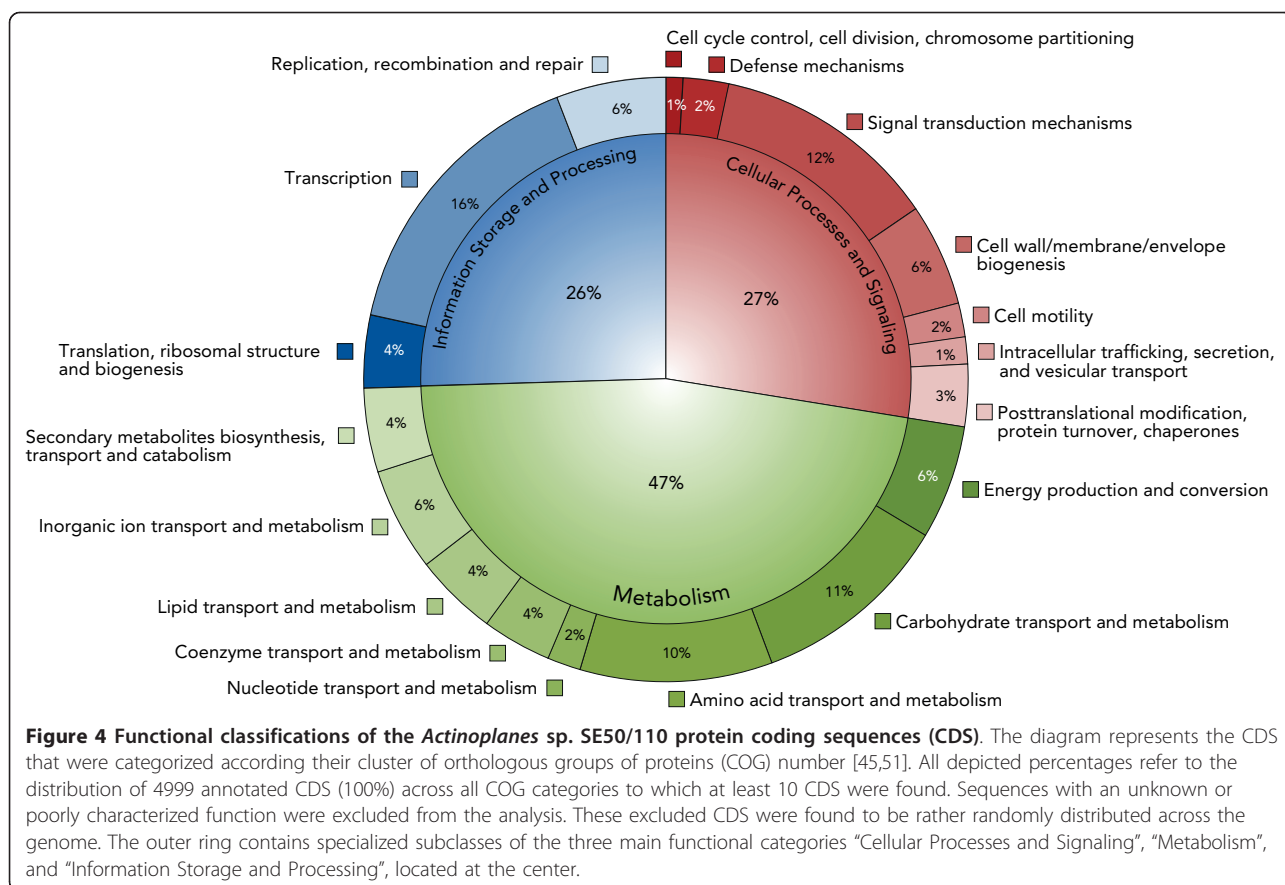
Bioinformatic classification of 4,999 CDS with an annotated COG-category revealed a strong emphasis (47%) on enzymes related to metabolism (Figure 4). In particular, *Actinoplanes* sp. SE50/110 features an emphasis on amino acid (10%) and carbohydrate metabolism (11%), which is in good accordance with the identification of at least 29 ABC-like carbon substrate

Figure 3 Plot of the complete *Actinoplanes* sp. SE50/110 genome. The genome consists of 9,239,851 base pairs and 8,270 predicted coding sequences. The circles represent from the inside: 1, scale in million base pairs; 2, GC skew; 3, G + C content (blue above- and black below genome average); 4, genes in backward direction; 5, genes in forward direction; 6, gene clusters and other sites of special interest. Abbreviations were used as follows: *oriC*, origin of replication; *dif*, chromosomal terminus region; *rrn*, ribosomal operon; NRPS, non-ribosomal peptide synthetase; PKS, polyketide synthase; AICE, actinomycete integrative and conjugative element.

Table 1 Features of the *Actinoplanes* sp. SE50/110 genome

Feature	Chromosome
Total size (bp)	9,239,851
G + C content (%)	71.32
No. of protein-coding sequences	8,270
No. of orphans	973
Coding density (%)	89.31
Average gene length (bp)	985
No. of rRNAs	6 × 16S-23S-5S
No. of tRNAs	99

importer complexes. Furthermore, 16% of the COG-classified CDS code for proteins involved in transcriptional processes, which suggests a high level of gene-expression regulation. This is especially relevant for the ongoing search for a regulatory element or - network controlling the expression of the acarbose biosynthetic gene cluster. Interestingly, the great proportion of transcriptional regulators is accompanied by a similar high percentage of proteins involved in signal transduction mechanisms (12%), which suggests a close connection between extracellular nutrient sensing and transcriptional regulation of uptake systems and degradation pathways. Comparatively, an analogous analysis of 4,431

Figure 4 Functional classifications of the *Actinoplanes* sp. SE50/110 protein coding sequences (CDS). The diagram represents the CDS that were categorized according their cluster of orthologous groups of proteins (COG) number [45,51]. All depicted percentages refer to the distribution of 4999 annotated CDS (100%) across all COG categories to which at least 10 CDS were found. Sequences with an unknown or poorly characterized function were excluded from the analysis. These excluded CDS were found to be rather randomly distributed across the genome. The outer ring contains specialized subclasses of the three main functional categories "Cellular Processes and Signaling", "Metabolism", and "Information Storage and Processing", located at the center.

annotated CDS from *Streptomyces coelicolor* revealed an even higher percentage of genes coding for enzymes related to metabolism (55%), a minor portion dedicated to signal transduction mechanisms (7%), and a highly similar amount (16%) of proteins involved in transcriptional processes. Finally, the genome of *Actinoplanes* sp. SE50/110 reveals a striking focus (27%) on cellular processes and signaling when compared to *S. coelicolor* (20%).

Besides these findings, only the genes for carbohydrate transport and -metabolism show another notable difference of more than 1% between *S. coelicolor* (13%) and *Actinoplanes* sp. SE50/110 (11%). Interestingly, 4% of the *Actinoplanes* sp. SE50/110 CDSs were found to be involved in secondary metabolite biosynthesis, a portion similar to the one found in the well-known producer *S. coelicolor* (5%) [52]. Taken together, these considerations lead to a new perception of the capabilities *Actinoplanes* sp. SE50/110 might offer, as for example, a new source of bioactive compounds. Furthermore and in contrast to *S. coelicolor*, the genome of *Actinoplanes* sp. SE50/110 hosts significantly more genes for signal transduction proteins. This might be one key to induce the expression of acarbose and novel secondary metabolite gene clusters by

appropriately composed cultivation media following the OSMAC (one strain, many compounds) approach [53]. These considerations are in good accordance with empirical knowledge gathered through long lasting media optimizations [Bayer HealthCare AG, personal communication].

Phylogenetic analysis of the *Actinoplanes* sp. SE50/110 16S rDNA reveals highest similarity to *Actinoplanes utahensis*

An unsupervised nucleotide BLAST [54] run of the 1509 bp long DNA sequence of the 16S rRNA gene from *Actinoplanes* sp. SE50/110 against the public non-redundant database (NCBI *nr/nt*) revealed high similarities to numerous species of the genera *Actinoplanes*, *Micromonospora* and *Salinispora*. Within the best 100 matches, the maximal DNA sequence identity was in the range of 100 - 96%. The coverage of the query sequence varied within this cohort between 100 - 97%. The hits with the highest similarity, based on the number of sequence substitutions were *A. utahensis* IMSNU 20044[T] (17 substitutions, 3 gaps) and *A. utahensis* IFO 13244[T] (16 substitutions, 3 gaps), both of which retrace to the type strain ([T]) *A. utahensis* ATCC 14539[T], firstly described first by Couch in 1963 [2]. The third hit to *A. palleronii*

IMSNU 2044[T] differs from *Actinoplanes* sp. SE50/110
by 24 substitutions and 5 gaps.

Based on the DNA sequences of the best 100 BLAST
hits, a phylogenetic tree was derived. A detailed view on
a subtree contains *Actinoplanes* sp. SE50/110 and 34 of
the most closely related species (Figure 5). This subtree
displays the derived phylogenetic distances between the
analyzed strains, represented by their distance on the x-
axis. From this analysis, it is evident that *A. utahensis* is
the nearest species to *Actinoplanes* sp. SE50/110 cur-
rently publicly known, followed by *A. palleronii* and *A.
awajiensis* subsp. *mycoplanecinus*. A second analysis
using the latest version of the ribosomal database pro-
ject [55] resulted in highly similar findings (data not
shown). Interestingly, *A. utahensis* and *Actinoplanes* sp.
SE50/110 form a subcluster within the *Actinoplanes*
genus although the different isolates originate from far
distant locations on different continents (Salt Lake City,
USA, North America and Ruiru, Kenya, Africa). In addi-
tion, it is noteworthy that *Actinoplanes* sp. SE50/110
was renamed several times and in the early 1990s this
strain was also classified as *A. utahensis* [49].

