Journal of Biotechnology xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biotechnology

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

# Improving the genome annotation of the acarbose producer *Actinoplanes* sp. SE50/110 by sequencing enriched 5′-ends of primary transcripts

Patrick Schwientek[a,b,1], Armin Neshat[c,1], Jörn Kalinowski[c], Andreas Klein[d], Christian Rückert[c], Susanne Schneiker-Bekel[a], Sergej Wendler[a], Jens Stoye[b,e], Alfred Pühler[a,*]

[a] *Genome Research of Industrial Microorganisms, Center for Biotechnology, Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany*
[b] *Genome Informatics Research Group, Faculty of Technology, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany*
[c] *Microbial Genomics and Biotechnology, Center for Biotechnology, Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany*
[d] *Bayer HealthCare AG, Friedrich-Ebert-Str. 475, 42117 Wuppertal, Germany*
[e] *Institute for Bioinformatics, Center for Biotechnology, Bielefeld University, Universitätsstraße 27, 33615 Bielefeld, Germany*

ARTICLE INFO

*Article history:*
Received 17 December 2013
Received in revised form 25 February 2014
Accepted 1 March 2014
Available online xxx

*Keywords:*
*Actinoplanes*
Acarbose production
Transcriptomics
Sequencing of 5′-ends of primary transcripts
Genome annotation improvement

ABSTRACT

*Actinoplanes* sp. SE50/110 is the producer of the alpha-glucosidase inhibitor acarbose, which is an economically relevant and potent drug in the treatment of type-2 diabetes mellitus. In this study, we present the detection of transcription start sites on this genome by sequencing enriched 5′-ends of primary transcripts. Altogether, 1427 putative transcription start sites were initially identified. With help of the annotated genome sequence, 661 transcription start sites were found to belong to the leader region of protein-coding genes with the surprising result that roughly 20% of these genes rank among the class of leaderless transcripts. Next, conserved promoter motifs were identified for protein-coding genes with and without leader sequences. The mapped transcription start sites were finally used to improve the annotation of the *Actinoplanes* sp. SE50/110 genome sequence. Concerning protein-coding genes, 41 translation start sites were corrected and 9 novel protein-coding genes could be identified. In addition to this, 122 previously undetermined non-coding RNA (ncRNA) genes of *Actinoplanes* sp. SE50/110 were defined. Focusing on antisense transcription start sites located within coding genes or their leader sequences, it was discovered that 96 of those ncRNA genes belong to the class of antisense RNA (asRNA) genes. The remaining 26 ncRNA genes were found outside of known protein-coding genes. Four chosen examples of prominent ncRNA genes, namely the transfer messenger RNA gene *ssrA*, the ribonuclease P class A RNA gene *rnpB*, the cobalamin riboswitch RNA gene *cobRS*, and the selenocysteine-specific tRNA gene *selC*, are presented in more detail. This study demonstrates that sequencing of enriched 5′-ends of primary transcripts and the identification of transcription start sites are valuable tools for advanced genome annotation of *Actinoplanes* sp. SE50/110 and most probably also for other bacteria.



## 1. Introduction

The pseudotetrasaccharide acarbose is an alpha-glucosidase inhibitor which is used as an oral drug in the treatment of diabetes mellitus type 2 (Schwientek et al., 2013). Worldwide the prevalence of diabetes type 2 is rapidly rising, with no immediate end of this development is in sight. According to the International Diabetes Federation, diabetes mellitus type 2 currently affects over 340 million people worldwide (IDF, 2013). Due to this, a continually increasing demand for diabetes drugs such as acarbose needs to be anticipated. Currently, acarbose is produced by industrial fermentation of yield-optimized *Actinoplanes* strains which were derived from the strain *Actinoplanes* sp. SE50/110. Originally, *Actinoplanes* sp. SE50 was isolated from a soil sample during a screening programme of the Bayer AG for new substances with inhibitory effects on glycoside hydrolases. It was observed that the natural variant *Actinoplanes* sp. SE50/110 produced elevated levels of acarbose (Truscheit et al., 1981). During the further strain development, *Actinoplanes* sp. SE50/110 was very

* Corresponding author. Tel.: +49 521 106 8750; fax: +49 521 106 89046.
*E-mail address:* Puehler@CeBiTec.Uni-Bielefeld.de (A. Pühler).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jbiotec.2014.03.013
0168-1656/

**Table 1**
RNA-sequencing results using 5′-enriched cDNA libraries prepared from *Actinoplanes* sp. SE50/110 cells.

| Condition | Sequenced reads | Mappable reads | Unique matches | Map rate |
|---|---|---|---|---|
| Mal-MM | 7,889,721 | 810,404 (10.27%) | 553,311 (7.01%) | 2.25× |
| Mal-MM-TE | 932,245 | 158,447 (17.00%) | 115,659 (12.41%) | 2.19× |
| Glc-CM | 220,625 | 109,589 (49.67%) | 34,492 (15.63%) | 3.34× |
| Total | 9,042,591 | 1,078,440 (11.93%) | 703,462 (7.78%) | 2.35× |

successfully optimized through conventional mutagenesis to increase the production of acarbose. However, this strategy seems to have reached its limits by now. As a first step towards rational designed pathways for the optimization of acarbose production, the complete genome sequence was determined and ribosomal, tRNA, and protein-coding genes were annotated (Schwientek et al., 2012). It should be mentioned that the high G + C-content of the *Actinoplanes* sp. SE50/110 genome leads to the formation of regions with very stable secondary structures, which represent a major challenge for sequencing (Schwientek et al., 2011). The annotated genome sequence of *Actinoplanes* sp. SE50/110 was subsequently used to analyze the cytosolic and extracellular proteome (Wendler et al., 2013). In order to gain first insights into the differential expression of genes involved in the acarbose biosynthesis under different growth conditions, high-throughput sequencing of cDNA libraries, commonly referred to as RNA-sequencing or RNA-seq, was performed recently (Schwientek et al., 2013).

