Journal of Biotechnology 251 (2017) 112–123

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Genome improvement of the acarbose producer *Actinoplanes* sp. SE50/110 and annotation refinement based on RNA-seq analysis

Timo Wolf[a], Susanne Schneiker-Bekel[b], Armin Neshat[a], Vera Ortseifen[b], Daniel Wibberg[b], Till Zemke[c], Alfred Pühler[b], Jörn Kalinowski[a,*]

[a] *Microbial Genomics and Biotechnology, Center for Biotechnology, Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany*
[b] *Senior Research Group in Genome Research of Industrial Microorganisms, Center for Biotechnology, Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany*
[c] *Product Supply, Bayer Pharma AG, Friedrich Ebert Str. 217-475, 42117 Wuppertal, Germany*

ARTICLE INFO

Keywords:
Actinoplanes
Acarbose
RNA-seq
Cis-regulatory elements
Attenuation
Secondary metabolite gene clusters

ABSTRACT

Actinoplanes sp. SE50/110 is the natural producer of acarbose, which is used in the treatment of diabetes mellitus type II. However, until now the transcriptional organization and regulation of the acarbose biosynthesis are only understood rudimentarily. The genome sequence of *Actinoplanes* sp. SE50/110 was known before, but was resequenced in this study to remove assembly artifacts and incorrect base callings. The annotation of the genome was refined in a multi-step approach, including modern bioinformatic pipelines, transcriptome and proteome data. A whole transcriptome RNA-seq library as well as an RNA-seq library enriched for primary 5′-ends were used for the detection of transcription start sites, to correct tRNA predictions, to identify novel transcripts like small RNAs and to improve the annotation through the correction of falsely annotated translation start sites. The transcriptome data sets were also applied to identify 31 *cis*-regulatory RNA structures, such as riboswitches or RNA thermometers as well as three leaderless transcribed short peptides found in putative attenuators upstream of genes for amino acid biosynthesis. The transcriptional organization of the acarbose biosynthetic gene cluster was elucidated in detail and fourteen novel biosynthetic gene clusters were suggested. The accurate genome sequence and precise annotation of the *Actinoplanes* sp. SE50/110 genome will be the foundation for future genetic engineering and systems biology studies.

1. Introduction

The genus *Actinoplanes* is assigned to the family *Micromonosporaceae* within the phylum of *Actinobacteria* (Ludwig et al., 2012). The species associated to this genus typically grow in thin hyphae, form spores and are characterized by their extraordinarily high G + C content of 69–73% (Vobis et al., 2012). *Actinoplanes* species and other rare actinomycetes exhibit a high potential for the production and discovery of active secondary metabolites. More than 120 antibiotics, like teicoplanin (Bardone et al., 1978), actaplanin (Debono et al., 1984), ramoplanin (Ciabatti et al., 1989) and friulimicins (Aretz et al., 2000) are already known to be produced by *Actinoplanes* spp. Nevertheless, the rare actinomycetes are highly potential candidates for the discovery of new biotechnologically relevant products (Okami and Hotta, 1988; Vobis, 2006).

The Gram-positive actinobacterium *Actinoplanes* sp. SE50/110 is known as a natural producer of acarbose (acarviosyl-1,4-maltose), a pharmaceutically relevant pseudo-tetrasaccharide (Truscheit et al., 1981). Acarbose inhibits alpha-glucosidases of the human intestine, by which the absorption of monosaccharides from starch-containing diets is reduced (Bischoff 1994). Due to this, acarbose is used for the treatment of diabetes mellitus type II as it supports the reduction of blood sugar levels (Creutzfeldt, 1988; Wehmeier and Piepersberg, 2004).