Comparative genome analysis reveals 50% unique genes in the *Actinoplanes* sp. SE50/110 genome

To date, seven full genome sequences belonging to the
family Micromonosporaceae are publicly available. Using
the comparative genomics tool EDGAR [60], a gene
based, full genome phylogenetic analysis of these strains
revealed a phylogeny comparable to the one inferred
from 16S rRNA genes (Figure 6). For comparison, some
industrially used *Streptomyces* and *Frankia* strains were
also included in the analysis. As expected, each genus
forms its own cluster. Interestingly, the genera *Micromo-
nospora*, *Verrucosispora* and *Salinispora* are more clo-
sely related to each other than to *Actinoplanes*, whereas
Streptomyces and *Frankia* are clearly distinct from the
whole Micromonosporaceae family. Based on this analy-
sis, the marine sediment isolate *Verrucosispora maris*
AB-18-032 is the closest sequenced species to *Actino-
planes* sp. SE50/110 currently publicly known with
2,683 orthologous genes, a G + C content of 70.9% and
a genome size of 6.67 MBases [61]. Comparative BLAST
analysis of conserved orthologous genes of all sequenced
Micromonosporaceae strains revealed prevalence for
being located in the upper half of the genome, near the
origin of replication (data not shown). The core genome
analysis revealed a total of 1,670 genes common to all
seven Micromonosporaceae strains, whereas the pan
genome consists of 18,189 genes calculated by the
EDGAR software [60]. An analysis of genes that exclu-
sively occur in the *Actinoplanes* sp. SE50/110 genome
revealed 4,122 singleton genes (49.8%).

The high quality genome sequence of *Actinoplanes* sp. SE50/110 corrects the previous sequence and annotation of the acarbose biosynthetic gene cluster

The acarbose biosynthetic (*acb*) gene cluster sequencing
was initiated [21] and successively expanded [17,24] by
classical Sanger sequencing. Until now, this sequence was
the longest (41,323 bp [GenBank:Y18523.4]) and best stu-
died contiguous DNA fragment available from *Actino-
planes* sp. SE50/110. However, with the complete, high
quality genome at hand, a total of 61 inconsistent sites
were identified in the existing acarbose gene cluster
sequence (Figure 7). Most notably, the deduced correc-
tions affect the amino acid sequence of two genes, namely
acbC, coding for the cytoplasmic 2-*epi*-5-*epi*-valiolone-
synthase, and *acbE*, translating to a secreted long chain
acarbose resistant α-amylase [17,24]. Because of two erro-
neous nucleotide insertions (c.1129_1130insG and
c.1146_1147insC) in *acbC* (1197 bp), the resulting frame-
shift caused a premature stop codon to occur, shortening
the actual gene sequence by 42 nucleotides. In *acbE* (3102
bp), the sequence differences are manifold, including mis-
matches, insertions and deletions, leading to multiple tem-
porary frameshifts and single amino acid substitutions.
Such differences occur in the mid part of the gene
sequence, ranging from nucleotide position 1102 to 2247.
These sequence corrections improved the similarity of the
α-amylase domain to its catalytic domain family.

Several genes of the acarbose gene cluster are also found in other locations of the *Actinoplanes* sp. SE50/110 genome sequence
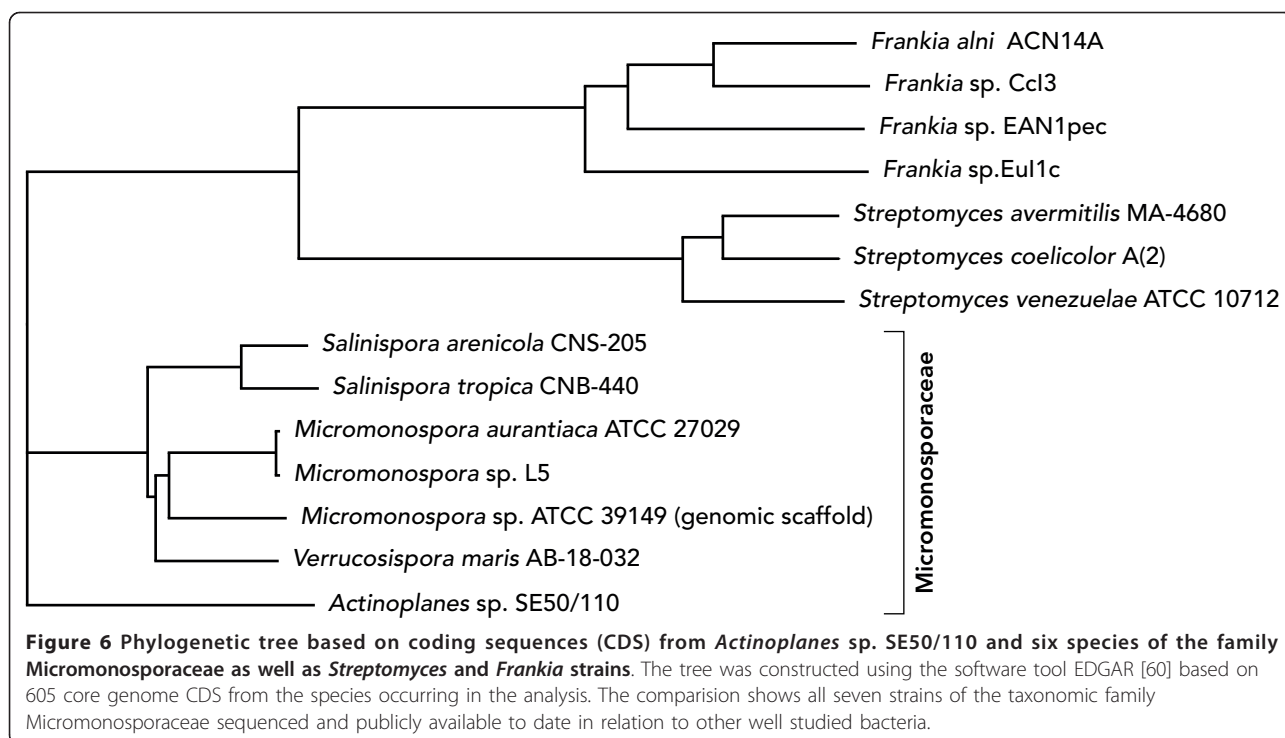
It is known that the copy number of genes can have
high impact on the efficiency of secondary metabolite
production [19,62-64]. It is therefore worthwhile to
study the genome wide occurrences of the genes
encoded within the acarbose biosynthetic gene cluster,
particularly with regard to import and export systems
and the assessment of possible future knock-out
experiments.

Our results show, that the *acb* gene cluster does not
occur in more than one location within the *Actinoplanes*
sp. SE50/110 genome. However, single genes and gene
sets with equal functional annotation and amino acid
sequence similarity to members of the *acb* cluster were
found scattered throughout the genome by BLASTP
analysis. Most notably, homologues to genes encoding
the first, second and fourth step of the valienamine moi-
ety synthesis of acarbose were found as a putative
operon with moderate similarities of 52% (Acpl6250 to
AcbC), 35% (Acpl6249 to AcbM) and 34% (Acpl6251 to
AcbL). Furthermore, one homologue for each of the
proteins AcbA (61% to Acpl3097) and AcbB (66% to
Acpl3096) was identified. In the arcabose gene cluster,
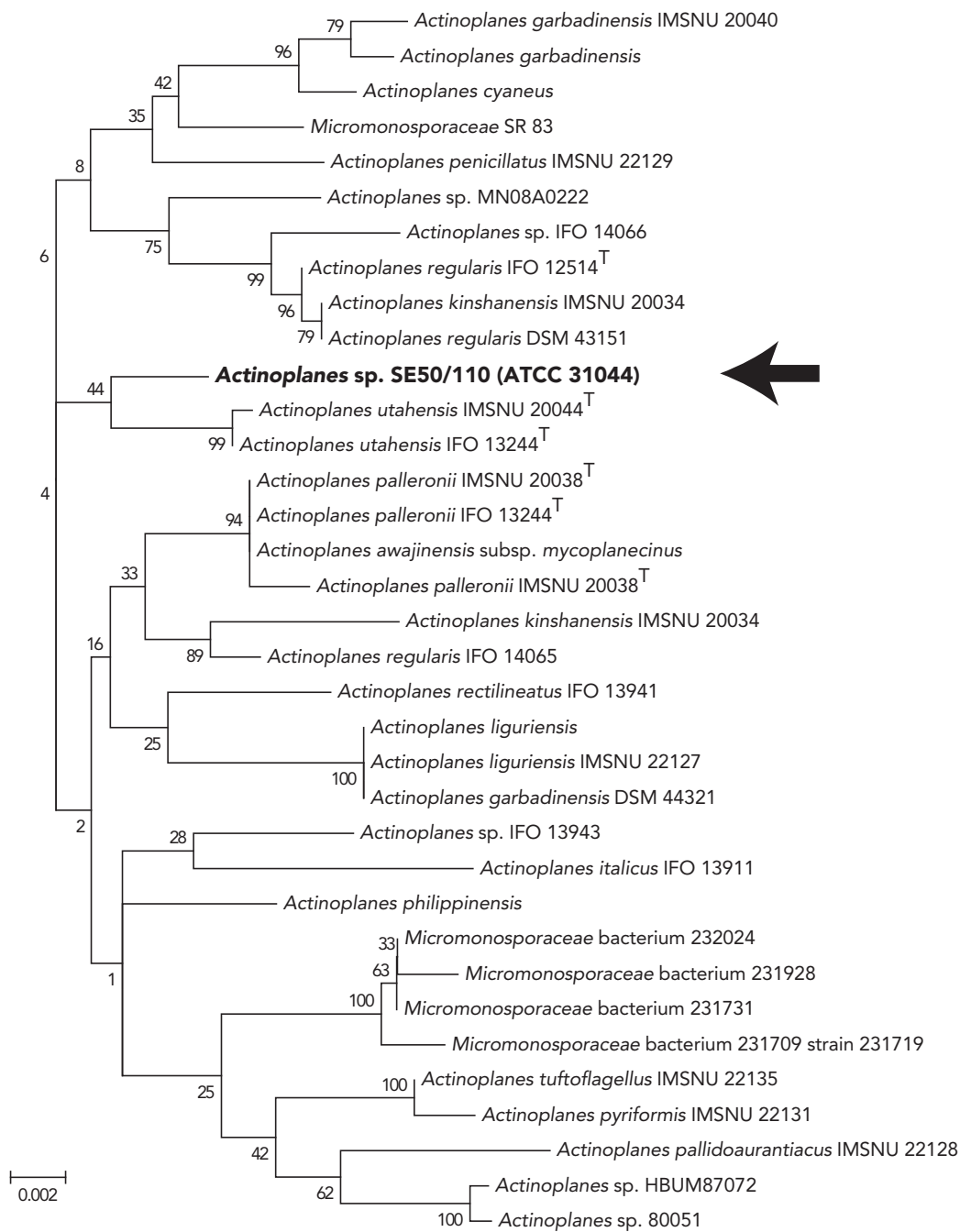
Figure 5 Phylogenetic tree based on 16S rDNA from *Actinoplanes* sp. SE50/110 and the 34 most closely related species. Shown is an excerpt of a phylogenetic tree build from the 100 best nucleotide BLAST hits for the *Actinoplanes* sp. SE50/110 16S rDNA. The shown subtree contains the 34 hits most closely related to *Actinoplanes* sp. SE50/110 (black arrow) with their evolutionary distances. The numbers on the branches represent confidence values in percent from a phylogenetic bootstrap test (1000 replications). The evolutionary history was inferred using the Neighbor-Joining method [56]. The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analyzed [57]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Jukes-Cantor method [58] and are in the units of the number of base substitutions per site. The analysis involved 100 nucleotide sequences of which 35 are shown. Codon positions included were 1st + 2nd + 3 rd + Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 1396 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 [59]. Bar, 0.002 nucleotide substitutions per nucleotide position.
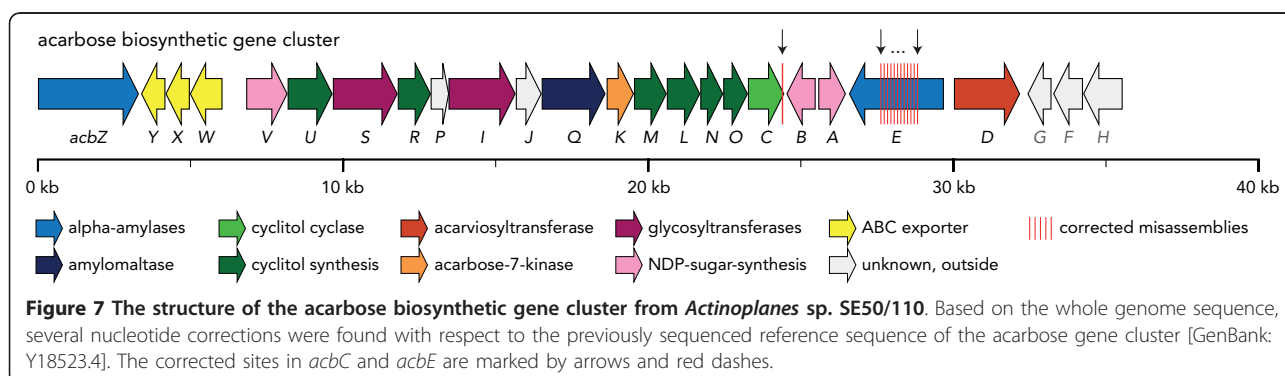
Figure 6 Phylogenetic tree based on coding sequences (CDS) from *Actinoplanes* sp. SE50/110 and six species of the family Micromonosporaceae as well as *Streptomyces* and *Frankia* strains. The tree was constructed using the software tool EDGAR [60] based on 605 core genome CDS from the species occurring in the analysis. The comparision shows all seven strains of the taxonomic family Micromonosporaceae sequenced and publicly available to date in relation to other well studied bacteria.

