High-resolution genome and transcriptome analysis of *Gluconobacter oxydans* 621H and growth-improved strains by next-generation sequencing

Inaugural Dissertation

submitted to

the Faculty of Mathematics and Natural Sciences

of the Heinrich Heine University

presented by

Angela Kranz

born in Alfeld/Leine

Jülich, September 2017

The thesis in hand has been performed at the

Institute of Bio- and Geosciences, IBG-1: Biotechnology, Research Centre Jülich GmbH, from May 2014 until September 2017 under the supervision of Prof. Dr. Michael Bott.

Printed with the permission of

the Faculty of Mathematics and Natural Sciences

of the Heinrich Heine University Düsseldorf

Examiner: **Prof. Dr. Michael Bott**
Institute of Bio- and Geosciences, IBG-1: Biotechnology
Research Centre Jülich GmbH

Co-examiner: **Jun.-Prof. Dr. Ilka Maria Axmann**
Institute of Synthetic Microbiology
Heinrich Heine University Düsseldorf

Date of oral examination: 26th February 2018

Results described in this thesis have been published with revision changes in the following original publications:

Kranz, A., Vogel, A., Degner, U., Kiefler, I., Bott, M., Usadel, B., and Polen, T. (2017). High precision genome sequencing of engineered *Gluconobacter oxydans* 621H by combining long nanopore and short accurate Illumina reads. **J Biotechnol 258: 197-205.**

doi: 10.1016/j.jbiotec.2017.04.016

Kranz, A., Busche, T., Vogel, A., Usadel, B., Kalinowski, J., Bott, M., and Polen, T. (2018). RNAseq analysis of α-proteobacterium *Gluconobacter oxydans* 621H. **BMC Genomics 19:24.**

doi: 10.1186/s12864-017-4415-x

Kranz, A., Steinmann, A., Degner, U., Mengus-Kaya, A., Matamouros, S., Bott, M., and Polen, T. (2018). Global mRNA decay analysis and 23S rRNA fragmentation in *Gluconobacter oxydans* 621H. **BMC Genomics (accepted).**

Table of contents

Abstract

The acetic acid bacterium *Gluconobacter oxydans* is an important organism used in industrial biotechnology. It is characterized by its exceptional ability to regio- and stereoselectively oxidize a broad range of substrates in the periplasm. However, a disadvantage of this bacterium is the low final biomass yield on sugar-containing complex media. Recently, metabolic engineering allowed construction of strain IK003.1 with a 60% increased biomass yield on glucose. Modifying metabolism in such a way may affect the genome stability and may cause suppressor mutations. Therefore, one aim of this thesis was the use of next-generation and nanopore sequencing to sequence the genomes of engineered and reference strains in order to detect mutations. Except for the introduced genetic alterations and one mobile element insertion, no further mutations were found in strain IK003.1 in comparison to the reference strains. This suggests that the constructed strain is quite stable and therefore well suited for further metabolic engineering efforts. Furthermore, the new sequencing results were used to update the reference genome sequence of *G. oxydans* 621H.

The second part of this thesis dealt with comprehensive RNAseq analysis to characterize the transcriptional landscapes of *G. oxydans*. This resulted in the detection of 2,449 transcription start sites (TSSs) and allowed to define the -10 region “nATnnn” and the -35 region “ttGnnn” as promoter consensus sequences. Analysis of 5´-UTRs also showed that 5% of all transcripts with an identified TSS are leaderless and 43% are longer than 100 nt. Furthermore, 971 potential novel transcripts were identified. 1,144 genes (41%) were found to be expressed monocistronically, whereas 1,634 genes (59%) belonged to 571 operons. Also, TSSs within operons indicated expression of 720 genes in 341 sub-operons.

The stability of mRNAs plays an important role in the post-transcriptional regulation of gene expression and can influence the production rate of proteins and growth of bacteria. Using DNA microarrays, we determined mRNA half-lives for 2,500 genes (95%) and analysed them based on a functional categorization. Furthermore, we observed instability of the 23S rRNA. Next-generation sequencing of rRNA isolated from enriched ribosomes revealed a distinct fragmentation pattern and indicated the presence of three fragmentation positions in three 23S rRNAs and four fragmentation positions in one 23S rRNA of *G. oxydans*.

Zusammenfassung

Das Essigsäurebakterium *Gluconobacter oxydans* ist ein wichtiger Organismus für die industrielle Biotechnologie. Es zeichnet sich besonders durch die Fähigkeit aus, ein breites Spektrum an Substraten im Periplasma regio- und stereoselektiv oxidieren zu können. Ein Nachteil dieses Bakteriums ist allerdings die geringe Biomasse-Ausbeute auf zuckerhaltigen Komplexmedien. In jüngster Zeit erlaubte Metabolic Engineering die Konstruktion des Stammes IK003.1, der eine um 60% gesteigerte Biomasse-Ausbeute auf Glucose-haltigem Medium zeigt. Solche Veränderungen im Metabolismus können die Genomstabilität beeinflussen und Suppressor-Mutationen hervorrufen. Ein Ziel dieser Arbeit war daher die Genom-Sequenzierung der konstruierten Stämme sowie von Referenz-Stämmen unter Verwendung von Next-Generation- und Nanopore-Sequenzierung, um Mutationen zu detektieren. Abgesehen von den eingeführten genetischen Veränderungen sowie einer strukturellen Variante verursacht durch ein mobiles genetisches Element, wurden keine zusätzlichen Mutationen im Stamm IK003.1 im Vergleich zu den Referenzstämmen gefunden. Dies weist darauf hin, dass der konstruierte Stamm ziemlich stabil ist und daher gut für weitere Stammentwicklungen geeignet ist. Weiterhin resultierten die Ergebnisse der Sequenzierungen in einer Aktualisierung der Referenzsequenz von *G. oxydans* 621H.

Der zweite Teil dieser Arbeit umfasste ausführliche RNA-Seq Analysen, um die transkriptionelle Landschaft von *G. oxydans* zu charakterisieren. Dies führte zu der Detektion von 2449 Transkriptionsstartpunkten (TSSs) und erlaubte die Definierung der -10 Region „nAtnnn“ und der -35 Region „ttGnnn“ als Promoter-Konsensusmotiv. Analyse der 5´-untranslatierten Regionen (5´-UTRs) zeigte außerdem, dass 5% aller Transkripte mit einem identifizierten TSS leaderless sind und dass 43% länger als 100 nt sind. Außerdem wurden 971 neue Transkripte identifiziert. 1144 Gene (41%) sind als Einzelgene exprimiert, während 1634 Gene (59%) 571 Operons zugeordnet werden konnten. TSSs innerhalb von Operons deuteten weiterhin auf die Expression von 720 Genen in 341 Sub-Operons hin.

Die Stabilität von mRNAs spielt eine große Rolle für die post-transkriptionelle Regulation von Genexpression und kann daher Einfluss auf die Produktionsrate von Proteinen und auf das Wachstum von Bakterien nehmen. Mittels DNA Microarrays konnten mRNA Halbwertszeiten für 2500 Gene (95%) bestimmt und hinsichtlich ihrer funktionellen Kategorisierung analysiert werden. Außerdem wurde eine Instabilität der 23S rRNA beobachtet. Next Generation Sequencing der aus angereicherten Ribosomen isolierten rRNA zeigte ein ausgeprägtes Fragmentierungsmuster und deutete die Anwesenheit von drei Fragmentierungspositionen in drei 23S rRNAs und vier Fragmentierungspositionen in einer 23S rRNA von *G. oxydans* an.

Abbreviations

CDS	Coding sequence
CO_2	Carbon dioxide
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
EDP	Entner-Doudoroff pathway
EMP	Embden-Meyerhof-Parnas pathway
et al.	*et alii*
$FADH_2$	Flavin adenine dinucleotide
mRNA	Messenger RNA
NADH	Nicotinamide adenine dinucleotide
NADPH	Nicotinamide adenine dinucleotide phosphate
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
PCR	Polymerase chain reaction
PPP	Pentose phosphate pathway
PQQ	Pyrroloquinoline quinone
RBS	Ribosome binding sites
RNA	Ribonucleic acid
RNAP	RNA polymerase
RNase	Ribonuclease
rRNA	ribosomal RNA
SBL	Sequencing-by-ligation
SBS	Sequencing-by-synthesis
TCA cycle	Tricarboxylic acid cycle
tRNA	transfer RNA
TSS	Transcription start site
UTR	Untranslated region

Further abbreviations not included in this section are according to international standards, for example listed in the author guidelines of the FEBS Journal.

1. Scientific context and key results of this thesis

1.1 *Gluconobacter oxydans* as a microbial cell factory for oxidative biotransformations – relevance of transcriptome and genome analysis

1.1.1 Characteristics and industrial use of *G. oxydans*

Gluconobacter oxydans 621H is a Gram-negative, strictly aerobic acetic acid bacterium. The cells of this α-proteobacterium are ellipsoidal to rod-shaped with a length of up to 4.2 µm (De Ley, 1961; Yamada et al., 1997). *G. oxydans* prefers sugar-containing habitats like flowers, fruits, and alcoholic beverages. By populating fruits, especially apples and pears, it causes bacterial rots with a characteristic browning of the fruits (De Ley, 1961). In accordance with its favorite niches, *G. oxydans* is cultivated in media with a high content of sugars or sugar alcohols, for example $_{D}$-glucose, $_{D}$-mannitol, or $_{D}$-fructose (De Ley and Schwings, 1984; Gosselé et al., 1980; Olijve and Kok, 1979; Uspenskaia and Loitsianskaia, 1979). Furthermore, the bacteria need yeast extract containing complex medium to reach higher cell densities since growth in defined minimal media is weak (Olijve and Kok, 1979). The optimal cultivation temperatures are between 25-30°C and the pH optimum is 5.5 – 6. However, *G. oxydans* is capable to grow even at an acidic pH of 2.5.

In 2005, the genome of *G. oxydans* 621H (DSM 2343) was sequenced for the first time using the Sanger method (Prust et al., 2005). It consists of one circular chromosome with a size of 2,702,173 base pairs (bp) and five plasmids (pGOX1: 163,186 bp; pGOX2: 26,568 bp; pGOX3: 14,547 bp; pGOX4: 13,223 bp; pGOX5: 2,687 bp). 2,664 protein-coding open reading frames (ORF) were predicted. 1,877 (70.5%) of them could be assigned to biological functions, 446 ORFs (16.9%) were annotated as hypothetical proteins based on sequence similarity to ORFs of related bacteria, and 341 ORFs (12.8%) were only found in the genome of *G. oxydans* (GenBank ID: CP0000004-CP000009, locus tag: GOXxxxx). Four copies of rRNA operons and 55 genes encoding tRNAs were annotated. Recently, NCBI relaunched the RefSeq microbial genomes database, which provides improved annotations of already published genomes by using the Prokaryotic Genome Annotation Pipeline, which allows standardizing of non-redundant reference sequences (Tatusova et al., 2015; Tatusova et al., 2016). This

update resulted in annotation of 2,713 protein-coding genes in the genome of *G. oxydans* (RefSeq ID: NC_006672-NC006677, locus tag: GOX_RSxxxx).

G. oxydans has the exceptional ability to incompletely oxidize a great variety of sugars, sugar alcohols, sugar acids, alcohols, and other substrates regio- and stereoselectively in the periplasm. Oxidation products are released almost completely *via* porins into the medium (Deppenmeier and Ehrenreich, 2009; Herrmann et al., 2004; Merfort et al., 2006a; Merfort et al., 2006b). This exceptional ability makes *G. oxydans* a highly interesting microorganism for industrial processes that involve oxidative biotransformations. Since the 1930s *G. oxydans* is used for the production of L-sorbose, a precursor of vitamin C (Pappenberger and Hohmann, 2014). Further products obtained with *G. oxydans* are dihydroxyacetone, a compound used in pharmaceutical products or as a tanning agent, or 6-amino-$_L$-sorbose, which is used for the synthesis of the antidiabetic drug Miglitol (Deppenmeier et al., 2002; Gupta et al., 2001; Herrmann et al., 2004; Mamlouk and Gullo, 2013; Saichana et al., 2015).

1.1.2 The carbon metabolism of *G. oxydans*

Characteristically, growth of *G. oxydans* in the laboratory with glucose or mannitol as carbon sources is divided into two phases. Both carbohydrates are favourable for biomass production of *G. oxydans*, however, mannitol is more expensive. During growth on mannitol-containing media, the major part of the carbon source is oxidized in the periplasm by the membrane-bound major polyol dehydrogenase SldAB to fructose and only a minor part is taken up by the cell into the cytoplasm (growth phase I). During growth phase II, fructose is partially converted to 5-ketofructose. A minor part of the fructose enters the cytoplasm, where it is phosphorylated to fructose-6-phosphate by the fructose kinase FrkA. Fructose-6-phosphate is converted to glucose-6-phosphate by glucose-6-phosphate isomerase GPI and enters the pentose phosphate pathway (PPP) or the Entner-Doudoroff (EDP) pathway (Richhardt et al., 2012). If glucose is used as carbon source, 90% is oxidized in growth phase I to gluconate in the periplasm by the membrane-bound glucose dehydrogenase GdhM (Hanke et al., 2013). Of the 10% glucose taken up into the cell, 9% are also oxidized to gluconate by the soluble glucose dehydrogenase GdhS in growth phase II. Therefore, only a small amount of glucose is phosphorylated to glucose-6-phosphate and then further metabolized by the cytoplasmic sugar metabolism (Figure 1).

Genome sequencing showed that the genes encoding 6-phosphofructokinase, succinate dehydrogenase, and succinyl-CoA synthetase are missing (Prust et al., 2005). Therefore, both the Embden-Meyerhof-Parnas (EMP) pathway and the tricarboxylic acid (TCA) cycle are incomplete. The only intact pathways for cytoplasmic sugar catabolism are the PPP and the EDP. Growth characterization of mutants lacking essential genes of either the PPP or the EDP revealed that neither of the two pathways is essential for *G. oxydans*. However, the PPP is favourable, whereas the EDP is dispensable (Richhardt et al., 2012, 2013). The restricted ability to oxidize carbohydrates in the cytoplasm and the periplasmic oxidation *via* membrane-bound dehydrogenases contribute to limited assimilation of carbohydrates into cell material and a low final biomass yield. This characteristic is unfavourable for the industrial use of *G. oxydans*, which requires an efficient generation of biomass for subsequent oxidative biotransformations.

Although many oxidoreductases encoded in the genome of *G. oxydans* have been characterized (Deppenmeier and Ehrenreich, 2009; Deppenmeier et al., 2002; Hekmat et al., 2003; Hölscher et al., 2007; Mientus et al., 2017; Schweiger et al., 2010) and several studies were performed regarding its metabolism and physiology (Hanke et al., 2013; Richhardt et al., 2012, 2013), little is known about global gene expression and the transcriptional landscapes. In the last years, RNAseq (see chapter 1.4.2) revealed a high complexity of bacterial transcriptomes and emphasized its importance for the expansion of application ranges of bacteria used in industry. Comprehensive analysis of transcriptomes may provide a wealth of information important for an increased application of *G. oxydans* as a microbial cell factory. Especially knowledge about transcription start sites (TSSs), promoters, ribosome binding sites (RBS), weakly expressed genes, and identification of suitable regions for chromosomal integrations of genes and operon constructs can be useful for metabolic engineering approaches.

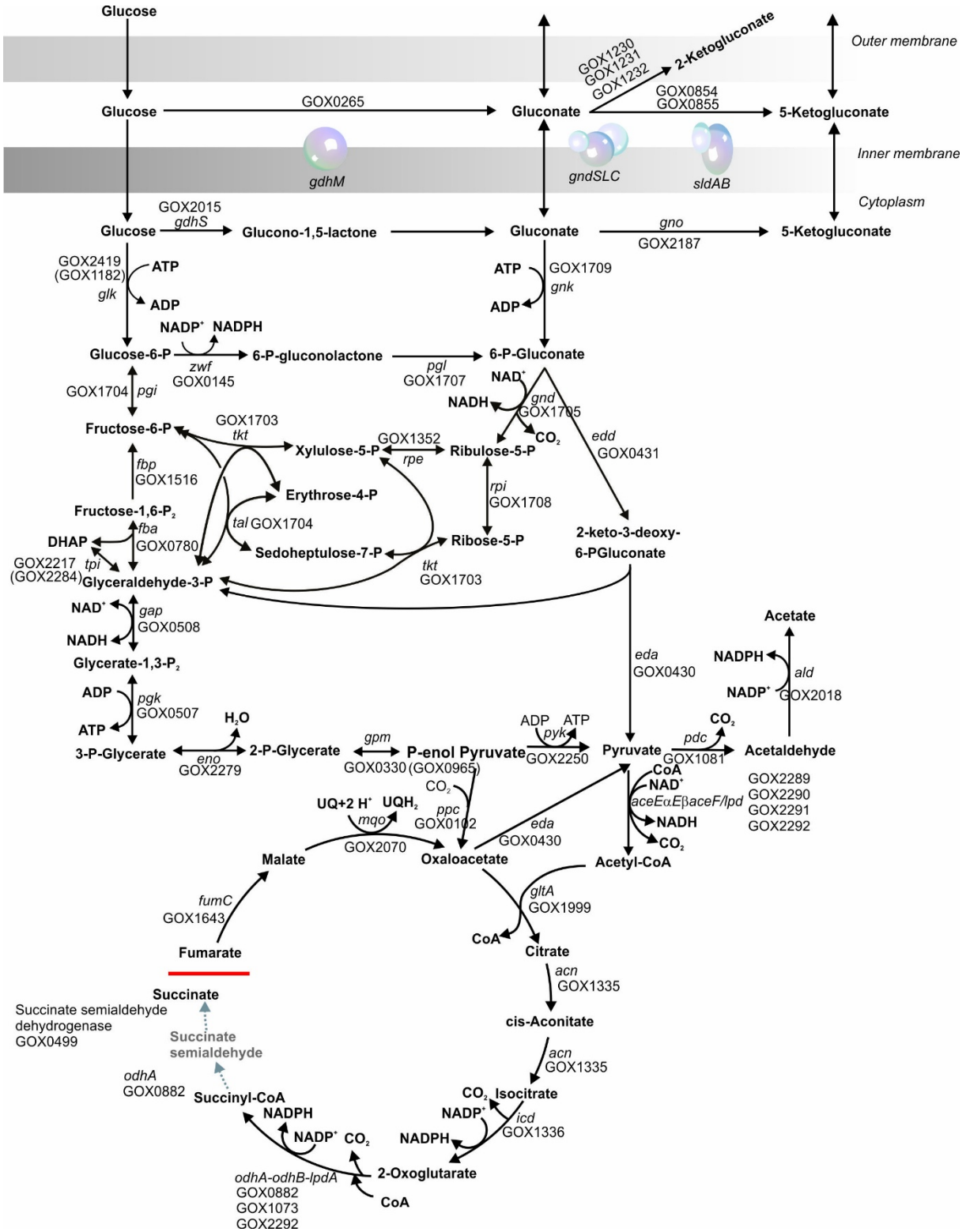

Figure 1. Scheme of the central carbon metabolism of *G. oxydans* with glucose as substrate (Richhardt et al., 2013, modified).

1.1.3 Metabolic engineering for increased biomass yield

Due to the incomplete periplasmic oxidation of carbohydrates, release of the resulting products into the medium and the restricted central carbon metabolism, *G. oxydans* reaches only very low biomass yields of 0.09 $g_{cdw}/g_{glucose}$ (Hanke et al., 2013). Therefore, its industrial use is associated with high costs for biomass generation compared to other bacteria such as *Escherichia coli* with a biomass yield of 0.49 $g_{cdw}/g_{glucose}$ (Soini et al., 2008). Recently, several metabolic engineering steps were performed to increase the biomass formation from glucose as carbon source as illustrated in Figure 2 (Kiefler et al., 2017):

(1) Completion of the TCA cycle by chromosomal integration of the succinate dehydrogenase genes (*sdhCDAB*) and the flavinylation factor (*sdhE*) from *Acetobacter pasteurianus* and of the succinyl-CoA synthetase genes (*sucCD*) from *Gluconacetobacter diazotrophicus* to enable the complete oxidation of acetyl-CoA to CO_2 with formation of NADH, NADPH, and UQH_2 as reductants for ATP synthesis via oxidative phosphorylation.

(2) Completion of the TCA cycle should lead to an increase in NADH formation. To re-oxidize NADH, the oxidative ability of *G. oxydans* was increased by genomic integration of a second NADH dehydrogenase gene (*ndh*) from *G. oxydans* DSM3504 (Kostner et al., 2015).

(3) Prevention of acetaldehyde and acetate formation from pyruvate by deletion of the gene for the pyruvate decarboxylase (*pdc*).

(4) Elimination of periplasmic and cytoplasmic gluconate formation by deletion of the genes coding for the PQQ-dependent membrane-bound glucose dehydrogenase (*gdhM*) and the soluble glucose dehydrogenase (*gdhS*).

In total, three different metabolically engineered strains were constructed: IK001 (Δ*upp* Δ*gdhS::sdhCDABE*), IK002.1 (Δ*upp* Δ*gdhS::sdhCDABE* Δ*pdc::ndh*), and IK003.1 (Δ*upp* Δ*gdhS::sdhCDABE* Δ*pdc::ndh* Δ*gdhM::sucCD*). The final strain, *G. oxydans* IK003.1, showed a 60% increased biomass yield on glucose (Kiefler et al., 2017). Alteration of glucose catabolism in such a drastic way may cause suppressor mutations and structural variants or may affect the genome stability. Thus, regular checking of the genome sequences of genetically modified strains is important to identify possible secondary genomic alterations at an early stage.

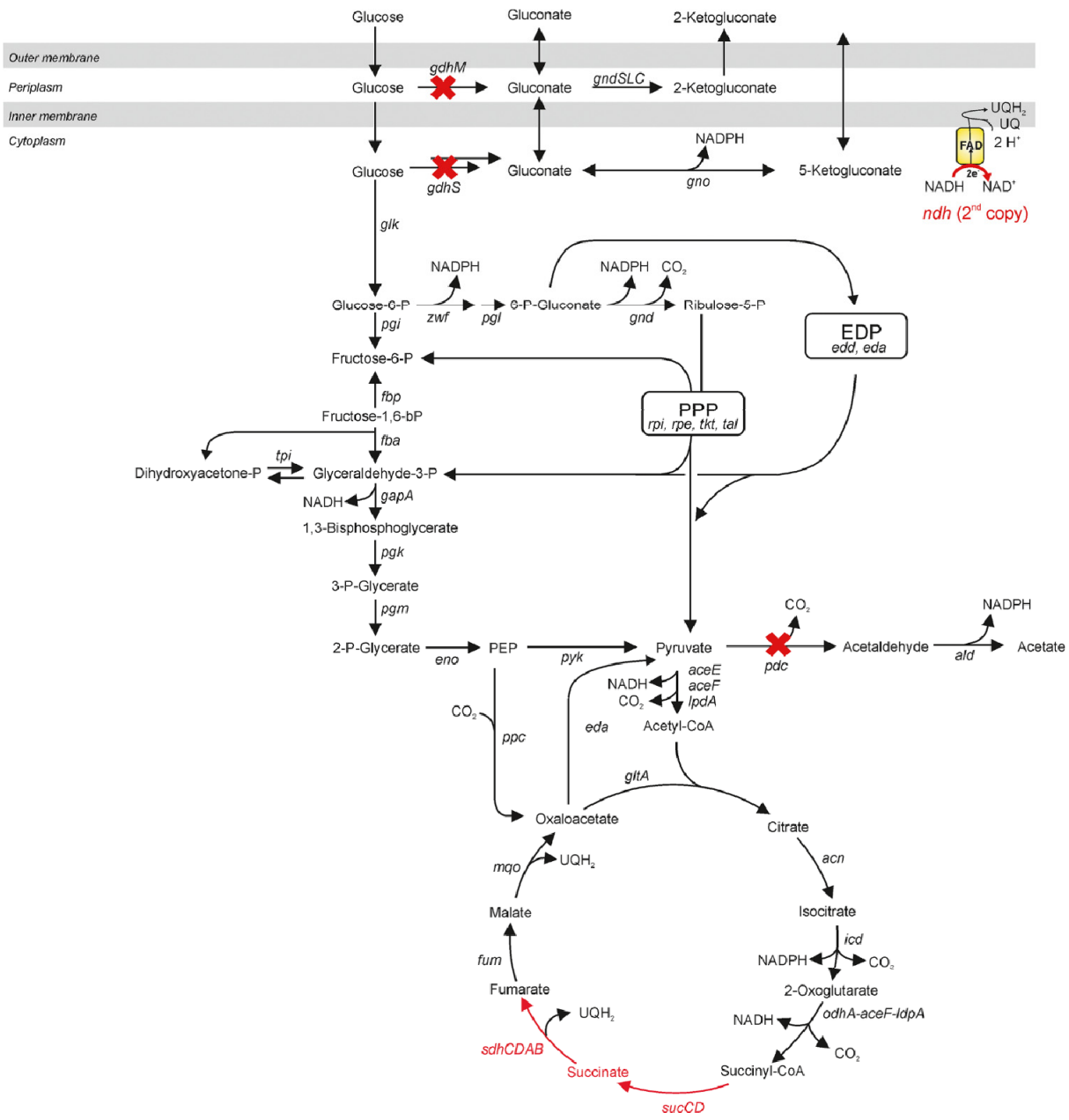

Figure 2. Scheme of the glucose metabolism of *G. oxydans*. Targets that were addressed during metabolic engineering for improved biomass yield are highlighted (Kiefler et al., 2017).

1.2 Key players of gene expression

Bacterial genomes usually contain more than 1,000 genes encoding proteins and non-coding RNAs, for example ribosomal RNAs, transfer RNAs, and regulatory RNAs. By transcription of coding sequences (CDS), mRNA is synthesized and then further translated into proteins. The amount of the different RNAs (transcriptome) and proteins (proteome) in a bacterial cell at a specific time point varies depending on growth phase, availability of nutrients in the media, and environmental conditions due to a complex

regulation of transcription and translation. This chapter focuses on major key players involved in regulation of gene expression.

Regulation of transcription initiation plays an important role for the expression of genes. Transcription initiation is enabled by binding of the RNA polymerase (RNAP) holoenzyme to specific sequences (promoters) upstream of the transcription start site (TSS). Every RNAP holoenzyme consists of five core-subunits (α_2, β, β´, and ω) and a σ-factor (Bae et al., 2015; Rothman-Denes, 2013). The α-dimer is responsible for assembly of the β- and β´-subunit and for the interaction with transcription factors, whereas the β- and β´-subunit form the active center of the RNAP (Busby and Ebright, 1994; Ebright, 2000). ω is the smallest subunit, which recruits the β´-subunit for the assembly with the $\alpha_2\beta$ complex (Mathew and Chatterji, 2006). The dissociable σ-factor is essential for recognition of the promoter sequence and unwinding of the DNA duplex (Figure 3). Usually, more than one σ-factor can be found in bacterial genomes. There is one primary σ-factor, which is responsible for the transcription of housekeeping genes, and a variable number of non-essential σ-factors (Paget and Helmann, 2003). Sigma factors of the σ^{70} family recognize the -35 and -10 regions upstream of the TSS and activate transcription of the downstream genes. The limited availability of free RNA polymerase complexes in the bacterial cell also controls gene expression (Ishihama, 2000). Depending on several conditions such as growth rate, most of the RNA polymerases are used for the transcription of genes encoding rRNAs and tRNAs, because they are needed in high amounts to maintain cellular functions at certain time points in the cell (Ishihama, 2000). Therefore, transcription of other genes needs to be regulated in differentiated ways by, for example, the promoter sequences. Promoters, which are almost identical to the ideal conserved consensus motif, usually allow a more efficient transcription than promoter sequences strongly deviating from the consensus sequence (Busby and Ebright, 1994). Therefore, these genes are higher expressed. Regulation of transcription occurs also by use of alternative σ-factors. Their abundance and activity can be affected by different stimuli, such as growth rate changes or environmental stress, and they can replace the housekeeping σ-factor in the RNAP holoenzyme (Maeda et al., 2000). Alternative σ-factors recognize different promoter sequences than the primary σ-factor and are therefore able to regulate expression of genes involved in, for example, stress responses.

Transcription can also be influenced by *cis*-regulatory elements, which are located in the 5'-untranslated regions (UTRs) of transcripts. These RNA elements are able to recognize temperature shifts or changes in the concentration of specific metabolites in

the cell. Expression of downstream genes is then regulated by formation of secondary RNA structures in the UTRs leading to premature termination of transcription or read-through. Alternatively, the RNA structures can influence translation by blocking access to the RBS (Mandal and Breaker, 2004).

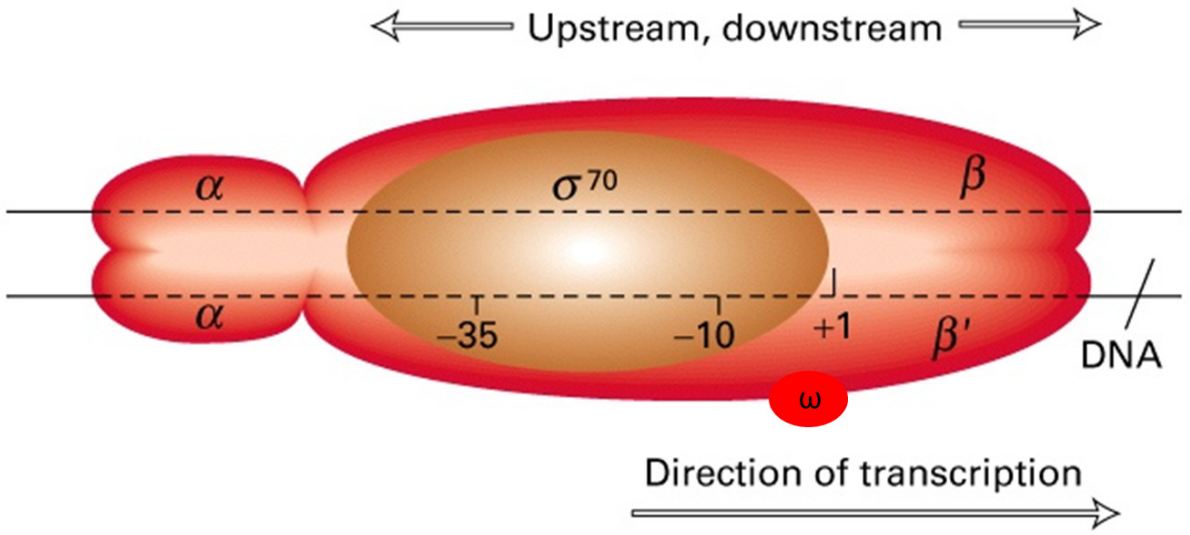

Figure 3. Schematic representation of the prokaryotic RNA polymerase holoenzyme with a σ^{70} subunit bound to DNA (Lodish et al., 2000, modified).

1.3 Half-lives of mRNAs affect transcript abundance

Besides the regulation of gene expression by σ-factors, promoter strengths, and *cis*-regulatory elements, mRNA half-lives also influence the abundance of transcripts. It is assumed that the steady-state level of transcripts in the cell at a specific time point is determined by the rates of mRNA synthesis and degradation. In bacteria, mRNA half-lives range from around 0.5 min to 30 min with a median half-life of ca. 5 min (Andersson et al., 2006; Bernstein et al., 2002; Hambraeus et al., 2003). In general, there is a correlation between mRNA half-lives and the cellular function of the encoded proteins. Transcripts of genes assigned to regulatory functions typically have relatively short half-lives, whereas transcripts of housekeeping genes (e.g. ion transport, cell envelope) are usually more stable (Mohanty and Kushner, 2016; Morey and Van Dolah, 2013). mRNA half-lives of genes are not constant, but can change during bacterial adaptation. Especially stress-induced growth rate reduction has a significant effect on the stability of

transcripts. For example, a mean half-life of transcripts of 5.8 min was observed during exponential growth in *Lactococcus lactis*, but increased to 19.4 min during carbon starvation (Redon et al., 2005). Similarly, an increase of the mean half-life correlated with a decreased growth rate during isoleucine starvation (Dressaire et al., 2011).

Key players involved in the degradation of mRNAs are several endo- and exonucleases such as the endonuclease RNase E, which is also involved in tRNA, rRNA, and sRNA processing (Apirion and Lassar, 1978; Li and Deutscher, 2002), the 5´→3´ exonuclease RNase J (Mathy et al., 2007), RNase III, which can cleave double-stranded RNA (Srivastava et al., 1992), or RNase II, a 3´→5´ exonuclease (Nossal and Singer, 1968). Stability of mRNAs can be affected by elements found in 5´- or 3´-UTRs, ribosome density on the mRNA, access to endonucleolytic cleavage sites, ribonuclease levels, or the location of the mRNA in the cell (Mohanty and Kushner, 2016). Since transcription and translation in bacteria is tightly coupled, ribosomes are usually bound in higher amounts to mRNAs and thereby prevent degradation by endonucleases (Braun et al., 1998). However, untranslated regions are ribosome-free and can easily be attacked by exonucleases. Therefore, formation of secondary structures of long 5´- or 3´-UTRs is a possible mechanism to stabilize transcripts. For example, it was shown that the 5´-UTR of the *ompA* gene is responsible for the long half-life (13 min) of the transcript. Absence of the 5´-UTR leader region reduces the half-life to 4 min (Emory et al., 1992). Also, riboswitches can control not only transcription or translation, but also mRNA decay (Deana and Belasco, 2005). For example, the *lysC* riboswitch in *E. coli* binds lysine and forms a secondary structure that inhibits translation initiation by blocking of the RBS and also exposes RNAse E cleavage sites enabling degradation (Caron et al., 2012). Availability of ribonucleases also influences mRNA decay. Expression of many RNases is controlled by auto-regulation or is stress-induced (Bardwell et al., 1989; Chen and Deutscher, 2010).

1.4 Next-generation sequencing provides important information for strain development

1.4.1 High-throughput sequencing methods

Until the mid-2000s the chain-termination sequencing method (1st generation), developed by Frederick Sanger in 1977 with further improvements by using capillary gel electrophoresis and semi-automation of the sequencing process in the early 90´s was the main method used for nearly every application that required determination of nucleotide sequences (Hunkapiller et al., 1991; Sanger et al., 1977; Swerdlow and Gesteland, 1990; Swerdlow et al., 1990). This technology was also used for the first blueprint of the human genome sequence in 2001, which comprises about 3×10^9 bp (Lander et al., 2001; Venter et al., 2001). The first bacterial genome of *Haemophilus influenzae* Rd was already sequenced in 1995 (Fleischmann et al., 1995). However, Sanger sequencing is very time- and labour-intensive and thus very expensive for such large sequencing projects. Therefore, high-throughput sequencing methods, which allow extreme parallelization and drastic cost reduction of the sequencing process were developed in the last two decades for sequencing of complete genomes and transcriptomes and have revolutionized this field.

Commonly used sequencing platforms of the 2nd generation are from Life Technologies (SOLiD, Ion Torrent), Roche (454), and Illumina, Inc. (MiSeq, HiSeq, and NextSeq). These technologies require template amplification prior to sequencing. In general, one distinguishes between sequencing-by-ligation (SBL) and sequencing-by-synthesis (SBS) approaches. SOLiD- and 454-sequencing uses amplification of templates by emulsion PCR (Dressman et al., 2003). In this process, adapters containing universal priming sites are ligated to ends of double-stranded DNA fragments, which are subsequently denatured and captured into droplets in an oil-aqueous emulsion. Every droplet contains one DNA molecule, a bead covered with oligos complementary to parts of the adapter sequences, primers, dNTPs, and DNA polymerases. The single stranded DNA fragments can bind to the complementary oligos on the bead surface. Starting from the free 3'-OH end of the oligo the reverse strand is generated, which is consequently attached to the bead. Afterwards the original single-stranded DNA fragment, which is released from the bead, can anneal to another oligo on the bead and a novel reverse strand is generated, which is also attached to the bead. These steps are repeated until the bead is covered with several thousand copies of the same DNA molecule (Figure 4).

After amplification, beads can be cross-linked to glass slides or used for sequencing in PicoTiterPlates (PTP).

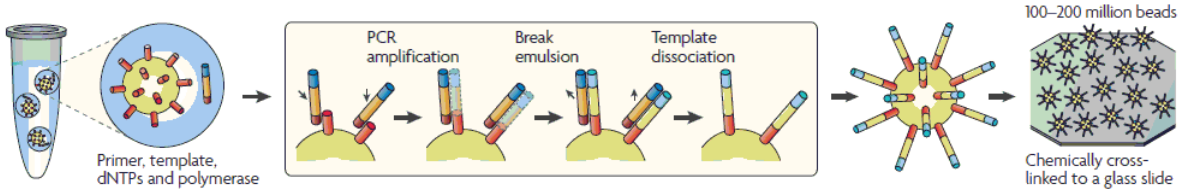

Figure 4. Emulsion PCR in oil-aqueous emulsion (Metzker, 2010).

SOLiD uses a four-colour SBL approach (Shendure et al., 2005). Here, four different fluorophore-labelled 1,2-probes are used for ligation to complementary positions of the template strand upstream of the sequencing primer (Figure 5A). They consist of two interrogation bases, three degenerate bases (nnn), and three universal bases (zz). After imaging, the ligated 1,2-probe are cleaved by removing the last three universal bases along with the linked fluorophore allowing ligation of the next 1,2-probe. Since the three degenerated and the two interrogation bases are not removed, the 1,2-probes are ligated at every 5th position of the template strand during the ligation cycles. To cover every position of the strand, the primer of the first round is stripped and the next sequencing primer is shifted by one position (*n*-1) during the 2nd primer round (Figure 5A). This resetting is repeated three more times. Due to this iteration, every nucleotide is interrogated twice (Figure 5B). To determine the sequence of the DNA based on the four-colour imaging, a two-base encoding scheme is used. Each of the four different fluorophores is associated with four dinucleotide sequences, for example AC, CA, GT, and TG are represented by a green dye (Figure 5C). As long as one of the bases is known, the determined color code can be converted to the nucleotide sequence (Figure 5D).

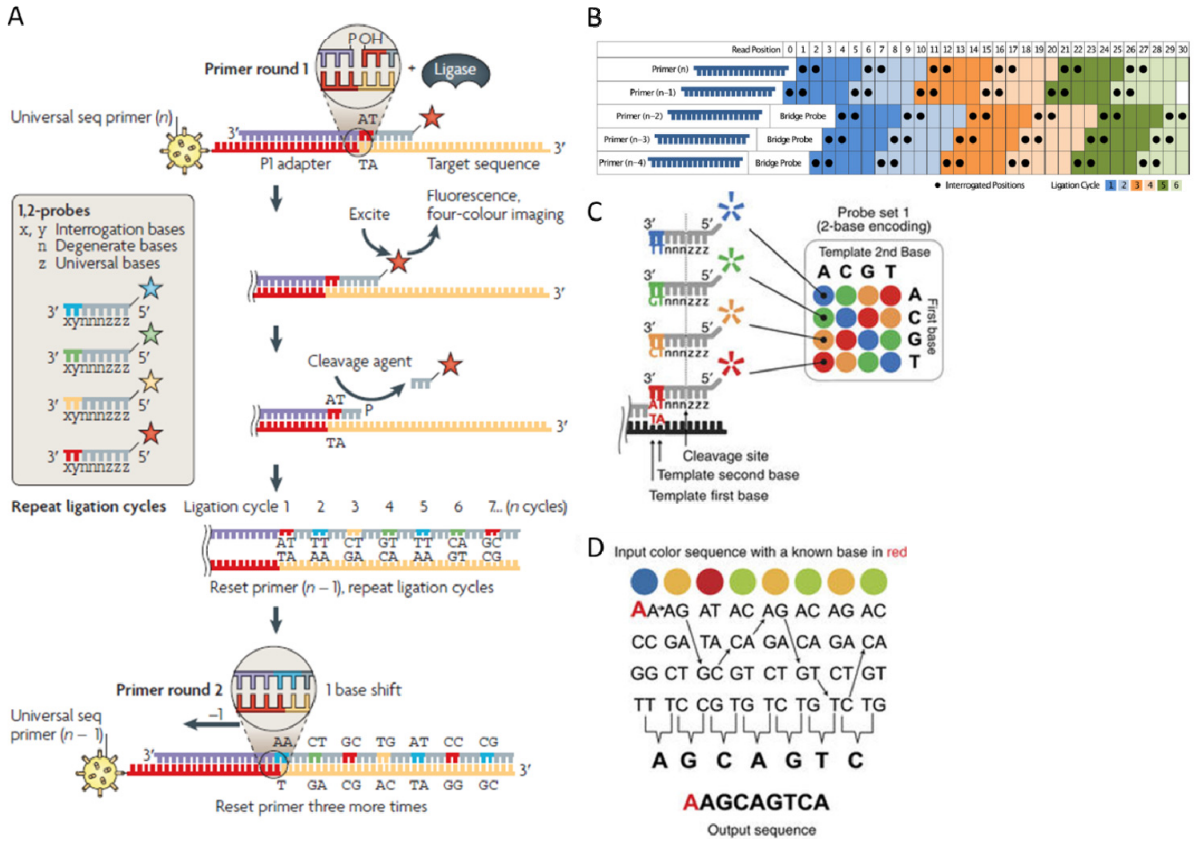

Figure 5. Sequencing-by-ligation approach used by SOLiD. (A) Ligation of 1,2-probes to template DNA sequence, four-colour imaging and iteration of the process (Metzker, 2010). (B) Schematic representation of the five primer rounds, which result in two-time interrogation of every base. (C) Two-base encoding scheme used for determination of the nucleotide sequence. (D) Example of an output sequence based on the colour code determined by four-colour imaging. The first base in red is known (www.appliedbiosystems.com).

The 454 system, which uses pyrosequencing, was the first next-generation sequencer (Margulies et al., 2005). Here, amplification of templates occurs also by emulsion PCR (Dressman et al., 2003). For sequencing (SBS), template-covered beads are given into a PTP with nucleotides and an enzyme cocktail. If a nucleotide is incorporated, an enzyme cascade is started resulting in a bioluminescence signal (Figure 6A). Pyrosequencing belongs to the single nucleotide addition approaches, where only one dNTP species is used *per* cycle of sequencing. Therefore, more than one identical dNTP can be incorporated resulting in a linear increase in light signal (Margulies et al., 2005). A disadvantage of this method is the occurrence of sequencing errors at homopolymer stretches since the increase in light signal is less linear for longer homopolymers (Figure 6B). Ion Torrent also uses emulsion PCR for amplification and SBS (Rothberg et al., 2011). It is similar to the 454 system. However, it uses no enzymatic cascade as signal, but detects the H^+ ions that are released when a dNTP is incorporated. No cameras, lasers or fluorescent dyes are necessary (Glenn, 2011).

Since the pH change is proportional to the number of nucleotides, it is possible that imprecision occurs during measuring of homopolymer stretches.

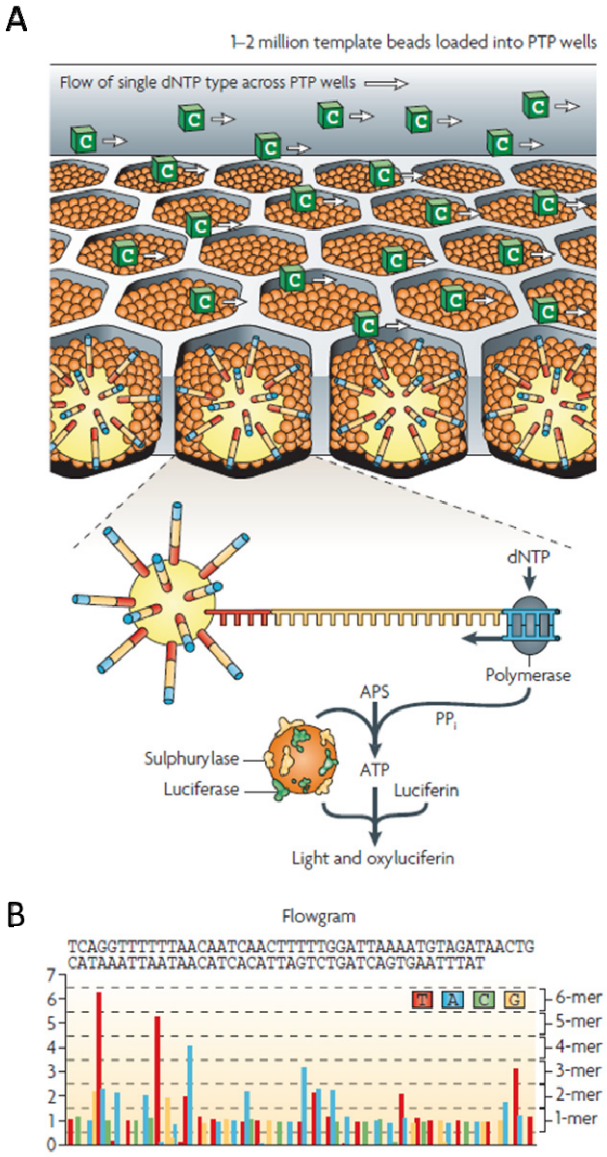

Figure 6. Pyrosequencing using the 454 system. (A) Amplification of DNA templates prior to sequencing occurs by emulsion PCR. Beads covered with million copies of single-stranded DNA templates are then loaded into PTP wells and dTNPs of one type (e.g. dCTP) are flowed across the wells in a specific sequential order allowing incorporation of complementary bases. Smaller beads with ATP sulphurylase and luciferase attached to them are also loaded into the wells. Incorporation of dNTPs results in the release of pyrophosphate (PPi). In the presence of adenosine 5´-phosphosulfate (APS) the ATP sulphurylase is able to convert PPi to ATP, which acts as a substrate for the luciferase-mediated conversion of luciferin to oxyluciferin. The light signal generated by this reaction is proportional to the amount of ATP. After degradation of unincorporated nucleotides and ATP by the apyrase, the reaction can restart with the next dNTP type. (B) The bioluminescence is imaged with a charge-coupled device and converted into the DNA sequence (Metzker, 2010).