The genome of *Actinoplanes* sp. SE50/110 was first sequenced combining the sequencing data generated by paired-end and whole-genome shotgun 454 pyro-sequencing strategies, followed by a manual assembly using PCR products. This first version of the genome (GenBank: CP003170) consisted of one circular chromosome with a size of 9,239,851 bp and a high mean G + C content of 71.32%. In total, 8270 protein coding sequences were predicted including the genes for the well-known acarbose biosynthetic gene cluster as well as four other putative secondary metabolite gene clusters (Schwientek et al., 2012). By sequencing enriched 5′-ends of primary transcripts, 1427 transcription start sites (TSS) were identified and used for a first improvement of the *Actinoplanes* sp. SE50/110 genome (Schwientek

* Corresponding author.
E-mail address: Joern.Kalinowski@Cebitec.Uni-Bielefeld.de (J. Kalinowski).

http://dx.doi.org/10.1016/j.jbiotec.2017.04.013
Received 1 December 2016; Received in revised form 11 April 2017; Accepted 14 April 2017
Available online 17 April 2017
0168-1656/

et al., 2014). Recently, while performing a genome resequencing for the evaluation of off-target effects in genome edited mutants, several single nucleotide polymorphisms (SNPs) were detected both in the parental wild type strain as well as the mutant strain (Wolf et al., 2016). Thereby, the necessity of a corrected genome sequence became apparent.

Sequence-specific motifs, like repeats and homopolymers or high G + C contents can lead to sequencing errors, which can result in misassemblies, assembly artifacts as well as incorrect base callings (Luo et al., 2012; Allhoff et al., 2013; Shin and Park, 2016). Especially pyrosequencing of high G + C content genomes is challenging and requires substantial manual work to finish these genomes. The high G + C content can lead to stable secondary structures, which result in low read coverages and consequently, to a high number of contiguous sequences and gaps in the genome (Frey et al., 2008; Schwientek et al., 2011). Although the protocol for sequencing the whole genome of *Actinoplanes* sp. SE50/110 was optimized through a modified sequencing chemistry, the finishing of the genome had to be performed manually. Therefore gaps between contiguous sequences were closed through Sanger sequencing of PCR amplicons (Schwientek et al., 2012).

Annotation by gene prediction software only detects a small set of non-coding and small RNAs, does not detect small genes, neglects untranslated regions (UTRs) and can be biased (Sorek and Cossart, 2010). Whole transcriptome analyses, like RNA sequencing (RNA-seq), can be used to refine and expand automatic gene prediction tools as well as unlock complete transcriptomic landscapes. RNA-seq has been used to reveal a high level complexity of bacterial transcriptomes, like antisense transcription, non-coding RNAs, *cis*-regulatory elements or alternative operon structures (Sharma et al., 2010; Croucher and Thomson, 2010; Guell et al., 2009; Filiatrault et al., 2010; Pfeifer-Sancar et al., 2013; Irla et al., 2015).

In this study, the genomic sequence of *Actinoplanes* sp. SE50/110 was updated through single nucleotide corrections and by removing assembly artifacts. Therefore the genome was resequenced on an Illumina MiSeq System using a paired-end PCR-free sequencing library. The automated annotation of the circular genome was improved through modern bioinformatics tools and refined through proteomics as well as high quality transcriptomic data. Separate RNA-seq data sets were generated for the analysis of the whole transcriptome and the primary transcriptome for a genome wide identification of transcription start sites. The comprehensive information was used to verify and correct the automatic tRNA prediction, to identify novel transcripts like coding sequences and small RNAs, to improve the annotation through the correction of wrongly annotated translation start sites and to analyze *cis*-regulatory elements like riboswitches and peptide leader structures, acting as attenuators. Moreover, the transcription organization of the acarbose biosynthetic gene cluster was deciphered, multiple novel biosynthetic gene clusters for secondary metabolites were identified and examined for transcription.