acbA and *acbB* are located adjacent to each other (Figure 7), and catalyze the first two sequential reactions needed for the formation of dTDP-4-keto-6-deoxy-D-glucose, another acrabose essential intermediate [21,28]. It is therefore interesting to note that the identified homologues to *acbA* and *acbB* were also found adjacent to each other in the context of a putative dTDP-rhamnose synthesis cluster (*acpl3095-acpl3098*). In *Mycobacterium smegmatis*, the orthologous dTDP-rhamnose biosynthetic gene cluster codes for mandatory proteins RmlABCD, involved in cell wall integrity and thus, cell survival [65]. Further genome analysis detected the ABC-transporters Acpl3214-Acpl3216 and Acpl5011-Acpl5013 with moderate similarity (28-49%) to the acarbose exporter complex AcbWXY. Both operons resemble the gene structure of *acbWXY* consisting of an

ABC-type sugar transport ATP-binding protein and two ABC-type transport permease protein coding genes. Therefore, considering that ABC-transporters often present a promiscuous substrate-specificity, it is possible that these structurally similar transporters also be involved in acarbose transport. Acpl6399, a homologue with high sequence similarities to the alpha amylases AcbZ (65%) and AcbE (63%) was found encoded within the maltose importer operon *malEFG*. For the remaining *acb* genes only weak (*acbV, acbR, acbP, acbJ, acbQ, acbK* and *acbN*) or no similarities (*acbU, acbS, acbI* and *acbO*) were found outside of the *acb* cluster by BLASTP searches using an e-value threshold of e^{-10}.

In contrast to previous findings [27], the extracellular binding protein AcbH, encoded within the *acbGFH* operon (Figure 7), was recently shown to exhibit high

Figure 7 The structure of the acarbose biosynthetic gene cluster from *Actinoplanes* sp. SE50/110. Based on the whole genome sequence, several nucleotide corrections were found with respect to the previously sequenced reference sequence of the acarbose gene cluster [GenBank: Y18523.4]. The corrected sites in *acbC* and *acbE* are marked by arrows and red dashes.

affinity to galactose instead of acarbose or its homologues [22]. This implicates that *acbGFH* does not directly belong to the acarbose cluster as was proposed by the carbophore hypothesis, which indicates that acarbose or acarbose homologues can be reused by the producer [17,66]. In order to search the *Actinoplanes* sp. SE50/110 genome for a new acarbose importer candidate, the *gacGFH* operon of a second acarbose gene cluster identified in *Streptomyces glaucescens* GLA.O has been used as query [67]. GacH was recently shown to recognize longer acarbose homologues but exhibits only low affinity to acarbose [68]. However, the search revealed rather weak similarities towards the best hit operon *acpl5404-acpl5406* with GacH showing 26% identity to its homologue Acpl5404. A consecutive search of the extracellular maltose binding protein MalE from *Salmonella typhimurium*, which has been shown to exhibit high affinity to acarbose [68], revealed 32% identity to its MalE homologue in *Actinoplanes* sp. SE50/110. Despite the low sequence similarities, these findings suggest that acarbose or its homologues are either imported by one or both of the above mentioned importers, or that the extracellular binding protein exhibits a distinct amino acid sequence in *Actinoplanes* sp. SE50/110 and can therefore not be identified by sequence comparison alone.

The *Actinoplanes* sp. SE50/110 genome hosts an integrative and conjugative element that also exists in multiple copies as an extrachromosomal element

Actinomycete integrative and conjugative elements (AICEs) are a class of mobile genetic elements possessing a highly conserved structural organization with functional modules for excision/integration, replication, conjugative transfer and regulation [69]. Being able to replicate autonomously, they are also said to mediate the acquisition of additional modules, encoding functions such as resistance and metabolic traits, which confer a selective advantage to the host under certain environmental conditions [70]. Interestingly, a similar AICE, designated pACPL, was identified in the complete genome sequence of *Actinoplanes* sp. SE50/110 (Figure 8). Its size of 13.6 kb and the structural gene organization are in good accordance with other known AICEs of closely related species like *Micromonospora rosario*, *Salinispora tropica* or *Streptomyces coelicolor* (Figure 8).
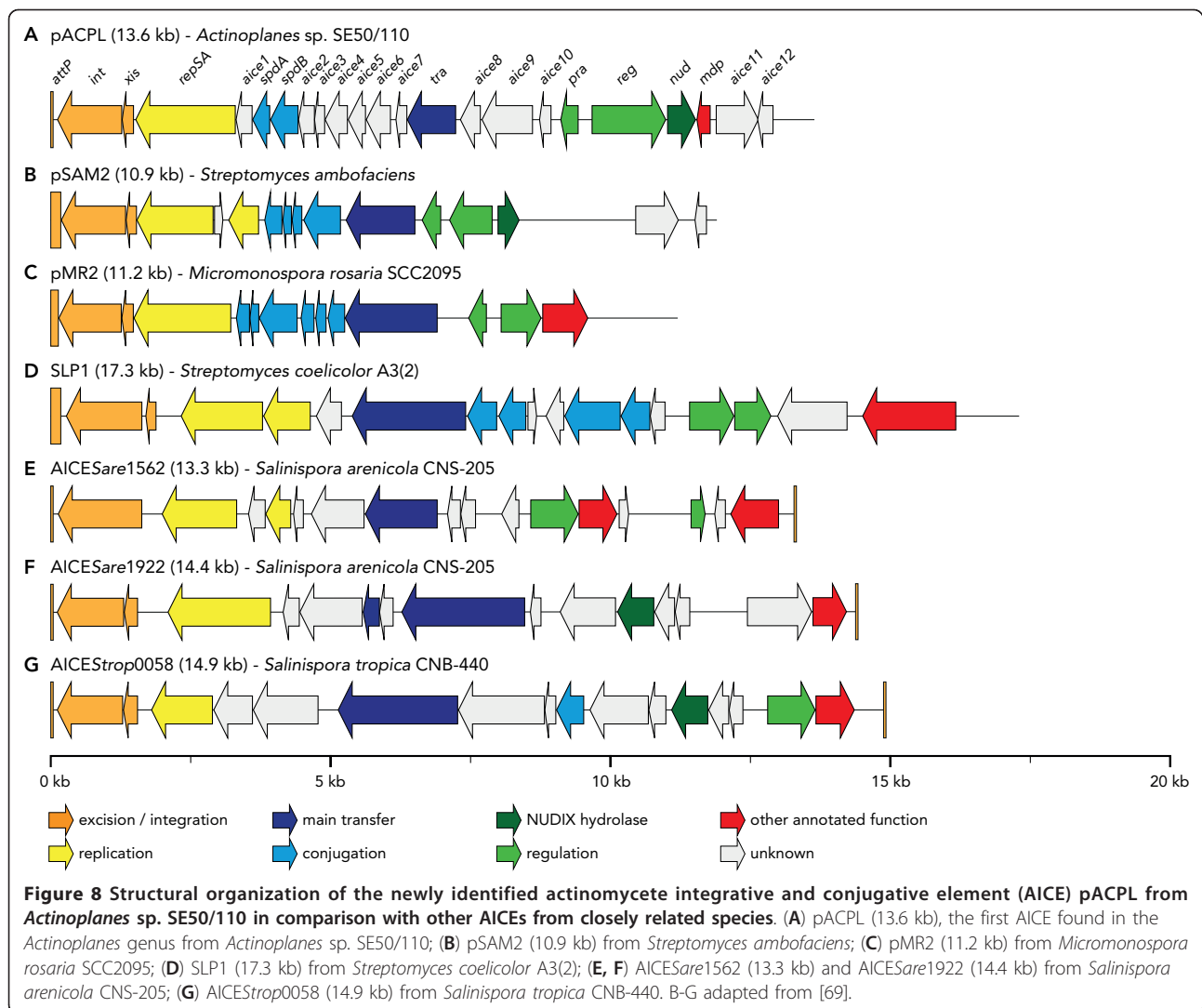
Most known AICEs subsist in their host genome by integration in the 3' end of a tRNA gene by site-specific recombination between two short identical sequences (*att* identity segments) within the attachment sites located on the genome (*attB*) and the AICE (*attP*), respectively [69]. In pACPL, the *att* identity segments are 43 nt in size and *attB* overlaps the 3' end of a proline tRNA gene. Moreover, the identity segment in *attP*