Sequencing experiments for this thesis were performed using Illumina´s SBS approach. Here, template preparation includes ligation of specific sequencing adapters (for more information on these sequencing adapters see section 1.4.2) to fragmented DNA and cluster generation on a solid glass surface (flow cell). Oligos complementary to parts of the adapters are covalently attached to the flow cell allowing binding of single-stranded template DNA (Figure 7A). A polymerase moves along the single strand generating the complementary strand. Subsequently, the template strand is washed away and the reverse strand bends and attaches to the oligo on the flow cell surface, which is complementary to the upper part of the adapter sequence on the other end of the strand. Polymerases generate then the complementary strand, which is identical to the original one. The double stranded DNA is denatured so that the single strand can again bend and attach to complementary oligos on the flow cell surface (Fedurco et al., 2006). After several rounds of this bridge amplification, 100-200 million clusters are generated, each one representing one DNA template (Figure 7A). After amplification, SBS is used to determine the nucleotide sequence. The double-stranded DNA is denatured and the reverse strands are washed away so that only the forward strands are first used for sequencing. A primer with a free 3´-OH group binds to a complementary part of the adapter allowing incorporation of dNTPs. In contrast to pyrosequencing and SOLiD sequencing, all four dNTPs can be used simultaneously. These dNTPs are 3´-blocked reversible terminators allowing incorporation of only one dye-labelled dNTP *per* cycle. These dNTPs are characterized by an attached fluorophore and a 3´-blocked terminator, which contains a cleavable group (3´-*O*-azidomethyl) attached to the 3´-oxygen (Figure 7B). Incorporation of further dNTPs by the polymerase is therefore inhibited. To determine the sequence of the DNA template, all four dye-labelled dNTPs are flowed simultaneously across the flow cell for incorporation. After imaging, the fluorophore is cleaved and the 3´-OH group is regenerated using tri(2-carboxyethyl)phosphine (TCEP) as reducing agent thereby enabling incorporation of the next dNTPs (Figure 7C) These sequencing cycles are repeated generating reads with a specific length. Usage of all four dNTPs simultaneously allows determination of the nucleotide sequences for all clusters (Figure 7D) (Turcatti et al., 2008).

In general, every DNA molecule flanked by adapters can be sequenced from both end generating so called paired-end reads.

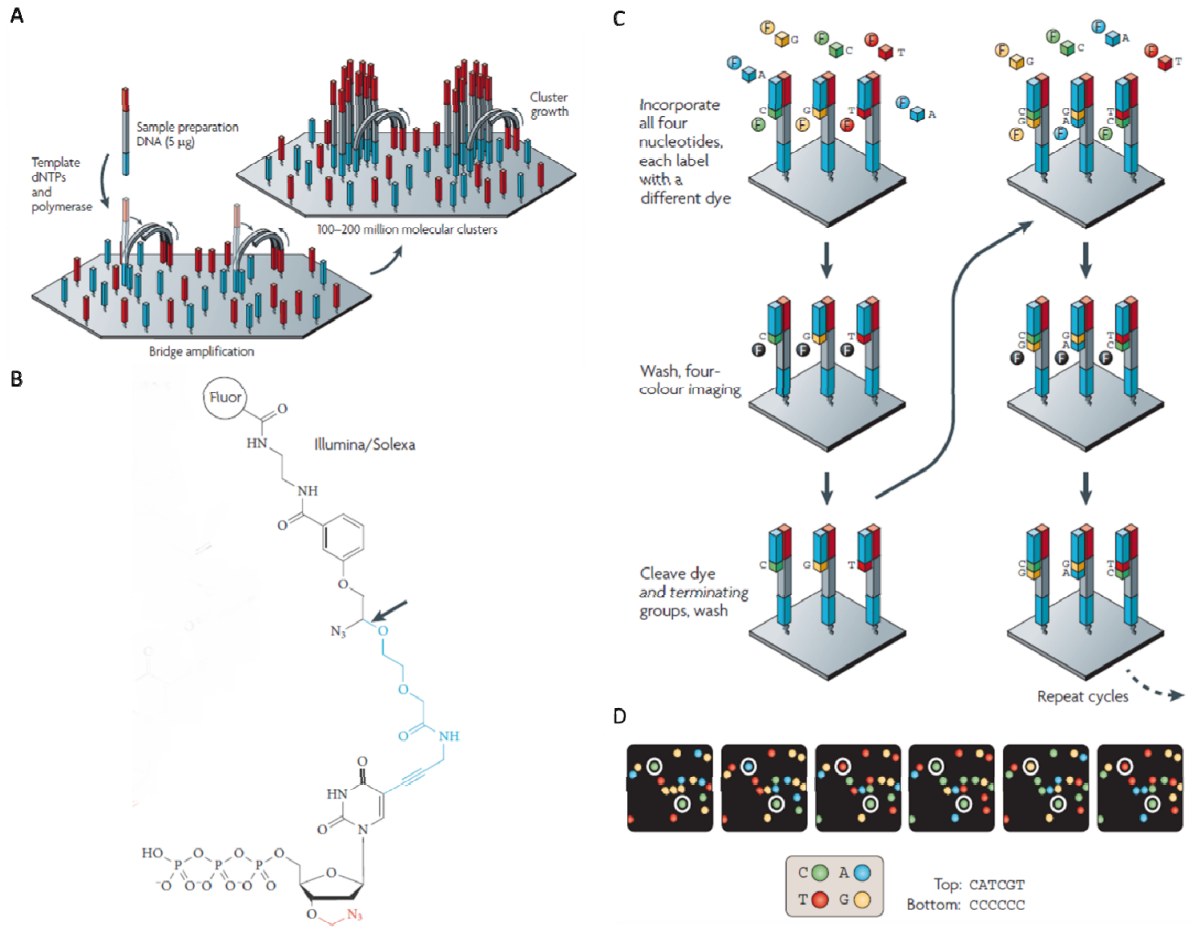

Figure 7. Amplification of DNA templates and SBS used by Illumina. (A) Cluster generation by bridge amplification. (B) Example of a 3´-blocked reversible terminator. (C) SBS using the dye-labelled and 3´-blocked dNTPs. After imaging, a cleavage step removes the fluorescent dyes and regenerates the 3´-OH group using the reducing agent TCEP, thereby enabling incorporation of the next 3´-blocked reversible terminator. (D) Four-colour images showing generation of sequencing data based on two different clusters (Metzker, 2010).

In the last ten years, platforms of the 3rd generation expanded the field of sequencing by allowing analysis of single DNA molecules. The first commercial single-molecule sequencer was the HeliScope from Helicos (Harris et al., 2008). However, high costs and short read-lengths hindered a broader use. The single-molecule real-time (SMRT) sequencer from PacBio is currently the most widely used technology (Eid et al., 2009). The principle of this method is the usage of picolitre wells with immobilized polymerases. Sequencing occurs by measuring the emitted light when a fluorophore is released from the γ-labelled dNTP during its incorporation. Since 2014, the nanopore sequencer MinION from Oxford Nanopore Technologies is available (Clarke et al., 2009). Here, a single molecule with a leader sequence passes a specific nanopore (e.g. α-

hemolysin) and the ionic current change is measured to determine nucleotide sequences (Figure 8). Read lengths with up to 50 kb and much more can be achieved. This makes nanopore sequencing a very good method for *de novo* assemblies or clarification of larger structural variants. The latter one is especially difficult with read lengths of up to 300 nt, which are achieved *via* Illumina´s sequencing approach. However, error rates between 7.5%-14.5% can occur in reads generated with the MinION (Jain et al., 2017). Therefore, hybrid approaches combining long nanopore and accurate short reads are used (Laver et al., 2015).

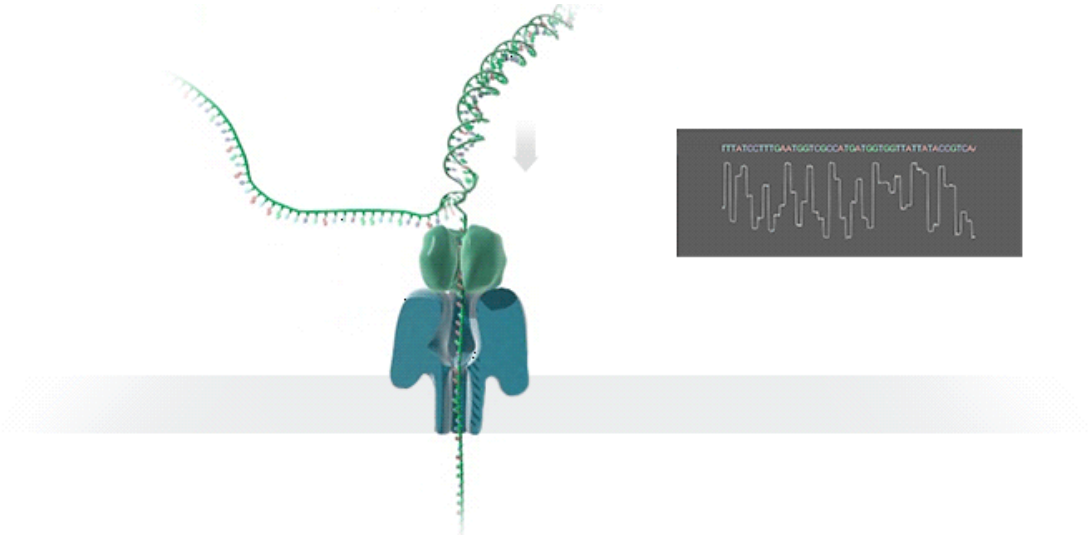

Figure 8. Schematic representation of a nanopore. When a DNA molecule stepwise passes the nanopore, the ionic current changes in dependence of the DNA sequence. The changes are measured and the resulting profile allows conclusions on the base composition (www.nanoporetech.com).

1.4.2 RNAseq allows high-resolution transcriptome analysis

Transcriptomics is a wide field comprising analysis of transcripts with a broad range of techniques. For relative mRNA quantification, microarrays and qRT-PCR can be used. Microarrays are hybridization-based and can be used genome-wide, whereas qRT-PCR is only applicable for quantification of a certain number of short cDNA fragments simultaneously. Both methods require information about the genome sequence. 5´- or 3´-RACE (Rapid Amplification of cDNA Ends) allow identification of transcript starts and ends. However, these methods are costly and not applicable for genome-wide studies.

Development and application of NGS improved the possibilities to comprehensively analyse transcriptomes (RNAseq). In contrast to microarrays and qRT-PCR, RNAseq allows single nucleotide resolution, quantification of transcripts and does not need knowledge about the genome sequence (Wang et al., 2009). Therefore, it is also very well suited to detect novel transcripts and operon structures. Preparation of whole transcriptomes for sequencing can be automated by usage of specific kits, for example from Illumina. Typically, 95% of the total RNA isolated from bacterial cells is ribosomal RNA. Therefore, depletion of rRNA is the most important step prior to library preparation and sequencing. A specific Ribo Zero rRNA removal kit from Illumina uses hybridization of rRNA to fragments, which are complementary to conserved regions of the rRNAs (Wang et al., 2009). Those fragments are bound to magnetic beads that are removed after hybridization. Afterwards, the RNA is fragmented in a further step. A great advantage of RNAseq methods is the availability of protocols to maintain strand-specificity of the RNA. This allows detection of antisense transcripts. One method applied by the TruSeq stranded mRNA library preparation kit from Illumina is the dUTP method (Hrdlickova et al., 2017). First strand cDNA synthesis is performed by using random hexamer priming (Figure 9A). During second strand synthesis, dUTPs instead of dTTPs are incorporated (Figure 9B). Afterwards, Y-shaped sequencing adapters are ligated to the double-stranded cDNA prior to cDNA amplification. Those Y-adapters are not complementary at their ends and consist of a p5 and a p7 part (Figure 9C). During cDNA amplification the 2nd strand with the dUTPs is degraded, so that only the first strand representing the reverse-antisense strand is amplified (Figure 9D). Degradation of the second strand results in loss of the Y-shape of the adapters. Therefore, the final double-stranded cDNA is flanked by the p5 adapter at one end and by the p7 adapter at the other end. Strand-specificity of the RNA can be retraced during paired-end sequencing due to the orientation of the adapters, because read 1 is always generated starting from p5 and read 2 from p7. Therefore, it is possible to detect antisense transcripts.

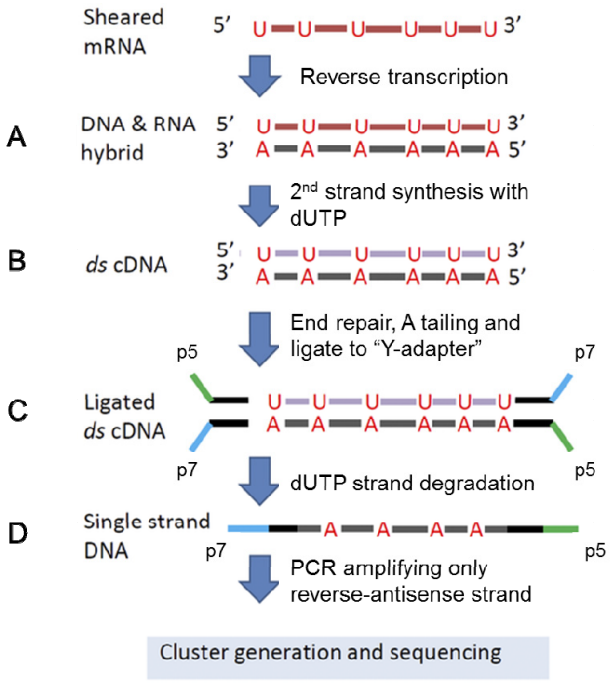

Figure 9. Workflow for the preparation of strand-specific whole transcriptome libraries using the dUTP method (Wang et al., 2011, modified).

In the last years, a new method for the comprehensive identification of transcription start sites (TSSs) by sequencing of the primary transcriptome has revealed a high complexity of bacterial transcriptomes. For example, widespread antisense transcription was observed and a high number of small regulatory RNAs was found (Mentz et al., 2013; Pinto et al., 2011; Raghavan et al., 2011; Sharma et al., 2010; Sorek and Cossart, 2010). Identification of TSSs allows a broad range of subsequent analysis. For example, 5´-UTRs of protein-coding genes can be analysed in order to identify regulatory elements (Güell et al., 2011; Sorek and Cossart, 2010). Furthermore, knowledge about the exact positions of transcription initiation can be used to identify promoter motifs, which are typically located 10 bp and 35 bp upstream of TSSs (Pribnow, 1975). Also, combination of whole and primary transcriptome data showed that operons can be divided into sub-operons due to the identification of internal TSSs within operon structures (Güell et al., 2011; Sharma et al., 2010). Generally, genes expressed as polycistronic transcripts are functionally related (Lawrence and Roth, 1996; Osbourn and Field, 2009). The first protocol to sequence 5´-ends of RNA species was developed for sequencing of the primary transcriptome of *Helicobacter pylori* (Sharma et al., 2010). Since then, several alterations were applied to the technique. For this thesis, a slightly modified method as originally used for the RNAseq analysis of

Corynebacterium glutamicum was used (Pfeifer-Sancar et al., 2013). This protocol also starts with the isolation of total RNA, depletion of rRNA, and fragmentation (Figure 10). The main difference between the preparation of whole transcriptome and primary transcriptome libraries is the digestion of processed transcripts with a terminator 5´-phosphate-dependent exonuclease. Thus, only primary transcripts with triphosphates at their 5´-ends should remain. To prepare the 5´-ends for RNA adapter ligation, the ends are then processed with a RNA 5´-polyphosphatase to obtain monophosphates. Direct ligation of the sequencing RNA adapter to the 5´-end allows preservation of strand-specificity. For reverse transcription of the adapter-ligated RNA, loop adapters with a 3´-NNNNNNN-tail are used. Then, the cDNA is amplified and the primary transcriptome library is sequenced in single-end mode so that every read should start with the transcription initiation site (Figure 10).

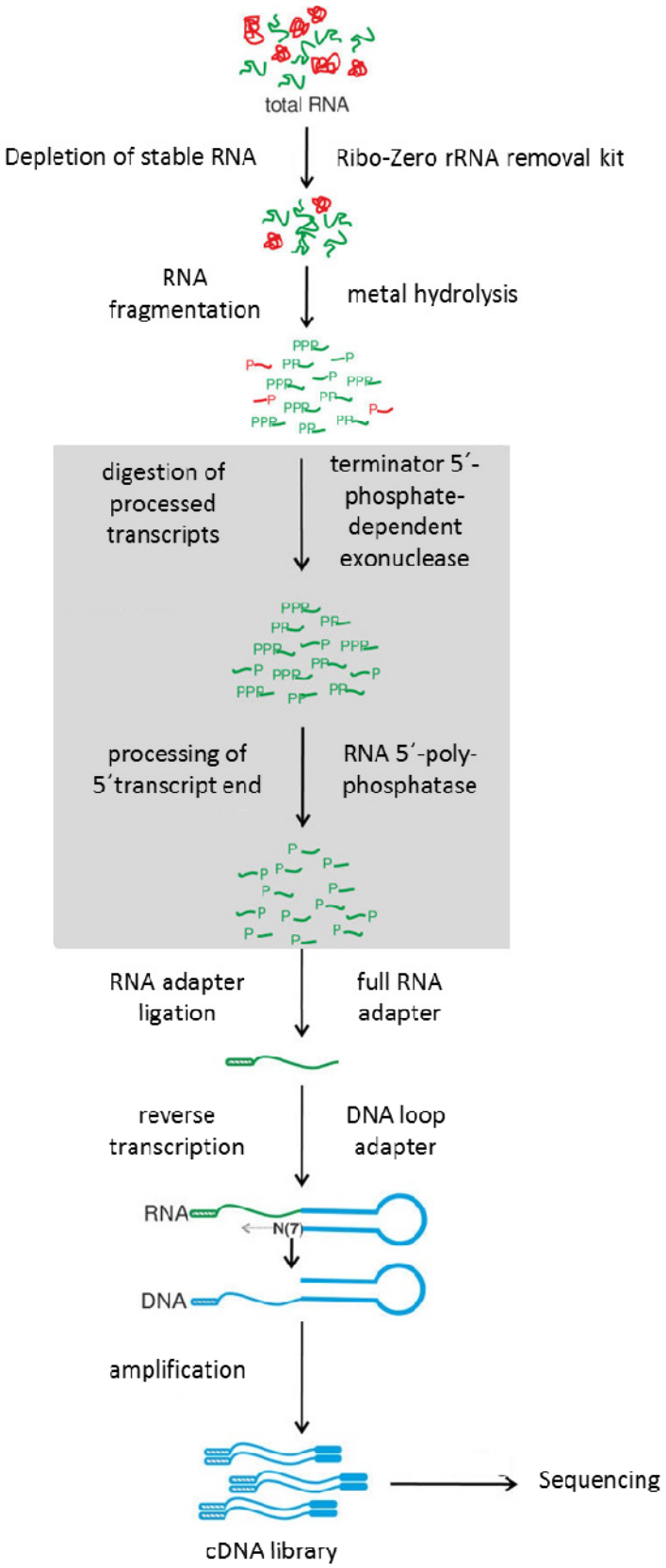

Figure 10. Experimental workflow for the preparation of libraries enriched for primary 5´-transcript ends (Pfeifer-Sancar et al., 2013, modified).

1.5 Aims of this thesis

The major aims of this thesis are the comprehensive analyses of the genomes and transcriptomes of *G. oxydans* 621H strains using Illumina sequencing and microarray studies. Three different studies should be performed to gain important information on genome stability, transcriptional organization, and mRNA stability.

Recently, metabolic engineering was performed to increase the biomass yield of *G. oxydans* on glucose as carbon source (Kiefler et al., 2017). Therefore, the incomplete TCA cycle was completed by chromosomal integration of missing genes, a second NADH was integrated into the genome, acetate formation was prevented, and periplasmic as well as cytoplasmic gluconate formation was eliminated. The final strain showed a 60% increased biomass yield. These steps may affect genome stability or cause suppressor mutations. To check for these possibilities, engineered strains should be sequenced and compared with reference strains. Furthermore, this approach should enable to update the reference sequence of the genome of *G. oxydans*.

The second part of this thesis should focus on transcriptome analysis by using RNAseq. Both whole transcriptome libraries and libraries enriched for 5´-transcript ends of *G. oxydans* cells grown under different conditions should be sequenced. This allows detection of TSSs, analysis of 5´-UTRs, identification of promoter and RBS motifs, and detection of novel transcripts including antisense transcription. Also, the organization of genes in primary operons and sub-operons should be analysed.

Abundance of transcripts in the cell is regulated by transcription and stability of mRNAs. To globally determine mRNA half-lives in *G. oxydans*, microarray studies should be performed. Analysis of the data obtained should correlate transcript stability with gene product function and expression values to identify possible bottlenecks in the metabolism of *G. oxydans*.

1.6 Key results of comprehensive genome and transcriptome analysis of *G. oxydans* strains

1.6.1 Genome stability of engineered pathway-restored strains

(Kranz et al. 2017, Journal of Biotechnology, chapter 2.1)

Recently, heterologous genes encoding succinate dehydrogenase and succinyl-CoA synthetase were chromosomally integrated and expressed in *G. oxydans* to complete the TCA cycle. Furthermore, several metabolic engineering approaches were performed to increase the utilization of glucose for biomass synthesis (see chapter 1.1.3). This resulted in a 60% increase in biomass yield of the final engineered strain *G. oxydans* IK003.1 (Kiefler et al., 2017). Changing the metabolism in such a drastic way may cause suppressor mutations and affect genome stability. Different kinds of mutations are known. Nucleotide variants with a maximal length of three nucleotides are called SNVs (single nucleotide variants; substitutions), MNVs (multiple nucleotide variants; substitutions), and InDels (insertions and deletions). Besides these smaller variants, structural variants are known which include deletions, insertions of novel sequences or mobile element insertions. Mutations could make engineered strains useless for further applications. Therefore, it is advisable to sequence engineered strains, especially those with major metabolic changes that affect growth.

Next-generation sequencing (NGS) is an ideal method to analyse engineered strains regarding the occurrence of small suppressor mutations. Especially Illumina´s sequencing-by-synthesis approach offers the best accuracy and quality. However, relatively short read lengths make the detection and clarification of larger structural variants difficult. Therefore, a combination of short and long reads can be advantageous (see chapter 1.4.1). Here, we used Oxford's Nanopore technology to generate long reads.

In our study, we sequenced the final engineered strain *G. oxydans* IK003.1 as well as its precursors IK001 and IK002.1 *via* NGS. As reference strains three different wild types from different sources (WT-DSMZ, WT-BM, and WT-E) were sequenced. WT-DSMZ was derived from the German Collection of Microorganism and Cell Cultures, WT-BM from our strain collection, and WT-E from Dr. Armin Ehrenreich (TU Munich). The latter one is the parental strain of *G. oxydans* Δ*upp*, a strain used for the construction of the IK series of strains, which was also sequenced. Sequence reads of every strain were mapped to the genome reference and analysed regarding the occurrence of mutations

and resulting amino acid changes. Compared to the reference strains, no additional suppressor mutations on the single nucleotide level occurred in the genomes of the engineered strains.

For the clarification of structural variants, the final engineered strain IK003.1 was additionally sequenced with the MinION (Oxford Nanopore) technology generating long reads with an average length of 18.9 kb. Both Illumina and Nanopore reads were combined in a hybrid assembly and then compared with the *G. oxydans* reference sequence. In addition to the known engineered modifications, three structural variants caused by mobile element insertions were found. Verification of them was done by Sanger sequencing and by mapping of the Illumina reads to a genome reference, which was adjusted by insertion of the structural variants. Two mobile element insertions found in an intergenic region and within a hypothetical protein (GOX_RS00080) were already present in the wild-type strain (WT-E), which was used for the construction of *G. oxydans* Δ*upp* and therefore also for the generation of the three engineered strains (IK001-IK003.1). Only one additional mobile element insertion within an intergenic region on pGOX2 was found in IK003.1. Overall, the final engineered strain was almost unaffected by mobile element insertions and no suppressor mutations occurred after circa 70 generations of strain handling including 28 hours in a bioreactor under controlled growth conditions. This high genome stability makes *G. oxydans* IK003.1 a suitable host for further metabolic engineering efforts.

1.6.2 High-quality update of the *G. oxydans* reference genome

(Kranz et al. 2017, Journal of Biotechnology, chapter 2.1)

A high-quality and well-annotated genome sequence is of great importance for analysis of both transcriptome and proteome data. Thus, re-sequencing of the already published genome of *G. oxydans* by using a combination of short high-quality Illumina reads and long nanopore reads is very well suited to update the genome sequence (see chapter 1.6.1). The genome of *G. oxydans* was first sequenced in 2005 using Sanger sequencing (Prust et al., 2005). This method can be inaccurate, especially at homopolymer stretches in DNA sequences.

To comparatively analyse sequencing results of engineered pathway-restored strains, we also sequenced wild-type *G. oxydans* strains from different sources (see chapter 1.6.1). Sequencing reads were mapped to the reference and variants were

reported. Overall, 160-165 variants were detected. 158 of them were present in all strains and 91 with very high frequencies of occurrence close to 100% were located in CDS. 78 of the variants in CDS resulted in frameshifts due to InDels, 7 variants resulted in amino acid exchanges, and 6 variants in CDS regions were synonymous substitutions. Since we uniformly found these variants in all strains, we assume that the corresponding regions in the reference genome are inaccurate and the variants represent the true reference sequence. Therefore, we updated the *G. oxydans* 621H reference genome sequence by introducing the nucleotide exchanges or InDels into the old reference. In case of the 78 InDels, which result in frameshifts in 64 genes, resulting new ORFs were searched to correct the genome annotation. Possible ORFs were used for BLASTP search and the resulting protein sequences were compared to the original annotation regarding their length and predicted function. 49 of these genes were recently annotated as frameshifted pseudogenes in an updated version of the original genome reference (Tatusova et al., 2016 Tatusova et al., 2015). Frameshifts found by mapping of the accurate Illumina reads to the reference sequence (Prust et al., 2005) resulted in correction of the original frameshifts and therefore to proper annotated ORFs with an annotated function. For nine genes the predicted function was not changed. It was only changed for DotG, which was re-annotated as a hypothetical protein. InDels in two neighbouring coding regions favoured merging of six hypothetical proteins to three without a change of the predicted protein function (Figure 11). These annotation changes were also introduced into the new reference genome sequence.

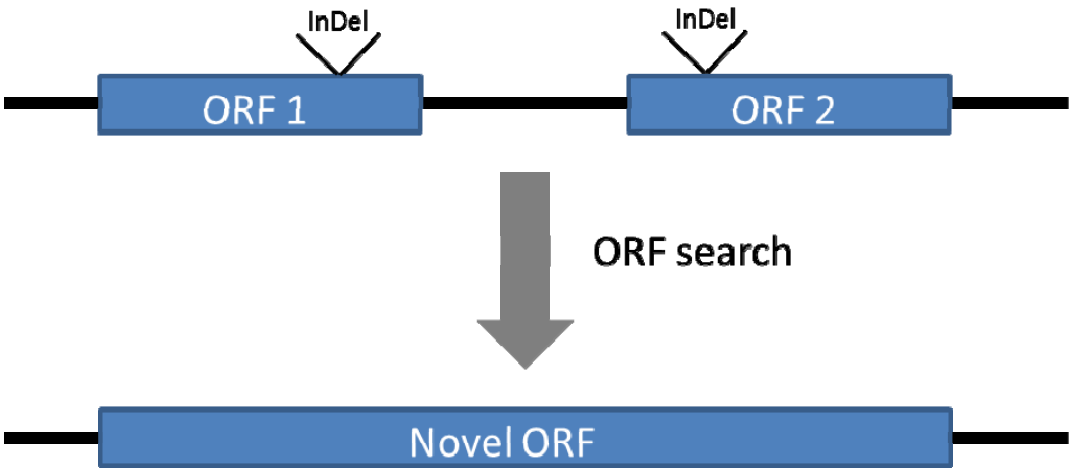

Figure 11. Scheme representing merging of two ORFs. InDels in two neighbouring ORFs resulted in frameshifts. ORF search lead to identification of a novel ORF spanning the region of the two original ORFs.

Additionally, sequencing of the engineered strain *G. oxydans* IK003.1 using long nanopore reads revealed another structural variant that was no mobile element insertion like the other three (see chapter 1.6.1). This structural variant is a 1,420 bp long sequence with 94% identity to a genomic region in *G. oxydans* DSM3504 (RefSeq ID: NZ_CP004373) and contains a transposase and the full coding sequence of a putative glyoxylase family protein. By using Sanger sequencing and re-mapping of Illumina reads to an adjusted reference sequence bearing the novel sequence, we could find this novel sequence in all wild-type and engineered strains. Therefore, we assumed that this sequence was missed during the first sequencing of the genome. The region is flanked by two identical mobile elements, which could have hindered a correct assembly of sequences. We also included the novel sequence with the additional transposase (GOX_RS13232) and the full CDS of the putative glyoxylase family protein (GOX_RS13233) into the update of the reference sequence (Figure 12).

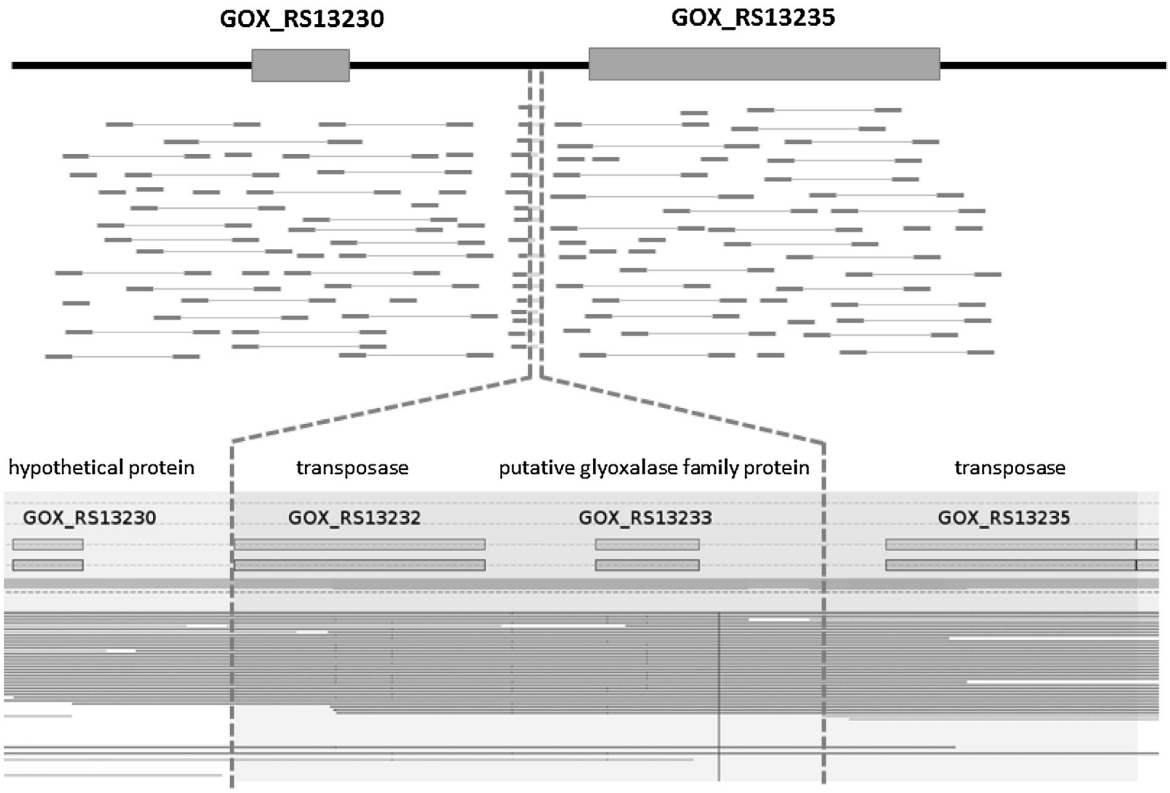

Figure 12. Scheme of read mapping and assembly at the GOX_RS13230-GOX_RS13235 locus. Short Illumina reads (upper panel) indicate a possible structural variant between GOX_RS13230 and GOX_RS13235 (unaligned read ends in light grey). *De novo* assembly using long nanopore reads resolved a novel transposon-flanked sequence insertion (lower panel). Two novel ORFs were identified encoding a transposase (GOX_RS13232) and a putative glyoxalase family protein (GOX_RS13233).

1.6.3 High-resolution transcriptome analysis of *G. oxydans*

(Kranz et al. 2018, BMC Genomics, chapter 2.2)

RNAseq has become an exceptional method to comprehensively study bacterial transcriptomes (chapter 1.4.2). Here, we sequenced the whole and primary transcriptome of *G. oxydans* to obtain as many transcripts and transcription start sites (TSSs) as possible. All identified TSSs were manually verified and classified according to their genomic context (Figure 13A). In total, 2,449 TSSs were detected. 134 of them belonged to genes for rRNAs, tRNAs, and RNase P. Of the remaining 2,315 TSSs, 994 belong to annotated genes and are located within a maximal distance of 300 nt upstream of the translational start site. They were named sense TSS (sTSS). 360 TSSs are also found upstream of annotated genes, but with a distance of >300 nt and <600 nt. They were therefore classified as putative TSSs (pTSSs). Also, 621 intragenic TSSs (iTSSs), which are located in sense orientation within coding regions and 683 antisense TSSs (asTSSs), which are located in antisense orientation within coding regions, were detected. 24 TSSs could not be assigned to any of these categories and were therefore classified as novel TSSs (nTSSs) belonging to possible novel transcripts. Since a 5´- or 3´-UTR of a neighbouring gene may overlap with a TSS already assigned to a category, some TSSs can be found in more than one category (Figure 13B). We also analysed nucleotide frequencies at the transcription initiation site for all TSSs (Figure 13C). For sTSSs and pTSSs, which are assigned to protein-coding genes, the most frequent initiation nucleotides are purines (65% and 75% A+G, respectively). This mean distribution was also observed in other bacteria (Kröger et al., 2012; Mendoza-Vargas et al., 2009; Schlüter et al., 2013) and was related to a relatively larger pool size of purine *versus* pyrimidine nucleotides increasing the transcription initiation rate in the cell (Buckstein et al., 2008). In contrast, the A+G frequency for iTSSs and asTSSs are lower (51% and 47%). The shift from 35% T+C for sTSSs to 51% T+C for asTSSs could reduce the rate of transcription initiation due to a smaller pool size of pyrimidine nucleotides, which could contribute to the overall tendency of lower antisense transcript levels in *G. oxydans* and other bacteria in comparison to the expression values of the respective sense transcripts.

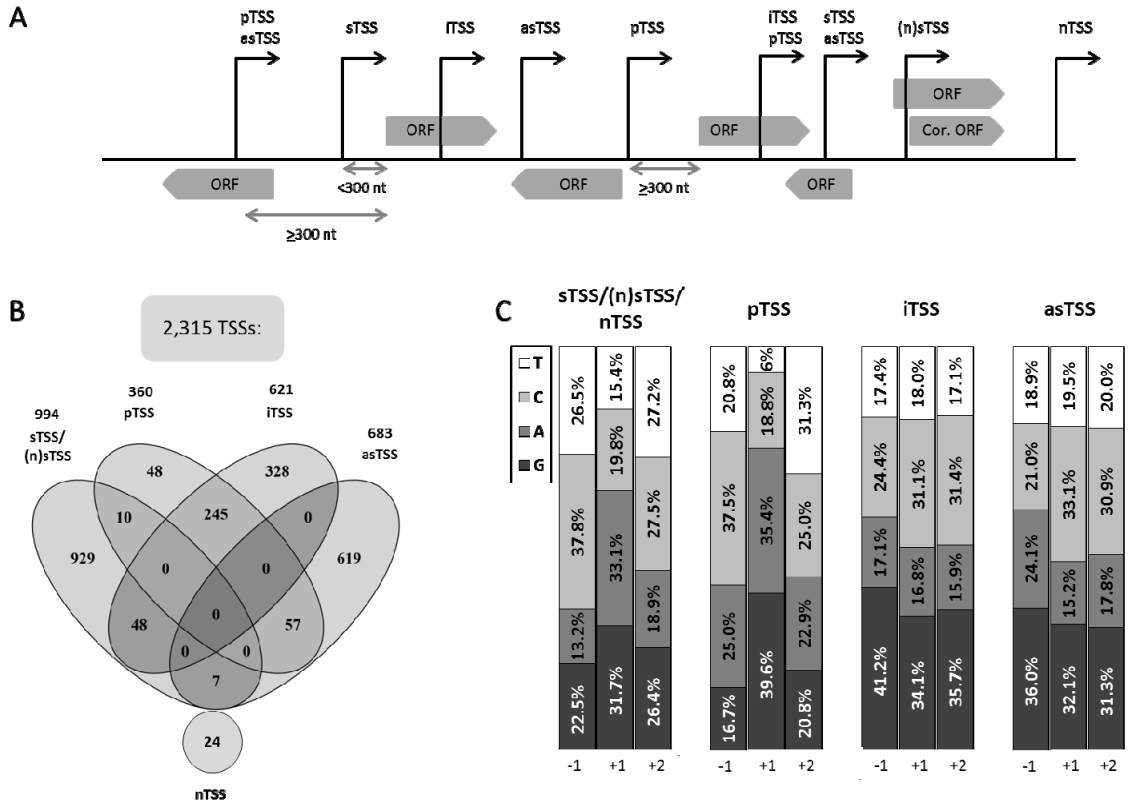

Figure 13. Classification of TSSs. (A) Schematic overview of categories used for classification of TSSs according to their genomic context. sTSS: sense TSS with an annotated gene downstream in a maximal distance of 300 nt. (n)sTSS: TSSs downstream of an ORF start, which were used to revise the translation start position (corrected ORF). pTSS: putative TSS assigned to annotated genes downstream with a minimal distance of 300 nt and a maximal distance of 600 nt. iTSS: intragenic TSS in sense orientation. asTSS: TSS in antisense orientation to coding regions. nTSS: intergenic TSS representing possible novel transcripts. Also, possible scenarios with TSSs associated with more than one category are shown. (B) Number and classification of detected TSSs. (C) Upper panel: Nucleotide distribution at the transcription initiating site +1 as well as at -1 and +2 based on the TSSs identified solely for the categories sTSS, pTSS, iTSS, and asTSS. The 10 TSSs assigned both to sTSS and pTSS were assumed to be sTSSs. Lower panel: nucleotide distribution at transcription initiation site +1 for the TSSs with the highest (top 10%) and lowest (low 10%) read start coverage.

In total, a TSSs was found for 1,073 (40%) of the 2,710 annotated protein-coding genes in the genome of *G. oxydans*. The number of TSS does not reflect the number of expressed genes, because many genes in bacteria are expressed polycistronically. Operon analysis based on whole transcriptome data indicated that 1,144 genes (41%) are expressed monocistronically, whereas 1,634 genes (59%) belong to 571 operons. Furthermore, 341 sub-operons with internal TSSs within operons were detected

comprising 720 genes. Taken the organisation in operons and sub-operons into account, the 1,073 identified TSSs control expression of 1,463 genes (54%) in total.

Based on TSSs detected by sequencing of the primary transcriptome it was possible to analyse 5´-UTRs of protein-coding genes. 62 (5%) of all mRNAs with a TSS were leaderless (≤3 nt distance), whereas 94% of all 5´-UTRs with a maximal length of 300 nt were longer than 10 nt. 43% of all 5'-UTRs were longer than 100 nt and shorter than 300 nt. These long leader sequences may contain regulatory elements such as riboswitches. Predictions of such elements can be found in the Rfam database (Nawrocki et al., 2015). For the *G. oxydans* genome, the FMN riboswitch, the glycine riboswitch, the SAM-II-riboswitch, and the TPP riboswitch were detected once in front of the expected genes downstream of the experimentally determined TSS. Further investigations are necessary to identify possible unknown regulatory elements in other long 5´-UTRs.

As expected, the most frequent initiation codon in *G. oxydans* is ATG (ca. 84%), followed by GTG (ca. 9%), TTG (ca. 3%), and CTG (ca. 2%). The translation start codon is known to influence translation initiation (Stenstrom et al., 2001). Other important factors are the RBS sequence and the distance between RBS and translation start codon (Makrides, 1996). Based on the RNAseq data, we identified and analysed RBSs in *G. oxydans*. The conserved motif “aGGAg”, which was found in 94 % of all analysed sequences, represents the reverse complement of the 3´-end of the 16S ribosomal RNA. Also, a mean spacing of 7.9 ± 2.8 nt between the end of the RBS and translation start codon was found. Information about initiation codon, RBS, and distance between RBS and initiation codon are very important and will be advantageous for further metabolic engineering approaches with *G. oxydans* including the development of expression systems.

We also searched for promoter motifs upstream of detected TSSs assigned to protein-coding genes and found the weakly conserved -10 region “nAtnnn” and the -35 region “ttGnnn” (Figure 14). In many bacteria, for example *Escherichia coli* or *C. glutamicum*, the -10 region “TATnnT” is highly conserved. The promoter motif of *G. oxydans* is based on TSS data from different primary transcriptome libraries generated from cells grown under standard and stress conditions. Therefore, the promoter motif does not solely represent the binding sites of σ^{70}, which is the sigma factor responsible for transcription of housekeeping genes (Hawley and McClure, 1983), but also binding sites of alternative sigma factors. Those alternative sigma factors regulate gene expression under different stress conditions (Güell et al., 2011; Paget and Helmann, 2003). Interestingly, highly abundant transcripts (top 5%) showed a highly

conserved “T” (90%) at the first position of the -10 region “Tatnnn”. Therefore, a simple search for conserved motifs based on all sequences upstream of a TSS assigned to protein-coding genes may not be sufficient for a comprehensive study of promoters. Further analyses are necessary to get deeper insights into promoter structures of *G. oxydans*. Such grouping and detailed analysis of promoters was recently performed for *C. glutamicum* (Albersmeier et al., 2017).

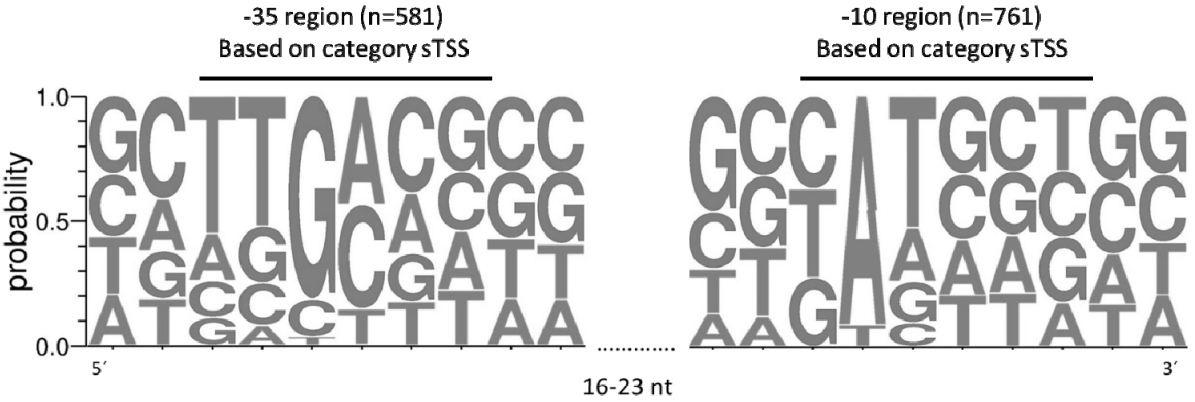

Figure 14. Promoter motifs found upstream of identified TSSs assigned to protein-coding genes.

Detection of TSSs also allowed the analysis of novel transcripts. Here, we distinguished between intragenic transcripts, antisense transcripts, and novel intergenic transcripts. In total, 18 novel transcripts in intergenic regions with a potential ORF were identified. A possible RBS was detected for nine of these 18 ORFs and homologous proteins were found for six of them. The other 12 novel transcripts without significant BLAST hits may represent novel small proteins or non-coding RNAs. Intragenic transcripts were detected for 12% of all protein-coding genes in *G. oxydans*. They may represent alternative mRNAs encoding smaller proteins, novel protein-coding genes or non-coding RNAs (Denoeud et al., 2007; Mitschke et al., 2011). For 313 out of 619 antisense TSSs, a corresponding transcript was found in the whole transcriptome data. Therefore, antisense transcripts were found for 11% of all protein-coding genes. It is assumed that these non-coding RNAs have regulatory roles in gene expression, for example by enabling transcription termination due to the formation of secondary structures or by blocking the RBS and therefore translation (Thomason and Storz, 2010). A GC-rich promoter motif was found upstream of the antisense TSSs with no similarity to the promoter motif found for protein-coding genes. It is assumed that this is due to the

location of most of these promoters within coding regions, where the GC content is higher (Bohlin et al., 2017). Also, it was shown that promoters of antisense transcripts are rarely conserved between or even within species and that their expression is usually lower than that of the sense transcripts (Nicolas et al., 2012; Raghavan et al., 2011; Shao et al., 2014). Due to their low expression, it is assumed that they are rather by-products of the transcription machinery (Raghavan et al., 2011).

To sum up, our RNAseq analysis provided deep insights into the transcriptional landscapes of *G. oxydans* and is an excellent basis for the further analysis of regulatory networks, for rational strain development, and for targeted gene expression in this bacterium.

1.6.4 Global mRNA decay analysis

(Kranz et al., BMC Genomics, chapter 2.3)

The stability of mRNA plays an important role in post-transcriptional regulation of gene expression and can influence the production rate of proteins. Therefore, we performed mRNA decay analysis in addition to transcriptome analysis in order to identify possible bottlenecks in the metabolism of *G. oxydans*. mRNA half-lives can be determined by treating bacterial cells with rifampicin to stop transcription (Lin et al., 2008). In our study, we performed DNA microarray analysis to estimate mRNA half-lives based on the ratios of three time points after rifampicin treatment (2 min, 5 min, and 10 min). The relative mRNA level changes were normalized *via* internal controls and then used for mRNA half-life calculation.