2. Methods

2.1. Cultivation conditions of Actinoplanes sp. SE50/110

Actinoplanes sp. SE50/110 (ATCC 31044) was grown in NBS complex medium (10 g/L glucose, 4 g/L peptone, 4 g/L yeast extract, 1 g/L $MgSO_4 \cdot 7H_2O$, 2 g/L KH_2PO_4, 5.2 g/L $K_2HPO_4 \cdot 3H_2O$) for subsequent DNA and RNA isolation. Minimal medium was supplemented with 2.4 C-mole of glucose, maltose or galactose as sugar sources for additional cultivation conditions for RNA isolation. All cultivations were carried out at 28 °C and 140 rpm (GFL shaking incubator 3032) in baffled polycarbonate flasks (Corning, Corning, NY, USA). The exact composition of the minimal medium as well as the cultivation conditions and the performed inoculation strategy are described elsewhere (Wendler et al., 2013). DNA was isolated directly from freshly grown cells as described previously (Wolf et al., 2016). For RNA isolation, 1 mL of cell suspension was centrifuged for 15 s at 16.000 *g* and immediately frozen in liquid nitrogen. Cell pellets were stored at −80 °C until the RNA was isolated.

2.2. Resequencing of Actinoplanes sp. SE50/110

For the resequencing of the genome, libraries were prepared using a TruSeq DNA PCR-Free Library Preparation Kit (550 bp target insert size). Subsequent paired-end sequencing was performed on an Illumina MiSeq System (Illumina, San Diego, CA, USA) using the MiSeq Reagent v3 Kit (Illumina, San Diego, CA, USA). The acquired read length was 2 × 300 bp. Base calling was performed with an in-house software package (Wibberg et al., 2016).

2.2.1. Sequence assembly and synteny inspection of Actinoplanes sp. SE50/110

The gsAssembler software (Newbler) v2.8 was applied to assemble the reads. R2cat (Husemann and Stoye, 2010) was used for synteny inspection by comparing the acquired contiguous sequences (contigs) to the published reference genome of *Actinoplanes* sp. SE50/110 (GenBank CP003170).

2.2.2. Read mapping, coverage analyses and genome improvement of Actinoplanes sp. SE50/110

Afterwards the reads were paired, trimmed (error probability limit of 0.01) and mapped (Geneious mapper with standard settings for medium sensitivity with a minimum overlap identity of 90% and up to 5 iterations) to the reference genome by using Geneious 9.1.3 (Kearse et al., 2012). Coverage analysis was performed with a minimal cut-off of 35 reads and Geneious default settings for all other parameters. The detected suspicious regions were manually inspected with the help of the assembly software package consed (Gordon et al., 1998) in the original genome assembly data used by Schwientek et al. (2012). All regions, in which fewer than 35 reads were mapped, were manually inspected for noticeable amounts of mismatches or gaps in the mapping. A variant calling was applied with settings for minimum coverage of 10, a minimum variant frequency of 0.9, a maximum variant p-value of 1×10^{-6} and Geneious standard settings for all other parameters. All confirmed rearrangements, insertions, deletions and substitutions, which were conducted in the genome sequence, were verified through re-mappings of the sequencing reads to the updated genome with the described parameters.

2.3. Genome annotation

The software pipeline prokka version 1.11 (Seemann 2014) was used for the automatic annotation of the updated *Actinoplanes* sp. SE50/110 genome sequence. In the next step, the data was imported into the annotation platform GenDB 2.0 (Meyer et al., 2003). A reference annotation based on a RefSeq annotation (NC_017803) and manually annotated genes as well as further manual annotation refinements were performed within GenDB 2.0.

2.4. Total RNA isolation and sequencing of cDNA libraries made from mRNA

RNA was isolated using a Qiagen RNeasy mini kit in combination with an RNase-free DNase kit (Qiagen, Hilden, Germany). Absence of DNA was proven by PCR with primers binding to genomic *Actinoplanes* sp. SE50/110 DNA. RNA quantity as well as quality were checked with a NanoDrop 1000 spectrometer (Peqlab, Erlangen, Germany) and an Agilent RNA 6000 Pico kit run on an Agilent Bioanalyzer 2100 (Agilent Technologies, Böblingen, Germany). RNA was isolated from *Actinoplanes* sp. SE50/110 shake flask cultures grown in complex media and three separate minimal media, supplemented with glucose, maltose and galactose. Of each cultivation condition, RNA was isolated from