is flanked by two 21 nt repeats containing two mismatches: GTCACCCAGTTAGT(T/C)AC(C/T)CAG. These exhibit high similarities to the arm-type sites identified in the AICE pSAM2 from *Strepomyces ambofaciens*. For pSAM2 it was shown that the integrase binds to these repeats and that they are essential for efficient recombination [71].

pACPL hosts 22 putative protein coding sequences (Figure 8). The integrase, excisionase and replication genes *int*, *xis* and *repSA* are located directly downstream of *attP* and show high sequence similarity to numerous homologues from closely related species. The putative main transferase gene *tra* contains the sequence of a FtsK-SpoIIIE domain found in all AICEs and *Streptomyces* transferase genes [69]. SpdA and SpdB show weak similarity to spread proteins from *Frankia* sp. CcI3 and *M. rosaria* where they are involved in the intramycelial spread of AICEs [72,73]. The putative regulatory protein Pra was first described in pSAM2 as an AICE replication activator [74]. On pACPL, it exhibits high similarity to an uncharacterized homologue from *Micromonospora aurantica* ATCC 27029. A second regulatory gene *reg* shows high similarities to transcriptional regulators of various *Streptomyces* strains whereas the downstream gene *nud* exhibits 72% similarity to the amino acid sequence of a NUDIX hydrolase from *Streptomyces* sp. AA4. In contrast, *mdp* codes for a metal dependent phosphohydrolase also found in various *Frankia* and *Streptomyces* strains.

Homologues to the remaining genes are poorly characterized and largely hypothetical in public databases although *aice4* is also found in various related species and shows akin to *aice1*, *aice2*, *aice5*, *aice6*, and *aice9*, high similarity to homologues from *M. aurantiaca*. Interestingly, homologues to *aice1* and *aice2* were only found in *M. aurantiaca*, whereas *aice3*, *aice7*, *aice8*, *aice10*, *aice11*, and *aice12*, seem to solely exist in *Actinoplanes* sp. SE50/110.

Based upon read-coverage observations of the AICE containing genomic region, an approximately twelve-fold overrepresentation of the AICE coding DNA sequences has been revealed (Figure 1). As only one copy of the AICE was found to be integrated in the genome, it was concluded that approximately eleven copies of the element might exist as circular, extrachromosomal versions in a typical *Actinoplanes* sp. SE50/110 cell. However, the number of extrachromosomal copies per cell might be even higher, as it is possible that a proportion of the AICEs was lost during DNA isolation. It should also be noted that the rather low G + C content of the AICE (65.56%) might have introduced a similar amplification bias during the library preparation as discussed above for the *rrn* operons. Nevertheless, these findings are of great interest, as they demonstrate the

Figure 8 Structural organization of the newly identified actinomycete integrative and conjugative element (AICE) pACPL from *Actinoplanes* **sp. SE50/110 in comparison with other AICEs from closely related species**. (**A**) pACPL (13.6 kb), the first AICE found in the *Actinoplanes* genus from *Actinoplanes* sp. SE50/110; (**B**) pSAM2 (10.9 kb) from *Streptomyces ambofaciens*; (**C**) pMR2 (11.2 kb) from *Micromonospora rosaria* SCC2095; (**D**) SLP1 (17.3 kb) from *Streptomyces coelicolor* A3(2); (**E, F**) AICE*Sare*1562 (13.3 kb) and AICE*Sare*1922 (14.4 kb) from *Salinispora arenicola* CNS-205; (**G**) AICE*Strop*0058 (14.9 kb) from *Salinispora tropica* CNB-440. B-G adapted from [69].

first native functional AICE for *Actinoplanes* spp. in general and imply the possibility of future genetic access to *Actinoplanes* sp. SE50/110 in order to perform targeted genetic modifications as done before for e.g. *Micromonospora* spp. [75]. The newly identified AICE may also improve previous efforts in the analysis of heterologous promoters for the overexpression of the lipopeptide antibiotic friulimicin in *Actinoplanes friuliensis* [76].

Four putative antibiotic production gene clusters were found in the Actinoplanes sp. SE50/110 genome sequence

Bioactive compounds synthetized through secondary metabolite gene clusters are a rich source for pharmacologically relevant products like antibiotics, immunosuppressants or antineoplastics [77,78]. Besides aminoglycosides, the majority of these metabolites are built up in a modular

fashion by using non-ribosomal peptide synthetases (NRPS) and/or polyketide synthases (PKS) as enzyme templates (for a recent review see [79]). Briefly, the nascent product is built up by sequential addition of a new element at each module it traverses. The complete sequence of modules may reside on one gene or spread across multiple genes in which the order of action of each gene product is determined by specific linker sequences present at proteins' N- and C-terminal ends [78,80].

For NRPSs, a minimal module typically consists of at least three catalytic domains, namely the andenylation (A) domain for specific amino acid activation, the thiolation (T) domain, also called peptidyl carrier protein (PCP) for covalent binding and transfer and the condensation (C) domain for incorporation into the peptide chain [78]. In addition, domains for epimerization (E), methylation (M) and other modifications may reside within a module. Oftentimes a thioesterase domain (Te)

is located at the C-terminal end of the final module, responsible for e.g. cyclization and release of the non-ribosomal peptide from the NRPS [81].

In case of the PKSs, an acyltransferase (AT) coordinates the loading of a carboxylic acid and promotes its attachment on the acyl carrier protein (ACP) where chain elongation takes place by a β-kethoacyl synthase (KS) mediated condensation reaction [79]. Additionally, most PKSs reduce the elongated ketide chain at accessory β-kethoacyl reductase (KR), dehydratase (DH), methyltransferase (MT) or enoylreductase (ER) domains before a final thioesterase (TE) domain mediates release of the polyketide [82].

Modular NRPS and PKS enzymes strictly depend on the activation of the respective carrier protein domains (PCP and ACP), which must be converted from their inactive apo-forms to cofactor-bearing holo-forms by a specific phosphopantetheinyl transferase (PPTase) [83]. The genome of *Actinoplanes* sp. SE50/110 hosts three such enzyme encoded by the genes *acpl842*, *acpl996*, and *acpl6917*.

In *Actinoplanes* sp. SE50/110, one NRPS (cACPL_1), two PKS (cACPL_2 & cACPL_3) and a hybrid NRPS/PKS cluster (cACPL_4) were found by gene annotation and subsequent detailed analysis using the antiSMASH pipeline [84]. The first of the identified gene clusters (cACPL_1) contains four NRPS genes (Figure 9A), hosting a total of ten adenylation (A), thiolation (T) and condensation (C) domains, potentially making up 10 modules. Thereof, three modules are formed by intergentic domains, which suggests a specific interaction of the four putative NRPS enzymes in the order *nrps1A-B-D-C*. Such interaction order is the only one that leads to the assembly of all domains into 10 complete modules - 9 minimal modules (A-T-C) and one module containing an additional epimerization domain. These considerations were corroborated by matching linker sequences, named short communication-mediating (COM) domains [78], found at the C-terminal part of NRPS1D and the N-terminal end of NRPS1C. Furthermore, this cluster shows high structural and sequential similarity to the SMC14 gene cluster identified on the pSCL4 megaplasmid from *Streptomyces clavuligerus* ATCC 27064 [85]. However, in SMC14 a homolog to *nrps1D* is missing, which leads to the speculation that *nrps1D* was subsequently added to the cluster as an additional building block. In fact, leaving *nrps1D* out of the assembly line would theoretically still result in a complete enzyme complex built from 9 instead of 10 modules. Based on the antiSMASH prediction, the amino acid backbone of the final product is likely to be composed of the sequence: Ala-Asn-Thr-Thr-Thr-Asn-Thr-Asn-Val-Ser (Figure 9A). Besides the NRPSs, the cluster also contains multiple genes involved in regulation and transportation

as well as two *mbtH*-like genes, known to facilitate secondary metabolite synthesis. In this regard, it is noteworthy that the occurrence of two *mbtH* genes in a secondary metabolite gene cluster is exceptional in that it has only been found once before in the teicoplanin biosynthesis gene cluster of *Actinoplanes teichomyceticus* [86].

The type-1 PKS-cluster cACPL_2 (Figure 9B) hosts 5 genes putatively involved in the synthesis of an unknown polyketide. The sum of the PKS coding regions adds up to a size of ~49 kb whereas all encoded PKSs exhibit 62-66% similarity to PKSs from various *Streptomyces* strains. However unlike the NRPS-cluster, no cluster structurally similar to cACPL_2 was found in public databases. Analysis of the domain and module architecture revealed a total of 10 elongation modules (KS-AT-[DH-ER-KR]-ACP) including 9 β-kethoacyl reductase (KR) and 8 dehydratase (DH) domains as well as a termination module (TE). However, an initial loading module (AT-ACP) could not be identified in the proximity of the cluster. To elucidate the most likely build order of the polyketide, the N- and C-terminal linker sequences were matched against each other using the software SBSPKS [87] and antiSMASH. Remarkably, both programs independently predicted the same gene order: *pks1E-C-B-A-D*.