In total, based on the four time points we could determine a half-life for 2,500 (95%) genes with a regression coefficient R^2>0.7. Generally, half-lives in *G. oxydans* range from 3 min to 25 min with an average half-life of 5.7 min. Comparison of mRNA half-lives with ORF lengths showed no correlation at all. This could reflect site specificity of endo- and exonucleases. If cleavage sites are evenly distributed across gene sequences, longer transcripts would be more unstable than shorter transcripts. Since mRNA abundance in the cell results from the ratio of transcription rate and degradation rate, we compared mRNA half-lives with expression values (FPKM = fragments per kilo base of gene per million fragments mapped) based on RNAseq data. FPKMs were used as expression values, because they are solely based on paired end reads, which could be mapped as a pair over the complete read length to the reference

gene (fragment). If a read of a pair overlaps between an annotated gene and the intergenic region, the read pair is ignored for calculation of the FPKM value. This allows normalization of reads mapped only to annotated genes (Trapnell et al., 2010). Similar to observations in other mRNA decay studies (Andersson et al., 2006), linear regression analysis showed a statistically significant slightly inverse relationship between transcript stability and abundance in *G. oxydans*. This may indicate that, in general, mRNA stability plays not the same important role as transcription in the regulation of the steady-state mRNA level in bacterial cells. Indeed, it seems reasonable that mRNA stability is rather important for the adaptation to environmental changes by enabling a quick turnover of highly expressed genes due to their low mRNA stability. However, this inverse relationship is not the case for all genes. There are also genes with high FPKM values and high stability and *vice versa*. For example, genes encoding the chaperones GroES and GroEL, which are required for proper protein folding, were highly expressed and their transcripts were stable. This probably reflects the importance of these genes for cellular functions. In contrast, certain transcripts exhibited also low abundance and stability. Among them are genes belonging to the amino acid metabolism, purine biosynthesis, and tryptophan biosynthesis. Low transcript abundances and short half-lives may result in low protein level and could therefore represent possible bottlenecks in the metabolism of *G. oxydans*.

Studies in other bacteria showed a correlation between mRNA stability and the function of gene products (Andersson et al., 2006; Bernstein et al., 2002; Hambraeus et al., 2003). Thus, we also assigned mRNA half-lives to functional categories and performed a statistical analysis to determine mean half-lives of functional categories that differed significantly from the overall average half-life (5.7 min). Categories with significantly shorter half-lives were ATP-proton motif force interconversion, transcription, fatty acid and phospholipid metabolism, nucleotide metabolism, and degradation of proteins and peptides. The subsets with significantly longer half-lives were sugar and alcohol degradation, DNA restriction and modification, cell motility, and ion homeostasis. In general, it is assumed that transcripts of genes involved in housekeeping functions have longer half-lives, whereas transcripts of genes with regulatory functions or involved in stress responses are less stable. Similar to other organisms (Bernstein et al., 2002), transcripts belonging to the functional categories amino acid synthesis and nucleotide biosynthesis are among the least stable in *G. oxydans*. This allows a fast adaptation of the transcription and translation machinery due to environmental changes. Also, cell

motility and ion homeostasis are cellular functions that need to be maintained independent of the growth condition.

Whereas genes assigned to energy metabolism typically have longer mRNA half-lives than average in other bacteria, the stability of these genes strongly varies in *G. oxydans*. Especially the genes encoding the H^+-translocating F_1F_o-ATP synthase (GOX_RS06715-GOX_RS06730 and GOX_RS07740-GOX_RS07760) exhibited short transcript half-lives ranging from 2.85 min to 3.72 min. *G. oxydans* possesses a second ATP synthase, which is an ortholog to Na^+-translocating ATP synthases. Transcripts assigned to this second ATP synthase snowed an average half-life of 5 min, but 80-fold lower expression values than the genes assigned to the H^+-translocating ATP-synthase. Comparative DNA microarray studies suggested that the Na+-translocating ATP synthase may play an important role during oxygen limitation (Hanke et al., 2012). Whether the low transcripts half-lives of the H^+-translocating ATP-synthase genes have any effect on the energy metabolism of G. oxydans need to be further investigated. Genes assigned to energy metabolism with the highest mRNA half-lives were those of triosephosphate isomerase (*tpi*, GOX_RS12375) with 12.4 min and of dihydroxyacetone kinase (*dhak*, GOX_RS12400) with 19.1 min. Neighbouring genes (GOX_RS12365-GOX_RS12400) show also quite high half-lives ranging from 9.8 to 19 min. Those genes encode a glycerol-3-phosphate dehydrogenase (GOX_RS12365), a hypothetical protein (GOX_RS12370), a ribose-5-phosphate isomerase (GOX_RS12380), and two of three components belonging to a ribose ABC transporter (GOX_RS12385). These genes form an operon. Two glycerol-3-phosphate dehydrogenases are encoded in the genome of *G. oxydans*. The second one (GOX_RS11720) besides GOX_RS12365 forms an operon together with the genes for a glycerol uptake facilitator protein (GOX_RS11725) and a glycerol kinase (GOX_RS11730) and is involved in glycerol uptake and metabolism. Since GOX_RS12365 is located closely to genes assigned to glycolysis, PPP, and ribose uptake, it is possible that this glycerol-3-phosphate dehydrogenase plays another role. Whether the high average half-life of this operon (GOX_RS12365-GOX_RS12400) may play a role in the energy metabolism of *G. oxydans*, need to be further investigated.

1.6.5 Fragmentation of 23S rRNA in *G. oxydans*

(Kranz et al., BMC Genomics, chapter 2.3)

During several experiments that required isolation of RNA, we observed an underrepresentation of the 23S rRNA band compared to the 16S rRNA band in agarose gels (Figure 15A). Therefore, we enriched ribosomes from bacterial cells and isolated the RNA bound to the ribosomes. First, the RNA was analysed by gel electrophoresis. It showed a distinct fragment pattern in agarose gels (Figure 15B). Three bands with different intensity ranging in size between 800 nt and 1,000 nt and two middle-sized fragments at 300 and 400 nt were detected. Differences in intensity of these bands were observed between exponential and stationary phase.

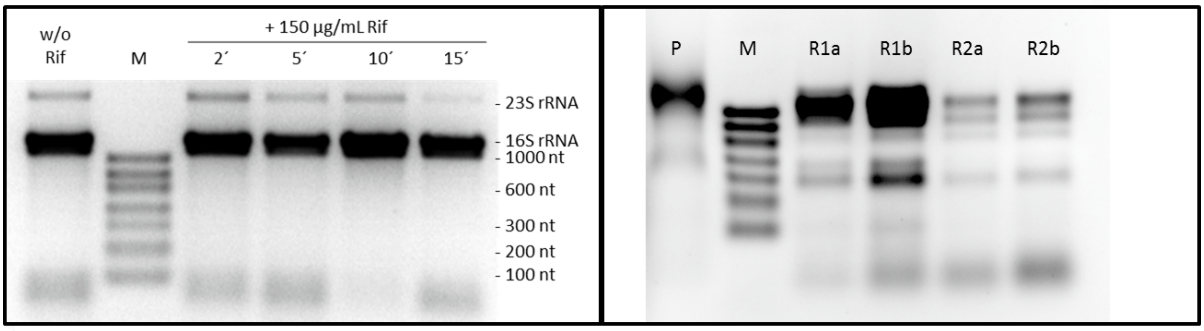

Figure 15. (A) Formaldehyde agarose gel analysis to inspect the quality of total RNA isolated from cells before and 2, 5, 10, and 15 min after addition of rifampicin (chapter 1.6.4). Bands corresponding to the 23S rRNA (2709 to 2711 nt) and 16S rRNA (1478 nt) are indicated. M = RiboRuler Low Range RNA Ladder. (B) Pattern of rRNA fragments isolated from enriched ribosomes. P, Protein fraction; M, RiboRuler low range ladder (Thermo Fisher Scientific); R1a, 1 µl (1,232 ng/µl) of RNA (exponential phase); R1b, 2 µl (1,232 ng/µl) of RNA (exponential phase); R2a, 1.5 µl (877 ng/µl) of RNA (stationary phase); R2b, 3 µl (877 ng/µl) of RNA (stationary phase).

To further analyse the rRNA fragments, we sequenced them *via* NGS and mapped the sequencing reads to the four 16S and 23S rRNA genes of *G. oxydans* (Figure 16). No possible fragmentation position was found for the 16S rRNA genes, because the read coverage was higher than 40 over the complete gene length. Three to four possible fragmentation positions with a coverage below 5% of the average coverage were visible in the mapping to the 23S rRNA genes. Three were present in all four 23S

rRNA genes (GOX_RS07780, GOX_RS02255, GOX_RS08565, and GOX_RS06970) and one additional position with lower coverage was found only in GOX_RS06970.

Since fragmentation of 23S rRNA is known from many bacteria, we compared 23S rRNA sequences from *G. oxydans* to those of selected bacteria. In general, one can distinguish between fragmentation of 23S rRNA due to the excision of intervening sequences (IVSs) or *via* plain cleavage. Also, fragmentation positions in the sequences are quite diverse and seldom showed similarities. Alignment of the 23S rRNA genes showed no known IVSs in the 23S rRNAs of *G. oxydans*. However, there are nucleotide differences present, which could belong to possible cleavage sites in the 23S rRNA of *G. oxydans*. Also, two nucleotide differences between GOX_RS06970 and the other three 23S rRNA genes are present and could enable the additional fourth fragmentation. The physiological role of fragmented 23S rRNA in bacteria is still under investigation. However, it was shown that integration of known IVSs in intact 23S rRNA genes of *E. coli* had no effect on growth behaviour (Gregory et al., 1996Gregory et al., 1996). Further experiments are necessary to analyse, whether fragmentation of 23S rRNA in *G. oxydans* has any consequences.

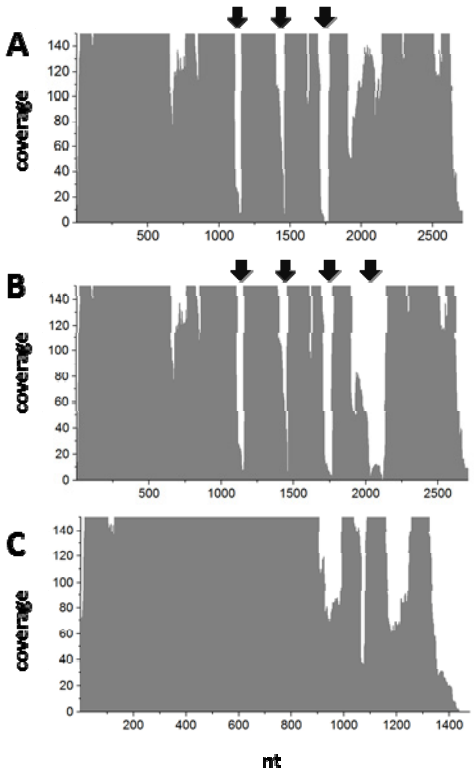

Figure 16. Detailed view of the coverage for 16S and 23S rRNA gene loci based on mapping of reads from RNA samples of enriched ribosomes isolated during the exponential growth phase of *G. oxydans*. Arrows indicate regions with a coverage <5% of the average coverage for the complete gene. (A) GOX_RS07780 (23S rRNA). The same coverage pattern was observed for GOX_RS02255 and GOX_RS08565 (not shown). (B) GOX_RS06970 (23S rRNA). (C) GOX_RS06955 (16S rRNA). The same coverage pattern was observed for GOX_RS02270, GOX_RS07765, and GOX_RS08550 (not shown).

1.7 Conclusions and Outlook

In this study, comprehensive genome and transcriptome analysis of *G. oxydans* strains were performed. Sequencing of genomically engineered strains in comparison to wild-type and reference strains showed that their genomes are very stable despite several drastic metabolic engineering steps. Therefore, the final engineered strain IK003.1 is a suitable host for applications and further efforts to improve growth and biomass yield. Furthermore, the data of the high precision genome sequencing of several wild-type strains resulted in an update of the genome reference sequence, which was used for the RNAseq analysis.

Sequencing of primary and whole transcriptomes of *G. oxydans* was performed to characterize transcriptional landscapes of this bacterium. Thereby, TSSs were identified, 5´-UTRs were analysed, consensus promoter and RBS motifs were identified, novel

transcripts including antisense transcripts were detected, and organization of genes in operons was resolved. These results expanded the knowledge about *G. oxydans* and will be very helpful for rational strain design. For example, the data provide information on promoters and their apparent strength or support identification of suitable intergenic regions that can be used as integration loci for gene constructs. Nevertheless, further analysis of the next-generation sequencing data is still necessary. For example, the -10 region “nAtnnn” of the consensus promoter motif differs, especially at the first position of the hexamer, which is in contrast to the known conserved motifs of other bacteria. Whether this is characteristic for *G. oxydans* or a result of the bioinformatic data analysis need to be further investigated.

Analysis of the global mRNA decay in *G. oxydans* by assigning mRNA half-lives to functional categories showed a good overall agreement with findings in other bacteria. For example, genes involved in fatty acid and phospholipid metabolism exhibited rather short mRNA half-lives, whereas longer mRNA half-lives were found for genes assigned to housekeeping functions. However, short mRNA half-lives of genes encoding the H^+-dependent ATP synthase may indicate a possible bottleneck in the energy metabolism of *G. oxydans*. In the central carbon metabolism, the FPKM expression values of TCA cycle genes were among the lowest and the mRNA half-lives of many TCA cycles genes were below the global mean. Both ATP synthase and TCA cycle gene expression should be considered in future metabolic engineering approaches to further improve growth and biomass yield of *G. oxydans*.

2. Publications

2.1 Genomic DNA sequencing of wild-type and engineered pathway-restored *G. oxydans* strains

Kranz, A., Vogel, A., Degner, U., Kiefler, I., Bott, M., Usadel, B., and Polen, T. (2017). High precision genome sequencing of engineered *Gluconobacter oxydans* 621H by combining long nanopore and short accurate Illumina reads. **J Biotechnol 258: 197-205.**

Author contributions:

AK performed genome sequencing with the Illumina platform, subsequent data analysis, updated the reference genome sequence, and wrote 60% of the manuscript. Shared first author AV performed sequencing with the MinION platform, subsequent data analysis, and wrote 15% of the manuscript. UD performed Sanger sequencing for validation of detected variants. IK cultivated engineered strains in the DASGIP cultivation system. MB supported the study. BU and TP designed and supervised the study. TP wrote 25% of the manuscript, and revised and finalized the manuscript.

Overall contribution AK: 70%

Journal of Biotechnology 258 (2017) 197–205

Contents lists available at ScienceDirect

Journal of Biotechnology

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

High precision genome sequencing of engineered *Gluconobacter oxydans* 621H by combining long nanopore and short accurate Illumina reads

Angela Kranz[a,d,1], Alexander Vogel[b,c,d,1], Ursula Degner[a,d], Ines Kiefler[a,d], Michael Bott[a,d], Björn Usadel[b,c,d], Tino Polen[a,d,*]

[a] Institute of Bio- and Geosciences IBG-1: Biotechnology, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany
[b] IBMG: Institute for Biology I, RWTH Aachen University, Worringer Weg 2, 52074 Aachen, Germany
[c] IBG-2 Plant Sciences, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany
[d] The Bioeconomy Science Center (BioSC), c/o Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

ARTICLE INFO

Keywords:
Metabolic engineering
Gluconobacter oxydans
MinION® nanopore device
Long reads library
Structural variants
Genome assembly

ABSTRACT

State of the art and novel high-throughput DNA sequencing technologies enable fascinating opportunities and applications in the life sciences including microbial genomics. Short high-quality read data already enable not only microbial genome sequencing, yet can be inadequately to solve problems in genome assemblies and for the analysis of structural variants, especially in engineered microbial cell factories. Single-molecule real-time sequencing technologies generating long reads promise to solve such assembly problems. In our study, we wanted to increase the average read length of long nanopore reads with R9 chemistry and conducted a hybrid approach for the analysis of structural variants to check the genome stability of a recombinant *Gluconobacter oxydans* 621H strain (IK003.1) engineered for improved growth. Therefore we combined accurate Illumina sequencing technology and low-cost single-molecule nanopore sequencing using the MinION® device from Oxford Nanopore. In our hybrid approach with a modified library protocol we could increase the average size of nanopore 2D reads to about 18.9 kb. Combining the long MinION nanopore reads with the high quality short Illumina reads enabled the assembly of the engineered chromosome into a single contig and comprehensive detection and clarification of 7 structural variants including all three known genetically engineered modifications. We found the genome of IK003.1 was stable over 70 generations of strain handling including 28 h of process time in a bioreactor. The long read data revealed a novel 1420 bp transposon-flanked and ORF-containing sequence which was hitherto unknown in the *G. oxydans* 621H reference. Further analysis and genome sequencing showed that this region is already present in *G. oxydans* 621H wild-type strains. Our data of *G. oxydans* 621H wild-type DNA from different resources also revealed in 73 annotated coding sequences about 91 uniform nucleotide differences including InDels. Together, our results contribute to an improved high quality genome reference for *G. oxydans* 621H which is available *via* ENA accession PRJEB18739.

1. Introduction

Advances in synthetic biology, where a variety of tools including novel DNA construction technologies, synthetic regulatory circuits and genetic parts for precise control of expression are developed, increasingly enable (multiplexed) microbial genome engineering for the construction of plasmidless, markerless recombinant strains carrying targeted short- and long-length chromosomal insertions, deletions, exchanges or rearrangements for fundamental and industrial purposes (Chiang et al., 2008; Chung et al., 2016; Hook et al., 2016; Phelan et al., 2016; Smanski et al., 2016; Tyo et al., 2009). The construction and application of such genomically extensively modified microbes also require high or at least sufficient stability of the engineered genome. Generally, the genome stability is affected by several naturally evolved specialized genetic elements including a number of mobile elements (Burrus and Waldor, 2004; Darmon and Leach, 2014; Nagy and Chandler, 2004). Among them, transposons are numerously present in most genomes and can move or copy a locus by integration reactions that are independent of homologous recombination. While transposons and other systems are increasingly applied for metabolic engineering of microbial cell factories to redirect carbon flux towards desired products and for screenings (Choi et al., 2006; Flagfeldt et al., 2009; Loeschcke

* Corresponding author.
E-mail address: t.polen@fz-juelich.de (T. Polen).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jbiotec.2017.04.016
Received 2 January 2017; Received in revised form 14 April 2017; Accepted 15 April 2017
Available online 19 April 2017

et al., 2013; Martinez-Garcia et al., 2014; Miyazaki and van der Meer, 2013; Nikel and de Lorenzo, 2013; Rawsthorne et al., 2006; Warner et al., 2009; Zobel et al., 2015), the mobile elements present in engineered hosts can generally interfere by causing structural variants in any intergenic as well as intragenic region. Also, non-canonical integration events are possible affecting the variability of clones with different productivities as well as aberrant morphological or growth characteristics (Schwarzhans et al., 2016). Furthermore, suppressor mutations on the single nucleotide level can arise, all together may affect outcome and phenotype of planned cell factories. This requires comprehensive and reliable sequencing of engineered genomes to confirm the intended genetic modifications and to reveal unintentional integrations, structural variants, and nucleotide changes.

In the last years, the next-generation sequencing (NGS) platforms changed the field of microbial genomics by enabling sequencing of numerous genomes in parallel for various applications. Among them are monitoring of genomes from laboratory strains to analyze accumulation of mutations over a long storage time (Ding et al., 2015; Lee et al., 2009; Shiwa et al., 2013), identification of beneficial mutations in strains after adaptive evolution for production of chemicals and amino acids (Hong et al., 2011; Mahr et al., 2015), and checking of engineered strains for possible suppressor mutations (Hochheim et al., 2016; Komati Reddy et al., 2015). Illumina's sequencing platform turned out to offer the best NGS quality and accuracy in real life performance making it highly interesting for all genomes and other DNA including ChIPseq/ChAPseq samples (Erguner et al., 2015; Loman et al., 2012; Ong et al., 2013; Pfeifer et al., 2016). However, its relatively short read length of up to 150 or 300 nt makes it very challenging or impossible to dissect and clarify structural variants detected in comparison to a reference or engineered genome, especially in case of transposon insertions and multicopy integration of heterologous genes (Choi et al., 2006; Rawsthorne et al., 2006; Tattini et al., 2015). Long read sequencing platforms with read length of several kb offer to solve such assembly problems (Goodwin et al., 2015; McCoy et al., 2014). To the whole research community the best accessible system is the recently released MinION® device, a single-molecule nanopore sequencer connected to a computer *via* USB, from Oxford Nanopore Technologies Ltd (Brown and Clarke, 2016). Studies using nanopore chemistry R9 or earlier showed the production of long reads with an average size of 6 kb up to 10 kb, yet error rates of 12%–35% were observed (Ashton et al., 2015; Madoui et al., 2015; Mikheyev and Tin, 2014; Quick et al., 2014). Nevertheless, these data have the potential to assist genome sequencing and assembly in hybrid approaches by combining long nanopore reads with short accurate Illumina reads. Error rates are currently and will be further decreased by improved nanopore chemistry and tools for error correction (Goodwin et al., 2015; Karlsson et al., 2015; Lu et al., 2016; Madoui et al., 2015).

In this study we tested a R9 chemistry MinION® library protocol to obtain improved nanopore read length with average size doubled up to about 20 kb and applied it in a hybrid approach with Illumina's short accurate reads for genome sequencing of an iteratively engineered *Gluconobacter oxydans* strain exhibiting improved growth (Kiefler et al., 2015, 2017). The α-proteobacterium *G. oxydans* is a gram-negative acetic acid bacterium used for a broad range of industrial applications due to its ability to oxidize a great variety of carbohydrates in the periplasm. Prominent biotransformation products among others are dihydroxyacetone, 5-ketofructose, and 2-keto-L-gulonic acid, which is used as precursor for vitamin C production (Ameyama et al., 1981; Deppenmeier et al., 2002; Gupta et al., 2001; Hekmat et al., 2003; Hölscher et al., 2009; Kosciow et al., 2016; Saito et al., 1997; Tkác et al., 2001; Wang et al., 2016). However, the low biomass yield of *G. oxydans* resulting from the high periplasmic oxidation of substrates combined with low carbon flux into the cytoplasmic metabolism is disadvantageous for a broader range of industrial applications (Greenfield and Claus, 1972; Hanke et al., 2013; Prust et al., 2005). According to the genome sequence, the glycolysis and the tricarboxylic acid (TCA) cycle are incomplete because of missing genes encoding phosphofructokinase, succinyl-CoA synthetase, and succinate dehydrogenase. This genome sequence was determined by Sanger sequencing with a whole-genome shotgun approach using plasmid and cosmid libraries already more than ten years ago and revealed a chromosome of 2.7 Mb containing 2432 open reading frames (ORFs) and five plasmids with 232 ORFs (Greenfield and Claus, 1972; Hanke et al., 2013; Prust et al., 2005). However, only recently heterologous genes for a succinate dehydrogenase, a succinyl-CoA synthetase, as well as a NADH dehydrogenase were chromosomally integrated and expressed to complete the TCA cycle and resulting in about 60% increased biomass yield (Kiefler et al., 2017). Enforcing the cytoplasmic glucose catabolism in *G. oxydans* in such a way may cause suppressor mutations and structural variants affecting the genome stability. Therefore, in this study we applied our modified nanopore protocol to check the genome sequence of the engineered *G. oxydans* strain after up to 70 generations of strain handling including process time in a controlled bioreactor using a hybrid approach. We found the engineered genome was very stable for up to 70 generations, making it a suitable engineered host for further metabolic engineering efforts. Surprisingly, a novel 1420 bp long transposon-flanked and ORF-containing sequence was revealed by the long read data. Further analysis and genome sequencing showed that this region, hitherto unknown in the *G. oxydans* 621H reference, is already present in *G. oxydans* 621H wild-type strains. Furthermore, accurate Illumina reads of *G. oxydans* 621H wild type DNA from different resources revealed in 73 annotated coding sequences about 91 uniform nucleotide differences including InDels which could not be revealed by the Sanger sequencing of the *G. oxydans* 621H reference. Together, the sequencing data of the hybrid approach contribute to an improved high quality genome reference of *G. oxydans* 621H.

2. Materials and methods

2.1. Cultivation of cells and DNA extraction

The bacterial strains used in this study are listed in Table 1. For extraction of genomic DNA from *G. oxydans* wild-type strains, cells were cultivated on mannitol medium containing 220 mM (4% w/v) mannitol, 5 g L^{-1} yeast extract, 1 g L^{-1} KH_2PO_4, 1 g L^{-1} $(NH_4)_2SO_4$, 2.5 g L^{-1} $MgSO_4 \times 7\ H_2O$, and 50 µg mL^{-1} cefoxitin as antibiotic. The initial pH of the medium is 6.0. Cultures were grown in 100 mL shaking flasks with three baffles containing 15 mL mannitol medium for about 16 h (30 °C, 140 rpm) and then harvested by centrifugation (4000 × *g*, 6 min). Genomic DNA from *G. oxydans* Δ*upp* and metabolically engineered strains were obtained from cells cultivated under pH- and oxygen-controlled conditions in the DASGIP cultivation system for 24 or 28 h as described (Kiefler et al., 2015, 2017).

Genomic DNA was isolated as described for *Corynebacterium glutamicum* (Eikmanns et al., 1994) with the following modifications: After harvesting 5 mL, cell pellets were washed in TE buffer and treated with 1 mL of TE buffer and 15 mg of lysozyme (37 °C, 1 h). Afterwards, 3 mL of lysis buffer (10 mM Tris-HCl, 400 mM NaCl, 2 mM EDTA; pH 8.2), 100 µl RNase A (10 mg/mL), 220 µl of 10% SDS, and 150 µl of Proteinase K were added to the mixture and incubated overnight at 37 °C. For precipitation of DNA, 2 mL of NaCl solution (6 M) was added, centrifuged (5000 × *g*, 30 min) and mixed with 2.5 volumes of ice-cold ethanol. Precipitated DNA was fished using a Pasteur pipette, washed with ethanol (75%) and resuspended in 200 µl of TE buffer. DNA concentrations in samples were determined using the Qubit system (Thermo Fisher Scientific) and quality-checked on agarose gels.

2.2. Library preparation and sequencing with the Illumina platform

Illumina sequencing libraries of nine samples analyzed in this study (Table 2) were prepared using the TruSeq DNA PCR-free sample preparation kit (Illumina) according to manufacturers' instructions.

Table 1
Strains used in this study.

G. oxydans strain	Relevant characteristics/genotype	Source/Reference
WT-E	DSM 2343 (ATCC 621H), wild type	Peters et al. (2013)
WT-BM	DSM 2343 (ATCC 621H), wild type	Herrmann et al. (2004)
WT-DSMZ	DSM 2343 (ATCC 621H), wild type from the German Collection of Microorganisms and Cell Cultures (DSMZ)	DSMZ, Germany
Δ*upp*	WT-E derivative with a deletion of *upp* (GOX0327/GOX_RS02795) coding for uracil phosphoribosyl-transferase	Peters et al. (2013)
IK001	Δ*upp*Δ*gdhS*:*sdhCDABE* Δ*upp* derivative with genomically integrated *sdhCDAB* (APA01_00310-00340) and *sdhE* (APA01_11050) from *Acetobacter pasteurianus* DSM3509 into the *gdhS* locus (GOX2015/GOX_RS11350)	Kiefler et al. (2017)
IK002.1	Δ*upp*Δ*gdhS*:*sdhCDABE*Δ*pdc*:*ndh* IK001 derivative with genomically integrated *ndh* from *G. oxydans* DSM3504 (GLS_c05650) into the *pdc* locus (GOX1081/GOX_RS06555)	Kiefler et al. (2017)
IK003.1	Δ*upp*Δ*gdhS*:*sdhCDABE*Δ*pdc*:*ndh*Δ*gdhM*:*sucCD* IK002.1 derivative with genomically integrated *sucCD* (GDI_2951-2952) from *Gluconacetobacter diazotrophicus* DSM5601 into the *gdhM* locus (GOX0265/GOX_RS02475)	Kiefler et al. (2017)

Briefly, 4 µg of genomic DNA was fragmented with the Biorupter® Pico (Diagenode) to an average fragment size of 550 bp. 2 µg of the fragmented DNA was end-repaired and size-selected using magnetic beads. After ligation of a single A nucleotide to the 3′ end, Illumina index PE adapters were ligated to the fragments. The indexed libraries were quantified *via* qPCR using the KAPA Library Quantification Kit (Peqlab). Normalized libraries (2 nM) were pooled, diluted to an average final concentration of 10 pM and paired-end sequenced on a MiSeq desktop sequencer (Illumina) with a read length of 2 × 150 or 2 × 250 bases.

2.3. Illumina reads processing, variant detection, and ORF search

An automated workflow for data analysis and variant detection was designed using tools of the CLC Genomics Workbench (Qiagen Aarhus A/S). The reads obtained in the MiSeq output were preprocessed by removing adapter sequences and by quality trimming to remove complete reads or sequence ends with a Phred quality value < 30 (Ewing and Green, 1998). Using default parameters, reads were mapped to the *G. oxydans* 621H reference genome (NC_006677) and the five plasmids pGOX1 to pGOX5 (NC_006672 to NC_006676). Non-specific matches were mapped randomly. To identify SNVs (single nucleotide variants), MNVs (multiple nucleotide variants), and InDels with a length of one to three nucleotides, the quality-based variant detection tool of the CLC Genomics Workbench was used. Default parameters were changed to call a variant when the quality score for the central nucleotide (variant) and five adjacent nucleotides was 20 or higher. Also, variants had to have a minimum frequency of 10% and coverage of ≥20. Variants found in regions with non-specific read alignments because of multicopy genes (*e.g.* transposases) or found within possible structural variants (SVs) were ignored. The individual variant lists of the nine samples were combined and analyzed using Excel (Microsoft) to check the occurrence and overlap of the detected variants in all samples.

For nucleotide changes in annotated genes the resulting amino acid change were reported. Genes with frameshifts due to InDels were further analyzed by extracting the new consensus sequence including 500 bp up- and downstream. With this appropriate new open reading frames (ORFs) were identified and confirmed by BLASTP search (Altschul et al., 1990).

2.4. Nanopore 20 kb library and MinION® flow cell preparation

MinION sequencing libraries with genomic DNA from strain IK003.1 were prepared using the Nanopore sequencing kit (R9) with a modified version of the genomic DNA (R9) protocol. Total genomic input DNA was upscaled to 5 µg in 150 µl of nuclease free water. Fragmentation was achieved by using a Covaris g-Tube at 4000 rpm for 120 s in an Eppendorf EP 5424 centrifuge. The sheared DNA was size-selected using a Sage Science BluePippin following the 20 kb high-pass protocol V3 with subsequent bead clean-up using an equal volume of Beckman Coulter AMPure XP beads. For improved clean-up recovery bead binding and elution time was increased to 15 min each on a Hula mixer at 37 °C. End repair was scaled up to adjust for the recovery of 2.5 µg. Adapter ligation was performed according to the manufacturer's instructions. For improved recovery of the Streptavidin selection step the incubation time for binding and elution was increased to 15 min, while the elution volume was decreased to 15 µl of elution buffer. In total 550 ng of pre-sequencing mix was obtained and stored on ice.

2.5. MinION® sequencing and reads processing

The Nanopore sequencing was performed on an Oxford Nanopore MinION Mk1b sequencer using a R9 flow cell. The flow cell was primed

Table 2
Read and mapping statistics of the Illumina reads obtained from genomic DNA of analyzed *G. oxydans* strains. Trimming and mapping was performed with CLC Genomics Workbench. The reference sequences of the chromosome (NC_006677) and the five plasmids pGOX1 to pGOX5 (NC_006672 to NC_006676) were obtained from NCBI.

strain	read length (bp)	reads (million)	reads after trimming (million)	mapped reads (million)	mapped reads (%)	reads mapped in pairs (%)	average genomic coverage
WT-DSMZ	2 × 150	2.81	2.38	2.36	99.16	95.34	112
WT-BM	2 × 150	2.39	1.97	1.96	99.14	95.94	91
WT-E	2 × 250	2.10	1.45	1.42	98.05	95.10	89
Δ*upp*	2 × 150	1.96	1.66	1.55	93.51	92.09	74
IK001	2 × 150	2.96	2.59	2.57	99.39	97.00	123
IK002.1	2 × 150	4.27	3.58	3.56	99.35	94.09	167
IK002.1[a]	2 × 150	2.66	1.86	1.85	99.63	95.56	83
IK003.1	2 × 150	3.68	3.26	3.24	99.26	97.31	155
IK003.1[b]	2 × 150	2.59	2.31	2.29	99.27	97.86	110

[a] After 24 h of growth under controlled conditions in a DASGIP bioreactor (Kiefler et al., 2017).
[b] After 28 h of growth under controlled conditions in a DASGIP bioreactor (Kiefler et al., 2017).

two times with 500 µl of priming mix (10 min). The library was prepared for loading by adding 75 µl of RBF1 and 63 µl of nuclease-free water to 12 µl of pre-sequencing mix. Upon completion of the priming, a total of 150 µl of prepared library was loaded using a P1000 pipette. The sequencing run was started and monitored *via* MinKNOW (v1.0.2) using the 48 h sequencing run protocol (FLO_MIN104). The base calling was performed in parallel using Metrichor (v1.107). Read sequences were extracted to FASTA format for quality-filtered 2D reads using poretools v0.5.1 (Loman and Quinlan, 2014).

2.6. Hybrid genome assembly and identification of structural variants

Short read assemblies for strain IK003.1 were computed using SPAdes v3.9.1 without error correction of raw data (Bankevich et al., 2012). Illumina short reads were trimmed using Trimmomatic v0.35 in paired-end mode using standard settings (Bolger et al., 2014). Success of trimming and general sequence data quality analysis was performed by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Hybrid assemblies were generated using Unicycler (https://doi.org/10.1101/096412) with SPAdes 3.9.1 and subsequent polishing with Pilon (Walker et al., 2014). Structural variants in IK003.1 were detected by genome-scale alignment of the hybrid assembly against the *G. oxydans* 621H reference genome (NC_006677) with plasmids pGOX1 to pGOX5 (NC_006672 to NC_006676) using LAST (version 712) (Frith et al., 2010). Based on the alignment the assembly was anchored to match the corresponding reference starts.

Gaps in the alignment of the query sequence were considered as potential insertions, while gaps in the alignment of the reference sequence were treated as potential deletions. Further verification of the initially detected variants was performed by mapping both, short read data and long read data, back to the hybrid assembly using bwa mem v0.7.15-r1140 and manual inspection of read depth and split read alignments for the candidate variants (Li and Durbin, 2010).

2.7. Validation of detected variants by Sanger sequencing

For validation of detected variants found by genome sequencing, the respective genomic regions 80–450 bp up- and downstream of 75 selected small variants (SNVs, MNVs, and InDels) and of all loci representing larger genomic alterations were amplified by PCR using the oligonucleotides listed in Tables S1 and S2, respectively. The resulting PCR products were Sanger sequenced (Eurofins MWG Operon) and analyzed to check the presence of the variants.

2.8. Data accessibility

The Illumina MiSeq, MinION® data and the resulting updates of the *G. oxydans* 621H reference are available in the European Nucleotide Archive under accession number PRJEB18739. Additional information, such as a Genome Browser, can be accessed *via* our online portal www.gluconobacterfactory.de.

3. Results

Recently, biomass yield of *G. oxydans* 621H was improved by 60% using heterologous genes for a succinate dehydrogenase, succinyl-CoA synthetase, and NADH dehydrogenase chromosomally integrated in *G. oxydans* 621H to complete the TCA cycle and improve energy metabolism (Kiefler et al., 2017). However, completing and enforcing the naturally incomplete cytoplasmic glucose catabolism in *G. oxydans* in such a way may cause suppressor mutations and structural variants possibly inactivating genes, thereby affecting the stability of the intended engineered genome on the long run. In this study, we checked this by sequencing genomes of the iteratively generated recombinant *G. oxydans* strain series and related references. For a combined approach using the Illumina and the MinION® nanopore platform, we tested a modified protocol to increase the average nanopore read length.

3.1. Sequencing and analysis of G. oxydans genomes using the Illumina platform

First we analyzed engineered strains and some WT references using the Illumina platform generating short accurate reads to screen for suppressor mutations and additional genomic alterations which may have emerged during or after the iterative strain construction. Therefore, genomic DNA samples of the engineered *G. oxydans* strains IK001, IK002.1, and IK003.1 as well as of the Δ*upp* reference were sequenced. To check the genome stabilities after integration of several genes in different loci, the genomic DNA of IK002.1 and IK003.1 was sequenced from cells cultivated in a bioreactor under defined conditions for 24 h and 28 h, respectively (Kiefler et al., 2017). Overall, this time point represents about 70 generations of strain handling and cultivation after obtaining desired clones in the strain construction procedure (Kiefler et al., 2017). As control we also sequenced the wild type reference (WT-E) of the *G. oxydans* Δ*upp* strain (Peters et al., 2013), which was used for the construction of the IK strains. As further controls, we also sequenced genomic DNA from cells of *G. oxydans* 621H wild type from our strain collection (WT-BM) and from the public resource (WT-DSMZ).

For minimizing DNA sequence bias, the sequencing libraries were prepared without additional amplification using the TruSeq DNA PCR-free sample preparation kit from Illumina and paired-end sequenced. The obtained sequencing reads were aligned to the *G. oxydans* reference genome (Prust et al., 2005), which was also subject to annotation and locus tag updates (GOXxxxx to GOX_RSxxxxx) during the NCBI update on RefSeq microbial genome resources (Tatusova et al., 2015). In summary of our Illumina results, 1.42–3.56 million paired-end reads

Table 3
Numbers of variants detected in wild-type and engineered *G. oxydans* genomes using Illumina sequencing. Illumina reads were mapped to the *G. oxydans* 621H reference sequences (NC_006677) and the five annotated plasmids pGOX1 to pGOX5 (NC_006672 to NC_006676). Variants exhibiting ≥10% frequency were taken into account.

Variant type	WT-DSMZ	WT-BM	WT-E	Δ*upp*	IK001	IK002.1	IK002.1[a]	IK003.1	IK003.1[b]
SNV	44	45	47	46	47	47	47	47	48
MNV	13	13	13	13	13	13	13	13	13
Insertion	59	59	60	59	59	59	59	59	59
Deletion	42	42	42	42	42	43	42	42	42
Total	158	159	162	160	161	162	161	161	162
Intergenic	67	67	69	68	68	69	68	68	69
Synonymous	6	7	7	7	7	7	7	7	7
Non-synonymous	7	7	8	7	8	8	8	8	8
Frameshift	78	78	78	78	78	78	78	78	78

[a] After 24 h of growth under controlled conditions in a DASGIP bioreactor (Kiefler et al., 2017).
[b] After 28 h of growth under controlled conditions in a DASGIP bioreactor (Kiefler et al., 2017).

were mapped to the reference sequence resulting in genomic coverages from 74-fold to 167-fold (Table 2). Based on these mappings, 160–165 variants with ≥10% frequency were detected in the nine samples (Table 3). 158 variants are always present in all samples and only 5 variants are present in only one or some samples. Thus, the number of detected variants in both engineered and wild-type *G. oxydans* genomes is very close. Approximately 100 variants are deletions or insertions of nucleotides, mostly located in homopolymer stretches. Since these variants were uniformly found in all genomes including the wild types, these variants are likely imprecision in the reference sequence where homopolymer stretches could not be fully resolved by the Sanger sequencing. 67–69 variants are located in intergenic regions (Table S3), while the remaining 91–93 variants are located in annotated coding DNA sequences (CDS). The 93 variants were analyzed for their frequencies and impact on the amino acid level (Table S4). 91 variants are present in all samples with ≥79% frequency, mostly ≥95% to 100% frequency. Only two variants in CDS regions exhibited relatively low frequency (12%–26%) suggesting cell heterogeneity. The latter two were not detected in WT-DSMZ and one not in WT-BM and Δ*upp*. 78 variants located in CDS regions represent frameshifts due to InDels of nucleotides and were uniformly found in all genomes. 7 variants in CDS regions result in amino acid changes, and 6 variants in CDS regions were synonymous substitutions. No additional variants were found in the engineered strains cultivated for up to 24 h or 28 h, suggesting the absence of any suppressor mutations despite several metabolic engineering steps, enforced cytoplasmic glucose metabolism and cultivation for several generations.

3.2. The modified 20 kb MinION® library protocol improved the nanopore read length

Next, we sequenced the final IK003.1 strain of the iteratively engineered *G. oxydans* IK strain series by nanopore sequencing after 28 h of growth under controlled conditions in a bioreactor (Kiefler et al., 2017). To obtain as long nanopore reads as possible, we modified a MinION® library protocol to improve the enrichment of fragments larger than 20 kb. In comparison to the standard 2D library preparation protocol performed for the initial Lambda Burn-In experiment (conducted with R7.3 chemistry), we could increase the N50 2D read length from 6780 bp to 19,822 bp (Fig. 1). We obtained 20,910 quality-filtered reads which exhibit a mean read length of 18,541 bp. A single R9 flow cell yielded a total of 387.7 Mb of data of which 131.5 Mb were 2D reads with a mean read length of 18,872 bp and a mean quality score of 12.34 (Table 4).

This data represent a theoretical genome coverage of 45-fold. For 1D the longest read was approximately 73 kb and for 2D approximately 65 kb. When aligned against the *G. oxydans* 621H reference genome using BLAST, both reads resulted in full length hits with 87% and with 93% identity, respectively. Alignment of all quality-filtered 2D reads to the updated *G. oxydans* 621H genome with BWA_MEM (v0.7.15-r1140, arXiv:1303.3997) determined an overall base identity of 87.74% including InDels (Fig. S1). In terms of overall yield per flow cell, a slight increase from 115 Mb for the Lambda sample to 132 Mb (both 2D) for our modified 20 kb protocol was obtained.

3.3. Hybrid assembly and detection of structural variants in IK003.1

Combining the long MinION nanopore reads with the high quality short Illumina reads enabled the assembly of the *G. oxydans* IK003.1 bacterial chromosome into a single contig. Genome-scale comparison of the Unicycler hybrid assembly with the *G. oxydans* 621H reference sequence (Prust et al., 2005) indicates a good overall agreement between both sequences and the presence of eight potential structural variants (Fig. S2). In addition to the three known genetically engineered modifications and the Δ*upp* locus, two mobile elements are translocated in the bacterial chromosome (Table 5). One mobile element is inserted in the intergenic region between GOX_RS06790 and GOX_RS06795 at nt 1,233,221 and shows 100% identity to the coding sequence of the annotated transposases GOX_RS06295 and GOX_RS12505, respectively. The second transposase insertion at nt 2,587,655 with 100% identity to GOX_RS06430 and GOX_RS13235 forms an inverted repeat together with GOX_RS13235 that frames a novel 1420 bp long sequence with 94% sequence identity to a genomic region of *G. oxydans* DSM3504. In this strain this sequence contains the partial gene of a putative coniferyl aldehyde dehydrogenase (CalB, GLS_c24760) and the full coding sequence of a putative glyoxalase family protein (GLS_c24750) (Kostner et al., 2015). This complex structural variant could be resolved by the hybrid approach using the long nanopore reads, where 25 long reads anchor the whole region with an overlap of at least 400 bp into the flanking regions of the chromosome (Figs. 2 and S3). Two further structural variants were detected as transposase insertions into another transposase gene in plasmid pGOX1 and into an intergenic region in plasmid pGOX2 (Table 5). The annotated plasmids with a size below of the 20 kb MinION® library size selection cut-off, namely pGOX3, pGOX4 and pGOX5, could not be assembled into single contigs using our hybrid approach. Alignment of the long-read data back to the assembly resulted in a very low coverage compared to the plasmids pGOX1 and pGOX2 as well as the chromosome. In contrast, with the short Illumina reads all five plasmids were fully covered in the IK003.1 sample (Fig. S2). Based on Illumina reads, in WT-E the plasmids pGOX1 (163.2 kB) and pGOX2 (26.6 kb) exhibited a similar average coverage as the chromosome, while pGOX3 (14.5 kB) and pGOX4 (13.2 kb) were covered 2-fold more and pGOX5 (2.7 kb) 5-fold more than the chromosome. In IK003.1 only plasmid pGOX1 exhibited a similar coverage as the chromosome, while pGOX2 was covered 3-fold more, pGOX3 5-fold more, pGOX4 6-fold more, and pGOX5 even 20-fold more. This partial absence (nanopore reads) and multiples in coverage of plasmid pGOX1 to pGOX5 obtained with nanopore reads from size-selected fragments and with Illumina reads confirm indirectly the presence of these DNA sequences as separate plasmid DNA, as already carefully annotated before (Prust et al., 2005).

To verify the additional four structural variants besides the known engineered structural variants, we amplified the relevant regions by PCR using genomic DNA of IK003.1 as well as Δ*upp* and WT-E (Table 5). Then the sequence of each PCR product was determined by Sanger sequencing. The results confirmed the insertions of the mobile elements between GOX_RS06790 and GOX_RS06795 and within GOX_RS00080 in all three strains. In contrast, the insertion between GOX_RS00895 and GOX_RS00900 was only found in the genome of IK003.1. The novel sequence with 94% identity to a region from *G. oxydans* DSM3504 and hitherto unknown in the *G. oxydans* 621H reference sequence was also confirmed in the genomes of all strains including the wild types. Furthermore, as an additional verification step, we adjusted the reference genome sequence with the eight genomic alterations. Using this as a reference sequence, mapping of the Illumina reads also supported the presence of the two structural variants at 1,233,221 and at 15,633 (pGOX1) in IK001 and IK002.1, respectively, while the detected structural variant at 13,804 (pGOX2) is absent in the other genomes (Table 5). Furthermore, the mapping also supported the presence of the novel 1420 bp transposon-flanked sequence, which was clearly detected only by the nanopore approach, in all engineered as well as all wild-type strains (Table 5). Taken together, the hybrid approach allowed accurate and comprehensive analysis of the engineered genome of IK003.1 and thereby detection of a hitherto unrecognized complex region in the genome of *G. oxydans* 621H. This region is similar to a genomic region of *G. oxydans* DSM3504. The results also indicate the potential for rapid and reliable sequencing of non-reference bacterial strains or extensively engineered strains as well as the reference grade *de-novo* assembly of such.