RNA-sequencing is an emerging approach which – in the few years since its initial application – has revolutionized the identification and quantification of entire transcriptomes (Wang et al., 2009; Westermann et al., 2012). The method offers multiple advantages in comparison to the other established approaches for transcriptome analysis such as Northern blots, RACE (rapid amplification of cDNA ends), microarrays or reverse-transcription qPCR. RNA-seq provides accurate and extensive characterization of transcriptomes with a single-base resolution (Ozsolak and Milos, 2011). As an efficient and universally applicable technique in the biotechnology toolbox, RNA-sequencing is especially suited for analyzing the usually uncharted transcriptional landscape and for improving bacterial genome annotation.

In this paper, we report on the assignment of transcription start sites by sequencing and mapping enriched 5′-ends of primary transcripts to the genome of *Actinoplanes* sp. SE50/110. The location and number of read starts of transcription start sites were used to determine the percentage of leaderless transcripts and to analyze conserved promoter motifs of genes with and without 5′-untranslated regions. It was expected that the identification and analysis of transcription start sites will help to improve the annotation of the *Actinoplanes* sp. SE50/110 genome by correcting translation start sites and by identifying novel protein-coding genes. Additionally, an important goal of this study was the identification of non-coding RNA genes in intergenic regions or of antisense RNA genes located within known genes.

## 2. Materials and methods

### 2.1. Cultivation of Actinoplanes sp. SE50/110 in different growth media

The exact media composition and the detailed procedure of the performed *Actinoplanes* sp. SE50/110 cultivations have been described previously (Schwientek et al., 2013). Three different growth media were used: a defined minimal medium (Mal-MM) with maltose as sole carbon source, the identical medium supplemented with trace elements (Mal-MM-TE), and a complex medium with yeast extract, casein peptone, and glucose as main carbon source (Glc-CM). *Actinoplanes* sp. SE50/110 cultures were grown for four days in four replicates for each of the Mal-MM and Mal-MM-TE conditions as well as in two replicates for the Glc-CM condition. For each cultivation, 1.5 ml aliquots of cells were centrifuged for 3 s at 20,000 × *g* and the supernatant discarded. The resulting pellets were frozen in liquid nitrogen and stored at −80 °C until further use.

### 2.2. Total RNA isolation from Actinoplanes sp. SE50/110

Prior to RNA isolation, for each condition the harvested cells of the replicates were pooled. The procedures for total RNA isolation with TRIzol reagent (Life Technologies, Darmstadt, Germany) and RNeasy® Mini Kit (QIAGEN, Hilden, Germany) have been described in detail earlier (Schwientek et al., 2013). The quality and quantity of the isolated RNA was assessed by PCR with gene-specific primers and with capillary gel electrophoresis on Agilent Bioanalyzer 2100 using the Agilent RNA 6000 Pico kit (Agilent Technologies, Böblingen, Germany).

### 2.3. Preparation of enriched 5′-ends cDNA libraries for high-throughput sequencing

The cDNA library preparations of enriched 5′-ends of primary transcripts were performed as described before (Schwientek et al., 2013). The only difference was that an additional digestion step using Terminator™ 5′-Phosphate-Dependent Exonuclease treatment (Epicentre Biotechnologies, Madison, WI, USA) was introduced after application of the RiboZero™ rRNA removal kit (Epicentre Biotechnologies, Madison, WI, USA). The cDNA library preparation was carried out by applying the TruSeq RNA sample prep kit (Illumina, San Diego, CA, USA). The libraries were then loaded on a single lane of an eight-lane flow cell and sequenced according to the manufacturer's instruction on a Genome Analyser *IIx* device (Illumina, San Diego, CA, USA) as a single-read sequencing workflow with 36 bp read length. The sequencing results are summarized in Table 1.

### 2.4. Mapping of short reads to the Actinoplanes sp. SE50/110 reference genome

In order to perform further analyses on the RNA-seq data, the generated sequencing reads had to be mapped to the *Actinoplanes* sp. SE50/110 reference genome. In a preprocessing step the last base of all sequence reads was trimmed because of its high error rate (Kircher et al., 2009). Subsequently, the reads were mapped to the *Actinoplanes* sp. SE50/110 genome (GenBank accession no. CP003170.1) using the exact mapping algorithm SARUMAN (Blom et al., 2011) and allowing a single mismatch to occur in the alignment of a read. For visualization of the mapped short reads, the JAVA software ReadXplorer (http://readxplorer.org/) was used (Hilker et al., submitted).

### 2.5. Determination of putative transcription start sites

To determine putative transcription start sites (TSS), first, for each strand and position of the *Actinoplanes* sp. SE50/110 genome, all mapped RNA-seq reads were counted. The sum of the absolute

Please cite this article in press as: Schwientek, P., et al., Improving the genome annotation of the acarbose producer *Actinoplanes* sp. SE50/110 by sequencing enriched 5′-ends of primary transcripts. J. Biotechnol. (2014), http://dx.doi.org/10.1016/j.jbiotec.2014.03.013