Just 15 kb downstream of cACPL_2, a second gene cluster (cACPL_3) containing a long PKS gene with various accessory protein coding sequences could be identified (Figure 9C). It shows some structural similarity to a yet uncharacterized PKS gene cluster of *Salinispora tropica* CNB-440 (genes *Strop_2768 - Strop_2777*). Besides the 3 elongation modules identified on *pks2A*, no other modular type 1 PKS genes were found in the proximity of the cluster. However, genes downstream of *pks2A* are likely to be involved in the synthesis and modification of the polyketide, coding for an acyl carrier protein (ACP), an ACP malonyl transferase (MAT), a lysine aminomutase, an aspartate transferase and a type 2 thioestrase. Especially type 2 thioestrases are often found in PKS clusters [88] like e.g., in the gramicidin S biosynthesis operon [89]. The presence of discrete ACP, MAT and two additional acetyl CoA synthetase-like enzymes is also typical for type 2 PKS systems [90] although no ketoacyl-synthase (KS_α) and chain length factor (KS_β) was found in this cluster [91].

Another 58 kb downstream of cACPL_3 a fourth secondary metabolite cluster (cACPL_4) was located (Figure 9D). It hosts both 3 NRPS and 3 PKS genes and may therefore synthesize a hybrid product as previously reported for bleomycin from *Streptomyces verticillus* [92], pristinamycin IIB from *Streptomyces pristinaespiralis* [93] and others [94]. N- and C-terminal sequence analysis of the two cluster types revealed the gene

Figure 9 The gene organization of the four putative secondary metabolite gene clusters found in the *Actinoplanes* sp. SE50/110 genome. (**A**) Non-ribosomal peptide synthetase (NRPS) cluster showing high structural and sequential similarity to the SMC14 gene cluster identified on the pSCL4 megaplasmid from *Streptomyces clavuligerus* ATCC 27064. (**B**) Large polyketide synthase (PKS) gene cluster exhibiting 62-66% similarity to PKSs from various *Streptomyces* strains. (**C**) A single PKS gene with various accessory genes showing some structural similarity to a yet uncharacterized PKS gene cluster of *Salinispora tropica* CNB-440. (**D**) Putative hybrid NRPS/PKS gene cluster with NRPS genes showing high similarity (63-76%) to genes from an uncharacterized cluster of *Streptomyces venezuelae* ATCC 10712 whereas the PKS genes exhibit highest similarity (63-66%) to genes scattered in the *Methylosinus trichosporium* OB3b genome.

orders *nrps2B-C-A* and *pks3A-B-C* as most likely. The prediction of the peptide backbone of the NRPS cluster resulted in the putative product dehydroaminobutyric acid (Dhb)-Cys-Cys. One could speculate that the PKSs are used prior to the NRPSs as *nrps2A* comes with a termination module (Te). However, two additional monomeric thioesterase (TE) and one enoylreductase (ER) domain containing genes do also belong to the cluster and may be involved in the termination and modification of the product. Notably, all three NRPS genes show high similarity (63-76%) to genes from an uncharacterized cluster of *Streptomyces venezuelae* ATCC 10712 whereas the PKS genes exhibit highest similarity (63-66%) to genes scattered in the *Methylosinus trichosporium* OB3b genome.

The four newly discovered secondary metabolite gene clusters broaden our knowledge of actinomycete NRPS and PKS biosynthesis clusters and represent just the tip of the iceberg of the manifold biosynthetical capabilities - apart from the well-known acarbose production - that *Actinoplanes* sp. SE50/110 houses. It remains to be

determined if all presented clusters are involved in industrially rewarding bioactive compound synthesis and how these clusters are regulated, because none of these metabolites were identified and isolated so far. These new gene clusters may also be used in conjunction with well-studied antibiotic operons, in order to synthesize completely new substances, as recently performed [95,96].

Conclusions

The establishment of the complete genome sequence of the acarbose producer *Actinoplanes* sp. SE50/110 is an important achievement on the way towards rational optimization of the acarbose production through targeted genetic engineering. In this process, the identified AICE may serve as a vector for future transformation of *Actinoplanes* spp. Furthermore, our work provides the first sequenced genome of the genus *Actinoplanes*, which will serve as the reference for future genome analysis and sequencing projects in this field. By providing novel insights into the enzymatic equipment of *Actinoplanes* sp. SE50/110, we

identified previously unknown NRPS/PKS gene clusters, potentially encoding new antibiotics and other bioactive compounds that might be of pharmacologic interest.

With the complete genome sequence at hand, we propose to conduct future transcriptome studies on *Actinoplanes* sp. SE50/110 in order to analyze differential gene expression in cultivation media that promote and repress acarbose production, respectively. Results will help to identify potential target genes for later genetic manipulations with the aim of increasing acarbose yields.

Methods

Cultivation of the *Actinoplanes* sp. SE50/110 strain
In order to isolate DNA, the *Actinoplanes* sp. SE50/110 strain was cultivated in a two-step shake flask system. Besides inorganic salts the medium contained starch hydrolysate as carbon source and yeast extract as nitrogen source. Pre culture and main culture were incubated for 3 and 4 days, respectively, on a rotary shaker at 28°C. Then the biomass was collected by centrifugation.

Preparation of genomic DNA
The preparation of genomic DNA of the *Actinoplanes* sp. SE50/110 strain was performed as previously published [30].

High throughput sequencing and automated assembly of the *Actinoplanes* sp. SE50/110 genome
The high throughput pyrosequencing has been carried out on a Genome Sequencer FLX system (454 Life Sciences). The subsequent assembly of the generated reads was performed using the Newbler assembly software, version 2.0.00.22 (454 Life Sciences). Details of the sequencing and assembly procedures have been described previously [30].

Construction of a fosmid library for the *Actinoplanes* sp. SE50/110 genome finishing
The fosmid library construction for *Actinoplanes* sp. SE50/110 with an average inset size of 40 kb has been carried out on isolated genomic DNA by IIT Biotech GmbH (Universitätsstrasse 25, 33615 Bielefeld, Germany). For construction in *Escherichia coli* EPI300 cells, the CopyControl™ Cloning System (EPICENTRE Biotechnologies, 726 Post Road, Madison, WI 53713, USA) has been used. The kit was obtained from Biozym Scientific GmbH (Steinbrinksweg 27, 31840 Hessisch Oldendorf, Germany).

Terminal insert sequencing of the *Actinoplanes* sp. SE50/110 fosmid library
The fosmid library terminal insert sequencing was carried out with capillary sequencing technique on a 3730xl DNA-Analyzer (Applied Biosystems) by IIT Biotech GmbH. The resulting chromatogram files were base called using the phred software [97,98] and stored in FASTA format. Both files were later used for gap closure and quality assessment.

Finishing of the *Actinoplanes* sp. SE50/110 genome sequence by manual assembly
In order to close remaining gaps between contiguous sequences (contigs) still present after the automated assembly, the visual assembly software package Consed [40,99] was utilized. Within the graphical user interface, fosmid walking primer and genome PCR primer pairs were selected at the ends of contiguous contigs. These were used to amplify desired sequences from fosmids or genomic DNA in order to bridge the gaps between contiguous contigs.

After the DNA sequence of these amplicons had been determined, manual assembly of all applicable reads was performed with the aid of different Consed program features. In cases where the length or quality of one read was not sufficient to span the gap, multiple rounds of primer selection, amplicon generation, amplicon sequencing and manual assembly were performed.

Prediction of open reading frames on the *Actinoplanes* sp. SE50/110 genome sequence
The potential genes were identified by a series of programs which are all part of the GenDB annotation pipeline [44]. For the automated identification of open reading frames (ORFs) the prokaryotic gene finders Prodigal [42] and GISMO [43] were primarily used. In order to optimize results and allow for easy manual curation, further intrinsic, extrinsic and combined methods were applied by means of the Reganor software [100,101] which utilizes the popular gene prediction tools Glimmer [102] and CRITICA [103].

Functional annotation of the identified open reading frames of the *Actinoplanes* sp. SE50/110 genome
The identified open reading frames were analyzed through a variety of different software packages in order to draw conclusions from their DNA- and/or amino acid-sequences regarding their potential function. Besides functional predictions, further characteristics and structural features have also been calculated.

Similarity-based searches were applied to identify conserved sequences by means of comparison to public and/or proprietary nucleotide- and protein-databases. If a significant sequence similarity was found throughout the major section of a gene, it was concluded that the gene should have a similar function in *Actinoplanes* sp. SE50/110. The similarity-based methods, which were used to annotate the list of ORFs are termed BLASTP [104] and RPS-BLAST [105].

Enzymatic classification has been performed on the basis of *enzyme commission* (EC) numbers [46,106]. These were primarily derived from the PRIAM database [107] using the PRIAM_Search utility on the latest version (May 2011) of the database. For genes with no PRIAM hit, secondary EC-number annotations were derived from searches against the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [108,109]. For further functional gene annotation, the *cluster of orthologous groups of proteins* (COG) classification system has been applied [45,51] using the latest version (March 2003) of the database [110].

To identify potential transmembrane proteins, the software TMHMM [111,112] has been utilized.

The software SignalP [113-115] was used to predict the secretion capability of the identified proteins. This is done by means of Hidden Markov Models and neural networks, searching for the appearance and position of potential signal peptide cleavage sites within the amino acid sequence. The resulting score can be interpreted as a probability measure for the secretion of the translated protein. SignalP retrieves only those proteins which are secreted by the classical signal-peptide-bound mechanisms.

In order to identify further *Actinoplanes* sp. SE50/110 proteins which are not secreted via the classical way, the software SecretomeP has been applied [116]. The underlying neural network has been trained with secreted proteins, known to lack signal peptides despite their occurrence in the exoproteome. The final secretion capability of the translated genes was derived by the combined results of SignalP and SecretomeP predictions.

To reveal polycistronic transcriptional units, proprietary software has been developed which predicts jointly transcribed genes by their orientation and proximity to neighboring genes (adopted from [117]). In light of these predictions, operon structures can be estimated and based upon them further sequence regions can be derived with high probability of contained promoter and operator elements.