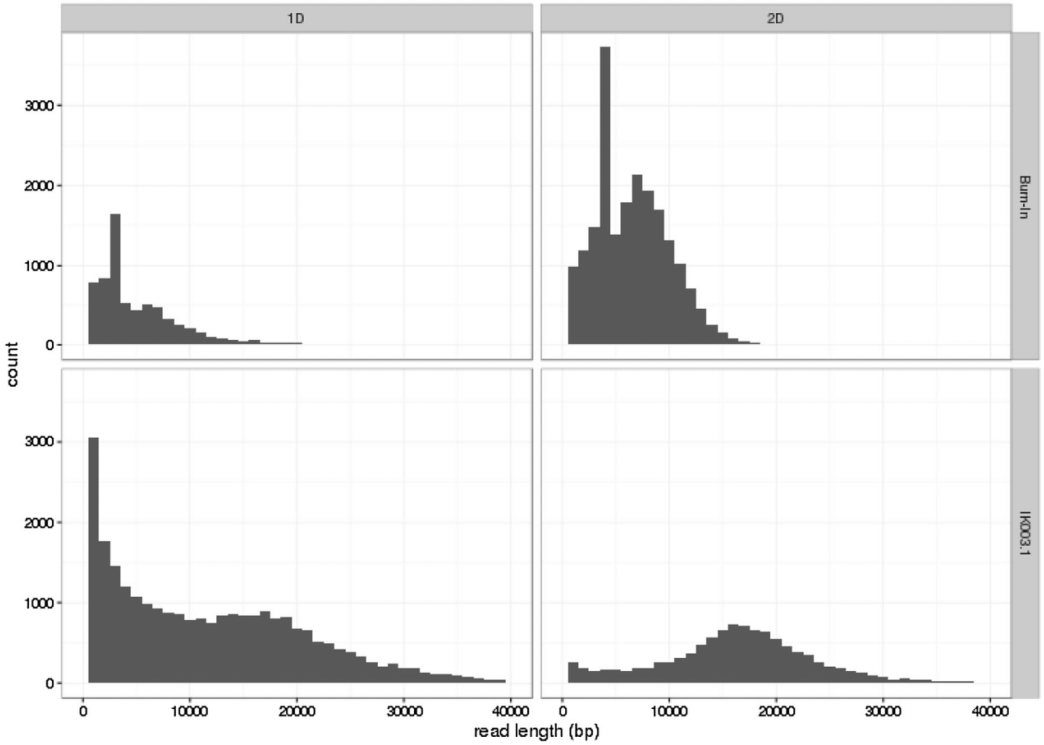

Fig. 1. Histograms of nanopore read lengths obtained with a standard protocol (Burn-In) compared to the modified 20 kb MinION library preparation protocol (IK003.1). Upper panels: Read length distribution for Oxford Nanopore Lambda Burn-In experiment. Lower panels: The modified protocol enabled improved enrichment of long reads as indicated by the average 2D read length of 18,872 bp, while 1D read length is dominated by small fragments.

Table 4
Reads statistics obtained with the modified 20 kb+ Oxford Nanopore MinION sequencing protocol (pass 2D quality-filtering).

	1D + 2D	2D only
Total reads	20,910	6970
Total bp	387,699,368	131,541,865
Mean length	18,541	18,872
Median length	17,958	18,327
Max. length	73,702	65,734
Length N50	19,515	19,822
Mean Qscore	8.36	12.34

3.4. Updates of the G. oxydans 621H reference genome sequence

The sequence data obtained in this study result in updates of the reference sequence of *G. oxydans* 621H. Analysis using the high-quality Illumina reads led to identification of 91 common variants in CDS regions with very high frequencies close to 100% in the genome of all sequenced strains. Of these, we selected 75 variants and all could be confirmed by Sanger sequencing. Therefore, these variants appear to be imprecision in the reference sequence originally determined by Sanger sequencing, which should be considered in an updated reference. Six synonymous and seven non-synonymous substitutions led to a nucleotide or amino acid change, while 78 InDels resulted in frameshifts within 64 genes (Tables 3 and S4). Among the latter, 49 genes were annotated as pseudo genes without deduced protein sequence

Table 5
Four additional structural variants detected in IK003.1 by long nanopore reads assembled and compared with the *G. oxydans* 621H reference sequence (NC_006677) and the five annotated plasmids pGOX1 to pGOX5 (NC_006672 to NC_006676). The structural variants could be verified by Sanger sequencing also in other strains (►) and/or mapping of Illumina reads to the adjusted reference sequence (▼). −) SV not detected; IGR) intergenic region; IG) intragenic.

Structural variant					Strain						
Genomic position	Region	Location	Inserted sequence	Comment	WT-DSMZ	WT-BM	WT-E	Δ*upp*	IK001	IK002.1	IK003.1
1,233,221	IGR	GOX_RS06790-GOX_RS06795	GOX_RS06295 or GOX_RS12505	mobile element insertion	–	–	►, ▼	►,▼	▼	▼	►, ▼
2,587,655	IGR	GOX_RS13230-GOX_RS13235	GOX_RS06430 or GOX_RS13235 + 1420 bp	mobile element and novel 1420 bp insertion	▼	▼	►, ▼	►, ▼	▼	▼	►, ▼
15,633 (pGOX1)	IG	GOX_RS00080	GOX_RS07905 or GOX_RS08680	mobile element insertion	–	–	►, ▼	►, ▼	▼	▼	►, ▼
13,804 (pGOX2)	IGR	GOX_RS00895-GOX_RS00900	GOX_RS00950 or GOX_RS06295	mobile element insertion	–	–	–	–	–	–	►, ▼

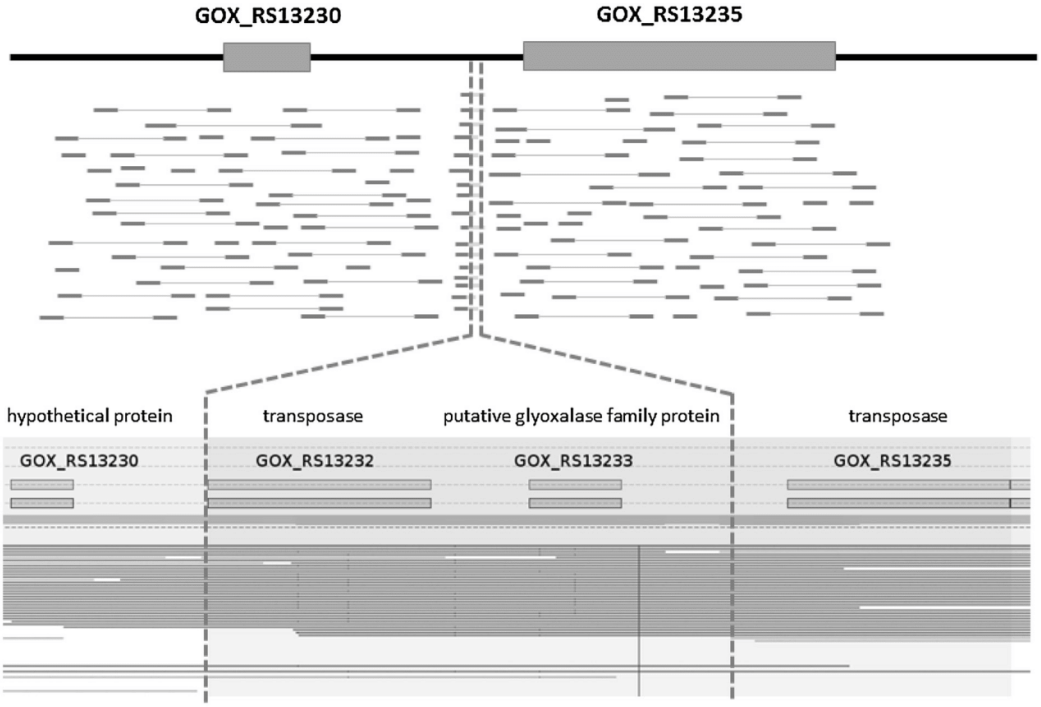

Fig. 2. Schema of read mapping and *de novo* assembly at the GOX_RS13230-GOX_RS13235 locus. Illumina reads mapped to the *G. oxydans* 621H reference sequence (Prust et al., 2005) indicate a structural variant (unaligned read ends in light gray; upper panel) that was resolved by *de novo* assembly using long nanopore reads spanning a novel 1420 bp transposon-flanked sequence insertion (lower panel). As a result of sequence analysis and ORF search GOX_RS13232 and GOX_RS13233 were added to the updated annotation of *G. oxydans* 621H.

(Tatusova et al., 2013). Eleven genes are annotated as hypothetical proteins (GOX_RS05325, GOX_RS07120, GOX_RS09465, GOX_RS10075, GOX_RS10885, GOX_RS11470, GOX_RS13575, GOX_RS13580, GOX_RS00055, GOX_RS13955, GOX_RS13970), and four genes are annotated as bacterial conjugation TrbI-like protein DotG (GOX_RS00555), DNA-directed RNA polymerase subunit beta (*rpoB*, GOX_RS03085), a DNA-binding response regulator (GOX_RS03130), and a histidine kinase (GOX_RS04330). To revisit the annotation of these 64 genes based on the corrected reference sequence, we performed search for new potential ORFs and subsequent BLASTP search. The resulting ORFs and their deduced protein sequence were reviewed and compared to the original annotation regarding their length and gene product function (Table S5). For all 49 pseudo genes affected, ORFs and a deduced protein sequence with significant BLAST hits and predicted protein function were found. Only for three genes no change in length and annotation was observed due to only a few codon changes (GOX_03085, GOX_RS04330, and GOX_RS10885). For six genes start or stop changed (GOX_RS00555, GOX_RS03130, GOX_RS05325, GOX_RS09465, GOX_RS10075, and GOX_RS11470), yet the protein annotation changed only for GOX_RS00555 (DotG), which is re-annotated to encode a hypothetical protein. InDels in the CDS region of six hypothetical proteins (GOX_RS07120, GOX_RS13575, GOX_RS13580, GOX_RS00055, GOX_RS13955, and GOX_RS13970) led to frameshifts favoring the merging of two neighboring ORFs to a single ORF, which still encodes a hypothetical protein in all cases. According to our detected differences to the original reference, we created a new consensus genome sequence with corresponding nucleotide and CDS changes. Additionally, the detected novel 1420 bp sequence, which was found by the nanopore approach in strain IK003.1, and then also in an all other strains by mapping of the Illumina reads to the adjusted reference sequence (Table 5), was also included in the update of the reference sequence. The data are accessible in ENA and *via* our internet portal www.gluconobacterfactory.de.

4. Discussion

In this study, we sequenced microbial genomes of engineered and wild-type *G. oxydans* strains. The latter are typically showing a low biomass yield due to the lack of several genes for central metabolic enzymes resulting in non-functionality or absence of glycolysis, tricarboxylic acid cycle, and glyoxylate shunt (Bringer and Bott, 2016; Prust et al., 2005). For the pathway-completed engineered *G. oxydans* strain IK003.1 exhibiting highly increased biomass formation on glucose (Kiefler et al., 2015, 2017), we combined accurate Illumina reads and nanopore-guided long reads to check for genome stability and large genomic alterations, including the known engineered heterologous gene arrangements. In recent studies, this combination of sequencing was also used and already revealed the high potential of this approach for closing gaps in assemblies of wild type references using short read data. The assemblies of long genomic features such as rRNA operons and transposable elements were much better resolved with long reads (Goodwin et al., 2015; Karlsson et al., 2015; Madoui et al., 2015). In particular, in the presence of complex genomic regions such as inverted repeats, the hybrid approach outperforms short read assemblies and alignments. Resolving such structures to reliably determine these genomic regions depends on the ability to fully span large repeats with potential novel sequence insertions. Therefore, an increased average read length supports the resolution of larger repetitive regions, novel sequence insertions and complex combinations of both as found in this study. With our modified 20 kb MinION® library protocol we significantly increased the average size of nanopore 2D reads with R9 chemistry to about 18.9 kb by improved enrichment of size-selected DNA fragments larger than 20 kb. This allowed to fully span and confirm the three engineered insertions of the heterologous genes

sdhCDABE, *ndh*, and *sucCD* in three different loci in the engineered IK003.1 strain, as well as to uncover a hitherto unrecognized transposase-flanked 1420 bp long sequence with 94% sequence identity to a genomic region of *G. oxydans* DSM3504 (Table 5, Figs. 2 and S3). As this sequence was also found in the reference strain Δ*upp* and in wild types, it is considered as unresolved before and is not a result of transposase activity during construction of IK003.1. Two other transposase insertions are also already present in Δ*upp* and in the wild types. The only strain-specific transposase insertion is in IK003.1 the insertion in plasmid pGOX2 into an intergenic region (Table 5). Whether this happened by chance or is a consequence of the pathway completion to counteract the enforced metabolism is currently unknown. Notably, the short-read plasmid coverage relative to the chromosome suggested that in the pathway-restored IK003.1 the copy numbers of the four plasmids pGOX2 to pGOX5 are increased compared to the reference WT-E. It needs to be determined whether increase in plasmid copy numbers are generally the case in *G. oxydans* 621H when IK003.1 or pathway-restored *G. oxydans* strains are constructed and whether this may affect the metabolism and the biomass yield. Overall, we found no non-canonical integration events. The genome of IK003.1 was almost unaffected by transposons under the conditions tested for approximately 70 generations of growth.

Besides structural variants such as mobile element insertions, suppressor mutations on the single nucleotide level may accumulate to counteract the enforced metabolism and therefore negatively affect the industrial use of such strains. However, we did not found suppressor mutations, yet we found SNPs and small InDels in all strains which can therefore be considered as unresolved by the Sanger sequencing of the wild type reference (Prust et al., 2005). These differences result in some updates of the *G. oxydans* 621H annotation. In our analysis we used the recent NCBI update of the *G. oxydans* 621H genome (NC_006677) and of the five plasmids (NC_006672 to NC_006676). Our mapping and analysis with high-quality Illumina reads revealed 91 common variants in CDS regions with high frequencies (> 79%) in all strains used in our study (Tables 3 and S4). We used these results to update the *G. oxydans* 621H reference. The six synonymous and seven non-synonymous nucleotide substitutions were inserted in the new consensus sequence. The 78 InDels detected in 64 genes were also inserted and the resulting changes were further analyzed on the gene and protein level (Tables S4 and S5). 76 InDels were observed in homopolymer stretches which could probably not be resolved by Sanger sequencing before. Since InDels in coding regions typically result in frameshifts, we checked the new resulting candidate ORFs, performed BLAST search with the deduced amino acid sequences and identified appropriate annotations. For 49 of 64 pseudo genes the frameshifts led to a full-length gene now with a deduced amino acid sequence and annotation as a predicted protein similar to already published RefSeq non-redundant proteins of other *Gluconobacter* species. 16 InDels were present as a combination in the same annotated coding region and overall did not change the frame of the gene. Therefore, these updated ORFs exhibit only minor deviation to the original annotated CDS and deduced protein sequences as indicated by the BLAST hits (Table S5). Besides the variants found by Illumina sequencing, the usage of long nanopore reads allowed the detection and resolution of a hitherto unrecognized transposase-flanked 1420 bp sequence. This novel sequence and all coding region changes were added to the new consensus genome sequence and annotation for downloading. However, as in the NCBI version, our update of gene starts and stops is by bioinformatic predictions. Therefore, it is possible that some are not the real gene starts and ends. Future studies using whole and differential RNAseq will help to further improve the *G. oxydans* 621H reference genome annotation as demonstrated for other microbes (Pfeifer-Sancar et al., 2013; Su et al., 2016).

In conclusion, with our modified 20 kb MinION® library protocol we could significantly increase the average size of nanopore 2D reads with R9 chemistry to about 18.9 kb. Longer reads will improve sequencing genomes of engineered cell factories with long inserts to check for genome stability and unintentional or non-canonical gene integrations, structural variants, as well as nucleotide changes including suppressor mutations. Since we found high genome stability of *G. oxydans* without suppressor mutations and suspicious structural variants in the presence of an enforced metabolism such as in IK003.1, it makes the TCA cycle-completed *G. oxydans* strain a suitable engineered host for further metabolic engineering efforts.

Acknowledgements

The scientific activities of the Bioeconomy Science Center were financially supported by the Ministry of Innovation, Science and Research within the framework of the NRW Strategy project BioSC (No. 313/323-400-002 13). The authors thank Armin Ehrenreich for kindly providing the wild-type reference strain of the Δ*upp* mutant.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbiotec.2017.04.016.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Ameyama, M., Shinagawa, E., Matsushita, K., Adachi, O., 1981. D-Fructose dehydrogenase of *Gluconobacter industrius*: purification, characterization, and application to enzymatic microdetermination of D-fructose. J. Bacteriol. 145, 814–823.
Ashton, P.M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., Wain, J., O'Grady, J., 2015. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat. Biotechnol. 33, 296–300.
Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.
Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.
Bringer, S., Bott, M., 2016. Central carbon metabolism and respiration in *Gluconobacter oxydans*. In: Matsushita, K., Toyama, H., Tonouchi, N., Okamoto-Kainuma, A. (Eds.), Acetic Acid Bacteria: Ecology and Physiology. Springer Japan, Japan, pp. 235–253.
Brown, C.G., Clarke, J., 2016. Nanopore development at Oxford Nanopore. Nat. Biotechnol. 34, 810–811.
Burrus, V., Waldor, M.K., 2004. Shaping bacterial genomes with integrative and conjugative elements. Res. Microbiol. 155, 376–386.
Chiang, C.J., Chen, P.T., Chao, Y.P., 2008. Replicon-free and markerless methods for genomic insertion of DNAs in phage attachment sites and controlled expression of chromosomal genes in *Escherichia coli*. Biotechnol. Bioeng. 101, 985–995.
Choi, Y.J., Bourque, D., Morel, L., Groleau, D., Miguez, C.B., 2006. Multicopy integration and expression of heterologous genes in *Methylobacterium extorquens* ATCC 55366. Appl. Environ. Microbiol. 72, 753–759.
Chung, M.E., Yeh, I.H., Sung, L.Y., Wu, M.Y., Chao, Y.P., Ng, I.S., Hu, Y.C., 2016. Enhanced integration of large DNA into *E. coli* chromosome by CRISPR/Cas9. Biotechnol. Bioeng. 114 (1), 172–183.
Darmon, E., Leach, D.R., 2014. Bacterial genome instability. Microbiol. Mol. Biol. Rev. 78, 1–39.
Deppenmeier, U., Hoffmeister, M., Prust, C., 2002. Biochemistry and biotechnological applications of *Gluconobacter* strains. Appl. Microbiol. Biotechnol. 60, 233–242.
Ding, Q., Chen, G., Wang, Y., Wei, D., 2015. Identification of specific variations in a non-motile strain of Cyanobacterium *Synechocystis* sp. PCC 6803 originated from ATCC 27184 by whole genome resequencing. Int. J. Mol. Sci. 16, 24081–24093.
Eikmanns, B.J., Thum-Schmitz, N., Eggeling, L., Ludtke, K.U., Sahm, H., 1994. Nucleotide sequence, expression and transcriptional analysis of the *Corynebacterium glutamicum gltA* gene encoding citrate synthase. Microbiology 140, 1817–1828.
Erguner, B., Ustek, D., Sagiroglu, M.S., 2015. Performance comparison of Next Generation sequencing platforms. Conf. Proc. IEEE Eng. Med. Biol. Soc. 2015, 6453–6456.
Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8, 186–194.
Flagfeldt, D.B., Siewers, V., Huang, L., Nielsen, J., 2009. Characterization of chromosomal integration sites for heterologous gene expression in *Saccharomyces cerevisiae*. Yeast 26, 545–551.
Frith, M.C., Hamada, M., Horton, P., 2010. Parameters for accurate genome alignment. BMC Bioinf. 11, 80.
Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M.C., McCombie, W.R., 2015. Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. Genome Res. 25, 1750–1756.
Greenfield, S., Claus, G.W., 1972. Nonfunctional tricarboxylic acid cycle and the mechanism of glutamate biosynthesis in *Acetobacter suboxydans*. J. Bacteriol. 112, 1295–1301.

Gupta, A., Singh, V.K., Qazi, G.N., Kumar, A., 2001. *Gluconobacter oxydans*: its biotechnological applications. J. Mol. Microbiol. Biotechnol. 3, 445–456.

Hölscher, T., Schleyer, U., Merfort, M., Bringer-Meyer, S., Gorisch, H., Sahm, H., 2009. Glucose oxidation and PQQ-dependent dehydrogenases in *Gluconobacter oxydans*. J. Mol. Microb. Biotechnol. 16, 6–13.

Hanke, T., Nöh, K., Noack, S., Polen, T., Bringer, S., Sahm, H., Wiechert, W., Bott, M., 2013. Combined fluxomics and transcriptomics analysis of glucose catabolism via a partially cyclic pentose phosphate pathway in Gluconobacter oxydans 621H. Appl. Environ. Microbiol. 79, 2336–2348.

Hekmat, D., Bauer, R., Fricke, J., 2003. Optimization of the microbial synthesis of dihydroxyacetone from glycerol with Gluconobacter oxydans. Bioproc. Biosyst. Eng. 26, 109–116.

Herrmann, U., Merfort, M., Jeude, M., Bringer-Meyer, S., Sahm, H., 2004. Biotransformation of glucose to 5-keto-D-gluconic acid by recombinant *Gluconobacter oxydans* DSM 2343. Appl. Microbiol. Biotechnol. 64, 86–90.

Hochheim, J., Kranz, A., Krumbach, K., Sokolowsky, S., Eggeling, L., Noack, S., Bocola, M., Bott, M., Marienhagen, J., 2016. Mutations in MurE, the essential UDP-N-acetylmuramoylalanyl-D-glutamate 2,6-diaminopimelate ligase of *Corynebacterium glutamicum*: effect on L-lysine formation and analysis of systemic consequences. Biotechnol. Lett. 39 (2), 283–288.

Hong, K.K., Vongsangnak, W., Vemuri, G.N., Nielsen, J., 2011. Unravelling evolutionary strategies of yeast for improving galactose utilization through integrated systems level analysis. Proc. Natl. Acad. Sci. U. S. A. 108, 12179–12184.

Hook, C.D., Samsonov, V.V., Ublinskaya, A.A., Kuvaeva, T.M., Andreeva, E.V., Gorbacheva, L.Y., Stoynova, N.V., 2016. A novel approach for *Escherichia coli* genome editing combining *in vivo* cloning and targeted long-length chromosomal insertion. J. Microbiol. Methods 130, 83–91.

Karlsson, E., Larkeryd, A., Sjodin, A., Forsman, M., Stenberg, P., 2015. Scaffolding of a bacterial genome using MinION nanopore sequencing. Sci. Rep. 5, 11996.

Kiefler, I., Bringer, S., Bott, M., 2015. SdhE-dependent formation of a functional *Acetobacter pasteurianus* succinate dehydrogenase in *Gluconobacter oxydans* – a first step toward a complete tricarboxylic acid cycle. Appl. Microbiol. Biotechnol. 99, 9147–9160.

Kiefler, I., Bringer, S., Bott, M., 2017. Metabolic engineering of *Gluconobacter oxydans* 621H for increased biomass yield. Appl. Microbiol. Biotechnol accepted.

Komati Reddy, G., Lindner, S.N., Wendisch, V.F., 2015. Metabolic engineering of an ATP-neutral Embden-Meyerhof-Parnas pathway in *Corynebacterium glutamicum*: growth restoration by an adaptive point mutation in NADH dehydrogenase. Appl. Environ. Microbiol. 81, 1996–2005.

Kosciow, K., Domin, C., Schweiger, P., Deppenmeier, U., 2016. Extracellular targeting of an active endoxylanase by a TolB negative mutant of *Gluconobacter oxydans*. J. Ind. Microbiol. Biotechnol. 43, 989–999.

Kostner, D., Luchterhand, B., Junker, A., Volland, S., Daniel, R., Büchs, J., Liebl, W., Ehrenreich, A., 2015. The consequence of an additional NADH dehydrogenase paralog on the growth of *Gluconobacter oxydans* DSM3504. Appl. Microbiol. Biotechnol. 99, 375–386.

Lee, J.H., Sung, B.H., Kim, M.S., Blattner, F.R., Yoon, B.H., Kim, J.H., Kim, S.C., 2009. Metabolic engineering of a reduced-genome strain of *Escherichia coli* for L-threonine production. Microb. Cell Fact. 8, 2.

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595.

Loeschcke, A., Markert, A., Wilhelm, S., Wirtz, A., Rosenau, F., Jaeger, K.E., Drepper, T., 2013. TREX: a universal tool for the transfer and expression of biosynthetic pathways in bacteria. ACS Synth. Biol. 2, 22–33.

Loman, N.J., Quinlan, A.R., 2014. Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics 30, 3399–3401.

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J., Pallen, M.J., 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 30, 434–439.

Lu, H., Giordano, F., Ning, Z., 2016. Oxford nanopore MinION sequencing and genome assembly. Genom. Proteom. Bioinform. 14, 265–279.

Madoui, M.A., Engelen, S., Cruaud, C., Belser, C., Bertrand, L., Alberti, A., Lemainque, A., Wincker, P., Aury, J.M., 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genom. 16, 327.

Mahr, R., Gätgens, C., Gätgens, J., Polen, T., Kalinowski, J., Frunzke, J., 2015. Biosensor-driven adaptive laboratory evolution of L-valine production in *Corynebacterium glutamicum*. Metab. Eng. 32, 184–194.

Martinez-Garcia, E., Aparicio, T., de Lorenzo, V., Nikel, P.I., 2014. New transposon tools tailored for metabolic engineering of gram-negative microbial cell factories. Front. Bioeng. Biotechnol. 2, 46.

McCoy, R.C., Taylor, R.W., Blauwkamp, T.A., Kelley, J.L., Kertesz, M., Pushkarev, D., Petrov, D.A., Fiston-Lavier, A.S., 2014. Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. PLoS One 9, e106689.

Mikheyev, A.S., Tin, M.M., 2014. A first look at the Oxford Nanopore MinION sequencer. Mol. Ecol. Resour. 14, 1097–1102.

Miyazaki, R., van der Meer, J.R., 2013. A new large-DNA-fragment delivery system based on integrase activity from an integrative and conjugative element. Appl. Environ. Microbiol. 79, 4440–4447.

Nagy, Z., Chandler, M., 2004. Regulation of transposition in bacteria. Res. Microbiol. 155, 387–398.

Nikel, P.I., de Lorenzo, V., 2013. Implantation of unmarked regulatory and metabolic modules in Gram-negative bacteria with specialised mini-transposon delivery vectors. J. Biotechnol. 163, 143–154.

Ong, F.S., Lin, J.C., Das, K., Grosu, D.S., Fan, J.B., 2013. Translational utility of next-generation sequencing. Genomics 102, 137–139.

Peters, B., Junker, A., Brauer, K., Muhlthaler, B., Kostner, D., Mientus, M., Liebl, W., Ehrenreich, A., 2013. Deletion of pyruvate decarboxylase by a new method for efficient markerless gene deletions in *Gluconobacter oxydans*. Appl. Microbiol. Biotechnol. 97, 2521–2530.

Pfeifer, E., Hünnefeld, M., Popa, O., Polen, T., Kohlheyer, D., Baumgart, M., Frunzke, J., 2016. Silencing of cryptic prophages in *Corynebacterium glutamicum*. Nucleic Acids Res. 44 (21), 10117–10131.

Pfeifer-Sancar, K., Mentz, A., Ruckert, C., Kalinowski, J., 2013. Comprehensive analysis of the *Corynebacterium glutamicum* transcriptome using an improved RNAseq technique. BMC Genom. 14, 888.

Phelan, R.M., Sachs, D., Petkiewicz, S.J., Barajas, J.F., Blake-Hedges, J.M., Thompson, M.G., Reider Apel, A., Rasor, B.J., Katz, L., Keasling, J.D., 2016. Development of next generation synthetic biology tools for use in *Streptomyces venezuelae*. ACS Synth. Biol. 6 (1), 159–166.

Prust, C., Hoffmeister, M., Liesegang, H., Wiezer, A., Fricke, W.F., Ehrenreich, A., Gottschalk, G., Deppenmeier, U., 2005. Complete genome sequence of the acetic acid bacterium *Gluconobacter oxydans*. Nat. Biotechnol. 23, 195–200.

Quick, J., Quinlan, A.R., Loman, N.J., 2014. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. GigaScience 3, 22.

Rawsthorne, H., Turner, K.N., Mills, D.A., 2006. Multicopy integration of heterologous genes, using the lactococcal group II intron targeted to bacterial insertion sequences. Appl. Environ. Microbiol. 72, 6088–6093.

Saito, Y., Ishii, Y., Hayashi, H., Imao, Y., Akashi, T., Yoshikawa, K., Noguchi, Y., Soeda, S., Yoshida, M., Niwa, M., Hosoda, J., Shimomura, K., 1997. Cloning of genes coding for L-sorbose and L-sorbosone dehydrogenases from *Gluconobacter oxydans* and microbial production of 2-keto-L-gulonate, a precursor of L-ascorbic acid, in a recombinant *G. oxydans* strain. Appl. Environ. Microbiol. 63, 454–460.

Schwarzhans, J.P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., Friehs, K., 2016. Non-canonical integration events in *Pichia pastoris* encountered during standard transformation analysed with genome sequencing. Sci. Rep. 6, 38952.

Shiwa, Y., Matsumoto, T., Yoshikawa, H., 2013. Identification of laboratory-specific variations of *Bacillus subtilis* strains used in Japan. Biosci. Biotechnol. Biochem. 77, 2073–2076.

Smanski, M.J., Zhou, H., Claesen, J., Shen, B., Fischbach, M.A., Voigt, C.A., 2016. Synthetic biology to access and expand nature's chemical diversity. Nat. Rev. Microbiol. 14, 135–149.

Su, Z., Zhu, J., Xu, Z., Xiao, R., Zhou, R., Li, L., Chen, H., 2016. A transcriptome map of *Actinobacillus pleuropneumoniae* at single-nucleotide resolution using deep RNA-seq. PLoS One 11, e0152363.

Tattini, L., D'Aurizio, R., Magi, A., 2015. Detection of genomic structural variants from next-generation sequencing data. Front. Bioeng. Biotechnol. 3, 92.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Ciufo, S., Li, W., 2013. Prokaryotic genome annotation pipeline. The NCBI Handbook [Internet], 2nd ed. National Center for Biotechnology Information (US), Bethesda (MD) December 10.

Tatusova, T., Ciufo, S., Federhen, S., Fedorov, B., McVeigh, R., O'Neill, K., Tolstoy, I., Zaslavsky, L., 2015. Update on RefSeq microbial genomes resources. Nucleic Acids Res. 43, D599–605.

Tkác, J., Navratil, M., Sturdik, E., Gemeiner, P., 2001. Monitoring of dihydroxyacetone production during oxidation of glycerol by immobilized *Gluconobacter oxydans* cells with an enzyme biosensor. Enzyme Microb. Technol. 28, 383–388.

Tyo, K.E., Ajikumar, P.K., Stephanopoulos, G., 2009. Stabilized gene duplication enables long-term selection-free heterologous pathway expression. Nat. Biotechnol. 27, 760–765.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963.

Wang, E.X., Ding, M.Z., Ma, Q., Dong, X.T., Yuan, Y.J., 2016. Reorganization of a synthetic microbial consortium for one-step vitamin C fermentation. Microb. Cell Fact. 15, 21.

Warner, J.R., Patnaik, R., Gill, R.T., 2009. Genomics enabled approaches in strain engineering. Curr. Opin. Microbiol. 12, 223–230.

Zobel, S., Benedetti, I., Eisenbach, L., de Lorenzo, V., Wierckx, N., Blank, L.M., 2015. Tn7-based device for calibrated heterologous gene expression in *Pseudomonas putida*. ACS Synth. Biol. 4, 1341–1351.

2.2 Transcriptome analysis of *G. oxydans* using RNAseq

Kranz, A., Busche, T., Vogel, A., Usadel, B., Kalinowski, J., Bott, M., and Polen, T. (2018). RNAseq analysis of *Gluconobacter oxydans* 621H. **BMC Genomics 19:24.**

Author's contributions:

AK carried out the experimental work, performed the substantial part of the data analysis, and wrote a draft of the manuscript. TB and JK developed the improved RNAseq protocol. AV estimated transcript abundances, included the results in the genome browser JBrowse, and made the RNAseq data publicly available. TB, JK, MB, and TP revised the manuscript. TP coordinated the study and finalized the manuscript.

Overall contribution AK: 90%

RNAseq analysis of *Gluconobacter oxydans* 621H

Angela Kranz[1,2], Tobias Busche[3], Alexander Vogel[2,4,5], Björn Usadel[2,4,5], Jörn Kalinowski[3], Michael Bott[1,2], and Tino Polen[1,2,*]

[1)] IBG-1: Biotechnology, Institute of Bio- and Geosciences, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany.

[2)] The Bioeconomy Science Center (BioSC), c/o Forschungszentrum Jülich GmbH, 52425 Jülich, Germany.

[3)] Center for Biotechnology (CeBiTec), Universität Bielefeld, Universitätsstr. 25, 33615 Bielefeld, Germany.

[4)] IBMG: Institute for Biology I, RWTH Aachen University, Worringer Weg 2, 52074 Aachen, Germany.

[5)] IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany.

ORCID:	Angela Kranz	orcid.org/0000-0002-8000-0400
ORCID:	Michael Bott	orcid.org/0000-0002-4701-8254
ORCID:	Tino Polen	orcid.org/0000-0002-0065-3007

*** for correspondence:**

Dr. Tino Polen
E-mail: t.polen@fz-juelich.de
Phone: +49 (0)2461 61 6205
Fax: +49 (0)2461 61 2710

Keywords: transcriptome
RNAseq
transcription start site
operons
antisense transcripts
Gluconobacter oxydans

Abstract

Background: The acetic acid bacterium *Gluconobacter oxydans* 621H is characterized by its exceptional ability to incompletely oxidize a great variety of carbohydrates in the periplasm. The metabolism of this □-proteobacterium has been characterized to some extent, yet little is known about its transcriptomes and related data. In this study, we applied two different RNAseq approaches. Whole transcriptomes were sequenced to identify expressed genes and operon structures. Primary transcriptomes enriched for 5´-ends of transcripts were sequenced to detect transcription start sites, which allow subsequent analysis of promoter motifs, ribosome binding sites, and 5´-UTRs.

Results: Sequencing of primary transcriptomes of *G. oxydans* revealed 2,449 TSSs, which were classified according to their genomic context followed by identification of promoter and ribosome binding site motifs, analysis of 5´-UTRs including validation of predicted *cis*-regulatory elements and correction of start codons. 1,144 (41%) of all genes were found to be expressed monocistronically, whereas 1,634 genes were organized in 571 operons. Together, TSSs and whole transcriptome data were also used to identify novel intergenic (18), intragenic (328), and antisense transcripts (313).

Conclusions: This study provides deep insights into the transcriptional landscapes of *G. oxydans.* The comprehensive transcriptome data, which we made publicly available, facilitate further analysis of promoters and other regulatory elements. This will support future approaches for rational strain development and targeted gene expression in *G. oxydans*. The corrections of start codons further improve the high quality genome reference and support future proteome analysis.

Background

The α-proteobacterium *Gluconobacter oxydans* 621H is a Gram-negative acetic acid bacterium, which is used for a broad range of industrial applications requiring regio-and stereoselective oxidations. This is due to the ability to incompletely oxidize a great variety of carbohydrates in the periplasm and the release of resulting products into the medium. Since the 1930s, it is especially used for the production of 2-keto-L-gulonic acid, a precursor for the vitamin C production [1-5]. Other biotransformation products are dihydroxyacetone, 6-amino-L-sorbose, or 5-ketogluconate [2, 6, 7]. The 2.9 Mb genome of *G. oxydans* consists of one circular chromosome and five plasmids [8]. Recently, MinION nanopore and Illumina read data revealed a novel 1,420 bp transposon-flanked and ORF-containing sequence and in 73 annotated coding sequences about 91 nucleotide differences resulting in an improved high quality genome reference [9]. 2,710 protein-coding sequences are annotated, including

31 membrane-bound dehydrogenases, which enable the periplasmic oxidation [8, 9]. Genome sequencing and annotation analysis revealed that genes encoding 6-phosphofructokinase, succinate dehydrogenase, and succinyl-CoA synthetase are missing. Therefore, the Embden-Meyerhof-Parnas (EMP) pathway and the tricarboxylic acid (TCA) cycle are incomplete [8]. Both the restricted ability to oxidize carbohydrates in the cytoplasm and the high activity of dehydrogenases in the periplasm as well as subsequent release of products into the medium result in a low final biomass yield on complex media with sugar or sugar alcohols such as mannitol or glucose as carbon source [2, 10, 11]. This is unfavourable for industrial biotransformation processes, as it increases the costs for the initially required biomass production.

The unorthodox metabolism of *G. oxydans* was studied to some extent by using mutational analysis, metabolic flux analysis, and DNA microarray experiments. These studies showed that the major part of the available glucose (90%) is already oxidized to gluconate in the periplasm [12]. Of the 10% glucose taken up by the cell, 9% is phosphorylated to glucose 6-phosphate and then predominantly metabolized *via* the pentose phosphate pathway (PPP), whereas 91% of the glucose is oxidized to gluconate by a soluble glucose dehydrogenase. Additionally, gluconate is taken up by the cell. 70% of the gluconate in the cytoplasm is oxidized to 5-ketogluconate and 30% is phosphorylated to 6-phosphogluconate [13, 14]. Mutational analysis of the mannitol metabolism also favored the PPP as essential for the cytoplasmic fructose metabolism [15]. Along with the information obtained by analysis of respiratory mutants [16, 17] and genome comparisons between different *G. oxydans* strains [18], the results of the metabolic studies provided the basis for metabolic engineering of *G. oxydans* 621H with the aim to improve the biomass yield, e.g. by complementing the incomplete pathways [19, 20]. In contrast to metabolism, current knowledge on global gene expression and transcriptional regulation is very restricted for *G. oxydans* [13, 15, 16]. Similarly, the availability of characterized promoters, which can be used for further rational strain development and targeted gene expression, is limited [21-25].

Revealing the complexity of bacterial transcriptomes by next-generation sequencing (NGS) *via* RNAseq has become the most efficient method to get detailed insights on the RNA level, thereby also providing important information for metabolic engineering of industrially used microbes [26, 27]. In contrast to other methods such as DNA microarrays and qPCR, RNAseq in principle allows absolute quantification of all expressed genes as well as single nucleotide resolution over the complete transcript length [28-31]. Strand-specific RNAseq approaches can be used to detect novel transcripts including antisense transcripts [31-33]. Also, uniquely mapped sequencing reads connecting two neighboring genes enable the detection of operon structures. This can be advantageous for identification of genes with related functions [34-36]. Another important RNAseq method is the sequencing of primary

transcriptomes by enrichment of native transcripts bearing a 5´-triphosphate group [27, 37]. Thereby, transcription start sites (TSSs) and respective promoter motifs, 5´-UTRs, ribosome binding sites (RBSs), leaderless transcripts, and *cis*-regulatory RNA elements such as riboswitches or RNA thermometers can be identified and analyzed [38-41].

In this study, we sequenced whole and primary transcriptomes of *G. oxydans* 621H under different conditions to obtain a broad range of expressed genes and TSSs. For the detection of TSSs, we used a protocol improved to distinguish between bona-fide TSSs and false positives due to inefficient digestion of non-primary transcripts. All sequencing data were used to analyze the operon and sub-operon structures, to detect new genes and antisense transcripts, to correct start codons, and to analyze further aspects.

Methods

Strain, media and cultivation conditions

In this study, *G. oxydans* wild type DSM 2343 (ATCC 621H) was used. *G. oxydans* was grown in complex medium (5 g L^{-1} yeast extract, 1 g L^{-1} KH_2PO_4, 1 g L^{-1} $(NH_4)_2SO_4$, 2.5 g L^{-1} $MgSO_4$ x 7 H_2O, and 50 µg mL^{-1} cefoxitin as antibiotic) with 220 mM (4% w/v) mannitol or 220 mM (4% w/v) glucose. Precultures were grown overnight in 100 mL shaking flasks with 15 mL medium, while main cultures were grown in 500 mL baffled shaking flasks containing 100 mL medium (140 rpm, 30°C). To potentially obtain as many transcripts as possible, bacterial cells were cultivated under non-stress conditions with mannitol or glucose as carbon source and harvested after reaching the exponential phase (OD_{600nm} 1.2-1.8), and under the following stress conditions: For oxygen limitation, the rotation of the shaker was stopped for 10 min. For heat shock, a fast temperature shift of the flask with medium from 30°C to 50°C was carried out in a water bath followed by cultivation at 50°C for 15 min. For salt stress, cells were exposed to 0.25 M NaCl for 30 min. For oxidative stress, after preliminary tests a concentration of 0.025 M H_2O_2 was chosen as supplement and cells were further cultivated for 30 min. After stress exposure, 1 mL of culture broth was harvested by centrifugation (10,000 x *g*; 30 sec). The cell pellet was immediately shock-frozen in liquid nitrogen and stored at -20°C until use for isolation of total RNA.

RNA isolation

Total RNA of *G. oxydans* 621H was isolated using TRIzol (Life Technologies). Frozen cell pellets were resuspended in 3 mL TRIzol reagent and 1 mL Rnase-free water. The cell suspension was aliquoted to four 1.5 mL tubes and cells were disrupted by bead-beating in two cycles (2 x 30 sec) using zirconia/silica beads (0.1 mm) and a Silamat device

(Ivoclar Vivadent). Afterwards, 200 µL of chloroform were added to the supernatant and shaken vigorously for 15 sec followed by centrifugation (12,000 x *g*; 15 min). The supernatant was transferred to a new tube, treated with 0.5 mL isopropanol, incubated at RT for 10 min and centrifuged (12,000 x *g*; 10 min). The RNA pellet was washed with 75% (v/v) ethanol, air-dried and resuspended in 50 µL of DEPC-treated water. The content of four tubes per sample were pooled and treated with 5 µL of DNase (Thermo Fisher Scientific) for 20 min (37°C). For purification of RNA, one volume of phenol-chloroform-isoamyl alcohol (25:24:1; PCI) was added to the sample, shaken and transferred to a Phase Lock Gel™ tube (Eppendorf AG), which allows better phase separation. After centrifugation (12,000 x *g*; 15 min), the supernatant was transferred to a new tube and treated with one volume of chloroform-isoamyl alcohol (24:1; CI) followed by centrifugation (12,000 x *g*; 15 min). Precipitation was performed by adding 1/10 volume of sodium acetate (3 M; pH 5.2) and 3 volumes of ethanol (~99%) to the supernatant and incubation at -20°C overnight. Afterwards, each sample was centrifuged (12,000 x *g*; 20 min) at 4°C. The pellet was washed two times with 75% (v/v) ethanol, air-dried and then dissolved in 30 µL of RNase-free water. RNA concentrations in samples were determined using a Qubit (Thermo Fisher Scientific) and checked for quality on formaldehyde agarose gels.

Construction of whole and primary transcriptome cDNA libraries

For depletion of rRNA, 5 µg or 2 x 5 µg of total RNA was treated with the Ribo-Zero magnetic kit for Gram-negative bacteria (Illumina). Afterwards, precipitation with ethanol was performed following the manufacturer's instructions. For preparation of whole transcriptome libraries, we used the TruSeq stranded mRNA sample preparation kit (Illumina) according to the manufacturer´s instructions, except that 5 µL of rRNA-depleted RNA was mixed with 13 µL of Fragment, Prime, Finish Mix and incubated at 94°C for fragmentation and priming (4 min). For primary 5'-end-enriched cDNA libraries, rRNA-depleted RNA samples obtained from 2 x 5 µg of total RNA were used. The preparation protocol has been described previously in detail [27]. In the present study, the experimental workflow was modified to strongly reduce the number of false positive 5'-ends which are non-primary. Therefore, RNA samples were denatured (95°C; 2 min) and immediately chilled on ice to destruct secondary structures. Then digestion with terminator 5'-phosphate-dependent exonuclease (TEN, Epicentre) was carried out at 30°C (60 min) and at 42°C (30 min). To flag non-digested non-primary transcripts still remaining, RNA samples were denatured (95°C; 2 min) followed by ligation of RNA 5'-index adapter (1 µL; 60 µM) 5'-CCCUACACGACGCUCUUCCGAUCGAGUACCCUAG (index underlined) to 5'-monophosphorylated ends (25°C; 120 min and 37°C; 30 min). Afterwards, the protocol was

continued with RNA 5'-polyphosphatase (RPP) treatment (Epicentre) to convert true primary 5′-triphosphate ends to 5′-monophosphate ends as described [27]. Ligation of the 5'-adapter to the converted 5′-monophosphate ends was performed as described for the index adapter. Reverse transcription with a stem-loop DNA adapter and library amplification was performed as described previously [27]. Prior to sequencing, 5'-enriched cDNA libraries were purified and size-selected for approximately 100-1000 nt *via* gel electrophoresis.

Next-generation sequencing of cDNA libraries

Sequencing libraries were quantified *via* qPCR using the KAPA Library Quantification Kit for Illumina libraries (Peqlab) or with an Agilent 2100 Bioanalyzer (Agilent Technologies) using a High Sensitivity DNA kit (Agilent Technologies). Sequencing of normalized libraries (10 pM) was carried out on a MiSeq desktop sequencer (Illumina) according to the manufacturer's protocol. For the whole transcriptome libraries, paired-end reads with a length of 2 x 75 bases were generated. Primary transcriptome libraries were sequenced in single read mode with a read length of 35 or 75 bases.

Read processing, mapping, and determination of transcript abundances

Read processing and mapping was carried out with the CLC Genomics Workbench (Qiagen Aarhus A/S). Reads were trimmed by removing adapter sequences using the *Trim Sequences* tool and filtered for Phred quality scores <30 [42]. Reads from primary transcriptome libraries containing the barcode sequence TACCCTAG at their 5´-ends indicated a false positive TSS and were removed from the read pool. Remaining reads were mapped to the *G. oxydans* 621H reference sequences updated recently by genome sequencing using high-quality Illumina and long nanopore reads [9]. Non-specific matches were mapped randomly.

Abundance of transcripts were determined by mapping quality and adapter-trimmed reads (Trimmomatic v0.36) to the published reference genome of *G. oxydans* using bowtie2 v2.2.7 [43, 44]. Cufflinks and cuffnorm were used to quantify transcript levels [45].