The software DNA mfold [118] has been used to calculate hybridization energies for the DNA sequences in order to identify secondary structures such as transcriptional terminators which indicate operon and gene ends, respectively. The involved algorithm computes the most stable secondary structure of the input sequence by striving to the lowest level in terms of Gibbs free energy (ΔG), a measure for the energy which is released by the formation of hydrogen bonds between the hybridizing base pairs.

Phylogenetic analyses

The 16S rDNA-based phylogenetic analysis was performed on DNA sequences retrieved by public BLAST [54] searches against the 16S rDNA of *Actinoplanes* sp. SE50/110. From the best hits, a multiple sequence alignment was built by applying the MUSCLE program [119,120] before deriving the phylogeny thereof by the MEGA 5 software [59] using the neighbor-joining algorithm [56] with Jukes-Cantor model [58].

Genome based phylogenetic analysis of *Actinoplanes* sp. SE50/110 in relation to other closely related species was performed with the comparative genomics tool EDGAR [60]. Briefly, the core genome of all selected strains was calculated and based thereon, phylogenetic distances were calculated from multiple sequence alignments. Then, phylogenetic trees were generated from concatenated core gene alignments.

Author details

[1]Senior research group in Genome Research of Industrial Microorganisms, Center for Biotechnology, Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany. [2]Genome Informatics Research Group, Faculty of Technology, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany. [3]Institute for Genome Research and Systems Biology, Center for Biotechnology, Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany. [4]Bayer HealthCare AG, Friedrich-Ebert-Str. 475, 42117 Wuppertal, Germany. [5]Bayer Technology Services GmbH, Friedrich-Ebert-Str. 475, 42117 Wuppertal, Germany. [6]Bergische Universität Wuppertal, Sportsmedicine, Pauluskirchstraße 7, 42285 Wuppertal, Germany. [7]Center for Biotechnology, Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany.

Authors' contributions

AP, KS and JK designed the study. KS was responsible for the cultivation and DNA isolation. RS and CR performed the library preparation and high-throughput sequencing. AP, UFW, RS and CR critically revised the manuscript. AP, UFW, JS and AK provided important intellectual contributions. PS carried out all work not contributed for by other authors and wrote the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Couch JN: *Actinoplane*, A New Genus of the Actinomycetales. *J Elisha Mitchell Scientific Soc* 1950, **66**:87-92.
2. Couch JN: **Some New Genera and Species of the Actinoplanaceae.** *J Elisha Mitchell Scientific Soc* 1963, **79**:53-70.
3. Lechevalier HA, Lechevalier MP: **The Actinomycetales.**Edited by: Jena: Veb G. Fisher 1970, 393-405.
4. Parenti F, Coronelli C: **Members of the Genus *Actinoplane* and their Antibiotics.** *Annu Rev Microbiol* 1979, **33**:389-411.

5. Varadaraj K, Skinner DM: **Denaturants or cosolvents improve the specificity of PCR amplification of a G + C-rich DNA using genetically engineered DNA polymerases.** *Gene* 1994, **140**:1-5.

6. McDowell DG, Burns NA, Parkes HC: **Localised sequence regions possessing high melting temperatures prevent the amplification of a DNA mimic in competitive PCR.** *Nucleic Acids Res* 1998, **26**:3340-3347.

7. Parenti F, Pagani H, Beretta G: **Lipiarmycin, a new antibiotic from *Actinoplane*. I. Description of the producer strain and fermentation studies.** *J Antibiot* 1975, **28**:247-252.

8. Debono M, Merkel KE, Molloy RM, Barnhart M, Presti E, Hunt AH, Hamill RL: **Actaplanin, new glycopeptide antibiotics produced by *Actinoplanes missouriensi*. The isolation and preliminary chemical characterization of actaplanin.** *J Antibiot* 1984, **37**:85-95.

9. Vértesy L, Ehlers E, Kogler H, Kurz M, Meiwes J, Seibert G, Vogel M, Hammann P: **Friulimicins: novel lipopeptide antibiotics with peptidoglycan synthesis inhibiting activity from *Actinoplanes friuliensi* sp. nov. II. Isolation and structural characterization.** *J Antibiot* 2000, **53**:816-827.

10. Cooper R, Truumees I, Gunnarsson I, Loebenberg D, Horan A, Marquez J, Patel M, Gullo V, Puar M, Das P: **Sch 42137, a novel antifungal antibiotic from an *Actinoplane* sp. Fermentation, isolation, structure and biological properties.** *J Antibiot* 1992, **45**:444-453.

11. Jung H-M, Jeya M, Kim S-Y, Moon H-J, Kumar Singh R, Zhang Y-W, Lee J-K: **Biosynthesis, biotechnological production, and application of teicoplanin: current state and perspectives.** *Appl Microbiol Biotechnol* 2009, **84**:417-428.

12. Frommer W, Puls W, Schäfer D, Schmidt D: **Glycoside-hydrolase enzyme inhibitors.** *German patent DE 2064092 (US patent 3,876,766)* 1975.

13. Frommer W, Junge B, Keup U, Mueller L, Schmidt D: **Amino sugar derivatives.** *German patent DE 2347782 (US patent 4,062,950)* 1977.

14. Frommer W, Puls W, Schmidt D: **Process for the production of a saccharase inhibitor.** *German patent DE 2209834 (US patent 4,019,960)* 1977.

15. Frommer W, Junge B, Müller L, Schmidt D, Truscheit E: **Neue Enzyminhibitoren aus Mikroorganismen.** *Planta Med* 1979, **35**:195-217.

16. IDF: *Diabetes Atlas.* 4 edition. Brussels, Belgium: International Diabetes Federation; 2009.

17. Wehmeier UF, Piepersberg W: **Biotechnology and molecular biology of the alpha-glucosidase inhibitor acarbose.** *Appl Microbiol Biotechnol* 2004, **63**:613-625.

18. Schedel M: **Weiße Biotechnologie bei Bayer HealthCare Product Supply: Mehr als 30 Jahre Erfahrung.** *Chemie Ingenieur Technik* 2006, **78**:485-489.

19. Baltz RH: **Strain improvement in actinomycetes in the postgenomic era.** *J Ind Microbiol Biotechnol* 2011, **38**:657-666.

20. Crueger A, Apeler H, Schröder W, Pape H, Goeke K, Piepersberg W, Distler J, Diaz-Guardamino Uribe PM, Jarling M, Stratmann A: **Acarbose (*acb*) Cluster from *Actinoplanes* sp. SE 50/110.** *International patent WO/1998/038313* 1998.

21. Stratmann A, Mahmud T, Lee S, Distler J, Floss HG, Piepersberg W: **The AcbC Protein from *Actinoplane* Species Is a C7-cyclitol Synthase Related to 3-Dehydroquinate Synthases and Is Involved in the Biosynthesis of the α-Glucosidase Inhibitor Acarbose.** *J Biol Chem* 1999, **274**:10889-10896.

22. Licht A, Bulut H, Scheffel F, Daumke O, Wehmeier UF, Saenger W, Schneider E, Vahedi-Faridi A: **Crystal Structures of the Bacterial Solute Receptor AcbH Displaying an Exclusive Substrate Preference for β-d-Galactopyranose.** *J Mol Biol* 2011, **406**:92-105.

23. Rauschenbusch E, Schmidt D: **Verfahren zur Isolierung von (O{4,6-Dideoxy-4[[1S-(1,4,6/5)-4,5,6-trihydroxy-3-hydroxymethyl-2-cyclohexen-1-yl]-amino]-α-D-glucopyranosyl)-(1 ⊠ 4)-O-α-D-glucopyranosyl-(1 ⊠ 4)-D-glucopyranose) aus Kulturbrühen.** *German patent DE 2719912 (Process for isolating glucopyranose compound from culture broths; US patent 4,174,439)* 1978.

24. Hemker M, Stratmann A, Goeke K, Schröder W, Lenz J, Piepersberg W, Pape H: **Identification, cloning, expression, and characterization of the extracellular acarbose-modifying glycosyltransferase, AcbD, from *Actinoplane* sp. Strain SE50.** *J Bacteriol* 2001, **183**:4484-4492.

25. Zhang C-S, Stratmann A, Block O, Brückner R, Podeschwa M, Altenbach H-J, Wehmeier UF, Piepersberg W: **Biosynthesis of the C(7)-cyclitol moiety of acarbose in *Actinoplane* species SE50/110. 7-O-phosphorylation of the initial cyclitol precursor leads to proposal of a new biosynthetic pathway.** *J Biol Chem* 2002, **277**:22853-22862.

26. Zhang C-S, Podeschwa M, Altenbach H-J, Piepersberg W, Wehmeier UF: **The acarbose-biosynthetic enzyme AcbO from *Actinoplane* sp. SE 50/110**

is a 2-epi-5-epi-valiolone-7-phosphate 2-epimerase. *FEBS Lett* 2003, **540**:47-52.

27. Brunkhorst C, Wehmeier UF, Piepersberg W, Schneider E: **The *acb* gene of *Actinoplane* sp. encodes a solute receptor with binding activities for acarbose and longer homologs.** *Res Microbiol* 2005, **156**:322-327.

28. Wehmeier UF: **The Biosynthesis and Metabolism of Acarbose in *Actinoplane* sp. SE 50/110: A Progress Report.** *Biocatal Biotransformation* 2003, **21**:279-284.

29. Wehmeier UF, Piepersberg W: **Enzymology of aminoglycoside biosynthesis-deduction from gene clusters.** *Meth Enzymol* 2009, **459**:459-491.

30. Schwientek P, Szczepanowski R, Rückert C, Stoye J, Pühler A: **Sequencing of high G + C microbial genomes using the ultrafast pyrosequencing technology.** *J Biotechnol* 2011, **155**:68-77.