Identification of transcription start sites (TSSs)

Detection of TSSs was done with libraries enriched for primary transcripts using ReadXplorer [46] with the following parameters: (i) Only single perfect mappings were considered. (ii) Minimum percent of coverage increase was set to 250% and minimum number of read starts to 20. (iii) A maximal distance of 600 nt upstream to the start codon was set to assign a TSS to the corresponding annotated ORF. (iv) A transcript was assumed leaderless, when its assigned TSS had a maximal distance of three nt to the start codon.

(v) TSSs, which could not be assigned to an ORF, were classified as indicators for possible novel transcripts. All automatically detected TSSs were checked manually and TSSs without a clear read start increase and unusual drops or increase of read coverage were removed.

The TSSs identified by ReadXplorer were classified according to the following categories allowing the occurrence of some TSS in more than one category (Figure 1A): (a) sTSS. TSSs assigned to an annotated ORF in sense orientation. On the one hand, this category includes TSSs with a downstream ORF within a range of 300 nt. On the other hand, it also includes TSSs, which lay within an ORF with a maximal distance of 200 nt downstream to the annotated start codon and which therefore could be used to correct the translation start codon position ((n)sTSS). The latter was checked by searching for a start codon in-frame to the annotated stop codon and by searching for a possible ribosome binding site (RBS) upstream of the possible start codon. Furthermore, it was verified, whether the mapping coverage of the whole transcriptome data matched the start of transcription as indicated by the corresponding TSSs. This was only possible at genomic positions where based on mappings no read-through from upstream genes occurs. (b) pTSS. Putative TSS assigned to an annotated gene. These are TSSs with a distance of more than 300 nt to the downstream gene. (c) iTSS. These intragenic TSSs lay within an annotated ORF in sense orientation. All iTSSs with a maximal distance of 300 nt to the end of the assigned gene, which were also classified as sTSSs, were discarded from the iTSS category. Also, (n)sTSSs (category a) located downstream of an annotated start codon without an alternative downstream in-frame start codon were included into this group. (d) asTSS. TSS located in antisense orientation to an ORF. To identify antisense transcripts associated to asTSSs, the whole transcriptome data were used. For every position, a minimal coverage of 15 was required and it was checked whether the possible novel antisense transcript can be extended downstream for at least 20 nt until the coverage at a position drops below 15. If the possible novel transcript was longer than 500 nt and had a mean coverage of >40, the cut-off coverage for the start of the transcript was set to 80. (e) nTSS. Intergenic TSS hitherto unassigned and potentially indicating novel RNA transcripts. In these cases, we checked the whole transcriptome data for mappings which could represent associated novel transcripts. Only data were considered further where nTSS and whole transcriptome mappings indicated a novel transcript. Potential ORFs were searched using the *Find Open Reading Frames* tool (CLC Genomics Workbench) and results were checked manually. Suitable ORF sequences were used for a Blastx search to identify homologous proteins in the NCBI reference proteins database (refseq_protein) [47].

It was possible that more than one TSS was associated to a gene (sTSS, pTSS). In these cases, the TSS exhibiting the highest number of read starts was assigned as primary

TSS, whereas all other valid TSSs were classified as secondary. For novel transcripts (iTSS, asTSS, nTSS), only primary TSSs were considered.

Identification of operons

For identification of polycistronic transcripts based on whole transcriptome data, ReadXplorer was used [46]. A minimal number of 10 spanning reads in sense orientation was required to combine neighboring genomic ORFs in the same transcript. Furthermore, TSS data were used to identify primary operons, with TSSs assigned to the first gene of an operon, and sub-operons, which are indicated by TSSs within primary operons.

Motif detection of promoter sequences

Promoter motifs were detected with the web-based tool *Improbizer* [48], which uses the expectation maximization (EM) algorithm. For each TSS, the 50 bases upstream were extracted and the -10 and -35 promoter motifs were searched within the sequences using default settings. The list of the 50 bp sequences used for this analysis was sorted according to the read counts starting with the highest coverage. Since we had no knowledge about consensus promoter motifs in *G. oxydans*, we used information about promoters already identified in other α-proteobacteria [49-52] to further analyze the *Improbizer* results with Excel (Microsoft). A maximal distance of 3 to 11 nt between the TSS and the -10 region was allowed, whereas the spacer length between the -10 and -35 regions was set to 16 to 23 nt.

Identification of ribosome binding sites (RBSs)

For identification of RBSs, all 5´-UTRs with a minimal length of 20 nt were analyzed. First, the frequencies of purines (G and A) were compared with the frequencies of pyrimidines (T and C) for every nucleotide position within the 20 nt long sequence upstream of the translation start codon. Sequences with an accumulation of purines (>55%) were extracted. The extracted sequences were used to search for a conserved RBS motif with *Improbizer* [48]. Resulting data were visualized with Origin (OriginLab) and WebLogo [53].

Results

Data generation and mapping statistics

Bacteria need to adapt to their environment by sensing environmental parameters and activation of appropriate regulatory programs, which typically involve the modulation of gene expression. To obtain a broad range of transcription start sites (TSSs) and transcripts

of genes from *G. oxydans* 621H, we analyzed total RNA from cells grown under non-stress (complex medium with glucose or with mannitol) and stress conditions (oxygen limitation, heat shock, oxidative stress by H_2O_2, salt stress by 0.25 M NaCl). The cDNA libraries were sequenced using a MiSeq sequencer (Illumina) and the data output of all libraries was combined in the following analysis. After quality-trimming of the MiSeq reads, 10.13 million reads of the primary and 55.76 million reads of the whole transcriptome libraries were obtained and used for the analysis (Table 1). 6.13 million (60.5%) reads of the libraries enriched for primary 5'-ends started with the barcode sequence TACCCTAG. These represented false positive primary 5'-ends, i.e. those originating from 5'-monophosphorylated mRNA, which was not degraded by the terminator 5'-phosphate–dependent exonuclease. These reads were therefore discarded from the TSS analysis. In total, 1.1 and 32.87 million reads from the primary and whole transcriptome libraries mapped uniquely to the *G. oxydans* 621H reference genome [9].

Detection of transcription start sites (TSSs) and revision of start codons

The mapping of reads from primary transcriptome libraries was used for the detection of TSSs by ReadXplorer [46]. All detected TSSs were manually inspected and, if necessary, compared with the mapping of whole transcriptome data. TSSs suggested yet having an uneven coverage gradient or no clear accumulation of read starts were manually removed. Of the 2,449 manually verified TSSs (Table S1), 134 belong to genes for rRNAs, tRNAs and RNase P (Table S2). The remaining 2,315 TSSs were classified according to their genomic context as described in detail in methods (Figure 1A). Since a neighboring ORF, its 5´-UTR or 3´-UTR, respectively, may overlap with a TSS already assigned to a category, some TSSs can be found in more than one category (Figure 1B). In general, it can be distinguished between TSSs belonging to annotated ORFs and TSSs that suggest the existence of further, not yet annotated ORFs. According to the classification rules applied, 994 TSSs were assigned to annotated ORFs (Table S3 and Table S4) and are located within a maximal distance of 300 nt upstream of the translation start codon (sense TSS, sTSS). 57 of them ((n)sTSS) were located downstream of an annotated ORF start and with a maximal distance of 200 nt (Table S4). Analysis of the mapping coverage of whole transcriptome data suggested that the originally annotated ORF start might be wrong and needs to be revised. Therefore, a possible translation start codon downstream of the detected TSS and in-frame to the annotated stop codon was searched to obtain the new ORF start and the deduced protein sequence. As an additional verification, the 20 nt-region upstream of new ORFs with a minimal 5´-UTR length of 20 nt were analyzed regarding the presence of a possible ribosome binding site (RBS). The maximal length difference between the revised shorter

amino acid sequence and the originally annotated one was 74 aa (Table S4). The maximal 5´-UTR length of the new ORFs was 191. After the revision of start codons the (n)sTSSs were treated as sTSSs in the further analysis.

48 of the 994 sTSSs also belong to the category intragenic TSS (iTSS), 7 to antisense TSSs (asTSS), and 10 to putative TSSs with downstream ORFs (pTSS). Altogether, 360 pTSSs upstream of annotated ORFs with a distance >300 nt and <600 nt were detected (Table S5). Besides the 10 pTSSs also assigned to sTSS, 245 pTSSs are additionally assigned to intragenic TSS (iTSS) and 57 were located in antisense orientation to an annotated ORF (asTSS). Altogether, 1,354 TSSs could be assigned to protein-coding ORFs (sTSS and pTSS). 10 of them belong to both categories and are therefore automatically assigned to two different ORFs (Figure 1B). However, it is more likely that these TSS are only related to the sTSSs data. Therefore, we considered 350 pTSSs and the 994 sTSSs for all following analysis, resulting in a total of 1,344 TSSs assigned to protein-coding ORFs. It is also possible that more than one TSS per ORF was detected. The TSS with the highest number of read starts was called primary TSS, whereas all other TSSs of the same gene are secondary TSSs. In total, we detected primary TSSs for 1,073 (40%) out of 2,710 annotated protein-coding ORFs in the genome of *G. oxydans* [8, 9]. 271 secondary TSSs were assigned to 227 ORFs with a maximal number of four TSSs per ORF.

TSSs belonging to possible novel transcripts were TSSs located within ORFs in sense orientation (iTSS), TSSs antisense to ORFs (asTSS), and TSSs assigned to new intergenic transcripts (nTSS). iTSSs include TSSs located within an ORF, which had a minimal distance of 200 nt to the ORF start. Also, those TSSs with a distance <200 nt to the start codon, which were not included in the category (n)sTSS, were assigned to iTSS. Of the 621 iTSS (Table S6), 328 were uniquely assigned to this category. Additional to the 7 asTSSs, which are also sTSSs, and the 57 asTSS also assigned to the category pTSS, 619 asTSSs were identified (Table S7). 24 nTSS were found in intergenic regions, suggesting the presence of possible novel genes not yet annotated (Table S8). For the following analysis, the nTSSs were assigned to the sTSSs.

The nucleotide frequencies of the TSSs uniquely assigned to the categories sTSS/(n)sTSS/nTSS, pTSS, iTSS, and asTSS differed to some extent (Figure 1C). The 10 sTSSs, which were also assigned to pTSS, were also solely assigned to sTSS for this analysis. In all 4 TSS categories G shows always almost highest frequency as initiating nucleotide (32%-40%), while only for sTSSs/(n)sTSSs/nTSSs and pTSSs A (33%-35%) shows second highest frequency as initiating nucleotide. For iTSSs and asTSSs C (31%-33%) shows second highest or highest frequency. Thus, given by frequencies the TSS categories differ and exhibit initiating nucleotide priority of A/G for sTSSs/(n)sTSSs/nTSSs and pTSSs, G/C for iTSSs, and C/G for asTSS. To check whether the TSS categories could

be distinguished further, we also analyzed the frequencies of the preceding nucleotide (position -1) and the following (position +2). Nucleotide frequencies at -1 and +2 of iTSSs are quite similar compared to that of +1. For asTSSs the nucleotide frequencies at +2 are also quite similar compared to that of +1, yet at -1 there is a change from C/G priority to G/A priority. Interestingly, for sTSSs/(n)TSSs/nTSSs and pTSSs there is a clear change of priority from A/G at +1 to C/T at both -1 and +2.

The distribution of read start coverages of the detected TSSs assigned to the 4 categories also differed to some extent (Figure 1C). The highest mean coverage (119) was observed for sTSSs/(n)sTSSs/nTSSs, followed by asTSSs (67), pTSSs (66), and iTSSs (47). If the top 10% of TSSs exhibiting the highest read start coverage are considered from each category, the highest mean value was observed for sTSSs (778). Mean of the top 10% from iTSSs (204), pTSSs (203) and asTSSs (382) exhibited 73% to 50% lower mean coverage compared to the top 10% from sTSSs (Figure 1C). For the top 10% by coverage, the nucleotide distributions at initiation position +1 exhibited an excess of A+G of 14% to 15% over A+G of the lowest 10% for the sTSS and asTSS group (Figure 2). These groups overall included the majority of high-coverage TSSs and therefore exhibited the highest coverage mean values. In contrast, for the pTSS and iTSS group, which overall exhibited the lowest coverages and mean, A+G of the top 10% was very similar to A+G of the lowest 10%. Furthermore, at the -1 position the already relative high average frequencies of C for sTSSs/(n)TSSs/nTSSs (37.8%) and for pTSSs (37.5%) as well as of G for iTSS (41.2%), were further increased on average among the top 10% to 40.6%, 40%, and 45.5%, respectively, while the frequency of G for asTSS (36%) was decreased by 28% to 25.8% (Figure 1C and Figure 2).

5´-UTRs and *cis*-regulatory elements

The 1,344 TSSs assigned to protein-coding genes were used for the analysis of 5´-UTRs (Figure 3). The 5´-UTR of 62 mRNAs (5%) is ≤3 nt and these were therefore classified as leaderless. With a length of 4-8 nt 24 transcripts (3%) have relatively short 5´-UTRs. It can be assumed that these do not contain a ribosome binding site (RBS). A relatively high number of short leaders with a length of 10-40 nt were observed (219; 16%). 427 (32%) transcripts contain leader sequences of 100-300 nt.

Long leader sequences may play a role in the translational regulation of their mRNAs. For example, they can contain *cis*-regulatory elements such as riboswitches that fold into secondary structures. Seven different regulatory regions were predicted in the genome of *G. oxydans*, which are listed in the Rfam database [54]. We used both the primary and whole transcriptome mapping data and compared them with the regions of the predicted regulatory elements (Table 2). For four predicted riboswitches a TSS was detected and the coverage

analysis of the whole transcriptome data indicated transcription termination for the FMN (6,000-fold coverage), glycine (2,000-fold coverage), SAM-II (1,500-fold coverage), and TPP riboswitch (1,300-fold coverage). Transcription termination is indicated by a higher read coverage for the 5´-UTR in contrast to a relatively low coverage for the assigned ORF (Figure 4). The remaining three predicted regulatory elements showing 130-fold and 106-fold coverage for the cobalamin and fluoride riboswitch, and 1,250-fold coverage for the ROSE element could only be supported by mapping of the whole transcriptome data, which, however, indicated no transcription termination since the respective downstream ORFs exhibited similar coverages.

Promoter motif in *G. oxydans* 621H

Initiation of transcription requires binding of the RNA polymerase holoenzyme to promoter motifs in the DNA sequence. Recognition of the promoter motifs is achieved by different sigma factors that are part of the RNA polymerase holoenzyme. σ^{70} (RpoD) is the primary sigma factor, which is essential for the transcription of the majority of genes during growth. The σ^{70} binding sites on the DNA can characteristically be found around positions -35 and -10 upstream of the TSS. The upstream regions (50 bp) of 808 primary TSSs, which were identified for genes with a 5´-UTR length <300 nt in *G. oxydans*, were used to search for conserved motifs with *Improbizer*. The weakly conserved -10 motif “nAtnnn” with a spacer of 3-11 nt to the TSS was found in 94% (761) of the sequences. For the -35 region, we allowed a spacer length of 16-23 nt to the -10 region and found the motif “ttGnnn” within 581 (72%) sequences (Figure 5, Table S9). The top 5% of transcripts with a high abundance under non-stress conditions also showed the conserved -35 region “ttGnnn” and a highly conserved “T” (90%) at the first position as well as a less conserved “a” (56%) at the second position of the predicted -10 region “Tatnnn”.

Consensus motif of ribosome binding sites in *G. oxydans* 621H

For the identification of the ribosome binding site (RBS) consensus motif, we used the 20 nt upstream of 973 protein-coding ORFs for which TSSs were assigned with a minimal 5´-UTR length of 20 bases. Typically RBSs are purine-rich. Therefore, we compared the frequencies of purines (G and A) and pyrimidines (T and C). Within the analyzed sequences, accumulation of A and G (>55%) was found 6-15 nt upstream of the translation start codon (Figure 6A). Analysis of these regions with *Improbizer* identified the conserved motif “aGGAg” (Figure 6B) in 913 sequences (94%) with a spacing of 3-14 nt (mean spacing 7.9±2.8 nt) to the start codon (Table S10). The preferred translation start codon is ATG (816; 84%), followed by GTG (86; 9%), TTG (29; 3%), and CTG (22; 2%).

Operon organizations in *G. oxydans* 621H

By using a combination of all whole transcriptome RNAseq data and TSS data, we analyzed the organization of genes in operons and differentiated monocistronic transcripts, primary operons, and sub-operons. Genes were assigned to a primary operon, when they could be joined by at least 10 spanning reads. If a TSS assigned to a protein-coding ORF (1,073) was located within a primary operon, it was assumed that this gene and all downstream genes of the primary operon form a sub-operon. In total, 1,144 monocistronic transcripts (41%) and 571 operons comprising of 1,634 (59%) genes were identified. Furthermore, 341 sub-operons were detected comprising 720 genes (Figure 7). Most of the operons (80%) comprise 2 or 3 genes. The largest operon comprises 14 genes coding for ribosomal proteins (GOX_RS02995-GOX_RS03060). Within this primary operon, 7 sub-operons with 2, 3, 5, 9, 10, 12, and 13 genes were found based on TSS data (Table S11). The encoded proteins of the 11 genes forming the second largest primary operon (GOX_RS11055-GOX_RS11105) exhibit diverse cellular functions (protein fate, amino acid metabolism, pantothenate and CoA biosynthesis, DNA replication, lipopolysaccharide synthesis, and nucleotide metabolism). Four sub-operons comprising 1, 3, 7, and 9 genes were identified within this primary operon. Altogether, we could find a TSS for 1,463 (54%) of the protein-coding genes expressed as monocistronic transcripts or as part of a transcriptional unit.

Identification of novel transcripts in *G. oxydans* 621H

Altogether, 971 TSSs belonging solely to novel transcripts were found in the primary transcriptome libraries. In this context, novel transcripts are based on TSSs that were not assigned to already annotated protein-coding ORFs. They can be classified according to their genomic context in 328 iTSSs (Table S6) in sense orientation within an ORF, 619 as TSSs (Table S7) in antisense orientation to an ORF, and 24 nTSSs (Table S8) located in intergenic regions (Figure 1B).

Comparison with the mapping coverage data of the whole transcriptome libraries downstream of the nTSSs supported the presence of novel intergenic transcripts, which were analyzed by ORF and subsequent BLAST search. 6 out of the 24 nTSSs represented alternative TSSs of the same transcript with two to three TSSs per gene. In total, 18 new ORFs were found ranging from 78 nt (26 aa) to 681 nt (227 aa) in length (Table S8). For 6 out of the 18 identified ORFs a homologous protein in other species was found by BLAST search. Two ORFs showed identity to sequences present in the genome of *G. oxydans*. One is a not yet annotated transposase with 100% identity to other mobile elements present in the genome of *G. oxydans*. The other one is a protein with a helix-turn-helix domain, which was

originally annotated in the genome sequence of *G. oxydans* 621H [8], but was removed later by NCBI's reference sequence updates [55]. Furthermore, we identified two hypothetical proteins, one MerR family transcriptional regulator, and a ParA family protein. For additional verification, we also searched for the RBS motif upstream of the translation start codons and could find a RBS for 9 out of the 18 novel transcripts (Table S8).

For 313 out of 619 asTSS, transcripts longer than 20 nt were found in the whole transcriptome data (Table S7). Promoter motif search revealed the presence of a -10 motif ("cctTCg") upstream of 299 asTSSs, but no -35 motif. 75% of asTSSs without a corresponding transcript in the whole transcriptome data had a read start coverage <42. This value is for the sTSSs at 69. Generally, antisense transcripts show lower expression values than sense transcripts. Therefore, it is possible that transcripts belonging to the remaining 306 asTSSs could not be detected within the whole transcriptome data due to very low coverage.

G. oxydans 621H RNAseq data in JBrowse

In order to establish a joint resource we incorporated the detected TSSs together with their expression strength and estimated gene expression levels (Table S12) for all samples into a publicly available JBrowse-based genome browser available *via* www.gluconobacterfactory.de. JBrowse offers to zoom and navigate through selected tracks representing data sets from individual samples [56]. For example, based on the updated underlying reference genome for *G. oxydans* 621H, a user could navigate and zoom to the ORFs GOX_RS13232 and GOX_RS13233, which are located in the 1,420 bp transposon-flanked region only recently revealed by nanopore sequencing [9]. These ORFs showed expression across all six growth conditions and thus provided further validation for these annotations in the updated genome reference (Table S12). Having the individual sample data sets available as corresponding tracks enables the user to independently investigate differences in expression levels and associated TSSs beyond the scope of the results presented here. Additionally, the graphical user interface provides an intuitive access to gene models, gene functions as well as direct retrieval of coding- and protein sequences.

Discussion

In the present study, we analyzed transcriptomes of *G. oxydans* 621H by RNA sequencing using the recently improved genome sequence as reference [8, 9]. For the comprehensive characterization of transcription start sites, promoter motifs, novel transcripts, transcript abundance, and transcriptional organization of genes, whole and primary

transcriptomes of *G. oxydans* were sequenced. To obtain a broad range of TSSs and transcripts, cells were grown under non-stress and stress conditions. Since the main aim of this study was an overall view on the transcriptional landscapes of *G. oxydans* 621H, we combined the sequencing data of the different conditions for the analysis. Solely transcript abundance based on whole transcriptome data were separately calculated for every condition. For the identification of TSSs, an improved protocol based on the method by [27] was used. By using this method, the number of false positive TSSs can be drastically reduced. Still, a manual inspection of automatically detected TSSs is necessary to remove TSSs within an uneven gradient of accumulated reads. In total, 2,449 TSSs were detected in *G. oxydans*, manually verified, and classified according to their genomic context. This data represent the basis for identification and analysis of 5´-UTRs, promoter motifs, RBSs, and operons, which were identified by analysis of whole transcriptome data.

Operon organization

59% of all genes in *G. oxydans* 621H are expressed polycistronically. In other bacteria, it was also shown by RNA sequencing that 60-90% of all genes are part of operons [27, 37, 57, 58]. Typically, genes belonging to operons have related functions [34, 36]. The most prominent example in *G. oxydans* is the largest operon consisting of genes encoding ribosomal proteins. Sequencing of primary transcriptomes using RNAseq also revealed the presence of sub-operons based on the detection of internal TSSs within operons for many bacteria. The first differential RNAseq approach focusing on the primary transcriptome of *Helicobacter pylori* revealed 337 primary operons exhibiting 126 sub-operons (37%) [37]. In other bacteria, the number of sub-operons was even higher. For example, in 616 primary operons 565 sub-operons (92%) were identified in *Corynebacterium glutamicum* [27]. The availability of internal TSSs is important for a more sophisticated regulation of gene expression [41]. Also, several experiments showed that the expression of genes as monocistronic or polycistronic transcripts may change depending on the growth condition [59, 60]. In *G. oxydans*, 571 operons exhibiting 341 sub-operons were detected. Taken the expression of genes in operons and sub-operons into account, all 1,073 TSSs assigned to protein-coding ORFs control the expression of 1,463 genes in total, which represents 54% of all annotated genes.

5´-UTRs and *cis*-regulatory elements

Since we had no knowledge about distances between TSSs and translation start codons in *G. oxydans*, we considered TSSs resulting in a maximal 5´-UTR length of 300 nt as sTSSs and classified all TSSs with a distance >300 nt and up to 600 nt to the start codon

as putative (pTSSs). 87% of the pTSSs were also classified as intragenic and antisense TSSs. 94% of all 5´-UTRs with a maximal length of 300 nt were longer than 10 nt. The length distribution showed a maximum from 10-40 nt (16%). This is in accordance with observations in other bacteria [27, 58]. 61 mRNAs were found to be leaderless. For 13 of the corresponding ORFs additional TSSs were found further upstream, indicating that they can also be transcribed with a 5'-UTR. Thus, 49 genes remain which presumably are transcribed exclusively leaderless. The sTSSs of 3 of these were also found in the categories asTSS and iTSS or were used to revise the translation start codon. In other bacteria, the number of leaderless transcripts is quite diverse with <0.5% in *Bacillus methanolicus,* 2.2% in *H. pylori*, 33% in *C. glutamicum*, and 47% in *Deinococcus deserti* [27, 37, 58, 61]. In *Sinorhizobium meliloti*, another □-proteobacterium, roughly 6% of all protein-coding genes were leaderless [52]. 57 of the leaderless protein-coding genes in *G. oxydans* have ATG as translation start codon, and only three and two exhibit GTG and TTG, respectively. In *Escherichia coli*, it was shown, that ATG is necessary for the translation of leaderless transcripts and that non-ATG start codons are inefficient [62, 63]. Analysis of leaderless transcripts in *Mycobacterium tuberculosis* showed that also the alternative GTG is sufficient for translation [64]. Moreover, *in silico* analysis of several other bacterial genomes confirmed GTG and TTG as possible start codons of leaderless transcripts [65]. A model for the translation of leaderless transcripts including the endoribonuclease MazF of the stress-induced *mazEF* toxin-antitoxin system was proposed for *E. coli* [66]. MazF is involved in alternative processing of 16S rRNA. This novel 16S rRNA molecule is part of a translation machinery that enables translation of leaderless transcripts and transcripts processed by MazF.

Besides the short-leadered and leaderless transcripts, a relatively high number of 5´-UTRs with lengths between 100 and 300 nt (43 %) were found. This is also the case in other □-proteobacteria [52] and could enable regulation by *cis*-regulatory elements. Predictions of such elements based on genome comparisons, experimental evidence, and prediction of secondary structures can be found in the Rfam database [54]. For four of the seven riboswitches predicted in the genome of *G. oxydans*, we could identify a TSS upstream and corresponding read mapping based on whole transcriptome data. The FMN riboswitch is located in the 5'-UTR of an operon composed of four genes encoding enzymes involved in riboflavin biosynthesis, i.e. riboflavin biosynthesis protein RibD (GOX_RS06030), riboflavin synthase subunit alpha (GOX_RS06035), bifunctional 3,4-dihydroxy-2-butanone 4-phosphate synthase/GTP cyclohydrolase (GOX_RS06040), and 6,7-dimethyl-8-ribizyllumazine synthase (GOX_RS06045). It has been suggested that the FMN riboswitch regulates gene expression in Gram-positive bacteria *via* transcription termination, whereas translational repression occurs in Gram-negative bacteria [67]. However, it was also shown

that FMN riboswitches in Gram-negative bacteria can influence both transcription and translation [68]. For *G. oxydans* grown in complex medium, the mapping of whole transcriptome data suggests transcription termination, since the 5´-UTR exhibited a 100-fold higher coverage than the ORF. Also, the absence of an intrinsic terminator in the 5´-UTR does not necessarily mean that transcription termination is not possible, because also riboswitches without clear terminator sequences can terminate transcription [69]. Upstream of ORFs encoding proteins of the glycine cleavage system (glycine cleavage system aminomethyltransferase T, GOX_RS06635; glycine cleavage system protein H, GOX_RS06640; glycine dehydrogenase, GOX_RS06645), the glycine riboswitch was predicted and our RNAseq results are in accordance with this prediction. It was shown that glycine typically leads to the activation of the downstream genes by binding to the riboswitch [70]. For *G. oxydans* grown in complex medium, the coverage reflecting the RNA level of the 5´-UTR upstream of GOX_RS06635 is significantly higher (60-fold) than the coverage of the ORFs downstream. The predicted SAM-II riboswitch is a *cis*-regulatory element found only in □-proteobacteria [71]. In *G. oxydans*, this riboswitch is located upstream of the ORF encoding O-succinylhomoserine sulfhydrylase (GOX_RS09595), an enzyme involved in methionine biosynthesis. Analysis of the whole transcriptome data suggested transcription termination in *G. oxydans*, because the 5´-UTR coverage is 20-fold higher than the coverage of the ORF. Additionally, computational prediction showed a stable terminator and antiterminator conformation for the SAM-II riboswitch [72]. Our RNAseq data are also in accordance with the predicted TPP riboswitch upstream of the phosphomethylpyrimidine synthase gene (GOX_RS12420). TPP-dependent riboswitches are known from all domains of life and can regulate expression of genes involved in thiamine biosynthesis by a variety of mechanisms [73, 74]. Whole transcriptome data suggested transcription termination in *G. oxydans*, since a 26-fold higher coverage was observed for the 5´-UTR compared to the ORF coverage. Moreover, other long 5´-UTRs in *G. oxydans* may also contain novel *cis*-regulatory elements.

Many transcripts with 5´-UTRs longer than 100 nt and shorter than 300 nt belong to the functional categories protein synthesis, energy metabolism, and amino acid metabolism. Besides the regulation of transcription and translation by riboswitches or RNA thermometers within 5´-UTRs, it is also possible that 5´-UTRs form stable stem-loop structures to protect transcripts from degradation [75]. A global mRNA decay analysis of *G. oxydans* revealed that some of the transcripts (e.g. those encoding translation initiation factor IF-3 (GOX_RS05240), zinc-dependent alcohol dehydrogenase (GOX_RS02720), and ribosomal proteins L2 (GOX_RS03040), S19 (GOX_RS03035), and L17 (GOX_RS02930)) with longer 5´-UTRs had half-lives longer than the average half-life of 5.8 min (Kranz et al., in preparation). This could indicate that the long 5´-UTRs play a role in the stabilization of these

transcripts. The presence of *cis*-regulatory elements or formation of stable stem-loop structures may explain the presence of 5´-UTRs longer than 100 nt in *G. oxydans*.

Start codons distribution and ribosomal binding sites

The most frequent translation initiation codon in *G. oxydans* is ATG (ca. 84%). GTG as initiation codon was found for ca. 9% of all protein-coding ORFs and only a minor part showed the less common codons TTG (ca. 3%) and CTG (ca. 2%). This is in accordance with findings in other bacteria, where ATG is also the most frequent initiation codon, whereas others show only small frequencies [76]. Experiments showed that the translation initiation codon as well as the downstream region have an effect on gene expression [77]. Other important factors that influence protein translation are the RBS sequence and the distance between RBS and translation start codon [78]. Based on our RNAseq data, we identified and analysed RBSs in *G. oxydans* 621H. The conserved motif “aGGAg”, which was found in 94% of all analysed sequences, represents the reverse complement of the 3´-end of the 16S ribosomal RNA. This fits very well to the findings in other bacteria [27, 58, 79]. Translation can be increased by using the optimal RBS, which is complementary to the 3´-end of the 16S ribosomal RNA [80]. Also, the spacing between the RBS and the start codon plays an important role for translation initiation. For *G. oxydans*, we found a mean spacing of 7.9 ± 2.8 nt, which matched the optimal spacing in *E. coli*, *C. glutamicum*, *Bacillus subtilis*, and other bacteria [81].

***G. oxydans* has a lax consensus promoter motif**

Conserved promoter motifs were determined upstream of TSSs, which were assigned to protein-coding ORFs. We found a weakly conserved -10 region “nAtnnn” with a highly conserved “A” at the 2nd position and a -35 region “ttGnnn” with a highly conserved “G” at position 3 of the hexamer. In many other bacteria, such as *E. coli*, *C. glutamicum*, or *B. subtilis*, the -10 region “TATnnT” is highly conserved, whereas the -35 region can be less conserved [27, 82, 83]. For the identification of TSSs in *G. oxydans*, we combined primary transcriptome libraries generated from bacterial cells grown under normal and stress conditions. Therefore, the promoter motif does not solely represent the σ^{70} binding sites on the DNA, because this sigma factor is essential for the transcription of housekeeping genes during regular growth [83]. Alternative sigma factors, which can regulate gene expression under stress conditions, recognize different promoter motifs [84, 85]. Prediction of promoter motifs in *Bradyrhizobium japonicum*, an □-proteobacterium, showed less conservation at the first position of the -10 region depending on the sigma factor, which is involved in recognition of the respective motif on the DNA [49]. This might explain the similar percentage of

occurrence of “t” (40%) and “c” (39%) at the first position of the -10 region and therefore the less conserved -10 region in *G. oxydans*. Four alternative sigma factors are annotated in the genome of *G. oxydans*. One of them encoded by GOX_RS03675 is associated to the heat shock response, whereas two encoded by GOX_RS07890 and GOX_RS13390 have a possible extracytoplasmic function (ECF). The latter ones can be activated in response to cell envelope stress or oxidative stress [86]. Growth under nitrogen-limitation could activate another sigma factor encoded by GOX_RS13390. Bacterial cells for the RNAseq experiments performed in this study were *inter alia* grown under heatshock and oxidative stress. Therefore, GOX_RS03675-, GOX_RS07890-, and GOX_RS13390-dependent genes, which have a different promoter motif than genes assigned to housekeeping functions, are very likely among all the genes for which sequences were analyzed.

Interestingly, when we used only the top 5% of transcripts exhibiting high abundance, the motif “Tatnnn” with a highly conserved “T” at the first position (90%) and a less conserved “a” (56%) at the second position was found in the -10 region. This could indicate that the simple search for conserved motifs with *Improbizer* using all sequences upstream of TSSs assigned to protein-coding ORFs can somehow distort the prediction of promoter motifs in *G. oxydans*. Therefore, additional grouping and detailed analysis of promoter motifs is necessary to get deeper insights into promoter structures in *G. oxydans*. Such an analysis was recently performed for *C. glutamicum* [38].

Description of novel intergenic transcripts in *G. oxydans*

In total, we could identify 18 transcripts in intergenic regions with potential ORFs and one or more assigned TSSs. These ORFs are not yet annotated in the genome of *G. oxydans* and might therefore represent novel transcripts. For six, a homologue was also found in other species and a possible RBS was found for nine. The other novel transcripts without significant BLAST hits and no similarity to sequences present in the Rfam database may represent novel small proteins or non-coding RNAs. In other bacteria, quite a high number of small RNAs are present. However, the comprehensive identification of small RNAs requires specific protocols, which were not applied in our study [69, 87]. Therefore, the analysis of small (non-coding) RNAs in *G. oxydans* requires further studies.

Novel intragenic and antisense transcripts

Besides the identification of novel intergenic transcripts, analysis of the primary transcriptome data also allowed the detection of 328 TSSs in sense orientation within ORFs and 619 TSSs in antisense orientation to ORFs in *G. oxydans*. Intragenic TSSs were detected for 12% of all protein-coding ORFs. Such a high or even higher number has also

been reported for other bacteria [37, 52, 58, 88, 89]. Their functional role is still not understood. However, it is possible that they represent alternative mRNAs encoding smaller proteins, novel protein-coding genes or non-coding RNAs with regulatory functions [88, 90].

For 313 out of the 619 antisense TSSs identified in *G. oxydans*, a transcript was found in the whole transcriptome data. The 313 antisense transcripts were identified for 310 protein-coding ORFs (11%). In other bacteria, antisense transcripts were detected for 5% to 50% of all genes [27, 37, 52, 59, 88, 89]. The physiological role of antisense transcripts was analyzed only for a small subset in few bacteria [91]. It is assumed that these non-coding RNAs have regulatory roles in gene expression, for example by enabling transcription termination due to the formation of secondary structures or by blocking the RBS and therefore translation [32]. Antisense transcripts are usually present at lower levels than the corresponding sense transcripts [38]. Our data also reflect this trend, because the number of read start coverage for many antisense transcripts was low in *G. oxydans*. This low expression might also limit the detection of the transcripts in the whole transcriptome data.

Nucleotide distributions at the transcription initiation sites

In *G. oxydans* the most frequent initiation nucleotides for sTSSs and pTSSs are purines (65% and 75% A+G), whose frequencies are even higher (75% A+G) among the top 10% of sTSSs according to coverage. This mean distribution was also observed in other bacteria [52, 92, 93] and was related to a relatively larger pool size of purine *versus* pyrimidine nucleotides supporting the transcription initiation rate in the cell [94]. In contrast, the frequency of purines as initiation nucleotides is much lower for iTSSs (51% A+G) and asTSSs (47% A+G). The shift from 35% T+C for sTSSs to 53% T+C for asTSSs (+51%) could reduce the overall rate of transcription initiation due to a smaller pool size of pyrimidine nucleotides [94], which could contribute to the overall tendency of lower antisense transcript levels in *G. oxydans* and other bacteria [69]. In accordance with this view, for every TSS category the frequencies of A+G at nucleotide position +1 for the top 10% by coverage are higher compared to A+G of the whole group. This is also reflected by the differences between the top 10% and the lowest 10%, especially for sTSSs and asTSSs. With 13.5% and 14.5%, these differences were relatively high for sTSSs and asTSSs, respectively. No difference or a low difference (3.1%) in the initiation nucleotides A+G between the top 10% and the lowest 10% according to coverage was observed for pTSSs and iTSS, respectively. This may reflect the lower scattering of the iTSS coverages and the lower mean coverage of the group. Also, the low number of pTSSs used for the analysis might contribute to the missing difference. Nevertheless, the higher A+G frequency at transcription initiation sites with higher read coverage supports the theory that a higher pool size of purine nucleotides is

related to increased transcription initiation rates. This way the intracellular purine pool could quickly affect or fine-tune gene expression independent of the regulation by, e.g. transcription factors. This could support fast adaptation of RNA levels, in particular for high-abundant RNAs, in response to environmental changes such as nutrient starvation, which likely result in a shortage of intracellular metabolites including purine nucleotides. Moreover, while the nucleotide frequencies at position +1 are much more similar to the nucleotide frequencies at position +2 than to the nucleotide frequencies at position -1, the nucleotide frequencies at +2 position could contribute similarly, thereby multiplying the outcome on transcription initiation frequencies. Among other key steps in the multistep processes of transcription, in the initiation stage the phosphodiester bond formation between the initial two NTPs is a key step that leads to a transition from the open complex to the initial transcribing complex that extends the RNA in the 5' to 3' direction [95]. Additionally, the nucleotide frequencies at the -1 position may have a hitherto unrecognized effect on transcription initiation rates, especially for the sTSSs.

Furthermore, only slightly conserved promoter motifs or none were identified for intragenic and antisense transcripts in *G. oxydans*. For antisense transcripts, a less conserved -10 region was identified, whereas no -35 region could be found. However, the GC-rich -10 region showed no similarity to the -10 region of the promoter motif identified for protein-coding ORFs. This could be due to the location of the TSSs and promoters, which are predominantly found in ORFs, where the average GC content is typically higher [96]. Also, it was shown that promoters of antisense transcripts are rarely conserved between or even within species [69, 97, 98]. The low levels of antisense transcripts and low conservation of their promoters may suggest that many antisense RNAs in bacteria are rather by-products of the transcript machinery and it was suggested that these pervasive antisense transcripts may serve as a basis for the evolution of novel regulatory RNAs [69].

Conclusion

In this study, we provided a comprehensive RNAseq analysis of the acetic acid bacterium *G. oxydans* using an improved RNAseq method. We identified more than 2,000 TSSs and classified them according to their genomic context. The data obtained allowed identification and analysis of promoter motifs, RBSs, 5´-UTRs and novel transcripts. Also, we were able to describe operon structures. Due to their exceptional metabolism and capabilities for oxidative biotransformations, acetic acid bacteria are of interest both for fundamental studies and for biotechnological applications. The transcriptome data obtained here opens up new possibilities for basic understanding and *Gluconobacter* strain development.

Additional files

Additional file 1:	Table S1; List of manually verified and categorized TSSs.
Additional file 2:	Table S2; List of TSSs assigned to rRNA, tRNA, and RNase P genes.
Additional file 3:	Table S3; List of sTSSs assigned to protein-coding ORFs.
Additional file 4:	Table S4; List of ORFs with corrected start codon ((n)sTSSs).
Additional file 5:	Table S5; List of pTSSs assigned to protein-coding ORFs.
Additional file 6:	Table S6; List of iTSSs.
Additional file 7:	Table S7; List of asTSSs.
Additional file 8:	Table S8; List of nTSSs and novel transcripts.
Additional file 9:	Table S9; List of primary sTSSs with -10 and -35 regions of promoter motifs predicted by *Improbizer* within 50 bases upstream of TSSs.
Additional file 10:	Table S10; List of ribosome binding site (RBS) motifs within 20 nt upstream of translation initiation codon.
Additional file 11:	Table S11; List of operons, sub-operons, and monocistronic transcripts.
Additional file 12:	Table S12; Transcript abundance determined by using cufflinks and cuffnorm.

Acknowledgements
The authors thank Ilka Maria Axmann for helpful discussion.

Availability of data and materials

The RNAseq data are publicly available in the European Nucleotide Archive under accession number PRJEB18739 or *via* the web portal www.gluconobacterfactory.de, the latter one also providing access to additional files including the updated genome reference with the revised start codons.

Funding

The study was supported by the scientific activities of the Bioeconomy Science Center, which were financially supported by the Ministry of Innovation, Science and Research within the framework of the NRW Strategy project BioSC (No. 313/323-400-002 13).

References

1. Bremus C, Herrmann U, Bringer-Meyer S, Sahm H: The use of microorganisms in L-ascorbic acid production. Journal of biotechnology 2006, 124(1):196-205.
2. Gupta A, Singh VK, Qazi GN, Kumar A: *Gluconobacter oxydans*: its biotechnological applications. Journal of molecular microbiology and biotechnology 2001, 3(3):445-456.
3. Pappenberger G, Hohmann H-P: Industrial Production of L-Ascorbic Acid (Vitamin C) and D-Isoascorbic Acid. In: Biotechnology of Food and Feed Additives. Edited by Zorn H, Czermak P. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014: 143-188.
4. Saito Y, Ishii Y, Hayashi H, Imao Y, Akashi T, Yoshikawa K, Noguchi Y, Soeda S, Yoshida M, Niwa M *et al*: Cloning of genes coding for L-sorbose and L-sorbosone dehydrogenases from *Gluconobacter oxydans* and microbial production of 2-keto-L-gulonate, a precursor of L-ascorbic acid, in a recombinant *G. oxydans* strain. Appl Environ Microb 1997, 63(2):454-460.
5. Wang EX, Ding MZ, Ma Q, Dong XT, Yuan YJ: Reorganization of a synthetic microbial consortium for one-step vitamin C fermentation. Microbial cell factories 2016, 15:21.
6. Ameyama M, Shinagawa E, Matsushita K, Adachi O: D-fructose dehydrogenase of *Gluconobacter industrius*: purification, characterization, and application to enzymatic microdetermination of D-fructose. Journal of bacteriology 1981, 145(2):814-823.
7. Herrmann U, Merfort M, Jeude M, Bringer-Meyer S, Sahm H: Biotransformation of glucose to 5-keto-D-gluconic acid by recombinant *Gluconobacter oxydans* DSM 2343. Applied microbiology and biotechnology 2004, 64(1):86-90.
8. Prust C, Hoffmeister M, Liesegang H, Wiezer A, Fricke WF, Ehrenreich A, Gottschalk G, Deppenmeier U: Complete genome sequence of the acetic acid bacterium *Gluconobacter oxydans*. Nature biotechnology 2005, 23(2):195-200.
9. Kranz A, Vogel A, Degner U, Kiefler I, Bott M, Usadel B, Polen T: High precision genome sequencing of engineered *Gluconobacter oxydans* 621H by combining long nanopore and short accurate Illumina reads. Journal of biotechnology 2017.
10. Deppenmeier U, Hoffmeister M, Prust C: Biochemistry and biotechnological applications of *Gluconobacter strains*. Applied microbiology and biotechnology 2002, 60(3):233-242.
11. Matsushita K, Toyama H, Adachi O: Respiratory chains and bioenergetics of acetic acid bacteria. Advances in microbial physiology 1994, 36:247-301.
12. Bringer S, Bott M: Central carbon metabolism and respiration in *Gluconobacter oxydans*. Berlin, Heidelberg, New York: Springer-Verlag; 2016.
13. Hanke T, Nöh K, Noack S, Polen T, Bringer S, Sahm H, Wiechert W, Bott M: Combined fluxomics and transcriptomics analysis of glucose catabolism *via* a partially cyclic pentose phosphate pathway in *Gluconobacter oxydans* 621H. Appl Environ Microb 2013, 79(7):2336-2348.
14. Richhardt J, Bringer S, Bott M: Role of the pentose phosphate pathway and the Entner-Doudoroff pathway in glucose metabolism of *Gluconobacter oxydans* 621H. Applied microbiology and biotechnology 2013, 97(10):4315-4323.
15. Richhardt J, Bringer S, Bott M: Mutational analysis of the pentose phosphate and Entner-Doudoroff pathways in *Gluconobacter oxydans* reveals improved growth of a Δ*edd* Δ*eda* mutant on mannitol. Applied and environmental microbiology 2012, 78(19):6975-6986.
16. Hanke T, Richhardt J, Polen T, Sahm H, Bringer S, Bott M: Influence of oxygen limitation, absence of the cytochrome bc(1) complex and low pH on global gene expression in *Gluconobacter oxydans* 621H using DNA microarray technology. Journal of biotechnology 2012, 157(3):359-372.
17. Richhardt J, Luchterhand B, Bringer S, Büchs J, Bott M: Evidence for a key role of cytochrome bo_3 oxidase in respiratory energy metabolism of *Gluconobacter oxydans*. Journal of bacteriology 2013, 195(18):4210-4220.

18. Kostner D, Luchterhand B, Junker A, Volland S, Daniel R, Büchs J, Liebl W, Ehrenreich A: The consequence of an additional NADH dehydrogenase paralog on the growth of *Gluconobacter oxydans* DSM3504. Applied microbiology and biotechnology 2015, 99(1):375-386.
19. Kiefler I, Bringer S, Bott M: SdhE-dependent formation of a functional *Acetobacter pasteurianus* succinate dehydrogenase in *Gluconobacter oxydans* - a first step toward a complete tricarboxylic acid cycle. Applied microbiology and biotechnology 2015, 99(21):9147-9160.
20. Kiefler I, Bringer S, Bott M: Metabolic engineering of *Gluconobacter oxydans* 621H for increased biomass yield. Applied microbiology and biotechnology 2017, 101(13):5453-5467.
21. Hu Y, Wan H, Li J, Zhou J: Enhanced production of L-sorbose in an industrial *Gluconobacter oxydans* strain by identification of a strong promoter based on proteomics analysis. Journal of industrial microbiology & biotechnology 2015, 42(7):1039-1047.
22. Kallnik V, Meyer M, Deppenmeier U, Schweiger P: Construction of expression vectors for protein production in *Gluconobacter oxydans*. Journal of biotechnology 2010, 150(4):460-465.
23. Merfort M, Herrmann U, Bringer-Meyer S, Sahm H: High-yield 5-keto-D-gluconic acid formation is mediated by soluble and membrane-bound gluconate-5-dehydrogenases of *Gluconobacter oxydans*. Applied microbiology and biotechnology 2006, 73(2):443-451.
24. Mientus M, Kostner D, Peters B, Liebl W, Ehrenreich A: Characterization of membrane-bound dehydrogenases of *Gluconobacter oxydans* 621H using a new system for their functional expression. Applied microbiology and biotechnology 2017, 101(8):3189-3200.
25. Shi L, Li K, Zhang H, Liu X, Lin J, Wei D: Identification of a novel promoter gHp0169 for gene expression in *Gluconobacter oxydans*. Journal of biotechnology 2014, 175:69-74.
26. Petzold CJ, Chan LJ, Nhan M, Adams PD: Analytics for Metabolic Engineering. Frontiers in bioengineering and biotechnology 2015, 3:135.
27. Pfeifer-Sancar K, Mentz A, Rückert C, Kalinowski J: Comprehensive analysis of the *Corynebacterium glutamicum* transcriptome using an improved RNAseq technique. BMC genomics 2013, 14:888.
28. Croucher NJ, Thomson NR: Studying bacterial transcriptomes using RNA-seq. Current opinion in microbiology 2010, 13(5):619-624.
29. Güell M, Yus E, Lluch-Senar M, Serrano L: Bacterial transcriptomics: what is beyond the RNA horiz-ome? Nature reviews Microbiology 2011, 9(9):658-669.
30. van Vliet AH: Next generation sequencing of microbial transcriptomes: challenges and opportunities. FEMS microbiology letters 2010, 302(1):1-7.
31. Wang Z, Gerstein M, Snyder M: RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics 2009, 10(1):57-63.
32. Thomason MK, Storz G: Bacterial antisense RNAs: how many are there, and what are they doing? Annual review of genetics 2010, 44:167-188.
33. Weirick T, Militello G, Muller R, John D, Dimmeler S, Uchida S: The identification and characterization of novel transcripts from RNA-seq data. Briefings in bioinformatics 2016, 17(4):678-685.
34. Osbourn AE, Field B: Operons. Cellular and molecular life sciences: CMLS 2009, 66(23):3755-3775.
35. Price MN, Arkin AP, Alm EJ: The life-cycle of operons. PLoS genetics 2006, 2(6):e96.
36. Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, Szekely LA, Koonin EV: Connected gene neighborhoods in prokaryotic genomes. Nucleic acids research 2002, 30(10):2212-2223.

37. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R *et al*: The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 2010, 464(7286):250-255.
38. Albersmeier A, Pfeifer-Sancar K, Rückert C, Kalinowski J: Genome-wide determination of transcription start sites reveals new insights into promoter structures in the actinomycete *Corynebacterium glutamicum*. Journal of biotechnology 2017.
39. Cohen O, Doron S, Wurtzel O, Dar D, Edelheit S, Karunker I, Mick E, Sorek R: Comparative transcriptomics across the prokaryotic tree of life. Nucleic acids research 2016, 44(W1):W46-53.
40. Filiatrault MJ: Progress in prokaryotic transcriptomics. Current opinion in microbiology 2011, 14(5):579-586.
41. Sorek R, Cossart P: Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nature reviews Genetics 2010, 11(1):9-16.
42. Ewing B, Green P: Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research 1998, 8(3):186-194.
43. Bolger AM, Lohse M, Usadel B: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014, 30(15):2114-2120.
44. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nature methods 2012, 9(4):357-359.
45. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology 2010, 28(5):511-515.
46. Hilker R, Stadermann KB, Doppmeier D, Kalinowski J, Stoye J, Straube J, Winnebald J, Goesmann A: ReadXplorer - visualization and analysis of mapped sequences. Bioinformatics 2014, 30(16):2247-2254.
47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. Journal of molecular biology 1990, 215(3):403-410.
48. Ao W, Gaudet J, Kent WJ, Muttumu S, Mango SE: Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. Science 2004, 305(5691):1743-1746.
49. Čuklina J, Hahn J, Imakaev M, Omasits U, Förstner KU, Ljubimov N, Goebel M, Pessi G, Fischer HM, Ahrens CH *et al*: Genome-wide transcription start site mapping of *Bradyrhizobium japonicum* grown free-living or in symbiosis - a rich resource to identify new transcripts, proteins and to study gene regulation. BMC genomics 2016, 17:302.
50. Malakooti J, Wang SP, Ely B: A consensus promoter sequence for *Caulobacter crescentus* genes involved in biosynthetic and housekeeping functions. Journal of bacteriology 1995, 177(15):4372-4376.
51. Ramírez-Romero MA, Masulis I, Cevallos MA, González V, Dávila G: The *Rhizobium etli* σ^{70} (SigA) factor recognizes a lax consensus promoter. Nucleic acids research 2006, 34(5):1470-1480.
52. Schlüter JP, Reinkensmeier J, Barnett MJ, Lang C, Krol E, Giegerich R, Long SR, Becker A: Global mapping of transcription start sites and promoter motifs in the symbiotic α-proteobacterium *Sinorhizobium meliloti* 1021. BMC genomics 2013, 14:156.
53. Crooks GE, Hon G, Chandonia JM, Brenner SE: WebLogo: a sequence logo generator. Genome research 2004, 14(6):1188-1190.
54. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J *et al*: Rfam 12.0: updates to the RNA families database. Nucleic acids research 2015, 43(Database issue):D130-137.
55. Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, Tolstoy I, Zaslavsky L: Update on RefSeq microbial genomes resources. Nucleic acids research 2015, 43(Database issue):D599-605.
56. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH: JBrowse: a next-generation genome browser. Genome research 2009, 19(9):1630-1638.

57. Guo J, Cheng G, Gou XY, Xing F, Li S, Han YC, Wang L, Song JM, Shu CC, Chen SW *et al*: Comprehensive transcriptome and improved genome annotation of *Bacillus licheniformis* WX-02. FEBS letters 2015, 589(18):2372-2381.
58. Irla M, Neshat A, Brautaset T, Rückert C, Kalinowski J, Wendisch VF: Transcriptome analysis of thermophilic methylotrophic *Bacillus methanolicus* MGA3 using RNA-sequencing provides detailed insights into its previously uncharted transcriptional landscape. BMC genomics 2015, 16:73.
59. Güell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kuhner S *et al*: Transcriptome complexity in a genome-reduced bacterium. Science 2009, 326(5957):1268-1271.
60. Koide T, Reiss DJ, Bare JC, Pang WL, Facciotti MT, Schmid AK, Pan M, Marzolf B, Van PT, Lo FY *et al*: Prevalence of transcription promoters within archaeal operons and coding sequences. Molecular systems biology 2009, 5:285.
61. de Groot A, Roche D, Fernandez B, Ludanyi M, Cruveiller S, Pignol D, Vallenet D, Armengaud J, Blanchard L: RNA sequencing and proteogenomics reveal the importance of leaderless mRNAs in the radiation-tolerant bacterium *Deinococcus deserti*. Genome biology and evolution 2014, 6(4):932-948.
62. Brock JE, Pourshahian S, Giliberti J, Limbach PA, Janssen GR: Ribosomes bind leaderless mRNA in *Escherichia coli* through recognition of their 5'-terminal AUG. Rna 2008, 14(10):2159-2169.
63. O'Donnell SM, Janssen GR: The initiation codon affects ribosome binding and translational efficiency in *Escherichia coli* of cI mRNA with or without the 5' untranslated leader. Journal of bacteriology 2001, 183(4):1277-1283.
64. Shell SS, Wang J, Lapierre P, Mir M, Chase MR, Pyle MM, Gawande R, Ahmad R, Sarracino DA, Ioerger TR *et al*: Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. PLoS genetics 2015, 11(11):e1005641.
65. Srivastava A, Gogoi P, Deka B, Goswami S, Kanaujia SP: In silico analysis of 5'-UTRs highlights the prevalence of Shine-Dalgarno and leaderless-dependent mechanisms of translation initiation in bacteria and archaea, respectively. Journal of theoretical biology 2016, 402:54-61.
66. Vesper O, Amitai S, Belitsky M, Byrgazov K, Kaberdina AC, Engelberg-Kulka H, Moll I: Selective translation of leaderless mRNAs by specialized ribosomes generated by MazF in *Escherichia coli*. Cell 2011, 147(1):147-157.
67. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS: Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. Nucleic acids research 2002, 30(14):3141-3151.
68. Hollands K, Proshkin S, Sklyarova S, Epshtein V, Mironov A, Nudler E, Groisman EA: Riboswitch control of Rho-dependent transcription termination. Proceedings of the National Academy of Sciences of the United States of America 2012, 109(14):5376-5381.
69. Raghavan R, Groisman EA, Ochman H: Genome-wide detection of novel regulatory RNAs in *E. coli*. Genome research 2011, 21(9):1487-1497.
70. Tezuka T, Ohnishi Y: Two glycine riboswitches activate the glycine cleavage system essential for glycine detoxification in *Streptomyces griseus*. Journal of bacteriology 2014, 196(7):1369-1376.
71. Corbino KA, Barrick JE, Lim J, Welz R, Tucker BJ, Puskarz I, Mandal M, Rudnick ND, Breaker RR: Evidence for a second class of S-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. Genome biology 2005, 6(8):R70.
72. Millman A, Dar D, Shamir M, Sorek R: Computational prediction of regulatory, premature transcription termination in bacteria. Nucleic acids research 2017, 45(2):886-893.
73. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS: Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. The Journal of biological chemistry 2002, 277(50):48949-48959.

74. Sudarsan N, Barrick JE, Breaker RR: Metabolite-binding RNA domains are present in the genes of eukaryotes. Rna 2003, 9(6):644-647.
75. Mohanty BK, Kushner SR: Regulation of mRNA Decay in Bacteria. Annual review of microbiology 2016, 70:25-44.
76. Villegas A, Kropinski AM: An analysis of initiation codon utilization in the Domain *Bacteria* - concerns about the quality of bacterial genome annotation. Microbiology 2008, 154(Pt 9):2559-2661.
77. Stenström CM, Holmgren E, Isaksson LA: Cooperative effects by the initiation codon and its flanking regions on translation initiation. Gene 2001, 273(2):259-265.
78. Makrides SC: Strategies for achieving high-level expression of genes in *Escherichia coli*. Microbiological reviews 1996, 60(3):512-538.
79. Shine J, Dalgarno L: Terminal-sequence analysis of bacterial ribosomal RNA. Correlation between the 3'-terminal-polypyrimidine sequence of 16-S RNA and translational specificity of the ribosome. European journal of biochemistry 1975, 57(1):221-230.
80. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A: Regulation of noise in the expression of a single gene. Nature genetics 2002, 31(1):69-73.
81. Vellanoweth RL, Rabinowitz JC: The influence of ribosome-binding-site elements on translational efficiency in *Bacillus subtilis* and *Escherichia coli in vivo*. Molecular microbiology 1992, 6(9):1105-1114.
82. Camacho A, Salas M: Effect of mutations in the "extended -10" motif of three *Bacillus subtilis* sigmaA-RNA polymerase-dependent promoters. Journal of molecular biology 1999, 286(3):683-693.
83. Hawley DK, McClure WR: Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nucleic acids research 1983, 11(8):2237-2255.
84. Browning DF, Busby SJ: The regulation of bacterial transcription initiation. Nature reviews Microbiology 2004, 2(1):57-65.
85. Paget MS, Helmann JD: The sigma70 family of sigma factors. Genome biology 2003, 4(1):203.
86. Staron A, Sofia HJ, Dietrich S, Ulrich LE, Liesegang H, Mascher T: The third pillar of bacterial signal transduction: classification of the extracytoplasmic function (ECF) sigma factor protein family. Molecular microbiology 2009, 74(3):557-581.
87. Mentz A, Neshat A, Pfeifer-Sancar K, Pühler A, Rückert C, Kalinowski J: Comprehensive discovery and characterization of small RNAs in *Corynebacterium glutamicum* ATCC 13032. BMC genomics 2013, 14:714.
88. Mitschke J, Georg J, Scholz I, Sharma CM, Dienst D, Bantscheff J, Voss B, Steglich C, Wilde A, Vogel J *et al*: An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. Proceedings of the National Academy of Sciences of the United States of America 2011, 108(5):2124-2129.
89. Mitschke J, Vioque A, Haas F, Hess WR, Muro-Pastor AM: Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in *Anabaena* sp. PCC7120. Proceedings of the National Academy of Sciences of the United States of America 2011, 108(50):20130-20135.
90. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J *et al*: Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. Genome research 2007, 17(6):746-759.
91. Sesto N, Wurtzel O, Archambaud C, Sorek R, Cossart P: The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. Nature reviews Microbiology 2013, 11(2):75-82.
92. Kröger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hébrard M, Händler K *et al*: The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. Proceedings of the National Academy of Sciences of the United States of America 2012, 109(20):E1277-1286.

93. Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juarez K, Contreras-Moreira B *et al*: Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. PloS one 2009, 4(10):e7526.
94. Buckstein MH, He J, Rubin H: Characterization of nucleotide pools as a function of physiological state in *Escherichia coli*. Journal of bacteriology 2008, 190(2):718-726.
95. Alhadid Y, Chung S, Lerner E, Taatjes DJ, Borukhov S, Weiss S: Studying transcription initiation by RNA polymerase with diffusion-based single-molecule fluorescence. Protein science : a publication of the Protein Society 2017, 26(7):1278-1290.
96. Bohlin J, Eldholm V, Pettersson JH, Brynildsrud O, Snipen L: The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. BMC genomics 2017, 18(1):151.
97. Nicolas P, Mader U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S *et al*: Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. Science 2012, 335(6072):1103-1106.
98. Shao W, Price MN, Deutschbauer AM, Romine MF, Arkin AP: Conservation of transcription start sites within genes across a bacterial genus. mBio 2014, 5(4):e01398-01314.
99. Oliveros JC: Venny. An interactive tool for comparing lists with Venn's diagrams. 2007-2015.

TABLES

Table 1. Sequencing and mapping statistics for all cDNA sequencing libraries.

Transcriptome	primary	whole
# reads [1)]	10.13	55.76
# reads [2)]	6.13	n.a.
mapped reads	1.26	42.77
unique matches	1.1	32.87

[1)] Sequence reads after trimming; [2)] sequence reads with barcode TACCCTAG at the 5´-end representing false positive TSSs. Values are given in million.

Table 2. Predicted *cis*-regulatory elements in *G. oxydans* 621H according to the Rfam database compared to RNAseq results.

Rfam prediction			Annotation		Primary	Whole
Description	**Accession**	**Start[a] End**	**Gene**	**Annotation**	**Start[b] Stop[d]**	**Start[c] Stop[d]**
FMN riboswitch	RF00050	1,075,971 1,076,128	GOX_RS06030	Riboflavin biosynthesis protein RibD	1,075,965 1,076,281	1,075,974 1,076,281
Glycine riboswitch	RF00504	1,200,190 1,200,279	GOX_RS06635	Glycine cleavage system, amino methyl-transferase T	1,200,192 1,200,436	1,200,201 1,200,436
SAM-II riboswitch	RF00521	1,829,621 1,829,542	GOX_RS09595	O-succinyl-homoserine sulfhydrylase	1,829,638 1,829,484	1,829,625 1,829,484
TPP riboswitch	RF00059	2,443,346 2,443,480	GOX_RS12420	Phosphomethyl-pyrimidine synthase	2,443,351 2,443,615	2,443,363 2,443,615
Cobalamin riboswitch	RF00174	1,111,882 1,111,673	GOX_RS06220	TonB-dependent receptor	-	1,111,858 1,111,529
ROSE element	RF00435	1,450,717 1,450,634	GOX_RS07835	Molecular chaperone Hsp20	-	1,450,712 1,450,641
Fluoride riboswitch	RF01734	152,387 152,452	GOX_RS00740	Camphor resistance protein CrcB	-	152,404 152,966

[a] Positions according to Rfam database were adjusted according to the updated genome reference [9]. European Nucleotide Archive accession number: PRJEB18739.
[b] Position of the TSS.
[c] Observed by manual inspection.
[d] End of the 5´-UTRs.

FIGURE LEGENDS

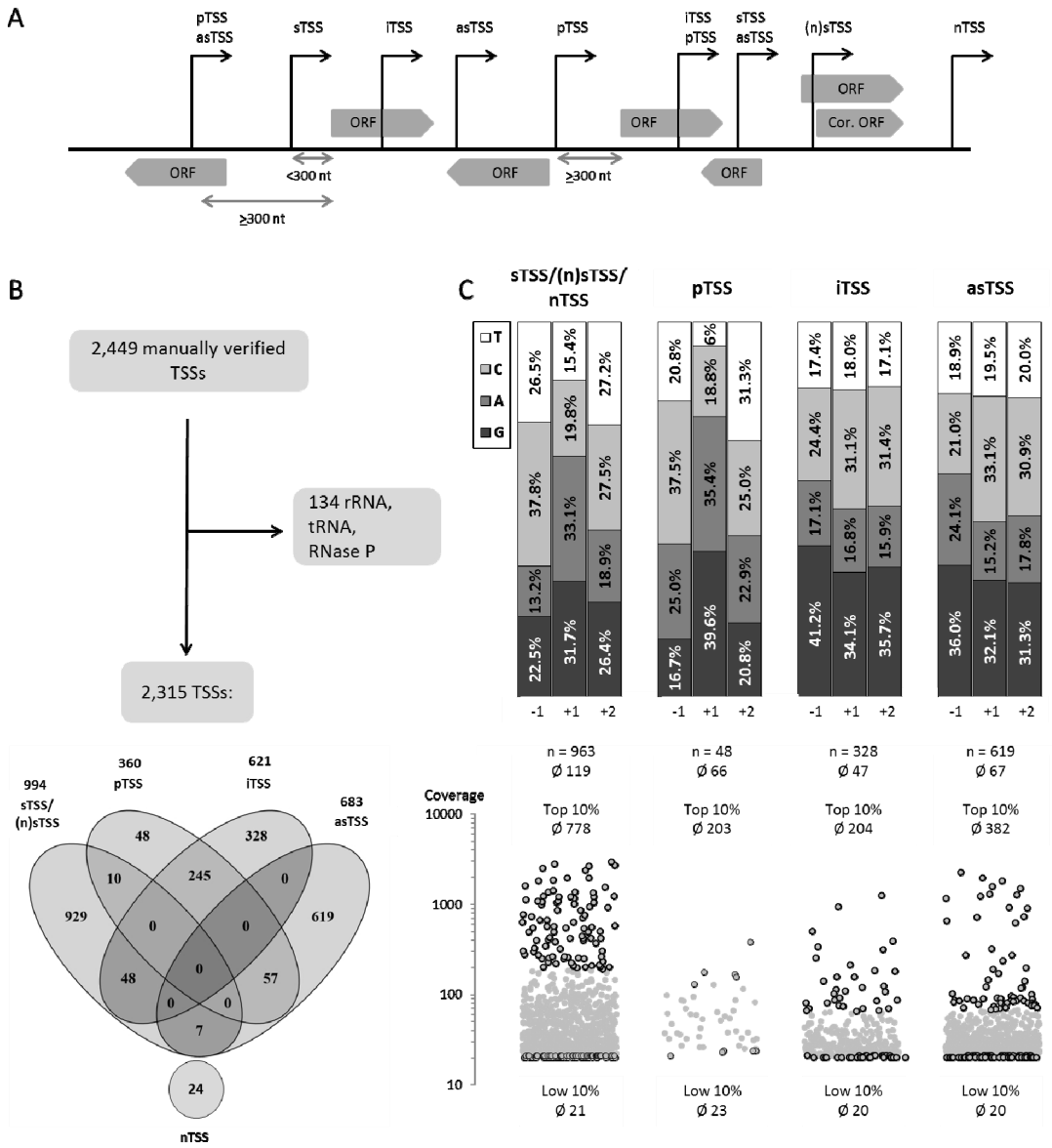

Figure 1 Classification of transcription start sites. (A) Schematic overview of categories used for classification of TSSs according to their genomic context. sTSS: Sense TSSs with a annotated ORF downstream in a maximal distance of 300 nt. (n)sTSS: TSSs downstream of an ORF start, which were used to revise the translation start position (corrected ORF). pTSS: Putative TSSs assigned to annotated ORFs downstream, yet with a minimal distance of 300 nt and a maximal distance of 600 nt. iTSS: Intragenic TSSs in sense orientation. asTSS: TSSs located antisense to ORFs or UTRs. nTSS: Intergenic TSS representing possible novel transcripts. Also, possible scenarios with TSSs associated to more than one category are shown. (B) Number and classification of detected TSSs. TSSs belonging to rRNA, tRNA, and RNase P genes as well as false positive TSSs were removed. 2,315 manually verified TSSs were considered for classification. The Venn diagram showing overlap between the categories was generated with Venny 2.1.0 [99]. (C) Upper panel: Nucleotide distribution at the transcription initiating site +1 as well as at -1 and +2 based on the TSSs identified solely for the categories sTSS/(n)sTSS/nTSS, pTSS, iTSS, and asTSS. The 10 TSSs assigned to both sTSS and pTSS were assumed to be sTSSs (see Results). Lower panel: Distribution of the read start coverage for TSSs assigned to the categories sTSS, pTSS, iTSS, and asTSSs. TSSs with the highest (Top 10%) and lowest coverage (Low 10%) are bold-framed. The

number (n) and the average coverage (∅) for all TSSs and the top as well as low 10% is given for each category.

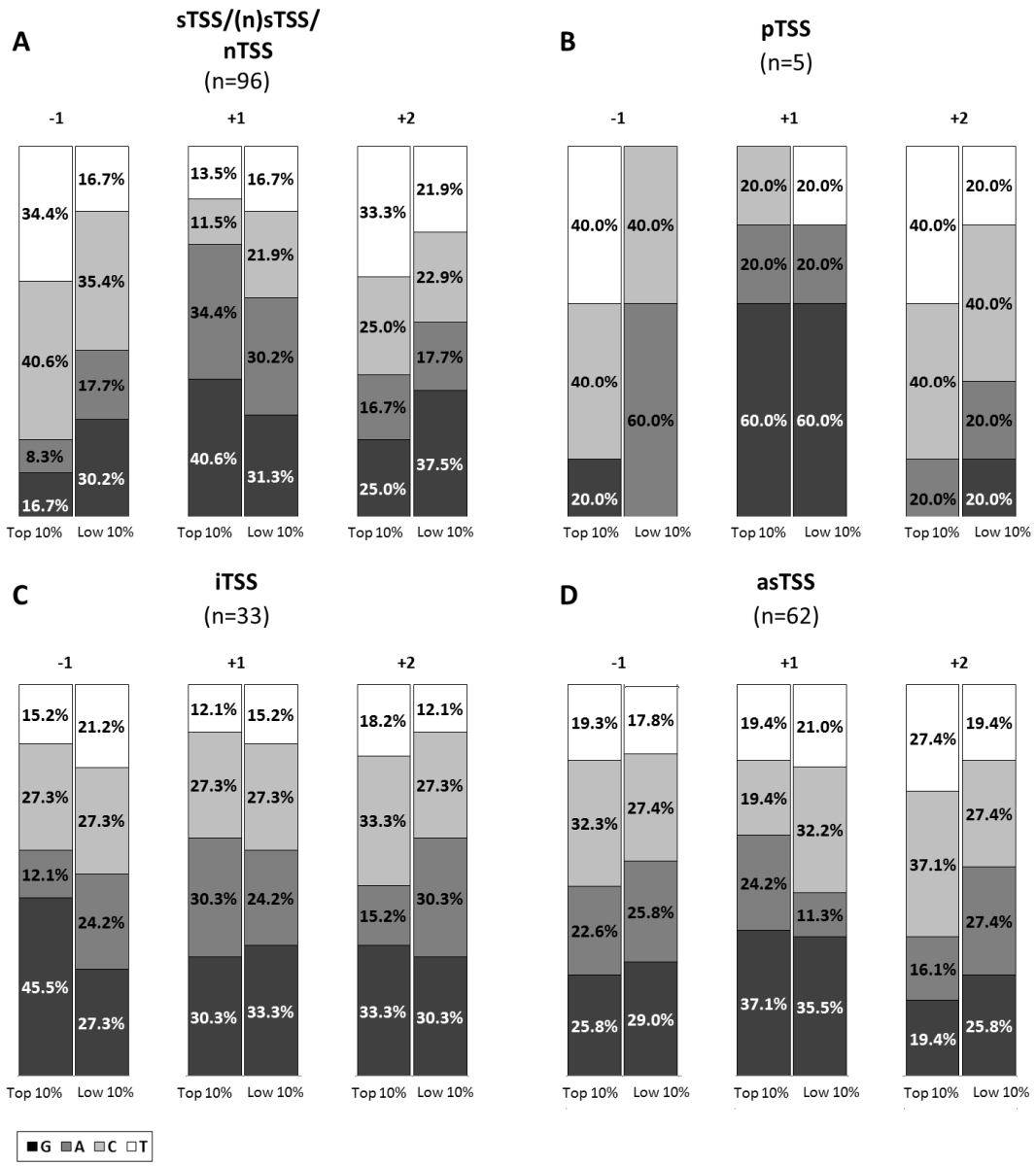

Figure 2 Nucleotide distribution at transcription initiation site +1, as well as at the -1 and +2 position for the TSSs with the highest (Top 10%) and lowest (Low 10%) read start coverage according to TSS categories.

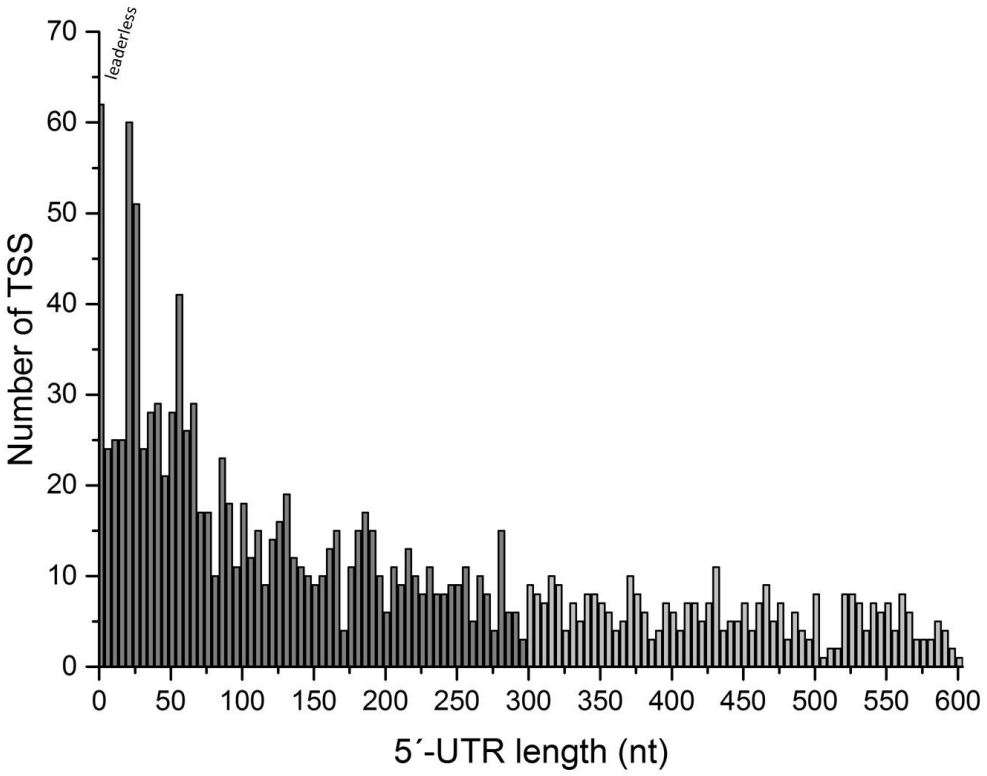

Figure 3 Number of TSSs assigned to annotated genes in correlation to the resulting 5´-UTR length. 1,344 primary and secondary TSSs were used for this analysis and grouped into 5 nt intervals. The 5´-UTR of 62 transcripts (5%) is ≤3 nt and they were therefore classified as leaderless. The 350 pTSSs belonging to 5´-UTRs with a length >300 nt are shown in light grey.

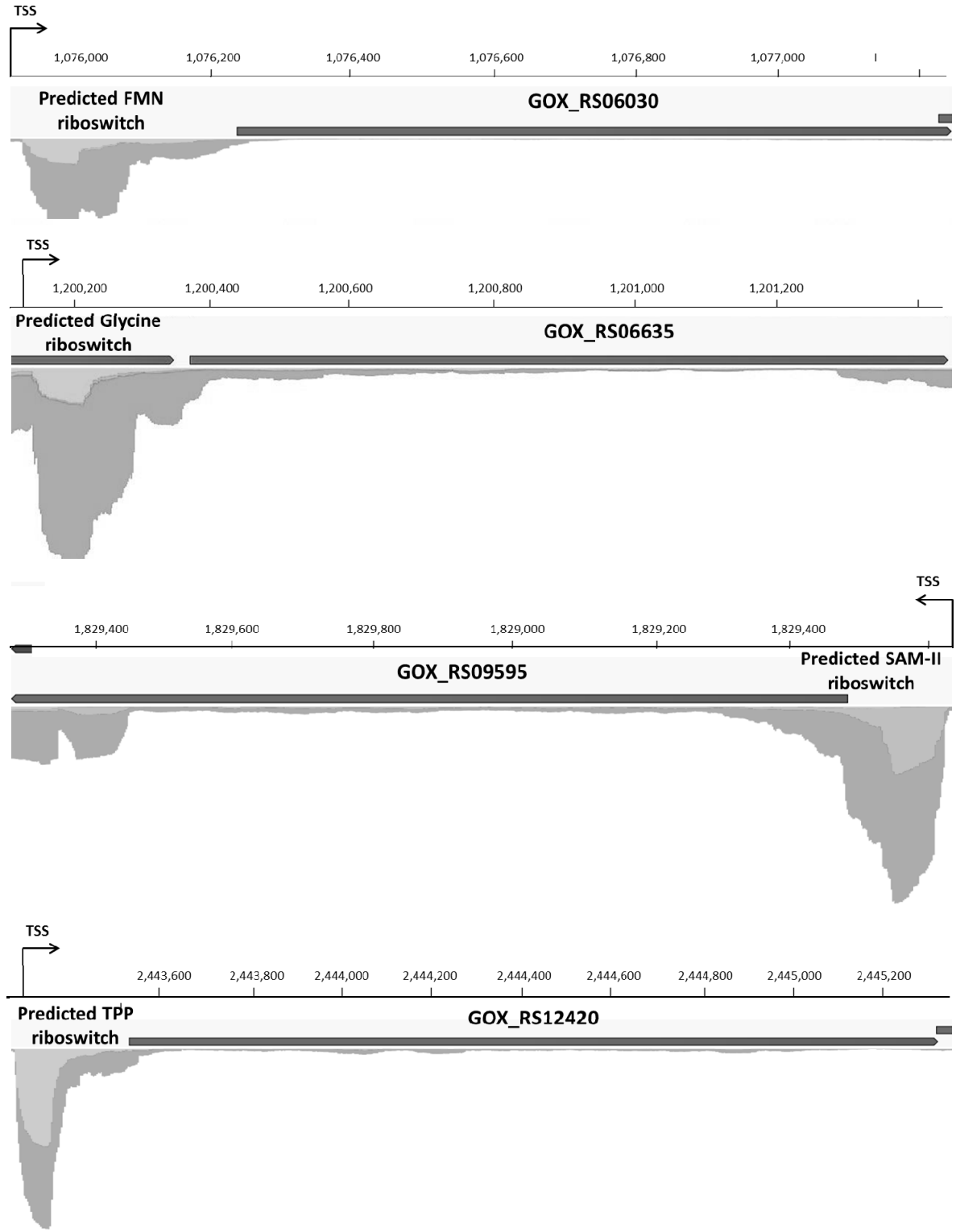

Figure 4 Mapping coverage of whole transcriptome data for the predicted FMN riboswitch (upstream of GOX_RS06030), predicted glycine riboswitch (upstream of GOX_RS06635), predicted SAM-II riboswitch (upstream of GOX_RS09595), and predicted TPP riboswitch (upstream of GOX_RS12420). Detailed positional data can be found in table 2.

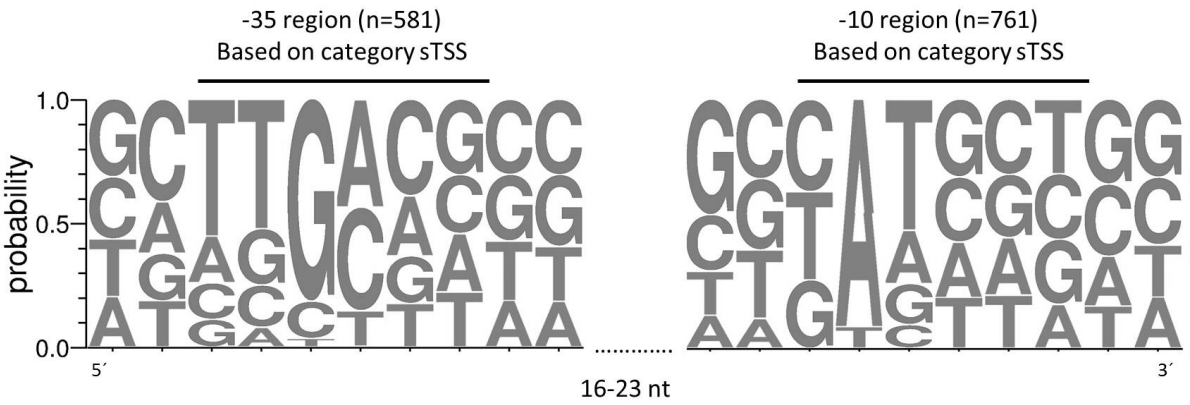

Figure 5 Promoter motifs found within sequences upstream of identified TSSs according to *Improbizer* [48]. Sequence logos were created with WebLogo [53]. Distribution of nucleotides within -10 and -35 regions is based on sequences from the category sTSS. Only primary TSSs were considered. 808 sequences were used for prediction of promoter motifs. A -10 region was identified for 761 sequences, a -35 region with a distance of 16-23 nt to the -10 region was found for 581 sequences.

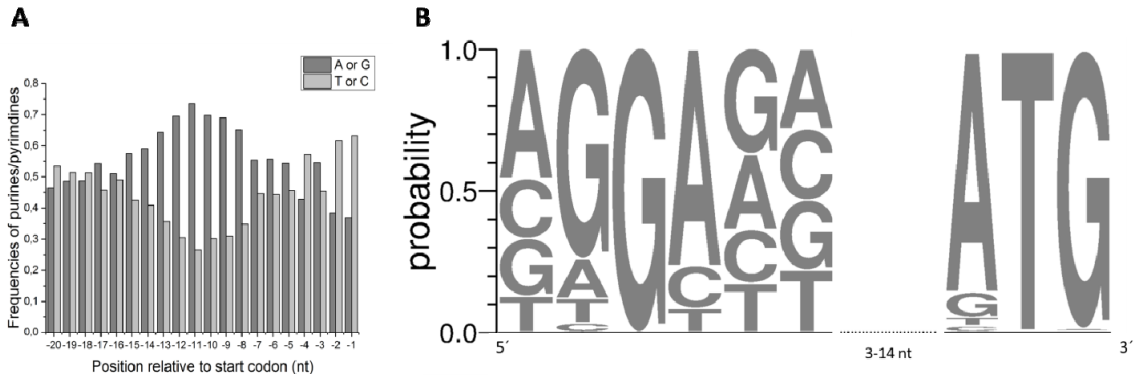

Figure 6 Identification and analysis of ribosome binding sites (RBSs) within 20 bases upstream of the translation start sites of 973 5´-UTRs with a minimal length of 20 nt. (A) Comparison of purine (A or G) *versus* pyrimidine (T or C) frequencies. (B) Nucleotide distribution within the RBSs identified by *Improbizer* [48] in 913 (94%) of the 5´-UTR sequences, a spacer of 3-14 nt, and the nucleotide distribution within the translation start codon. The motif logo was designed with WebLogo [53].

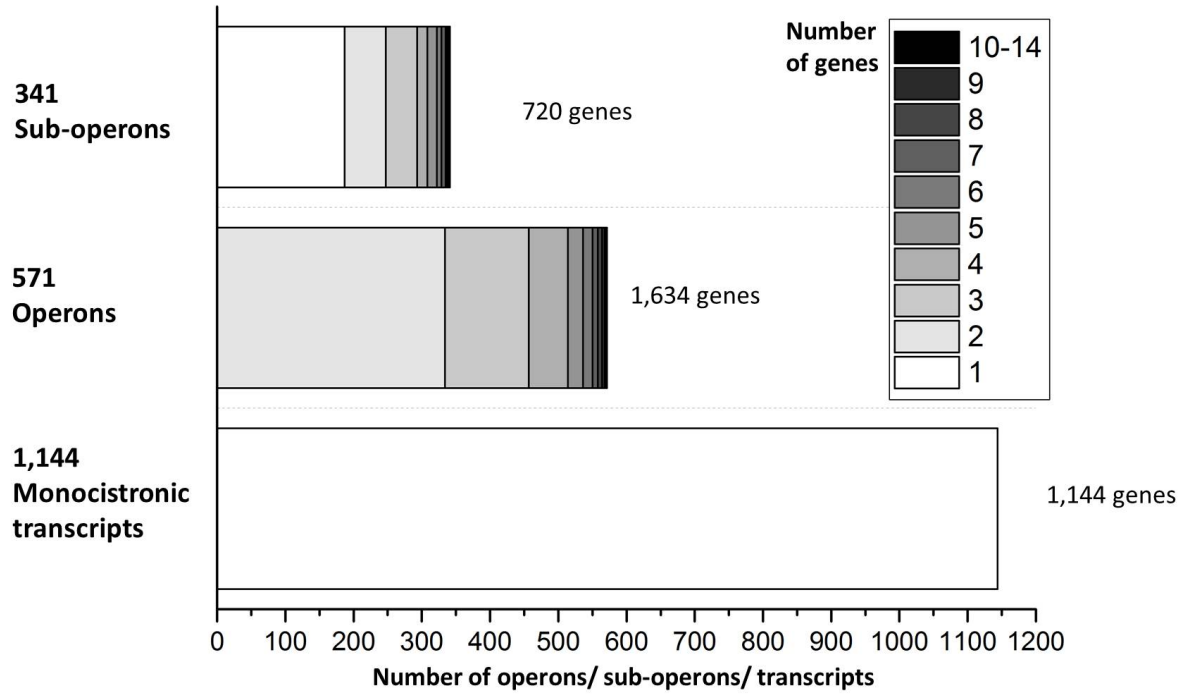

Figure 7 Analysis and number of monocistronic transcripts, operons, and sub-operons identified in *G. oxydans*. The number of genes in operons is gray-color coded as indicated.

2.3 Global mRNA decay analysis and analysis of fragmented 23S rRNA in *G. oxydans*

Kranz, A., Steinmann, A., Degner, U., Mengus-Kaya, A., Matamouros, S., Bott, M., and Polen, T. (2018). Global mRNA decay analysis and 23S rRNA fragmentation in *G. oxydans* 621H. **BMC Genomics (accepted).**

Author's contributions:

AK performed RNA sequencing, analysed the RNAseq and mRNA decay data, wrote major parts of the manuscript, and contributed to the finalization of the manuscript. AS contributed to mRNA decay data processing and analysis. UD carried out cell cultivations and DNA microarray experiments. SM developed the protocol for ribosome enrichment. UD and SM purified ribosomes. AMK performed LC-MS/MS analysis for identification of proteins. MB revised and improved the manuscript and data interpretation. TP designed and supervised the study, performed mRNA decay data processing, prepared and uploaded the microarray data to GEO, wrote parts of the manuscript, and revised and finalized the whole manuscript.

Overall contribution AK: 70%

Global mRNA decay and 23S rRNA fragmentation in *Gluconobacter oxydans* 621H

Angela Kranz[1,2], Andrea Steinmann[1,2], Ursula Degner[1], Aliye Mengus-Kaya[1], Susana Matamouros[1], Michael Bott[1,2] and Tino Polen[1,2,*]

[1)] IBG-1: Biotechnology, Institute of Bio- and Geosciences, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

[2)] The Bioeconomy Science Center (BioSC), c/o Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

ORCID:	Angela Kranz	orcid.org/0000-0002-8000-0400
ORCID:	Michael Bott	orcid.org/0000-0002-4701-8254
ORCID:	Tino Polen	orcid.org/0000-0002-0065-3007

Running title: mRNA decay and 23S rRNA in *Gluconobacter*

*** for correspondence:** Dr. Tino Polen
E-mail: t.polen@fz-juelich.de
Phone: +49 (0)2461 61 6205
Fax: +49 (0)2461 61 2710

Keywords: *Gluconobacter oxydans*
mRNA decay
ATP synthase
tricarboxylic acid cycle
ribosome
23S rRNA fragmentation
intervening sequence

Abstract

Background: The use of *Gluconobacter oxydans* 621H in biotechnological production processes such as whole-cell biotransformations is typically limited due to a low biomass yield. This is attributed to *G. oxydans*' incomplete central carbon metabolism and incomplete oxidation of a wide range of carbohydrates and alcohols, whose corresponding products are secreted almost completely into the medium. In previous studies the biomass yield of *G. oxydans* was substantially increased by metabolic engineering. However, there is still room for further improvement. This may also require more knowledge about gene expression in *G. oxydans* to uncover still unrecognized potential bottlenecks in the naturally evolved partially incomplete metabolism. To gain further insights into the physiology of *G. oxydans* we here analyzed the global mRNA decay.

Results: Using DNA microarrays we measured the time-dependent RNA level changes in the presence of rifampicin and estimated the mRNA half-lives by linear regression. Overall, the mRNA half-lives ranged from 3 min to 25 min with a global mean of 5.8 min. Linear regression analysis also showed a statistically significant ($p < 0.0001$) inverse relationship between transcript stability and abundance in *G. oxydans*. The transcripts encoding GroES and GroEL required for the proper folding of proteins ranked at the top among transcripts exhibiting both high abundance and long half-lives. The transcripts encoding the H^+-dependent F_1F_o-ATP synthase involved in energy metabolism ranked among the transcripts with the shortest mRNA half-lives. In the central carbon metabolism the FPKM expression values of TCA cycle and anaplerotic genes were among the lowest and the mRNA half-lives of many TCA cycles genes were below the global mean. Furthermore, quality control of mRNA decay samples indicated considerable instability of the 23S rRNA full-length transcripts. Further analysis by affinity purification of ribosomes and analysis of the associated RNA by RNAseq revealed a fragmentation pattern and new cleavage regions in 23S rRNAs, yet no known intervening sequences were found in *G. oxydans*.

Conclusions: The inverse relationship of transcript abundance and half-lives, which was also observed for *G. oxydans* in this study, supports the general view that overall mRNA stability does not play a major role to obtain high mRNA abundance. Rather a quick mRNA turnover for fast adaptation of the cell in case of environmental changes is made possible by shorter half-lives of highly expressed genes. In contrast to data from other bacteria, the transcripts of the H^+-dependent ATP synthase exhibited rather short half-lives which could be or become a bottleneck in *G. oxydans*. Also, expression of TCA cycle and anaplerotic genes could generally be limiting factors in *G. oxydans*, which should be considered in future

metabolic engineering approaches to further improve growth and biomass yield of *G. oxydans*.

Background

Gluconobacter oxydans is a Gram-negative strictly aerobic acetic acid bacterium industrially used for oxidative biotransformation of carbohydrates. Important biotransformation products are e.g. L-sorbose, a precursor for vitamin C production, 5-ketofructose, dihydroxyacetone, and 6-amino-L-sorbose [1-6]. The beneficial ability of *G. oxydans* is the incomplete oxidation of a variety of substrates (e.g. sugars and sugar alcohols) in the periplasm by membrane-bound dehydrogenases and release of resulting products into the cultivation medium [7-9]. Correspondingly, only a small amount of substrate is taken up by the cell and channeled into the cytoplasmic metabolism to be used for growth [10]. Furthermore, sequencing of the genome revealed the absence of genes for enzymes of the central metabolism, such as 6-phosphofructokinase, succinyl-CoA synthetase, and succinate dehydrogenase [11]. Accordingly, the Embden-Meyerhof-Parnas pathway (glycolysis) and tricarboxylic acid (TCA) cycle are incomplete. Both the periplasmic oxidation and the restrained cytoplasmic sugar metabolism contribute to limited assimilation of carbohydrates into cell material and therefore to a low final biomass yield. Industrial use of *G. oxydans* for oxidative biotransformations is therefore costly. To overcome these hindrances, metabolic engineering was performed to complete the TCA cycle by introducing the heterologous genes *sdhCDABE* and *sucCD* into the genome together with prevention of periplasmic and cytoplasmic glucose oxidation by simultaneous deletion of membrane-bound and soluble glucose dehydrogenase [12, 13]. Furthermore, the NADH oxidation capacity was increased by introducing an additional NADH dehydrogenase from *G.* oxydans DSM3504 [14]. These steps lead to an increase of biomass yield by up to 60%, thereby reducing the glucose cost for biomass formation. Although this is already very advantageous for industrial applications, still unrecognized bottlenecks in *G. oxydans'* naturally evolved partially incomplete metabolism might exist.

Therefore, in this study we conducted genome-wide mRNA decay analysis to get further insights into the physiology of *G. oxydans*. In living cells, the abundance of mRNAs is a result of the balance between gene expression and degradation of mRNAs. Global mRNA decay analysis in prokaryotes was already described, for example, for the model microorganisms *Escherichia coli* [15], *Bacillus subtilis* [16], and *Mycobacterium tuberculosis* [17]. These studies used rifampicin for inhibition of transcription at different time points during growth to measure the changes of relative mRNA levels using DNA microarrays. These changes reflect the degradation of transcripts and allow the calculation of mRNA half-lives [15, 18]. In bacteria, mRNA half-lives typically range from around 1 min or shorter up to

30 min [15, 16, 19]. Generally, a correlation between the half-lives of transcripts and the cellular function of the encoded proteins was observed in these studies. Transcripts associated to housekeeping functions such as cell envelope and ion transport exhibited relatively long mRNA half-lives, whereas genes involved in stress response and in regulatory functions exhibit a faster transcript turnover to adapt to environmental changes in a short time [20, 21]. Likewise, mRNA half-lives are not fixed and can be affected by, for example, changes of the growth rate when environmental conditions change [22, 23]. Since mRNA half-lives are affected by different growth conditions, regulation of their stability and degradation is quite diverse. It is dependent on secondary structures of the 5´ and 3´ untranslated regions, posttranscriptional modifications such as polyadenylation, abundance of ribonucleases and the presence of cleavage sites recognized by ribonucleases, interaction with small regulatory RNAs, as well as location of the mRNAs in the cell [20, 24].