31. Galibert F, Finan TM, Long SR, Puhler A, Abola P, Ampe F, Barloy-Hubler F, Barnett MJ, Becker A, Boistard P, Bothe G, Boutry M, Bowser L, Buhrmester J, Cadieu E, Capela D, Chain P, Cowie A, Davis RW, Dreano S, Federspiel NA, Fisher RF, Gloux S, Godrie T, Goffeau A, Golding B, Gouzy J, Gurjal M, Hernandez-Lucas I, Hong A, Huizar L, Hyman RW, Jones T, Kahn D, Kahn ML, Kalman S, Keating DH, Kiss E, Komp C, Lelaure V, Masuy D, Palm C, Peck MC, Pohl TM, Portetelle D, Purnelle B, Ramsperger U, Surzycki R, Thebault P, Vandenbol M, Vorholter FJ, Weidner S, Wells DH, Wong K, Yeh KC, Batut J: **The composite genome of the legume symbiont *Sinorhizobium melilot*.** *Science* 2001, **293**:668-672.

32. Kalinowski J, Bathe B, Bartels D, Bischoff N, Bott M, Burkovski A, Dusch N, Eggeling L, Eikmanns BJ, Gaigalat L, Goesmann A, Hartmann M, Huthmacher K, Krämer R, Linke B, McHardy AC, Meyer F, Möckel B, Pfefferle W, Pühler A, Rey DA, Rückert C, Rupp O, Sahm H, Wendisch VF, Wiegräbe I, Tauch A: **The complete *Corynebacterium glutamicu* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins.** *J Biotechnol* 2003, **104**:5-25.

33. Schneiker S, dos Santos VAM, Bartels D, Bekel T, Brecht M, Buhrmester J, Chernikova TN, Denaro R, Ferrer M, Gertler C, Goesmann A, Golyshina OV, Kaminski F, Khachane AN, Lang S, Linke B, McHardy AC, Meyer F, Nechitaylo T, Puhler A, Regenhardt D, Rupp O, Sabirova JS, Selbitschka W, Yakimov MM, Timmis KN, Vorholter F-J, Weidner S, Kaiser O, Golyshin PN: **Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium *Alcanivorax borkumensi*.** *Nat Biotech* 2006, **24**:997-1004.

34. Tauch A, Trost E, Tilker A, Ludewig U, Schneiker S, Goesmann A, Arnold W, Bekel T, Brinkrolf K, Brune I, Götker S, Kalinowski J, Kamp P-B, Lobo FP, Viehoever P, Weisshaar B, Soriano F, Dröge M, Pühler A: **The lifestyle of *Corynebacterium urealyticu* derived from its complete genome sequence established by pyrosequencing.** *J Biotechnol* 2008, **136**:11-21.

35. Trost E, Götker S, Schneider J, Schneiker-Bekel S, Szczepanowski R, Tilker A, Viehoever P, Arnold W, Bekel T, Blom J, Gartemann K-H, Linke B, Goesmann A, Pühler A, Shukla SK, Tauch A: **Complete genome sequence and lifestyle of black-pigmented *Corynebacterium aurimucosu* ATCC 700975 (formerly *C. nigrican* CN-1) isolated from a vaginal swab of a woman with spontaneous abortion.** *BMC Genomics* 2010, **11**:91.

36. Krause A, Ramakumar A, Bartels D, Battistoni F, Bekel T, Boch J, Böhm M, Friedrich F, Hurek T, Krause L, Linke B, McHardy AC, Sarkar A, Schneiker S, Syed AA, Thauer R, Vorhölter F-J, Weidner S, Pühler A, Reinhold-Hurek B, Kaiser O, Goesmann A: **Complete genome of the mutualistic, N2-fixing grass endophyte *Azoarcu* sp. strain BH72.** *Nat Biotechnol* 2006, **24**:1385-1391.

37. Schneiker S, Perlova O, Kaiser O, Gerth K, Alici A, Altmeyer MO, Bartels D, Bekel T, Beyer S, Bode E, Bode HB, Bolten CJ, Choudhuri JV, Doss S, Elnakady YA, Frank B, Gaigalat L, Goesmann A, Groeger C, Gross F, Jelsbak L, Jelsbak L, Kalinowski J, Kegler C, Knauber T, Konietzny S, Kopp M, Krause L, Krug D, Linke B, Mahmud T, Martinez-Arias R, McHardy AC, Merai M, Meyer F, Mormann S, Muñoz-Dorado J, Perez J, Pradella S, Rachid S, Raddatz G, Rosenau F, Rückert C, Sasse F, Scharfe M, Schuster SC, Suen G, Treuner-Lange A, Velicer GJ, Vorhölter F-J, Weissman KJ, Welch RD, Wenzel SC, Whitworth DE, Wilhelm S, Wittmann C, Blöcker H, Pühler A, Müller R: **Complete genome sequence of the myxobacterium *Sorangium cellulosu*.** *Nat Biotechnol* 2007, **25**:1281-1289.

38. Gartemann K-H, Abt B, Bekel T, Burger A, Engemann J, Flügel M, Gaigalat L, Goesmann A, Gräfen I, Kalinowski J, Kaup O, Kirchner O, Krause L, Linke B, McHardy A, Meyer F, Pohle S, Rückert C, Schneiker S, Zellermann E-M, Pühler A, Eichenlaub R, Kaiser O, Bartels D: **The genome sequence of the tomato-pathogenic actinomycete *Clavibacter michiganensi* subsp.**

michiganensi NCPPB382 reveals a large island involved in pathogenicity. *J Bacteriol* 2008, **190**:2138-2149.

39. Vorhölter F-J, Schneiker S, Goesmann A, Krause L, Bekel T, Kaiser O, Linke B, Patschkowski T, Rückert C, Schmid J, Sidhu VK, Sieber V, Tauch A, Watt SA, Weisshaar B, Becker A, Niehaus K, Pühler A: **The genome of *Xanthomonas campestri* pv. *campestri* B100 and its use for the reconstruction of metabolic pathways involved in xanthan biosynthesis.** *J Biotechnol* 2008, **134**:33-45.

40. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** *Genome Res* 1998, **8**:195-202.

41. Chain PSG, Grafham DV, Fulton RS, FitzGerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, Cole JR, Ding Y, Dugan S, Field D, Garrity GM, Gibbs R, Graves T, Han CS, Harrison SH, Highlander S, Hugenholtz P, Khouri HM, Kodira CD, Kolker E, Kyrpides NC, Lang D, Lapidus A, Malfatti SA, Markowitz V, Metha T, Nelson KE, Parkhill J, Pitluck S, Qin X, Read TD, Schmutz J, Sozhamannan S, Sterk P, Strausberg RL, Sutton G, Thomson NR, Tiedje JM, Weinstock G, Wollam A, Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium, Detter JC: **Genome project standards in a new era of sequencing.** *Science* 2009, **326**:236-237.

42. Hyatt D, Chen G-L, LoCascio P, Land M, Larimer F, Hauser L: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *BMC Bioinforma* 2010, **11**:119.

43. Krause L, McHardy AC, Nattkemper TW, Pühler A, Stoye J, Meyer F: **GISMO-gene identification using a support vector machine for ORF classification.** *Nucleic Acids Res* 2007, **35**:540-549.

44. Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, Pühler A: **GenDB-an open source genome annotation system for prokaryote genomes.** *Nucleic Acids Res* 2003, **31**:2187-2195.

45. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29**:22-28.

46. Webb EC: *Enzyme Nomenclature 1992: Recommendations of the NCIUBMB on the Nomenclature and Classification of Enzymes.* 1 edition. Academic Press, San Diego, California; 1992.

47. Zawilak-Pawlik A, Kois A, Majka J, Jakimowicz D, Smulczyk-Krawczyszyn A, Messer W, Zakrzewska-Czerwińska J: **Architecture of bacterial replication initiation complexes: orisomes from four unrelated bacteria.** *Biochem J* 2005, **389**:471-481.

48. Hendrickson H, Lawrence JG: **Mutational bias suggests that replication termination occurs near the *di* site, not at Ter sites.** *Mol Microbiol* 2007, **64**:42-56.

49. Mehling A, Wehmeier UF, Piepersberg W: **Nucleotide sequences of streptomycete 16S ribosomal DNA: towards a specific identification system for streptomycetes using PCR.** *Microbiol (Reading, Engl)* 1995, **141**(Pt 9):2139-2147.

50. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25**:955-964.

51. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278**:631-637.

52. Bentley SD, Chater KF, Cerdeno-Tarraga A-M, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang C-H, Kieser T, Larke L, Murphy L, Oliver K, O'Neil R, Rabbinowitsch E, Rajandream M-A, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA: **Complete genome sequence of the model actinomycete *Streptomyces coelicolo* A3 (2).** *Nature* 2002, **417**:141-147.

53. Höfs R, Walker M, Zeeck A: **Hexacyclinic acid, a Polyketide from *Streptomyce* with a Novel Carbon Skeleton.** *Angew Chem Int Ed* 2000, **39**:3258-3261.

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

55. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM: **The Ribosomal Database Project: improved alignments and new tools for rRNA analysis.** *Nucleic Acids Res* 2009, **37**:D141-D145.

56. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**:406-425.

57. Felsenstein J: **Confidence limits on phylogenies: an approach using the bootstrap.** *Evolution* 1985, **3**:783-791.

58. Jukes TH, Cantor CR: **Evolution of protein molecules.** *Mammalian Protein Metabolism* New York: Academic Press; 1969, 21-132.

59. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24**:1596-1599.

60. Blom J, Albaum SP, Doppmeier D, Pühler A, Vorhölter F-J, Zakrzewski M, Goesmann A: **EDGAR: a software framework for the comparative analysis of prokaryotic genomes.** *BMC Bioinforma* 2009, **10**:154.

61. Roh H, Uguru GC, Ko H-J, Kim S, Kim B-Y, Goodfellow M, Bull AT, Kim KH, Bibb MJ, Choi I-G, Stach JEM: **Genome Sequence of the Abyssomicin- and Proximicin-Producing Marine Actinomycete *Verrucosispora mari* AB-18-032.** *J Bacteriol* 2011, **193**:3391-3392.

62. Baltz RH: **New genetic methods to improve secondary metabolite production in *Streptomyce*.** *J Ind Microbiol Biotechnol* 1998, **20**:360-363.

63. Baltz RH: **Genetic methods and strategies for secondary metabolite yield improvement in actinomycetes.** *Antonie Van Leeuwenhoek* 2001, **79**:251-259.

64. Olano C, Lombó F, Méndez C, Salas JA: **Improving production of bioactive secondary metabolites in actinomycetes by metabolic engineering.** *Metab Eng* 2008, **10**:281-292.