Here, we measured temporal RNA level changes in *G. oxydans* 621H in response to rifampicin and estimated the mRNA half-lives. This analysis also revealed apparent instability of the 23S rRNA. Further analysis uncovered that the 23S rRNA in ribosomes is fragmented in *G. oxydans*, yet does not contain known intervening sequences.

Methods

Bacterial strain and cultivation conditions

In this study, the wild type *Gluconobacter oxydans* 621H strain DSM 2343 from the German Collection of Microorganisms and Cell Cultures (DSMZ) was used. The cells were cultivated in mannitol medium containing 220 mM (4% w/v) mannitol, 5 g L^{-1} yeast extract, 1 g L^{-1} KH_2PO_4, 1 g L^{-1} $(NH_4)_2SO_4$, 2.5 g L^{-1} $MgSO_4$ x 7 H_2O, and 50 µg mL^{-1} cefoxitin as antibiotic. Cells were grown in 500 mL shaking flasks with three baffles containing 50 mL of the mannitol medium (30°C, 140 rpm). Cell growth in liquid culture was followed by measuring the optical density at 600 nm (OD_{600}) using a spectrophotometer. If rifampicin was applied, it was added to a cell culture at OD_{600} of 0.6 to 0.8 from a stock (50 mg mL^{-1} in methanol) to obtain the final concentration as indicated, while the appropriate volume of methanol without rifampicin was added to the control culture. For isolation of total RNA from cells and purification of ribosomes, cells were harvested at indicated OD_{600} as described and stored at -20°C until use [10].

Isolation of RNA

For determination of mRNA half-lives by DNA microarray analysis, total RNA was isolated from harvested cells as described [10]. RNA fraction from ribosomes was isolated by phenol-chloroform-isoamyl alcohol (25:24:1) and chloroform-isoamyl alcohol (24:1) extractions followed by ethanol precipitation [25]. RNA concentrations were assessed photometrically using a Nanodrop ND-1000, and fluorometrically using a Qubit® 2.0 device and Qubit® RNA BR Assay Kit according to the manufacturer's instructions (Life Technologies). RNA quality and band pattern was checked and visualized on formaldehyde agarose gels as described [26].

DNA microarray analysis

The microarray analysis aimed at the determination of the mRNA level changes in *G. oxydans* after addition of rifampicin. The analysis was performed three times by independent biological replicates. *G. oxydans* cells were harvested for isolation of total RNA directly before (t0) and 2 min (t2), 5 min (t5), 10 min (t10) as well as 15 min (t15) after addition of rifampicin. Each sample after addition of rifampicin (tx) was compared to t0 samples. For pairwise comparisons, the A mix and the B mix of the Agilent Spike-In Kit (Agilent Technologies) was used to spike t0 and tx RNA samples. The synthesis of labeled cDNA samples were carried out as described [10]. Custom-made 4 x 44K DNA microarrays for genome-wide gene expression analysis were obtained from Agilent Technologies and were designed using Agilent's eArray platform (https://earray.chem.agilent.com/earray). The array design comprises oligonucleotides for the annotated protein-coding genes and structural RNA genes of the genome from *G. oxydans* 621H (CP000009, and CP000004 to CP000008), *Bacillus subtilis* str. 168, *Corynebacterium glutamicum* ATCC 13032, and *Escherichia coli* K12, as well as Agilent's control spots. After hybridization according to the manufacturer's instructions, the arrays were washed using Agilent's wash buffer kit. Subsequently, the fluorescence of DNA microarrays was determined at 532 nm (Cy3-dUTP) and 635 nm (Cy5-dUTP) at 5 µm resolution with a GenePix 4000B laser scanner and GenePix Pro 6.0 software (Molecular Devices). Raw data files of fluorescence images were saved in TIFF format followed by quantitative image analysis (GenePix Pro 6.0) using the corresponding Agilent's gene array list (GAL) file. The results were saved as GPR file (GenePix Pro 6.0).

Data normalization and calculation of mRNA half-lives

For calculation of mRNA half-lives, first the ratio of median values (GenePix Pro 6.0) reflecting the relative mRNA level changes were normalized for each array hybridization separately using a normalization factor. This factor was calculated for each hybridization based on the log base 2 of the non-normalized ratio of median values of the Spike-In

(+)E1A_r60_1 and (+)E1A_r60_a20 RNAs serving as internal controls and each having 32 array spots randomly scattered. Both RNAs are present in A mix and B mix in a 1:1 A/B ratio (Agilent Technologies). Accordingly, the normalization factor was calculated that the log base 2 of the factor-normalized ratio of median values of the 1:1 control RNAs is 0 on average. Subsequently, the ratio of median value for each gene was normalized with the calculated factor. All microarray data including the normalized ratio of medians were stored for further analysis, quality filtering and mRNA half-life calculation in the in-house DNA microarray database [27]. A normalized ratio value was included in the calculation of the average of the triplicates for each time point if the following quality filter was fulfilled by the spot data (GenePix Pro 6.0): i) Flags ≥0 and ii) signal/noise ≥3 for Cy5 (F635Median / B635Median) or Cy3 (F532Median / B532Median). The resulting data matrix with the average values of the four time points was used to calculate the mRNA half-lives and R^2 in Excel (Microsoft) by linear fit as described for *E. coli* data [18]. For further analysis, only genes where each average ratio is based on three quality-filtered values from the triplicates and with $R^2 > 0.7$ were considered for further analysis and functional grouping using the assigned product functions [11]. Significant differences between functional groups were identified *via* a one-way ANOVA test using Excel (Microsoft). Subsequently, a *post hoc* t-test was performed to identify mean mRNA half-lives of functional groups, which differ significantly from the overall average half-life. *p* values were adjusted using Bonferroni correction [28].

Determination of mRNA abundance by RNAseq

Bacterial cells were grown in mannitol-containing medium as described above and harvested at an OD_{600} of 1.4. Total RNA was isolated and 5 µg of total RNA were treated with the Ribo-Zero magnetic kit for Gram-negative bacteria (Illumina) for depletion of rRNA. Afterwards, ethanol precipitation was performed according to manufacturer's instructions. For preparation of the whole transcriptome libraries, we used the TruSeq stranded mRNA sample preparation kit (Illumina) according to manufacturer´s instructions, yet 5 µL of rRNA-depleted total RNA were mixed with 13 µL of Fragment, Prime, Finish Mix and incubated at 94°C for 4 min for fragmentation and priming. The library was quantified *via* qPCR with the KAPA Library Quantification Kit (Peqlab) and sequenced on a MiSeq desktop sequencer (Illumina), generating paired-end reads with a read length of 75 bp. Sequencing reads were trimmed and strand-specifically mapped to the genome reference (CP000009) and the five plasmids pGOX1 to pGOX5 (CP000004 – CP000008) using the RNAseq analysis tool of the CLC Genomics Workbench (Qiagen Aarhus A/S) to determine absolute FPKM values [29]. Only unique mapped reads with ≤1% mismatches were considered for this analysis. Linear

regression analysis of transcript abundances and mRNA half-lives was performed with GraphPad Prism 7.00 using default settings.

Purification of ribosomes

For preparation of lysates, frozen cell pellets were thawed and resuspended to 0.2 g mL^{-1} in lysis buffer (70 mM KCl, 10 mM $MgCl_2$, 10 mM Tris-HCl, pH 7.4). Resuspended cells were disrupted in a French press (SLM Aminco) at 15,000 psi (3 passages) using 10 mL of cell suspension. Remaining intact cells and debris were sedimented by centrifugation (57,400 *g*; 20 min; 4°C) followed by filtering through a 0.22 µm filter. The protein concentration was determined by Bradford assay. If required, the volume was adjusted by addition of lysis buffer to obtain a concentration of 1.5 to 4.5 $mg_{protein}$ mL^{-1}. Ribosomes were purified based on the use of a strong anion exchange monolithic column [30]. Therefore, 0.5 mL of prepared cell lysate was injected into an ÄKTA pure FPLC system equipped with fraction collector F9-C (GE Healthcare Life Sciences) to two CIM® QA-0.34 mL monolithic disks (BIA separations) pre-equilibrated in lysis buffer. The chromatography was performed at a flow rate of 2 mL min^{-1} using lysis buffer without (buffer A) and with 1 M NaCl (buffer B). The disks were washed with 5 column volumes of buffer A and sequentially eluted with 7 column volumes 40%, 56% and 100% of buffer B. During the step-wise elution the online chromatogram was manually inspected for upcoming peaks indicating elution of protein to manually collect from the start to the end of a peak into one elution fraction. The collected fractions were further analyzed using mass spectrometry to identify proteins.

Protein identification

For identification of proteins in relevant elution fractions, proteins were precipitated with trichloroacetic acid (10%) and resuspended in 50 µL of digestion buffer followed by tryptic digest using the Trypsin Singles Proteomics Grade Kit according to the manufacturer's instructions for solution digestion in a volume of 100 µL (Sigma-Aldrich). After tryptic digest, the peptides were separated chromatographically on a nanoLC Eksigent ekspert™ 425 system (Sciex) coupled with a quartz emitter Tip (New Objective) to a TripleTof™ 6600 mass spectrometer (Sciex). Digested samples were loaded on a pre-column (ChromXP C18-3 µm, 350 µm x 0.5 mm, Sciex) for desalting and enrichment using a flow of 3 µL/min (10 min) of buffer A (0.1% formic acid in HLPC grade water). The separation of peptides followed on an analytical column (ChromXP 3C18-CL-120, 0.075 x 150 mm, Sciex) with a gradient method (125 min) using buffer A and buffer B (0.1% formic acid in acetonitrile) at 40°C and a flow of 300 nL/min. The gradient conditions were 5% of buffer B for 1 min, 5%-9% for 9 min, 9%-20% for 50 min, 20%-40% for 40 min, 40%-80% for 5 min

and 80% for 4 min. The mass spectrometer was operated with a “top 50” method: Initially, a 250 ms survey scan (TOF-MS mass range 400-1,500 amu, high resolution mode) was collected from which the top 50 precursor ions were automatically selected for fragmentation, whereby each MS/MS event (mass range 100-1,700 amu, high sensitivity mode) consisted of a 75 ms fragment ion scan. The source and gas settings were 2,200 V spray, 40 psi curtain gas, 6 psi ion source gas, and 75°C interface heater. For peptide (99% confidence) and protein (1% FDR) identification the mass spectrometry data obtained were processed with ProteinPilot™ (V4.5 beta, Sciex) using the Paragon algorithm.

rRNA sequencing

RNA purified from ribosomes was sequenced to analyze the 23S rRNA fragments. Sequencing libraries were generated using the TruSeq stranded mRNA sample preparation kit (Illumina) according to manufacturer´s instructions, yet without the fragmentation step. cDNA libraries were quantified and sequenced as described above. Sequencing reads were processed and mapped to both the four 23S rRNA genes (GOX0221, GOX1159, GOX1319, and GOX1467) and 16S rRNA genes (GOX0224, GOX1156, GOX1316, and GOX1464) using CLC Genomics Workbench (Qiagen Aarhus A/S). The default mapping parameters were changed to consider only reads, which mapped over their complete length with an identity of at least 99%. The coverage *per* base was extracted and manually inspected in Excel (Microsoft) to identify regions with a coverage <5% of the average gene coverage and visualized with Origin (OriginLab).

Results

mRNA half-lives ranged from 3 min to 25 min

In *G. oxydans*, the glycolysis and the TCA cycle are incomplete due to the absence of genes for 6-phosphofructokinase, succinyl-CoA synthetase, and succinate dehydrogenase [11]. Also, only a very small amount of substrate is taken up by the cell (~10% for glucose), which then is only partially channeled (~5% of total glucose) into the cytoplasmic metabolism to be used for biomass formation [10]. Because of this apparent low usage of metabolism and that *G. oxydans* can almost not grow in defined minimal media, we were especially interested in the mRNA half-lives to check whether possibly still unrecognized and unexpected bottlenecks in *G. oxydans'* naturally evolved partially incomplete metabolism may exist.

Typically, mRNA half-lives are determined in rifampicin experiments. Rifampicin inhibits transcription initiation, thereby enabling to measure the time-dependent decrease of

mRNA levels by turnover in the absence of mRNA synthesis. Depending on the rifampicin concentration and the sensitivity of the host cells, growth is transiently stopped keeping most cells relatively intact or cells can be killed accompanied by cell lysis and loss of RNA for analysis. The sensitivity of *G. oxydans* towards rifampicin was not reported yet. Therefore, we first tested the influence of rifampicin at different concentrations on the growth of *G. oxydans* in liquid culture. Based on data reported for *E. coli*, *B. subtilis*, *M. tuberculosis* and *Sulfolobus* sp. we focused on a range of up to 250 µg/mL [15-17, 19]. In the presence of 50 and 100 µg/mL, growth was fully stopped within 30 min, very similar to the growth when adding 150 µg/mL. In contrast, with 200 and 250 µg/mL a significant drop of cell density was observed within 30 min suggesting a higher extent of cell damage or lysis (Figure 1A). We therefore chose 150 µg/mL of rifampicin to analyze global mRNA decay.

For comparison of transcriptomes, total RNA was isolated from cells directly before addition of rifampicin and 2, 5, 10, and 15 minutes after. Formaldehyde agarose gels were used to size-separate and visualize the isolated RNA for assessing RNA qualities according to the rRNA bands (Figure 1B). The 23S rRNA band was much weaker compared to the 16S rRNA band, and the latter showed a double band character. Also, the 23S rRNA band almost disappeared in the course of the sampling times after 15 min, suggesting a processing of the expected mature full length 23S rRNA transcripts comprising 2,709 to 2,711 nt. These results suggest that for *G. oxydans* the rRNAs should better not be used for ratio data normalization in DNA microarray analysis, as it was done in other studies where the rRNAs of the host were demonstrably considered as stable enough to be used for data normalization in such experiments. We therefore used Spike-In RNA mixtures A and B (Agilent Technologies) which contain several RNA transcripts for which known ratios between mix A and mix B can be expected. For the pairwise DNA microarray comparisons of transcriptomes corresponding to the time point before addition of rifampicin (t0) and after (t2, t5, t10, and t15), either Mix A or B were added to the RNA samples before cDNA synthesis (Figure 1C). Spike-In RNA transcripts present in a ratio of 1:1 between mix A and B were used to calculate for each microarray hybridization the normalization factor by which all ratios obtained in the respective hybridization experiments were normalized. The normalized ratio data were used to calculate the average for each time point each from three independent biological replicate experiments. The mRNA half-lives can be estimated *via* linear regression of the log ratios over time. Based on the 4 time points of the analysis and filtering for R^2 >0.7, we obtained half-life values for 1,193 transcripts. That is 44% of all protein-coding genes [11] and in the range of fractions reported for *B. subtilis* (35%), *E. coli* (53%), *M. tuberculosis* (53%), and *S. acidocaldarius* (70%) [15-17, 19].

Overall, the calculated mRNA half-lives of the 1,193 genes range from approximately 3 min for a metalloprotease (GOX2034) and subunits of a F_1F_o-ATP synthase (GOX1110-

GOX1112, GOX1311-GOX1314) to approximately 25 min for glycyl-tRNA synthetase subunit alpha (GOX1744) and a putative sugar/polyol transporter (GOX0354) (Table S1). The mean of the calculated half-lives is 5.8 min and the median is 5.2 min, which is also reflected by the high number of transcripts exhibiting half-lives between 3 and 7 minutes (Figure 1D).

mRNA half-life, ORF length and transcript abundance

Correlation of mRNA half-lives to lengths of ORFs exhibited no dependency in *G. oxydans* (Figure 2A), as was reported for *E. coli* and which is consistent with the site specificity of mRNA decay determinants [15, 31]. To investigate the relationship between mRNA half-life and expression level, we sequenced the transcriptome (complex medium with mannitol) and determined absolute expression values (FPKM) *via* normalization of mapped reads. Similar to observations in *E. coli* [15] and the archaea *Sulfolobus solfataricus* [19], where mRNAs exhibiting higher transcript abundance were less stable, linear regression analysis also showed a statistically significant ($p < 0.0001$) inverse relationship between transcript stability and abundance in *G. oxydans* (Table S2 and Figure 2B). So, many lower expressed genes showed longer apparent mRNA half-lives and *vice versa*. However, certain transcripts exhibited both high abundance and long half-life. Such transcripts include the ORFs encoding co-chaperonin GroES (GOX1901) and chaperonin GroEL (GOX1902), 50S ribosomal protein L17 (GOX0355), and coenzyme PQQ synthesis protein PqqA (GOX0987), as well as a number of putative/hypothetical proteins (GOX0352, GOX1332, GOX1424, GOX1787). Similarly, certain transcripts exhibited both low abundance and short half-life. Such transcripts include ORFs encoding, for example, threonine synthase (GOX1868), imidazole glycerol phosphate synthase subunit HisF (GOX0483) and phosphoribosyl-ATP pyrophosphatase (GOX0484), anthranilate synthase component I (GOX2286), phosphoribosylformylglycinamidine synthase PurL (GOX2301), peptidyl-prolyl *cis/trans* isomerases (GOX0301, GOX1677), as well as some putative/hypothetical proteins. The expression of these genes could represent bottlenecks, since low transcript abundance and short half-life may yield a low protein level which could additionally contribute to bottlenecks in the metabolism of *G. oxydans*.

mRNA half-lives and gene product function

In previous studies, correlations between stability of mRNAs and functions of the gene products were observed [15, 16, 19]. For *G. oxydans*, we also assigned mRNA half-lives to annotated functional categories [11]. According to these categories comprising 1,647 grouped genes in total, half-life estimations were obtained for 863 genes (52%). One-sided ANOVA test revealed significant differences between groups ($p = 0.000012$). Subsequent *post hoc* analysis identified 5 categories with significantly shorter mean half-lives and 4 with significantly longer mean half-lives than the overall average half-life (Table 1). The subset with significantly shorter mean half-lives contains the group of ATP-proton motive force interconversion, transcription, fatty acid and phospholipid metabolism, nucleotide metabolism, and degradation of proteins and peptides. Categories with significantly longer half-lives are sugar and alcohol degradation, DNA restriction and modification, cell motility,

and ion homeostasis. Since the *post hoc* analysis only focuses on mean values, we also looked at the distribution of the estimated mRNA half-lives to identify outliers within functional categories (Figure S1). Generally, the distribution shows that 99% of all analyzed transcripts exhibited half-lives shorter than 15 min. Outliers, which were observed in six categories belong to the top 1% of genes with the longest mRNA half-lives. The category DNA restriction and modification is the only group with a significantly longer mean half-life compared to global mean containing such an outlier. This is a putative restriction endonuclease (GOX1507) with a half-life (16.7 min) two times longer than the mean half-life (8.1 min) of this group. Further outliers were found in the categories protein folding and stabilization, transport, aminoacyl-tRNA biosynthesis, and ribosome assembly. Their mean half-lives range from 5.1 min to 6.5 min (Table 1) and do not differ significantly from that of other categories or global mean. Category protein folding and stabilization exhibited average mean half-life of 6.5 min, which is slightly longer than the mean half-life for all genes (5.8 min), yet among those genes, the transcripts of the two chaperonins GroES (GOX1901) and GroEL (GOX1902) notably exhibited long half-lives with 16.3 min and 19.6 min, respectively. A putative sugar/polyol transporter (GOX0354) exhibiting the longest observed transcript half-life (24.7 min) belongs to category transport. Also, a ribose ABC transporter ATP-binding protein (GOX2220) exhibited a relatively long transcript half-life (16.6 min) compared to the mean of this category (5.9 min). The glycyl-tRNA synthesase subunit α (GOX1744), exhibited the second longest transcript half-life (24.5 min), whereas the transcript for subunit β exhibited 7.2 min (GOX1743), yet R^2 is only 0.64 and did not fulfill the >0.7 criterium. Another outlier was found in category cell envelope. The average half-life of this group is 5.4 min, yet the transcript of the capsule polysaccharide export ATP-binding protein (GOX1486) was almost 4-fold longer (19.2 min). mRNA half-lives of category ribosome assembly were between 3 min and 9.7 min, except for the transcript of the 50S ribosomal protein L17 (GOX0355), which exhibited a relatively long half-life of 16.6 min.

Half-lives of mRNAs assigned to the central carbon metabolism

Since *G. oxydans* has an unusual central metabolism, we were especially interested in the stability of transcripts encoding enzymes of the glycolysis, pentose phosphate pathway (PPP), Entner-Doudoroff pathway (EDP), and TCA cycle including pyruvate metabolism. Therefore, we mapped the estimated mRNA half-lives to these pathways [32] and ignored less important genes also assigned to these categories [11]. Overall, mRNA half-lives range from only 3.6 min for *aceEα* (GOX2289) encoding pyruvate dehydrogenase E1 component subunit α to 12.4 min for one of two annotated triosephosphate isomerases (GOX2217), which is assigned to the glycolysis (Figure 3). Also, according to FPKM values this *tpi* transcript showed almost the lowest value in the central carbon metabolism (Figure 3,

Table S2). This apparent very low expression might be compensated by the very long mRNA half-life. Besides the transcript for *tpi*, which belongs to the top 2% with the longest mRNA half-lives overall, another outlier in category glycolysis (Figure S1) is the transcript for dihydroxyacetone kinase (GOX2222), which is not directly linked to the central carbon metabolism. However, its half-life of 19.1 min makes it one of the most stable transcripts in *G. oxydans*.

The transcripts of the PPP genes exhibited relatively short half-lives only ranging in a narrow range from 3.8 min to 5.1 min. Interestingly, besides the transcript of *rpi* (4.4 min) encoding the ribose-5-phosphate isomerase in the PPP, the transcript of GOX2218 also encodes a ribose-5-phosphate isomerase and exhibited a 3-fold longer half-life (13.3 min). The transcripts of the two genes of the EDP exhibited a half-life of 5 min (*edd*) and 8.3 min (*eda*), the latter belonging to the top 12% of stable transcripts. Transcripts assigned to the pyruvate metabolism exhibited half-lives from 3.6 min to 6 min. Among the enzymes of the TCA cycle, three transcripts notably exhibited half-lives longer than the average. The aconitase transcript (GOX1335) is relatively stable (9.3 min) in comparison to other transcripts and is among the top 10% of the long-lived transcripts overall in *G. oxydans*. However, based on FPKM values reflecting transcript levels, the TCA cycle genes overall exhibited the lowest apparent expression levels in the central metabolism compared to glycolysis, PPP and EDP (Figure 3, Table S2). Also, expression of *ppc* and *eda* encoding phosphoenolpyruvate carboxylase and KDPG aldolase, respectively, appear to be rather low. These enzymes catalyze anaplerotic reactions, which are important to replenish intermediates that have been extracted from a functional TCA cycle for biosynthesis. In summary, except for a few outliers, the genes belonging to the central carbon metabolism exhibited a relatively broad range from shorter to longer mRNA half-lives and the very low FPKM expression values of some candidate genes may indicate potential bottlenecks.

Revisiting the linear regression for estimation of mRNA half-lives

Although the proportion of *G. oxydans* transcripts (genes) with estimated half-lives (44%) is in the range of other mRNA decay studies (35 - 70%) [15-17, 19], there is no half-life estimation for the remaining genes which still represent a significant proportion of all genes. The only reason is that the selected criterion R^2 >0.7 was not fulfilled when linear regression was done along with all sampling times. Such uniform handling and analysis on a global scale may not be adequate for all genes. For example, genes with shorter half-lives may exhibit the progress in RNA level decrease only at earlier sampling times, while the differences (ratios) at later sampling times remain more or less the same for several reasons including limits of the DNA microarray technology. Therefore, many genes exhibiting such time-dependent ratio data will very likely not fulfill a well-intentioned R^2 criterion for filtering

results of linear regression. Depending on the range of the absolute sampling times it is legitimate to check the outcome of half-live estimations by linear regression calculations with a focus on the earlier time points and omitting later ones. According to that, we omitted the last time point (15 min) from half-live calculation for all the remaining genes where R^2 was <0.7. Of these 1,446 remaining genes, half-live estimations were obtained based on the first three time points for 1,307 transcripts with an R^2 >0.7 (Table S1). Together with the 1,193 transcripts mentioned above, this resulted in mRNA half-life estimations for 95% of the protein-coding genes of *G. oxydans*. Furthermore, when omitting the last two time points from linear regression analysis of the 139 remaining genes, further mRNA half-life estimation for 134 genes were obtained, while transcripts of 5 genes exhibited a negative half-life value by calculation and never fulfilled criterion R^2 >0.7 (Table S1). Although this estimation of half-lives for remaining transcripts may increasingly include uncertainty because of a reduced number of data points, it appears to make sense at least for a part of the genes to classify their transcripts as likely short-lived, medium- or long-lived. For example, based on all 4 time points, transcript of GOX1113 encoding F_1F_o-ATP synthase subunit A exhibited estimated half-live of 4.54 min with R^2 0.69, which therefore was close, yet did not pass the R^2 filter criterion. The transcript of GOX1113 was the only one of the genes encoding the subunits of an F_1F_o-ATP synthase (GOX1110-GOX1113, GOX1311-GOX1314), while the other transcripts exhibited R^2 >0.7 and passed the filter. Based on the first 3 time points, data for GOX1113 passed with R^2 of 0.96 and the estimated half-life is 2.4 min. This half-life better fits to the results for GOX1110-GOX1112 and GOX1311-GOX1314 than 4.54 min, whose transcripts mainly exhibited half-lives from 2.85 min to 3.12 min (and one with 3.72 min). Moreover, when calculating the RNA half-lives for all genes of this F_1F_o-ATP synthase based on the first three time points of the analysis (2, 5 and 10 min), the estimated RNA half-live values are even lower (approximately 30%) and in a narrow range from 1.92 min to 2.52 min. This suggests that some half-lives of the 1,193 transcripts may be over-estimated when using an overall time span of 15 min for *G. oxydans*, even when R^2 >0.7 is fulfilled. However, we included the results for the remaining 55% of all genes (Table S1) mentioned above into the calculation of mean half-lives of the categories to check the impact (Table 1). The global mean of the half-lives did not change (5.8 min), yet the means of categories changed somewhat and the overall range of the half-lives changed from 2.8 - 24.7 min to 1.8 – 79 min, with one outlier (273 min) for GOX0386 encoding DNA-directed RNA polymerase subunit β. The estimated RNA half-life of RNA polymerase subunit α (GOX0356) was 52 min and 6 min for subunit β‘ (GOX0385).

Mapping analysis revealed fragmentation regions in 23S rRNAs from *G. oxydans*

The visualization of total RNA isolated for global mRNA decay analysis indicated that the full-length 23S rRNA from *G. oxydans* was low abundant and very unstable (Figure 1B). This could be a potential limitation for the growth of *G. oxydans*. Alternatively, it could be a further 23S rRNA processing which results in some fragments which can be found in the ribosomes. 23S rRNA fragmentation is quite widespread in bacteria including α-proteobacteria [33, 34]. We therefore wanted to check whether the 23S rRNA in the ribosomes from *G. oxydans* is fragmented and if so, in which regions of the 23S rRNA the cleavage sites are located. We chromatographically enriched ribosomes from *G. oxydans* cells grown under regular condition without rifampicin. We analyzed two time points, one in the exponential growth phase (OD_{600} 1 after 4.25 h of cultivation) and one in the early stationary phase (OD_{600} = 2.5 after 9.5 h of cultivation). Cell pellets were used to obtain crude protein extracts which volumes were adjusted to set the range of the protein concentration for the chromatography runs on the monolithic disks. According to the elution profile, we always obtained four major peaks termed P1 to P4 (Figure 4A). Together, *per* run the four elution fractions of the peaks typically comprised 85% to 95% of the total protein applied to the column (Table 2). According to nanoLC-based mass spectrometry, 933 *G. oxydans* proteins were identified overall in the four protein fractions (Table S3). Already 449 to 865 proteins were identified in peak P1 and P2, yet the overall content of ribosomal proteins was rather low in peak P1 and P2. The highest abundance of the ribosomal proteins was found in peak P3 (56% elution buffer) as judged by SDS gel analysis and MALDI-ToF mass spectrometry (data not shown) and the overall high numbers of detected peptides *per* ribosomal protein in the nanoLC-MS/MS analysis (Table S3). This profile of the chromatogram and elution of ribosomes in the third peak is typical for this procedure and already described elsewhere for other bacteria [30]. Moreover, by far the largest amount of RNA could be isolated from the protein fraction of peak P3, which is expected when ribosomes are enriched in peak P3 (Table 2).

According to the genome annotation, the mature full-length 16S and 23S rRNA from *G. oxydans* is expected at 1,478 nt and 2,709 / 2,710 / 2,711 nt, respectively [11]. However, in the RNA isolated from peak P3 a mature full-length 23S rRNA could not be detected in the formaldehyde agarose gel analysis at ~2,710 nt, yet the RNA samples exhibited a specific pattern of smaller fragments (Figure 4B). The pattern shows three fragments at approximately 1,500 nt, 900 nt and 800 nt, and two shorter fragments at approximately 300 and 400 nt. By size the longest fragment of ~1,500 nt corresponds to the 16S rRNA. To narrow the regions of fragmentation down, the RNA was then sequenced *via* next-generation sequencing and the paired-end reads were mapped to the sequences of the 16S and 23S rRNA gene loci from *G. oxydans* (Figure 5). To identify possible fragmentation positions

based on the mapping coverage we searched for regions with a coverage of <5% of the total average mapping coverage of a gene loci. With RNA samples from the exponential phase, we found for the 23S rRNA genes three regions with such a low or almost no coverage suggesting fragmentation sites in GOX0221, GOX1319, and GOX1467 exhibiting a gene length of 2,710 nt and 2,711 nt, respectively (Figure 5A). The same three regions were also found in the shortest 23S rRNA locus from *G. oxydans* (GOX1159; 2,709 nt), yet notably a forth potential fragmentation region was suggested according to the mapping coverage (Figure 5B). The three extremely low coverage regions uniformly found for all four 23S rRNA genes were at positions 1,139-1,160; 1,456-1,465; and 1,716-1,770 (Figure 5, Tables S4-S7). Start and stop of these regions differed by only one nucleotide for some 23S rRNA genes. The fourth region only found for GOX1159 was found at position 2,022-2,136 (Table S5). For the 16S rRNA genes we did not found such a low coverage region indicating the presence of the mature full-length transcripts (Figure 5C, Tables S8-S11). With the RNA samples from the stationary phase, the Illumina read mappings showed very similar results (Tables S12-S19). Only for the second and forth fragmentation position a difference of 55 nt (1,401-1,465) and 31 nt (1,991-2,136) was observed for the length of the very low coverage region.

Comparison of *G. oxydans* 23S rRNA fragmentation with other 23S rRNAs

Fragmentation of 23S rRNA was observed in several bacteria [33]. Here, we compared the 23S rRNA of *G. oxydans* with selected 23S rRNA sequences from other bacteria including *E. coli* where the 23S rRNA is not fragmented and IVSs are absent (Figures 6 and S2). For *Salmonella typhimurium* [35], *Rhodobacter sphaeroides*, *Bradyrhizobium japonicum*, *Rhodopseudomonas palustris* [34, 36], *Rhizobium leguminosarum*, and *Agrobacterium radiobacter* [37], it was shown that fragmentation of 23S rRNA occurs by RNA cleavage to remove IVSs. For example, according to the literature, IVSs can be found at the positions 131-168, 543-550, and 1,176 relative to the *E. coli* sequence (Figure S2). Close to the first very low coverage region in the mapping for *G. oxydans* an IVS is present in *S. typhimurium* and *R. sphaeroides*. These regions only partially overlap with the relevant *G. oxydans* region and the sequence similarity is low (Figure 6A). For all bacteria shown here and independent of fragmentation, the sequences are quite diverse in this region. For *A. radiobacter* and *R. leguminosarum*, fragmentation without the presence of IVS was observed close to position 1,500. Close to this region we also observed fragmentation of 23S rRNA in *G. oxydans* (Figure 6B). Sequence similarities among the three bacteria in this region are approximately 70%. The third fragmentation position close to 1,750 appears to be present only in *G. oxydans* 23S rRNA (Figure 6C). Sequence similarities between rRNAs with and without fragmentation in this region are quite

high, yet there are a few differences in *G. oxydans* compared to the other bacteria. The fourth region with a very low coverage was found only for GOX1159 and not for the other three 23S rRNA copies of *G. oxydans*. In this region, GOX1159 differs only at the nucleotide position 2,034 (a C for T) and 2,170 (the C is missing in GOX1159) compared to the three other 23S rRNAs in *G. oxydans*. This suggests that these positions in the transcript of GOX1159 are specifically relevant for binding and/or cleavage by the processing RNase, likely RNase III as described for others [34, 38]. This fourth fragmentation position is also not known from the other bacteria (Figure 6D). Altogether, no known IVSs described in the literature were found in the 23S rRNAs from *G. oxydans*.

Discussion

In this study, we globally estimated mRNA half-lives in *G. oxydans* 621H and thereby we also found that the native full-length 23S rRNA transcript was unstable. Subsequently, we found 23S rRNA fragments in fractions of enriched ribosomes and almost no full-length 23S rRNA. This indicates further 23S rRNA processing in *G. oxydans*. mRNA decay plays an important role in the metabolism of nucleic acids in both prokaryotic and eukaryotic cells and also affects fluctuations in protein synthesis and growth [39, 40]. The abundance of mRNAs in cells is a result of both syntheses during gene expression and degradation over time resulting in some stability. For *E. coli* and the archaeon *S. solfataricus*, an inverse relationship between transcript abundance and half-lives was observed [15, 19]. We also observed such an inverse relationship for *G. oxydans* based on FPKM values roughly reflecting transcript abundance. Many highly abundant transcripts are among the least stable and transcripts with longer half-lives are less abundant. This also supports the view that overall mRNA stability does not play a major role to obtain high mRNA abundance and that rather a quick mRNA turnover for fast adaptation of the cell in case of environmental changes is made possible by shorter half-lives of highly expressed genes. For example, transcripts of such highly expressed genes which are growth-related can be decayed fast due to their short half-lives in case of cell cycle arrest [19]. Nevertheless, as in other studies, not all transcripts follow the inverse relationship between abundance and stability reflected by the half-life. For example, in *G. oxydans* the transcripts of the molecular chaperones GroES and GroEL exhibited high abundance as well as long half-lives compared to the average. GroES and GroEL are generally required for the proper folding of proteins [41]. For a cell it is of advantage to maintain higher transcript levels by both the transcription rates and longer half-lives of (essential) genes, whose availability is required independent from environmental conditions. Relationships between gene functions and mRNA stability were

already described [15, 19, 21, 42, 43]. Generally, it is assumed that transcripts of genes whose expression need to be changed rapidly exhibit shorter half-lives, whereas transcripts of genes involved in housekeeping functions exhibit longer half-lives. In *G. oxydans*, the general functional categories ATP-proton motive force interconversion, fatty acid and phospholipid metabolism, RNA metabolism, nucleotide metabolism, ribosome assembly, and tRNA metabolism are among the groups exhibiting the shortest mean mRNA half-lives. This could allow fast adaptations by reducing the cell's capacities to save energy and resources upon environmental changes detrimental for growth. For example, the capacity of fatty acid and consequently phospholipid metabolism affects the cell size of *E. coli* in response to nutrient availability and could be likewise in *G. oxydans* and others [44]. Also, the high mRNA stability of genes involved in cell motility and ion homeostasis is reasonable [19]. Such major cellular functions need to be maintained under many conditions.

In *E. coli*, genes involved in degradation of macromolecules exhibited shorter mRNA half-lives [15]. Likewise, in *G. oxydans* most of the functional categories for degradation of molecules exhibited half-lives shorter than the global mean (DNA degradation, degradation of polymers, degradation of proteins and peptides). Only category sugar and alcohol degradation notably exhibited a significantly longer mean half-life. In view of *G. oxydans*' typical habitats such as sugary, alcoholic and acidic environments in flowers, fruits, beer and wine, a broader range of substrates may be sufficiently available for longer periods and fast adaptation to the absence of substrates is less important. Also, the longer mRNA half-lives should support protein synthesis of the encoded enzymes. This would contribute to *G. oxydans*' exceptional ability to incompletely oxidize a broad range of sugars and alcohols in the periplasm from which *G. oxydans* generate the majority of the energy *via* cytochrome bo_3 oxidase extruding protons which are used by F_1F_o-ATP synthase to generate ATP [45]. The estimated mRNA half-lives of *cyoBACD* (GOX1911-14) encoding the subunits of the cytochrome bo_3 oxidase range above the global mean (5.8 min) from 6.2 min to 8.8 min, which does not point to an obvious bottleneck. In contrast, the mRNA half-lives of the F_1F_o-type ATP synthase encoded by the two operons *atpBEFF*' (GOX1110-13) and *atpHAGDC* (GOX1310-14) with a mean of only 3.2 min ranked among the shortest half-lives in *G. oxydans*. *G. oxydans* possesses a second F_1F_o-type ATP synthase (GOX2167-75) which transcripts exhibited a mean of 5.1 min. The first one is an ortholog of the ATP synthases of *Acetobacter pasteurianus* IFO 3283–01, *Gluconacetobacter diazotrophicus* Pal 5 and other α-proteobacteria, while the latter one is an ortholog of Na^+-translocating F_1F_o-ATP synthases present always in addition to the H^+-translocating ATP synthase in the archaea *Methanosarcina barkeri* and *M. acetivorans*, in a number of marine and halotolerant bacteria and in pathogenic *Burgholderia* species [46-48]. The Na^+-dependent ATP synthase is absent in *A. pasteurianus* and *G. diazotrophicus*, pointing to an acquisition of this operon by

G. oxydans via lateral gene transfer. Comparative DNA microarray analysis of oxygen limitation in *G. oxydans* revealed approximately 2-fold decreased expression of the H^+-dependent ATP synthase and 2- to 3-fold increased expression of the Na^+-dependent one [49]. This expression pattern suggests that Na^+ ions might play an important role as coupling ions under oxygen limitation and the H^+-dependent ATP synthase with the very short mRNA half-lives under regular conditions. Actually, one would expect that genes of the energy metabolism including ATP synthases are among the medium or most stable transcripts due to their classification as housekeeping genes [50]. Indeed, for *E. coli* the mRNA half-lives of genes associated to energy metabolism (mean 6.3 min) rank significantly in the top 3 with longest half-lives among all groups reported with mean values ranging from 3.8 min to 6.4 min [15]. In M9 medium the *E. coli* ATP synthase transcripts exhibited a mean half-life of 6.2 min which is twice as long compared to *G. oxydans* with 3.2 min. This 2-fold difference in time becomes much more impact when also considering the inverse difference of the related doubling times of *E. coli* (~60 to 70 min) and *G. oxydans* (~100 to 110 min). Therefore, the H^+-dependent ATP synthase and its very short mRNA half-lives in *G. oxydans* could be a bottleneck. Experimental results already revealed that the cytochrome bo_3 oxidase is a limiting factor, since plasmid-based overexpression of *cyoBACD* led to increased growth rates and growth yields, both in the wild type and its Δ*cyoBACD* mutant [45]. Both, the H^+-dependent ATP synthase and also the cytochrome bo_3 oxidase could be bottlenecks in *G. oxydans* at some point, for example in growth-improved strains [13, 14, 45].

The mRNA half-lives associated with the central carbon metabolism did not differ significantly from the global mean (5.8 min). This is attributed to partially high variation of half-lives within groups. Glycolysis (6.7 min) and EDP pathway (6.7 min) exhibited a somewhat longer mean half-life, whereas pyruvate metabolism (5.1 min) exhibited a somewhat shorter mean half-life. Nevertheless, for the glycolysis most of the genes have mRNA half-lives shorter than global mean, yet outliers increase the mean of the group. In *Sulfolobus* sp. transcripts of central metabolic pathways are generally also decayed rapidly [19], while there is also partially high variation of half-lives associated with the central carbon metabolism in *E. coli* [15]. In contrast, in *S. cerevisiae* transcripts encoding the enzymes that participate in the central metabolism are characteristically among those that live the longest [43]. In *G. oxydans*, the outliers are triosephosphate isomerase (GOX2217) with 12.4 min and dihydroxyacetone kinase (GOX2222) with 19.1 min. Interestingly, transcripts of the adjacent genes, which are related to glycerol and ribose metabolism, were very stable with half-lives ranging from 9.8 to 19 min. These genes encode a glycerol-3-phosphate dehydrogenase (GOX2215), a hypothetical protein (GOX2216) with a conserved aldolase class I superfamily domain, ribose-5-phosphate isomerase B (GOX2218) and components of a ribose ABC transporter (GOX2219-20). Besides GOX2215, GOX2088 also encodes a

glycerol-3-phosphate dehydrogenase. This gene forms a possible operon with a glycerol uptake facilitator protein (GOX2089) and a glycerol kinase (GOX2090), all most likely involved in glycerol uptake and catabolism. Therefore, GOX2215 may have another role, since it is located closely to genes associated to glycolysis, PPP, and ribose uptake. Analysis of the PPP by deletion of *gnd* using transcriptome analysis showed a significant upregulation of GOX2217-GOX2222 in comparison to the reference strain and lead to the assumption that this compensates the reduced activity of the PPP to still enable synthesis of nucleic acids and amino acids [32]. In our study, absolute expression values for these genes were relatively low. The high mRNA stability of genes belonging to this putative operon may indicate the need to generally increase the abundance of these transcripts or the encoded enzymes to account for a lack of building blocks. However, studies on the promoters and protein levels are required to assess the effect of the high mRNA stabilities.

Transcripts related to the pyruvate dehydrogenase and pyruvate decarboxylase exhibited relatively short half-lives (5.3 min), whereas the FPKM expression values belong to the top 20% of highly expressed genes. This resembles the inverse relationship between mRNA stability and transcript abundance to allow a fast decay of transcripts involved in energy metabolism during halt of growth. Half-lives of transcripts associated to the TCA cycle ranged from 5.3 to 9.3 min, yet most of them scattered around the average half-life (5.8 min). Only the aconitase transcript (GOX1335) is quite stable (9.3 min). FPKM expression values of the TCA cycle genes are the lowest in the central carbon metabolism. In view of the inverse relationship between mRNA stability and transcript abundance, the mRNA half-lives of TCA cycles genes appear to be rather short. The low expression might be a result of the incompleteness of the TCA cycle and according to carbon flux analysis its low usage [10]. Moreover, the relative low expression value of the anaplerotic genes *ppc* and *eda* might contribute to a strongly limited TCA cycle capacity already in the wildtype or possibly at least in growth-improved strains of *G. oxydans*.

Fragmentation of 23S rRNA in ribosomes was hitherto unrecognized in *Gluconobacter*, although this phenomenon is already well known in several bacteria including α-proteobacteria [33, 34]. In some cases, the rRNA processing steps lead to the removal of segments termed intervening sequences (IVS). Phenotypically, fragmentation of 23S rRNA appeared to be silent and processing of IVSs was not required for the production of functional ribosomes in *E. coli* [51]. In a first attempt, our Illumina sequencing and mapping analysis of the RNA fragments isolated from enriched ribosomes from *G. oxydans* revealed three possible fragmentation regions in all four 23S rRNAs and a fourth fragmentation position which was found only in the 23S rRNA transcript of GOX1159. In the γ-proteobacterium *S. typhimurium* fragmentation occurs by excision of IVS by RNase III [35, 51]. In many α-proteobacteria, for example *Rhizobiaceae*, *Brayrhizobiaceae*, and

Rhodobacteraceae, IVS processing at the 5´-end of 23S rRNAs leads to a ~130 nt and a ~2.6 nt fragment [34, 36, 37]. In *G. oxydans*, we did not find known IVS at this position or elsewhere according to the mapping analysis and sequence comparisons with other 23S rRNAs. Also, we did not observe fragments at approximately 130 nt. However, *G. oxydans* exhibits a potential fragmentation region at nt 1,456-1,465 which is within the same fragmentation region of 23S rRNA from *A. radiobacter* and *R. leguminosarum*. The other three fragmentation regions from *G. oxydans* are not present in 23S rRNAs from other α-proteobacteria. One of these is found at nt 1,139-1,160 within a region without conservation, where also one IVS is present in the 23S rRNAs of *S. typhimurium* and *R. sphaeroides*. These regions exhibit no or only partial similarity to the region of *G. oxydans*. The third fragmentation region at nt 1,716-1,770 is in a highly conserved region according to the alignment, yet a fragmentation in other bacteria was not reported yet. Although highly conserved, single nucleotide differences can enable or disable cleavage by RNases, as can be also seen by the fourth fragmentation position only found in GOX1159. This region contains two specific single nucleotide variations which could be responsible for recognition and cleavage by RNase. Since these variations are located far within the fragmentation region and not in the flanking regions, the results suggest further rRNA processing of the fragments after cleavage. Secondary maturation pathways, which enable further processing of 23S rRNA 5'- and 3'- ends after RNase III cleavage in an early step, was already suggested for α-proteobacteria [34].

Conclusion

In conclusion, the mRNA decay data from *G. oxydans* showed many similarities to the results obtained in other bacteria, yet also exhibited some specific differences. Overall, the inverse relationship of transcript abundance and mRNA stability also supports the view that mRNA stability does not play a major role to obtain high mRNA abundance. Rather, a quick mRNA turnover for fast adaptation of the cell to environmental changes is made possible by the shorter half-lives of highly expressed genes. The significantly longer half-lives of transcripts involved in sugar and alcohol degradation might reflect that in *G. oxydans*' typical habitats a broader range of substrates may be sufficiently available for longer periods and fast adaptation to the absence of substrates is less important. In this view, the longer mRNA half-lives could support synthesis of the encoded enzymes contributing to *G. oxydans*' exceptional ability to incompletely oxidize a broad range of sugars and alcohols in the periplasm. In contrast, the very short-lived H^+-dependent ATP synthase transcripts could be a bottleneck in *G. oxydans* at some point, together with the relatively low expression of TCA cycle genes and anaplerotic genes, which exhibited the lowest values in the central carbon

metabolism. Expression of ATP synthase, TCA cycle and anaplerotic genes should be considered in future metabolic engineering approaches to further improve growth and biomass yield of *G. oxydans*. In this context, the consequences of 23S rRNA fragmentation on growth and fitness of *G. oxydans* are unknown. Our Illumina sequencing and mapping analysis of the RNA fragments isolated from enriched ribosomes revealed three new possible fragmentation regions which were not already known from 23S rRNA fragmentation in other bacteria. Although fragmentation of 23S rRNA appeared to be phenotypically silent, it could be relevant under some conditions or for some aspects not resulting in an obvious growth phenotype. Further studies are needed to unravel the multistep process of fragmentation and exact cleavage sites in the 23S rRNAs from *G. oxydans*.