65. Li W, Xin Y, McNeil MR, Ma Y: *rml* and *rml* **genes are essential for growth of mycobacteria.** *Biochem Biophys Res Commun* 2006, **342**:170-178.

66. Piepersberg W, Diaz-Guardamino Uribe PM, Stratmann A, Thomas H, Wehmeier U, Zhang CS: **Developments in the Biosynthesis and Regulation of Aminoglycosides.** *Microbial Secondary Metabolites: Biosynthesis, Genetics and Regulation* Kerala, India: Research Signpost; 2002, 1-26.

67. Rockser Y, Wehmeier UF: **The *ga*-gene cluster for the production of acarbose from *Streptomyces glaucescen* GLA.O: identification, isolation and characterization.** *J Biotechnol* 2009, **140**:114-123.

68. Vahedi-Faridi A, Licht A, Bulut H, Scheffel F, Keller S, Wehmeier UF, Saenger W, Schneider E: **Crystal Structures of the Solute Receptor GacH of *Streptomyces glaucescen* in Complex with Acarbose and an Acarbose Homolog: Comparison with the Acarbose-Loaded Maltose-Binding Protein of *Salmonella typhimuriu*.** *J Mol Biol* 2010, **397**:709-723.

69. te Poele EM, Bolhuis H, Dijkhuizen L: **Actinomycete integrative and conjugative elements.** *Antonie Van Leeuwenhoek* 2008, **94**:127-143.

70. Burrus V, Waldor MK: **Shaping bacterial genomes with integrative and conjugative elements.** *Res Microbiol* 2004, **155**:376-386.

71. Raynal A, Friedmann A, Tuphile K, Guerineau M, Pernodet J-L: **Characterization of the *att* site of the integrative element pSAM2 from *Streptomyces ambofacien*.** *Microbiol (Reading, Engl)* 2002, **148**:61-67.

72. Kataoka M, Kiyose YM, Michisuji Y, Horiguchi T, Seki T, Yoshida T: **Complete Nucleotide Sequence of the *Streptomyces nigrifacien* Plasmid, pSN22: Genetic Organization and Correlation with Genetic Properties.** *Plasmid* 1994, **32**:55-69.

73. Grohmann E, Muth G, Espinosa M: **Conjugative plasmid transfer in gram-positive bacteria.** *Microbiol Mol Biol Rev* 2003, **67**:277-301.

74. Sezonov G, Hagège J, Pernodet JL, Friedmann A, Guérineau M: **Characterization of *pr*, a gene for replication control in pSAM2, the integrating element of *Streptomyces ambofacien*.** *Mol Microbiol* 1995, **17**:533-544.

75. Hosted TJ Jr, Wang T, Horan AC: **Characterization of the *Micromonospora rosari* pMR2 plasmid and development of a high G + C codon optimized integrase for site-specific integration.** *Plasmid* 2005, **54**:249-258.

76. Wagner N, Osswald C, Biener R, Schwartz D: **Comparative analysis of transcriptional activities of heterologous promoters in the rare actinomycete *Actinoplanes friuliensi*.** *J Biotechnol* 2009, **142**:200-204.

77. Challis GL, Ravel J, Townsend CA: **Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains.** *Chem Biol* 2000, **7**:211-224.

78. Hahn M, Stachelhaus T: **Selective interaction between nonribosomal peptide synthetases is facilitated by short communication-mediating domains.** *Proc Natl Acad Sci USA* 2004, **101**:15585-15590.

79. Meier JL, Burkart MD: **The chemical biology of modular biosynthetic enzymes.** *Chem Soc Rev* 2009, **38**:2012-2045.

80. Yadav G, Gokhale RS, Mohanty D: **Computational Approach for Prediction of Domain Organization and Substrate Specificity of Modular Polyketide Synthases.** *J Mol Biol* 2003, **328**:335-363.

81. Felnagle EA, Jackson EE, Chan YA, Podevels AM, Berti AD, McMahon MD, Thomas MG: **Nonribosomal Peptide Synthetases Involved in the Production of Medically Relevant Natural Products.** *Mol Pharm* 2008, **5**:191-211.

82. Du L, Lou L: **PKS and NRPS release mechanisms.** *Nat Prod Rep* 2010, **27**:255-278.

83. Copp JN, Neilan BA: **The Phosphopantetheinyl Transferase Superfamily: Phylogenetic Analysis and Functional Implications in Cyanobacteria.** *Appl Environ Microbiol* 2006, **72**:2298-2305.

84. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R: **antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences.** *Nucleic Acids Res* 2011, **39**:W339-46.

85. Medema MH, Trefzer A, Kovalchuk A, van den Berg M, Müller U, Heijne W, Wu L, Alam MT, Ronning CM, Nierman WC, Bovenberg RAL, Breitling R, Takano E: **The Sequence of a 1.8-Mb Bacterial Linear Plasmid Reveals a Rich Evolutionary Reservoir of Secondary Metabolic Pathways.** *Genome Biol Evol* 2010, **2**:212-224.

86. Baltz RH: **Function of MbtH homologs in nonribosomal peptide biosynthesis and applications in secondary metabolite discovery.** *J Ind Microbiol Biotechnol* 2011, **38**:1747-1760.

87. Anand S, Prasad MVR, Yadav G, Kumar N, Shehara J, Ansari MZ, Mohanty D: **SBSPKS: structure based sequence analysis of polyketide synthases.** *Nucleic Acids Res* 2010, **38**:W487-W496.

88. Kotowska M, Pawlik K, Butler AR, Cundliffe E, Takano E, Kuczek K: **Type II thioesterase from *Streptomyces coelicolo* A3(2).** *Microbiology* 2002, **148**:1777-1783.

89. Krätzschmar J, Krause M, Marahiel MA: **Gramicidin S biosynthesis operon containing the structural genes *grs* and *grs* has an open reading frame encoding a protein homologous to fatty acid thioesterases.** *J Bacteriol* 1989, **171**:5422-5429.

90. Dreier J, Khosla C: **Mechanistic Analysis of a Type II Polyketide Synthase. Role of Conserved Residues in the β-Ketoacyl Synthase Chain Length Factor Heterodimer.** *Biochemistry* 2000, **39**:2088-2095.

91. Wawrik B, Kerkhof L, Zylstra GJ, Kukor JJ: **Identification of Unique Type II Polyketide Synthase Genes in Soil.** *Appl Environ Microbiol* 2005, **71**:2232-2238.

92. Shen B, Du L, Sanchez C, Edwards DJ, Chen M, Murrell JM: **The biosynthetic gene cluster for the anticancer drug bleomycin from *Streptomyces verticillu* ATCC15003 as a model for hybrid peptide-polyketide natural product biosynthesis.** *J Ind Microbiol Biotechnol* 2001, **27**:378-385.

93. Mast Y, Weber T, Gölz M, Ort-Winklbauer R, Gondran A, Wohlleben W, Schinko E: **Characterization of the "pristinamycin supercluster" of *Streptomyces pristinaespirali*.** *Microb Biotechnol* 2011, **4**:192-206.

94. Du L, Sánchez C, Shen B: **Hybrid peptide-polyketide natural products: biosynthesis and prospects toward engineering novel molecules.** *Metab Eng* 2001, **3**:78-95.

95. Melançon CE, Liu H-W: **Engineered biosynthesis of macrolide derivatives bearing the non-natural deoxysugars 4-epi-D-mycaminose and 3-n-monomethylamino-3-deoxy-D-fucose.** *J Am Chem Soc* 2007, **129**:4896-4897.

96. Oh T-J, Mo SJ, Yoon YJ, Sohng JK: **Discovery and molecular engineering of sugar-containing natural product biosynthetic pathways in actinomycetes.** *J Microbiol Biotechnol* 2007, **17**:1909-1921.

97. Ewing B, Green P: **Base-calling of automated sequencer traces using phred II. Error probabilities.** *Genome Res* 1998, **8**:186-194.

98. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred I. Accuracy assessment.** *Genome Res* 1998, **8**:175-185.

99. Gordon D, Desmarais C, Green P: **Automated finishing with autofinish.** *Genome Res* 2001, **11**:614-625.

100. McHardy AC, Goesmann A, Pühler A, Meyer F: **Development of joint application strategies for two microbial gene finders.** *Bioinformatics* 2004, **20**:1622-1631.

101. Linke B, McHardy AC, Neuweger H, Krause L, Meyer F: **REGANOR: a gene prediction server for prokaryotic genomes and a database of high quality gene predictions for prokaryotes.** *Appl Bioinformatics* 2006, **5**:193-198.

102. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER.** *Nucleic Acids Res* 1999, **27**:4636-4641.

103. Badger JH, Olsen GJ: **CRITICA: coding region identification tool invoking comparative analysis.** *Mol Biol Evol* 1999, **16**:512-524.

104. Coulson A: **High-performance searching of biosequence databases.** *Trends Biotechnol* 1994, **12**:76-80.

105. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH: **CDD: a database of conserved domain alignments with links to domain three-dimensional structure.** *Nucleic Acids Res* 2002, **30**:281-283.

106. Bairoch A: **The ENZYME database in 2000.** *Nucleic Acids Res* 2000, **28**:304-305.

107. Claudel-Renard C, Chevalet C, Faraut T, Kahn D: **Enzyme-specific profiles for genome annotation: PRIAM.** *Nucleic Acids Res* 2003, **31**:6633-6639.

108. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28**:27-30.

109. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006, **34**:D354-D357.

110. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinforma* 2003, **4**:41.

111. Sonnhammer EL, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences.** *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:175-182.

112. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305**:567-580.

113. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Protein Eng* 1997, **10**:1-6.

114. Nielsen H, Krogh A: **Prediction of signal peptides and signal anchors by a hidden Markov model.** *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:122-130.

115. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340**:783-795.

116. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S: **Non-classical protein secretion in bacteria.** *BMC Microbiol* 2005, **5**:58.

117. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J: **Operons in *Escherichia col*: genomic analyses and predictions.** *Proc Natl Acad Sci USA* 2000, **97**:6652-6657.

118. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31**:3406-3415.

119. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792-1797.

120. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinforma* 2004, **5**:113.