Additional files

Additional file 1: Supplement Figure S1 and Figure S2; group histograms mRNA half-lives (S1) and 23S rRNA sequence alignment (S2).

Additional file 2: Table S1; mRNA decay data.

Additional file 3: Table S2; FPKM expression values and mRNA half-lives.

Additional file 4: Table S3; Proteins identified in chromatographic elution fractions.

Additional file 5: Table S4-S11; rRNA mapping coverage in exponential phase.

Additional file 6: Table S12-S19; rRNA mapping coverage in early stationary phase.

Acknowledgements

We thank Alexander Vogel for deposition of the RNAseq data in the ENA archive and at the *Gluconobacter* portal www.gluconobacterfactory.de.

Availability of data and materials

The microarray data are accessible in NCBI's Gene Expression Omnibus through accession number GSE103428. The RNAseq data are publicly available in the European Nucleotide Archive under accession number PRJEB18739.

Funding

The study was supported by the scientific activities of the Bioeconomy Science Center which were financially supported by the Ministry of Innovation, Science and Research within the framework of the NRW Strategy project BioSC (No. 313/323-400-002 13).

Competing interests

The authors declare that they have no competing interests.

References

1. Ameyama M, Shinagawa E, Matsushita K, Adachi O: D-fructose dehydrogenase of *Gluconobacter industrius*: purification, characterization, and application to enzymatic microdetermination of D-fructose. J Bacteriol. 1981, 145(2):814-823.
2. Gupta A, Singh VK, Qazi GN, Kumar A: *Gluconobacter oxydans*: its biotechnological applications. J Mol Microb Biotech. 2001, 3(3):445-456.
3. Hekmat D, Bauer R, Fricke J: Optimization of the microbial synthesis of dihydroxyacetone from glycerol with *Gluconobacter oxydans*. Bioproc Biosyst Eng. 2003, 26(2):109-116.
4. Saito Y, Ishii Y, Hayashi H, Imao Y, Akashi T, Yoshikawa K, Noguchi Y, Soeda S, Yoshida M, Niwa M *et al*: Cloning of genes coding for L-sorbose and L-sorbosone dehydrogenases from *Gluconobacter oxydans* and microbial production of 2-keto-L-gulonate, a precursor of L-ascorbic acid, in a recombinant *G. oxydans* strain. Appl Environ Microb. 1997, 63(2):454-460.
5. Tkac J, Navratil M, Sturdik E, Gemeiner P: Monitoring of dihydroxyacetone production during oxidation of glycerol by immobilized *Gluconobacter oxydans* cells with an enzyme biosensor. Enzyme Microb Tech. 2001, 28(4-5):383-388.
6. Wang EX, Ding MZ, Ma Q, Dong XT, Yuan YJ: Reorganization of a synthetic microbial consortium for one-step vitamin C fermentation. Microb Cell Fact. 2016, 15:21.
7. Mamlouk D, Gullo M: Acetic Acid bacteria: physiology and carbon sources oxidation. Indian J Microbiol. 2013, 53(4):377-384.
8. Mientus M, Kostner D, Peters B, Liebl W, Ehrenreich A: Characterization of membrane-bound dehydrogenases of *Gluconobacter oxydans* 621H using a new system for their functional expression. Appl Microbiol Biotechnol. 2017, 101(8):3189-3200.
9. Pappenberger G, Hohmann HP: Industrial production of L-ascorbic Acid (vitamin C) and D-isoascorbic acid. Adv Biochem Eng Biotechnol. 2014, 143:143-188.
10. Hanke T, Nöh K, Noack S, Polen T, Bringer S, Sahm H, Wiechert W, Bott M: Combined fluxomics and transcriptomics analysis of glucose catabolism via a partially cyclic pentose phosphate pathway in *Gluconobacter oxydans* 621H. Appl Environ Microb. 2013, 79(7):2336-2348.
11. Prust C, Hoffmeister M, Liesegang H, Wiezer A, Fricke WF, Ehrenreich A, Gottschalk G, Deppenmeier U: Complete genome sequence of the acetic acid bacterium *Gluconobacter oxydans*. Nat Biotechnol. 2005, 23(2):195-200.
12. Kiefler I, Bringer S, Bott M: SdhE-dependent formation of a functional *Acetobacter pasteurianus* succinate dehydrogenase in *Gluconobacter oxydans*--a first step toward a complete tricarboxylic acid cycle. Appl Microbiol Biot. 2015, 99(21):9147-9160.
13. Kiefler I, Bringer S, Bott M: Metabolic engineering of *Gluconobacter oxydans* 621H for increased biomass yield. Appl Microbiol Biotechnol. 2017.
14. Kostner D, Luchterhand B, Junker A, Volland S, Daniel R, Büchs J, Liebl W, Ehrenreich A: The consequence of an additional NADH dehydrogenase paralog on the growth of *Gluconobacter oxydans* DSM3504. Appl Microbiol Biotechnol. 2015, 99(1):375-386.
15. Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN: Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. P Natl Acad Sci USA. 2002, 99(15):9697-9702.
16. Hambraeus G, von Wachenfeldt C, Hederstedt L: Genome-wide survey of mRNA half-lives in *Bacillus subtilis* identifies extremely stable mRNAs. Mol Genet Genomics. 2003, 269(5):706-714.

17. Rustad TR, Minch KJ, Brabant W, Winkler JK, Reiss DJ, Baliga NS, Sherman DR: Global analysis of mRNA stability in *Mycobacterium tuberculosis*. Nucleic Acids Res. 2013, 41(1):509-517.
18. Lin PH, Singh D, Bernstein JA, Lin-Chao S: Genomic analysis of mRNA decay in *E. coli* with DNA microarrays. Method Enzymol. 2008, 447:47-64.
19. Andersson AF, Lundgren M, Eriksson S, Rosenlund M, Bernander R, Nilsson P: Global analysis of mRNA stability in the archaeon *Sulfolobus*. Genome Biol. 2006, 7(10):R99.
20. Mohanty BK, Kushner SR: Regulation of mRNA Decay in Bacteria. Annu Rev Microbiol. 2016, 70:25-44.
21. Morey JS, Van Dolah FM: Global analysis of mRNA half-lives and de novo transcription in a dinoflagellate, *Karenia brevis*. PLoS One. 2013, 8(6):e66347.
22. Dressaire C, Picard F, Redon E, Loubiere P, Queinnec I, Girbal L, Cocaign-Bousquet M: Role of mRNA stability during bacterial adaptation. PLoS One. 2013, 8(3):e59059.
23. Takayama K, Kjelleberg S: The role of RNA stability during bacterial stress responses and starvation. Environ Microbiol. 2000, 2(4):355-365.
24. Lalaouna D, Simoneau-Roy M, Lafontaine D, Masse E: Regulatory RNAs and target mRNA decay in prokaryotes. Biochim Biophys Acta. 2013, 1829(6-7):742-747.
25. Polen T, Rittmann D, Wendisch VF, Sahm H: DNA microarray analyses of the long-term adaptive response of *Escherichia coli* to acetate and propionate. Appl Environ Microbiol. 2003, 69(3):1759-1774.
26. Sambrock J, Fritsch EF, Maniatis T: Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.; 1989.
27. Polen T, Wendisch VF: Genomewide expression analysis in amino acid-producing bacteria using DNA microarrays. Appl Biochem Biotechnol. 2004, 118(1-3):215-232.
28. Bland JM, Altman DG: Multiple significance tests: the Bonferroni method. BMJ. 1995, 310(6973):170.
29. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010, 28(5):511-515.
30. Trauner A, Bennett MH, Williams HD: Isolation of bacterial ribosomes with monolith chromatography. PLoS One. 2011, 6(2):e16273.
31. Belasco JG, Nilsson G, von Gabain A, Cohen SN: The stability of *E. coli* gene transcripts is dependent on determinants localized to specific mRNA segments. Cell. 1986, 46(2):245-251.
32. Richhardt J, Bringer S, Bott M: Mutational analysis of the pentose phosphate and Entner-Doudoroff pathways in *Gluconobacter oxydans* reveals improved growth of a Δ*edd* Δ*eda* mutant on mannitol. Appl Environ Microb. 2012, 78(19):6975-6986.
33. Evguenieva-Hackenberg E: Bacterial ribosomal RNA in pieces. Mol Microbiol. 2005, 57(2):318-325.
34. Zahn K, Inui M, Yukawa H: Divergent mechanisms of 5' 23S rRNA IVS processing in the α-proteobacteria. Nucleic Acids Res. 2000, 28(23):4623-4633.
35. Burgin AB, Parodos K, Lane DJ, Pace NR: The excision of intervening sequences from *Salmonella* 23S ribosomal RNA. Cell. 1990, 60(3):405-414.
36. Zahn K, Inui M, Yukawa H: Characterization of a separate small domain derived from the 5' end of 23S rRNA of an α-proteobacterium. Nucleic Acids Res. 1999, 27(21):4241-4250.
37. Selenska-Pobell S, Evguenieva-Hackenberg E: Fragmentations of the large-subunit rRNA in the family *Rhizobiaceae*. J Bacteriol. 1995, 177(23):6993-6998.
38. Evguenieva-Hackenberg E, Klug G: RNase III processing of intervening sequences found in helix 9 of 23S rRNA in the alpha subclass of Proteobacteria. J Bacteriol. 2000, 182(17):4719-4729.

39. Fan J, Yang X, Wang W, Wood WH, 3rd, Becker KG, Gorospe M: Global analysis of stress-regulated mRNA turnover by using cDNA arrays. Proc Natl Acad Sci U S A. 2002, 99(16):10611-10616.
40. McAdams HH, Arkin A: Stochastic mechanisms in gene expression. Proc Natl Acad Sci U S A. 1997, 94(3):814-819.
41. Ishii N: GroEL and the GroEL-GroES Complex. Subcell Biochem. 2017, 83:483-504.
42. Selinger DW, Saxena RM, Cheung KJ, Church GM, Rosenow C: Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. Genome Res. 2003, 13(2):216-223.
43. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO: Precision and functional specificity in mRNA decay. P Natl Acad Sci USA. 2002, 99(9):5860-5865.
44. Yao Z, Davis RM, Kishony R, Kahne D, Ruiz N: Regulation of cell size in response to nutrient availability by fatty acid biosynthesis in *Escherichia coli*. Proc Natl Acad Sci U S A. 2012, 109(38):E2561-2568.
45. Richhardt J, Luchterhand B, Bringer S, Buchs J, Bott M: Evidence for a key role of cytochrome bo_3 oxidase in respiratory energy metabolism of *Gluconobacter oxydans*. J Bacteriol. 2013, 195(18):4210-4220.
46. Dibrova DV, Galperin MY, Mulkidjanian AY: Characterization of the N-ATPase, a distinct, laterally transferred Na^+-translocating form of the bacterial F-type membrane ATPase. Bioinformatics. 2010, 26(12):1473-1476.
47. Müller V, Grüber G: ATP synthases: structure, function and evolution of unique energy converters. Cell Mol Life Sci. 2003, 60(3):474-494.
48. Saum R, Schlegel K, Meyer B, Müller V: The F_1F_o ATP synthase genes in *Methanosarcina acetivorans* are dispensable for growth and ATP synthesis. FEMS Microbiol Lett. 2009, 300(2):230-236.
49. Hanke T, Richhardt J, Polen T, Sahm H, Bringer S, Bott M: Influence of oxygen limitation, absence of the cytochrome bc(1) complex and low pH on global gene expression in *Gluconobacter oxydans* 621H using DNA microarray technology. J Biotechnol. 2012, 157(3):359-372.
50. Martens M, Dawyndt P, Coopman R, Gillis M, De Vos P, Willems A: Advantages of multilocus sequence analysis for taxonomic studies: a case study using 10 housekeeping genes in the genus *Ensife*r (including former *Sinorhizobium*). Int J Syst Evol Microbiol. 2008, 58(Pt 1):200-214.
51. Gregory ST, O'Connor M, Dahlberg AE: Functional *Escherichia coli* 23S rRNAs containing processed and unprocessed intervening sequences from *Salmonella typhimurium*. Nucleic Acids Res. 1996, 24(24):4918-4923.

TABLES

Table 1 Mean mRNA half-lives of functional categories according to gene product functions.

Values were obtained using the half-life data from 1,193 genes for which R^2 was >0.7 when using all 4 time points of the analysis. Values given in parenthesis were obtained when data from 1,307 further genes were included for which R^2 was >0.7 when using the first 3 time points (2, 5, and 10 min) and omitting the last one (15 min).

Functional category/ Pathway [1]	Number of genes assigned[1]	with a calculated mRNA half-life	proportion %	half-life min
ATP-proton motive force interconversion	17	8 (17)	47.1 (100)	3.1** (4)
DNA degradation	9	3 (9)	33.3 (100)	4.4 (4.7)
Transcription	13	7 (13)	53.8 (100)	4.4* (29.1)
Mono/dioxygenase	10	5 (10)	50.0 (100)	4.9 (4.7)
Fatty acid and phospholipid metabolism	36	26 (35)	72.2 (97.2)	4.9* (5.9)
Biosynthesis and degradation of polymers	11	4 (11)	36.4 (100)	4.9 (6.1)
RNA metabolism	23	12 (23)	52.2 (100)	4.9 (4.4)
Pyruvate metabolism	12	9 (11)	75.0 (91.7)	5.1 (5.1)
Nucleotide metabolism	59	37 (59)	62.7 (100)	5.0* (6.0)
Degradation of proteins and peptides	51	30 (51)	58.8 (100)	5.1* (5.1)
Signal transduction	30	11 (30)	36.7 (100)	5.1 (4.8)
Ribosome assembly	79	45 (67)	57.0 (84.8)	5.1 (6.8)
tRNA metabolism	69	8 (14)	11.6 (20.3)	5.1 (4.8)
DNA replication	26	19 (26)	73.1 (100)	5.2 (5.1)
Antibiotics resistance	11	6 (11)	54.5 (100)	5.2 (4.8)
Cell division	27	17 (27)	63.0 (100)	5.2 (4.6)
Amino acid metabolism	120	79 (120)	65.8 (100)	5.2 (5.0)
Central intermediary metabolism	33	17 (33)	51.5 (100)	5.2 (4.9)
Cell envelop	134	75 (133)	56.0 (99.3)	5.4 (4.9)
Regulatory functions	98	38 (96)	38.8 (98)	5.4 (5.1)
Aminoacyl-tRNA biosynthesis	33	27 (33)	81.8 (100)	5.5 (5.3)
Translation factors	20	16 (20)	80.0 (100)	5.6 (5.4)
Unknown function	41	25 (41)	61.0 (100)	5.6 (5.4)
Pentose phosphate pathway (PPP)	13	10 (13)	76.9 (100)	5.6 (6.7)
Electron transport	53	37 (53)	69.8 (100)	5.8 (5.5)
Biosynthesis of cofactors	85	45 (85)	52.9 (100)	5.8 (5.4)
Transport	232	96 (232)	41.4 (100)	5.9 (5.2)
DNA repair	34	7 (34)	20.6 (100)	5.9 (5.2)
DNA recombination	18	8 (18)	44.4 (100)	6.1 (5.4)
Tricarboxylic acid cycle (TCA cycle)	8	6 (8)	75.0 (100)	6.4 (5.8)
Uncharacterized oxidoreductase	66	28 (66)	42.4 (100)	6.5 (5.6)
Protein folding and stabilization	28	26 (28)	92.9 (100)	6.5 (6.3)
Adaptations to atypical conditions	19	12 (19)	63.2 (100)	6.5 (6.1)
Detoxification	30	13 (28)	43.3 (93.)	6.6 (5.4)
Entner-Doudoroff pathway (EDP)	2	2	100	6.7
Glycolysis	17	13 (17)	76.5 (100)	6.7 (6.3)
Sugar and alcohol degradation	25	14 (24)	56.0 (96)	7.3* (6)
DNA restriction and modification	8	6 (8)	75.0 (100)	8.1* (7)
Cell motility	41	14 (41)	34.1 (100)	8.7** (6.4)
Ion homeostasis	6	2 (6)	33.3 (100)	8.9** (6.7)

[1)] Functional categories are based on assigned gene product functions [11].
[2)] Functional categories with significantly shorter or longer mean half-lives than the overall mean half-life are highlighted by ** ($p < 0.001$) or * ($p < 0.05$).

Table 2 Amount of protein and RNA obtained from the four peaks of the chromatography to enrich ribosomes.

G. oxydans cells were collected in the exponential phase (exp.) and in the early stationary phase (stat.). Values for protein and RNA are given in µg.

phase	protein	peak	protein	RNA
exp.	1,030	P1	267	5.3
		P2	519	1.3
		P3	160	25.8
		P4	6	6.6
stat.	1,475	P1	457	0.2
		P2	678	4.9
		P3	130	37.3
		P4	14	1.6

FIGURE LEGENDS

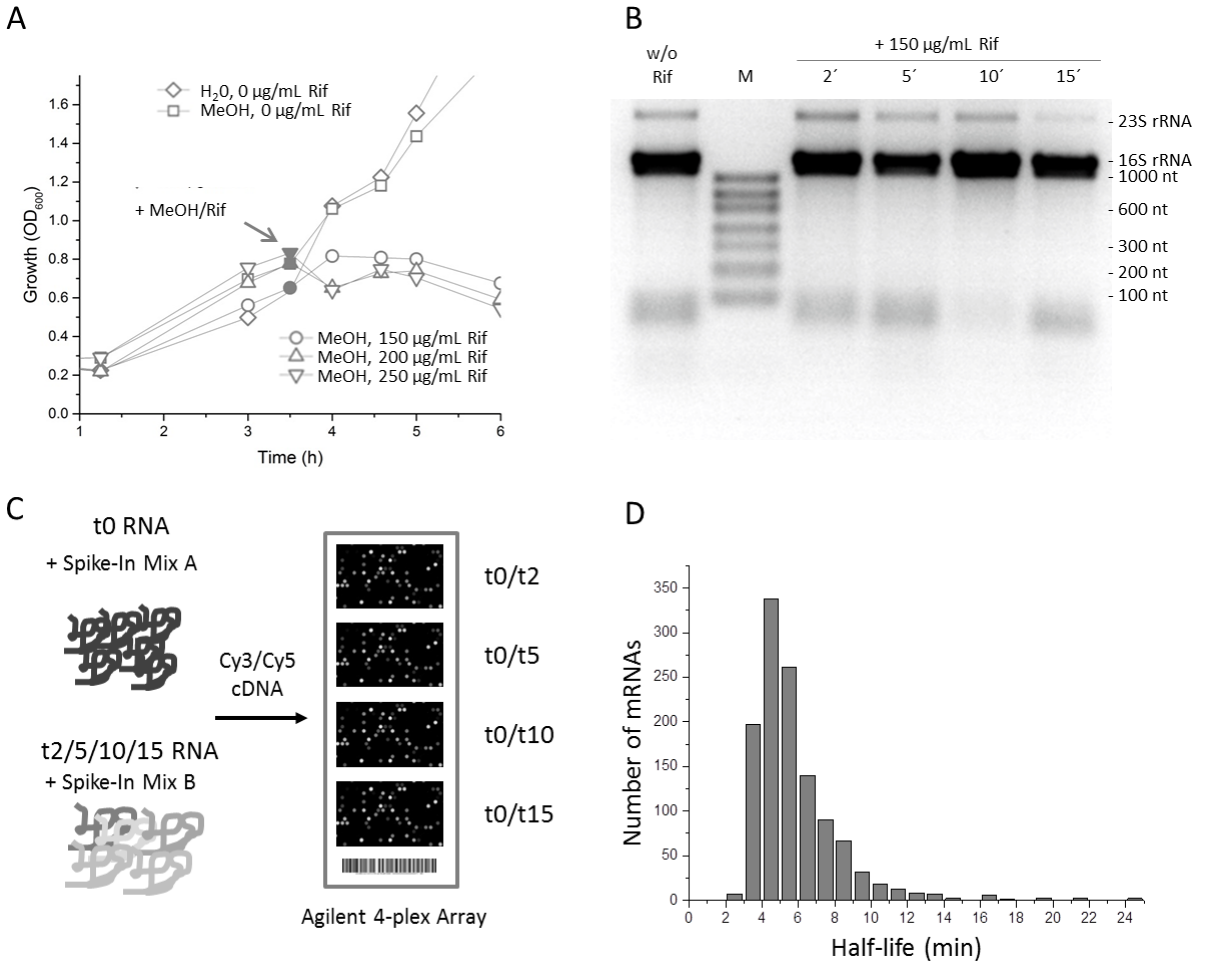

Figure 1 Overview of global mRNA decay analysis.

A) Growth of *G. oxydans* 621H in the absence and presence of different concentrations of rifampicin.

B) Formaldehyde agarose gel (1.8%) analysis to inspect quality of total RNA isolated from *G. oxydans* cells before and 2, 5, 10, as well as 15 min after addition of 150 µg/mL of rifampicin. Aliquots of 1 µg of total RNA were loaded. Bands corresponding to the 23S rRNA (2,709 to 2,711 nt) and 16S rRNA (1478 nt) are indicated. M = RiboRuler Low Range RNA Ladder.

C) Schema of the experimental design. RNA isolated at the given time points were mixed with Spike-In RNA Mix A or B and used for cDNA synthesis and Cy3 or Cy5 labeling. cDNA synthesized from RNA isolated from cells before addition of rifampicin (t0) and 2, 5, 10 and 15 min after addition of 150 µg/mL rifampicin were pairwise mixed and hybridized on Agilent 4-plex arrays. mRNA half-lives were calculated based on three biological replicates including dye-swap labeling.

D) Histogram showing the distribution of calculated mRNA half-lives of 1,193 (44%) genes from *G. oxydans* where R^2 was >0.7.

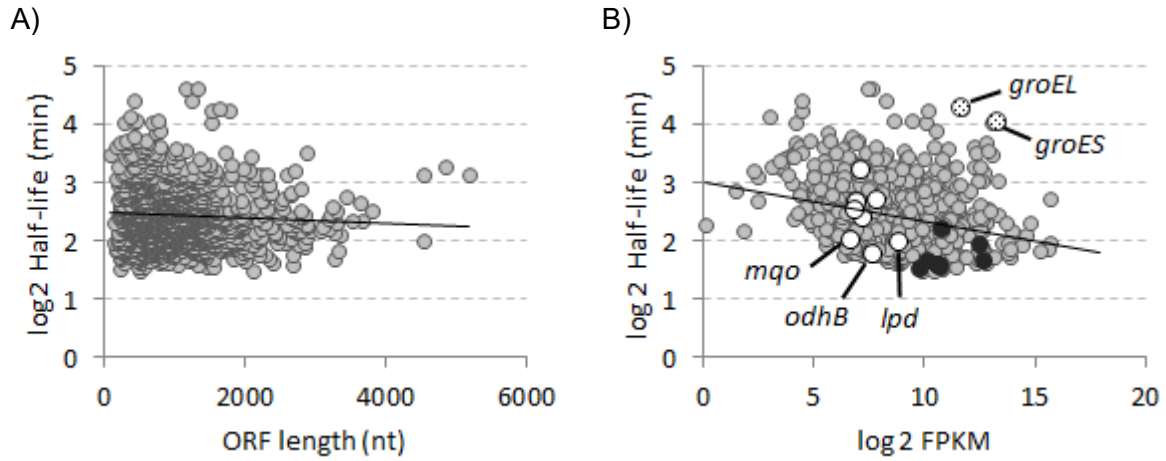

Figure 2 Correlation of mRNA half- lives with ORF length and with FPKM expression values.

(A) Scatterplot comparing mRNA half-lives and ORF lengths (R -0.06).

(B) Scatterplot comparing mRNA half-lives and FPKM values. Linear regression analysis showed a statistically significant ($p < 0.0001$) negative correlation ($R = -0.26$) between abundance of transcripts and stability. Plots are related to the half-life data obtained for 1,193 transcripts (●) with $R^2 > 0.7$ based on the 4 time points of the analysis. H^+-dependent F_1F_o-type ATP synthase genes (●) *atpBEFF*' (GOX1110-13) and *atpHAGDC* (GOX1310-14) belong to the genes with the shortest mRNA half-lives in *G. oxydans*. Transcripts of TCA cycle genes (○) exhibited low expression values and short mRNA half-lives, in particular *mqo* (GOX2070) encoding malate:quinone oxidoreductase, *odhB* (GOX1073) encoding dihydrolipoamide succinyl transferase (E2) of 2-oxoglutarate dehydrogenase, and *lpd* (GOX2292) encoding dihydrolipoamide dehydrogenase. Transcripts of the molecular chaperones GroES (GOX1901) and GroEL (GOX1901) exhibited high FPKM values as well as long half-lives (⊙).

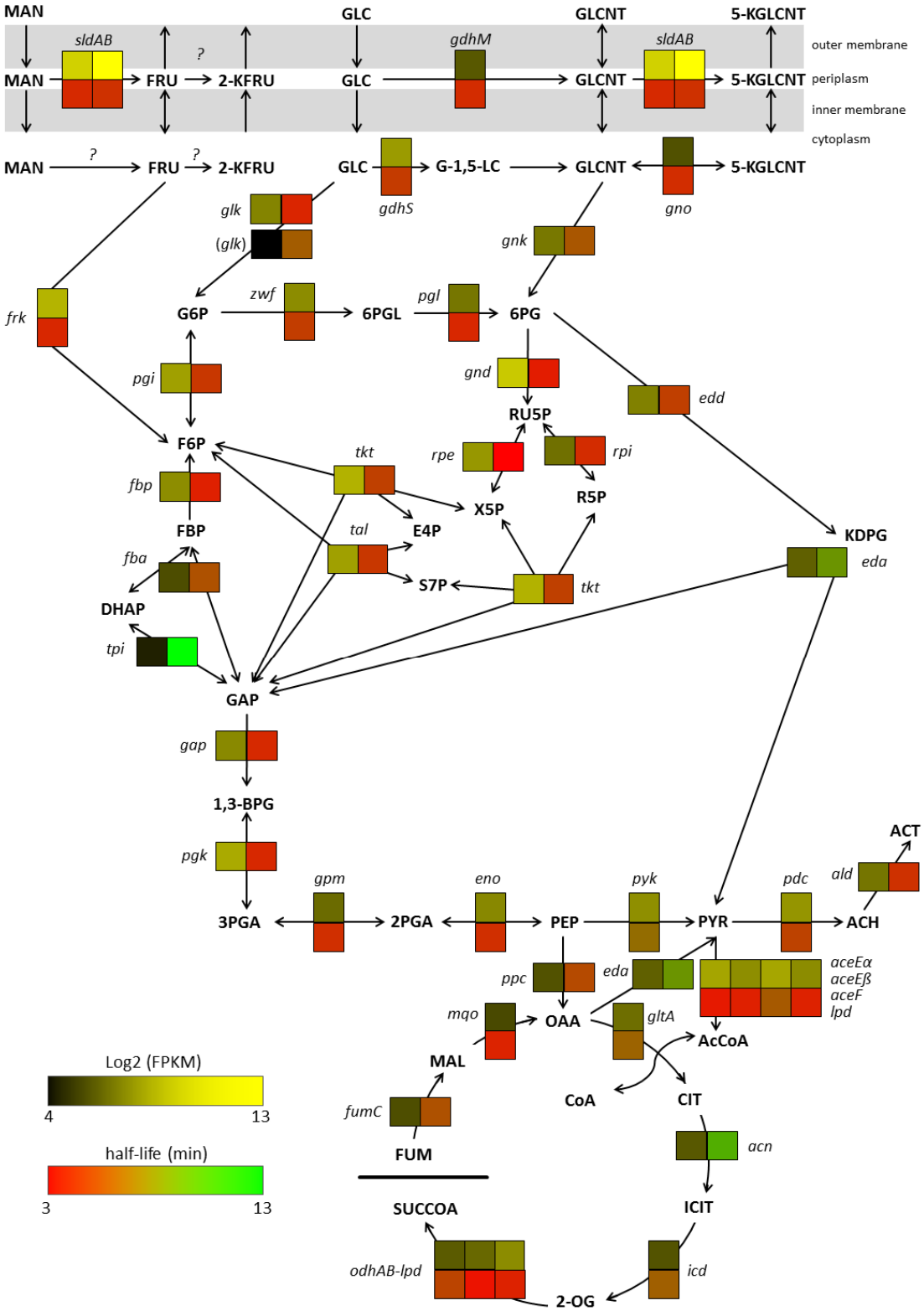

Figure 3 mRNA half-lives and FPKM expression values for genes of the central carbon metabolism.

Boxes representing a half-life in minutes were colored according to the red-green gradient ranging from 3 to 13 min based on the results obtained with all 4 time points of the analysis, except for *glk* (GOX1182), *gno*, *mqo*, *odhB*, and *rpe* where R^2 was only >0.7 using the first three time points (Table S1).

Genes/Enzymes: *aceEα*, pyruvate dehydrogenase E1 component alpha subunit (GOX2289); *aceEß*, pyruvate dehydrogenase E1 component beta subunit (GOX2290); *aceF*, dihydrolipoamide acetyltransferase component of pyruvate dehydrog enase (GOX2291); *acn*, aconitate hydratase (GOX1335); *ald*, aldehyde dehydrogenase (GOX2018); *eda*, KDPG aldolase (GOX0430); *edd*, phosphogluconate dehydratase (GOX0431); *eno*, phosphopyruvate hydratase (GOX2279); *fba*, fructose-bisphosphate aldolase (GOX0780); *fbp*, fructose 1,6-bisphosphatase (GOX1516); *frk*, fructokinase (GOX0284); *fumC*, fumarate hydratase (GOX1643); *gap*, glyceraldehyde 3-phosphate dehydrogenase (GOX0508); *gdhM*, membrane-bound glucose dehydrogenase (GOX0265); *gdhS*, soluble glucose dehydrogenase (GOX2015); *glk*, glucokinase (GOX2419); (*glk*), putative glucokinase (GOX1182); *gltA*, citrate synthase (GOX1999); *gnd*, 6-phosphogluconate dehydrogenase-like protein (GOX1705); *gnk*, gluconokinase (GOX1709); *gno*, gluconate 5-dehydrogenase (GOX2187); *gpm* phosphoglyceromutase (GOX0330); *icd*, isocitrate dehydrogenase (GOX1336); *lpd*, dihydrolipoamide dehydrogenase (GOX2292); *mqo*, malate:quinone oxidoreductase (GOX2070); *odhA*, 2-oxoglutarate dehydrogenase E1 component (GOX0882); *odhB*, dihydrolipoamide succinyltransferase E2 component of 2-oxoglutarate dehydrogenase complex (GOX1073); *pdc*, pyruvate decarboxylase (GOX1081); *pgk*, phosphoglycerate kinase (GOX0507); *pgl*, 6-phosphogluconolactonase (GOX1707); *ppc*, putative phosphoenolpyruvate carboxylase (GOX0102); *pyk*, pyruvate kinase (GOX2250); *rpe*, ribulose-phosphate 3-epimerase (GOX1352); *rpi*, ribose 5-phosphate isomerase (GOX1708); *sldA*, D-sorbitol dehydrogenase subunit SldA (GOX0854); *sldB*, D-sorbitol dehydrogenase subunit SldB (GOX0855); *tal*/*pgi*, bifunctional transaldolase (GOX1704); *tkt*, transketolase (GOX1703); *tpi*, triosephosphate isomerase (GOX2217); *zwf*, glucose-6-phosphate 1-dehydrogenase (GOX0145).

Metabolites: 1,3-BPG, 1,3-Bisphosphoglycerate; 2-KFRU, 2-Ketofructose; 2PGA, 2-OG, 2-Oxogularate; 2-Phosphogylcerate; 3PGA, 3-Phosphoglycerate; 5-KGLCNT, 5-Ketogluconate; 6PG, 6-Phosphogluconate; 6PGL, 6-Phosphogluconolactone; ACH, Acetaldehyde; ACT, Acetate; AcCoA, Acetyl coenzyme A; CIT, Citrate; CoA, Coenzyme A; DHAP, Dihydroxyacetone phosphate; E4P, Erythrose 4-phosphate; F6P, Fructose 6-phosphate; FBP, Fructose 1,6-bisphosphate; FRU, Fructose; FUM, Fumarate; G-1,5-LC, Glucono-1,5-lactone; G6P, Glucose 6-phosphate; GAP, Glyceraldehyde 3-phosphate; GLC, Glucose; GLCNT, Gluconate; ICIT, Isocitrate; KDPG, 2-keto-3-deoxy-6-phosphogluconate; MAL, Malate; MAN, Mannitol; OAA, Oxaloacetate; PEP, Phosphoenolpyruvate; PYR, Pyruvate; R5P, Ribose 5-phosphate; RU5P, Ribulose 5-phosphate; S7P, Sedoheptulose 7-phosphate; SUCCoA, Succinyl coenzyme A; X5P, Xylulose 5-phosphate.

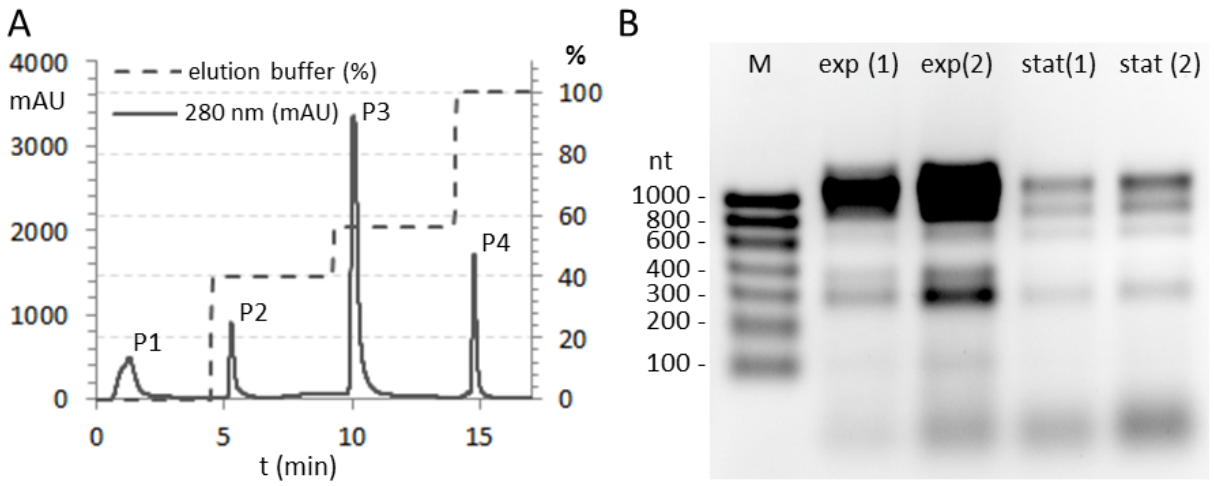

Figure 4 Chromatogram of the enrichment of ribosomes (A) and formaldehyde agarose gel analysis of RNA obtained from ribosomes (B).

A) Aliquots of *G. oxydans* cell extracts were loaded onto the column and ribosomes were isolated following a stepwise elution (% elution buffer). During elution the online chromatogram was manually inspected for upcoming peaks according to the absorbance at 280 nm (mAU) to manually collect from the start to the end of a peak into one elution fraction. The four peaks indicating the elution of protein fractions are labeled by P1, P2, P3, and P4.

B) Visualized RNA samples were isolated from peak P3 (A) containing enriched ribosomes from exponential phase (exp) or from early stationary phase (stat) each in two independent biological replicates (1, 2). M: RiboRuler low range ladder. Gel loads were 1.5 µg of RNA for exp (1), stat (1) as well as stat (2) and 3 µg of RNA for exp (2).

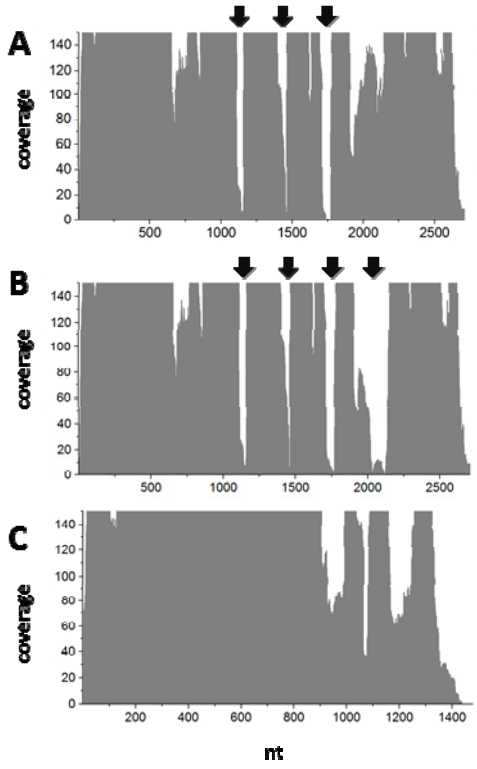

Figure 5 Graphical representation of the mapping coverage for 23S (A, B) and 16S rRNA (C).

The mapping is based on the Illumina reads obtained with RNA samples isolated from enriched ribosomes (peak P3) from cells at the exponential phase. The y axes are zoomed to a maximum of 150 to better illustrate the low coverage regions. Arrows indicate regions with a coverage <5% of the average coverage for the gene.

(A) Coverage of the 23S rRNA locus GOX1319 (2,711 nt). The same coverage pattern was also obtained for two other 23S rRNA loci, namely GOX0221 and GOX1467 (2,710 nt).

B) The mapping coverage of the 23S rRNA locus GOX1159 (2,709 nt) suggested an additional fragmentation site.

(C) 16S rRNA locus GOX1156. The same coverage pattern was observed for the other 16S rRNA loci, namely GOX0224, GOX1316, and GOX1464, respectively.

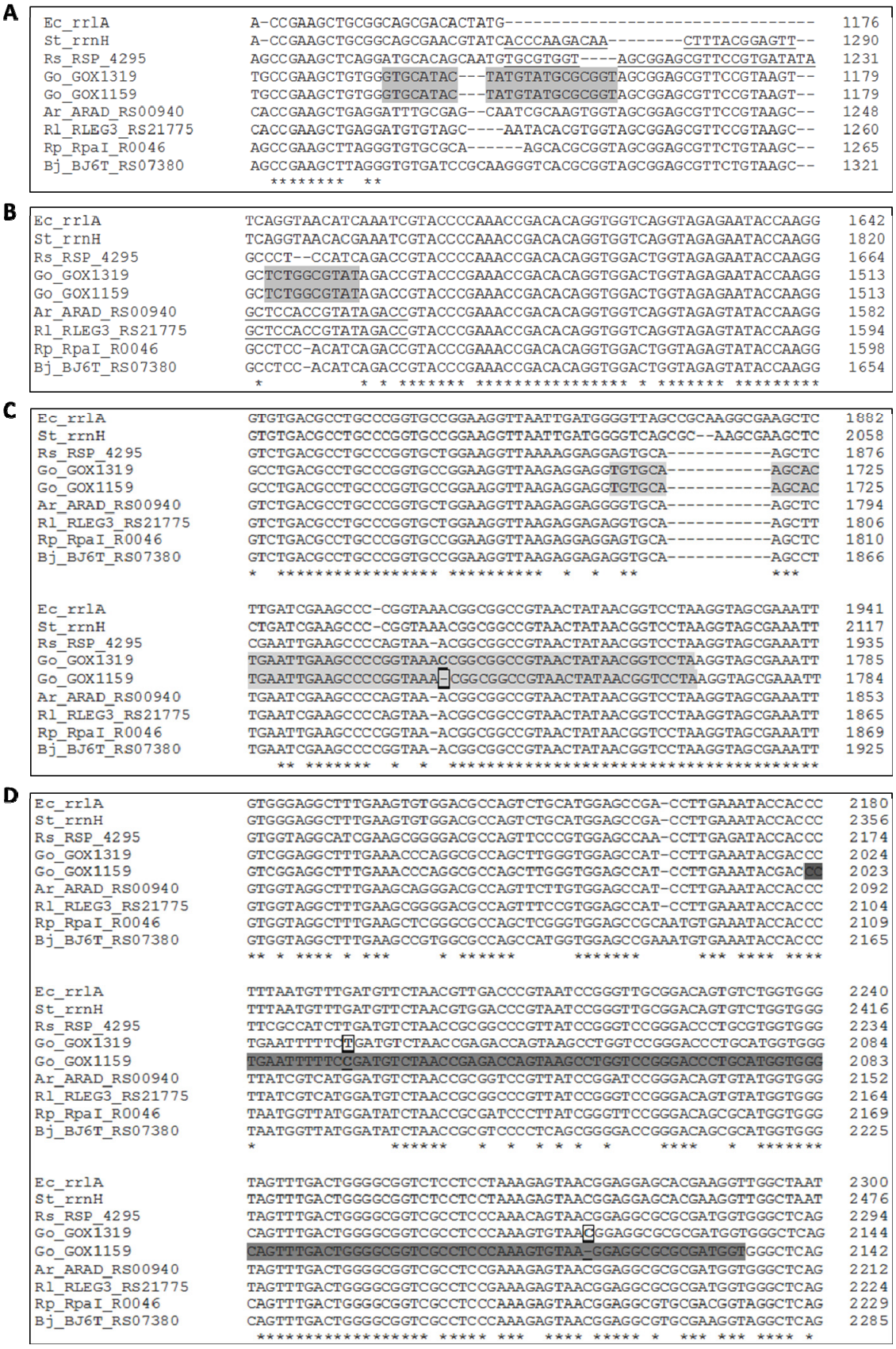

Figure 6 Sequence alignment of selected regions of 23S rRNA genes.

Sequences used are from *E. coli* MG1655 (Ec_rrlA), *Salmonella typhimurium* LT2 (St_rrnH), *Rhodobacter sphaeroides* 2.4.1 (Rs_RSP_4295), *G. oxydans* 621H (Go_GOX1319 and

Go_GOX1159), *Agrobacterium radiobacter* K84 (Ar_ARAD_RS00940), *Rhizobium leguminosarum bv. trifolii* WSM1689, *Rhodopseudomonas palustris* TIE-1 (Rp_rpaI_R0046), and *Bradyrhizobium japonicum* USDA 6 (Bj_BJ6T_RS07380). The alignments shows the regions of the 23S rRNA genes of *G. oxydans* which exhibited read coverage of <5% of the averaged gene coverage (Figure 5). Fragmentation positions in rRNAs known or assumed for other bacteria are underlined according to the literature.

(A) The first region at nucleotides 1,139-1,160 which is identical in all for all four *G. oxydans* gene copies.

(B) The second region found at the nucleotides 1,456- 1,465 for all four gene copies (Figure 5A and 5B).

(C) The third region found at the positions 1,716-1,771 for all four gene copies (Figure 5A and 5B).

(D) The fourth region found at the positions 2,022- 2,134 only for GOX1159 (Figure 5B). For A-C, the region is highlighted in light grey, for D it is shown in darker grey. Differences between the rRNA sequences GOX1319 and GOX1159 are underlined and bolded. Boxes at these positions show the base identity in GOX0221 and GOX1467, which are identical. The complete alignment can be found in the supplementary data (Figure S2).

3. References

Albersmeier, A., Pfeifer-Sancar, K., Rückert, C., Kalinowski, J. (2017) Genome-wide determination of transcription start sites reveals new insights into promoter structures in the actinomycete *Corynebacterium glutamicum*. *Journal of Biotechnology* 257: 99-109.

Andersson, A.F., Lundgren, M., Eriksson, S., Rosenlund, M., Bernander, R., Nilsson, P. (2006) Global analysis of mRNA stability in the archaeon *Sulfolobus*. *Genome Biology* 7: R99.

Apirion, D., Lassar, A.B. (1978) A conditional lethal mutant of *Escherichia coli* which affects the processing of ribosomal RNA. *The Journal of biological Chemistry* 253: 1738-1742.

Bae, B., Feklistov, A., Lass-Napiorkowska, A., Landick, R., Darst, S.A. (2015) Structure of a bacterial RNA polymerase holoenzyme open promoter complex. *eLife* 4.

Bardwell, J.C., Régnier, P., Chen, S.M., Nakamura, Y., Grunberg-Manago, M., Court, D.L. (1989) Autoregulation of RNase III operon by mRNA processing. *The EMBO Journal* 8: 3401-3407.

Bernstein, J.A., Khodursky, A.B., Lin, P.H., Lin-Chao, S., Cohen, S.N. (2002) Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *P Natl Acad Sci USA* 99: 9697-9702.

Bohlin, J., Eldholm, V., Pettersson, J.H., Brynildsrud, O., Snipen, L. (2017) The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics* 18: 151.

Braun, F., Le Derout, J., Régnier, P. (1998) Ribosomes inhibit an RNase E cleavage which induces the decay of the rpsO mRNA of *Escherichia coli*. *The EMBO Journal* 17: 4790-4797.

Buckstein, M.H., He, J., Rubin, H. (2008) Characterization of nucleotide pools as a function of phy siological state in *Escherichia coli*. *Journal of Bacteriology* 190: 718-726.

Busby, S., Ebright, R.H. (1994) Promoter structure, promoter recognition, and transcription activation in prokaryotes. *Cell* 79: 743-746.

Caron, M.P., Bastet, L., Lussier, A., Simoneau-Roy, M., Massé, E., Lafontaine, D.A. (2012) Dual-acting riboswitch control of translation initiation and mRNA decay. *P Natl Acad Sci USA* 109: E3444-3453.

Chen, C., Deutscher, M.P. (201 0) RNase R is a highly unstable protein regulated by growth phase and stress. *Rna* 16: 667-672.

Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S., Bayley, H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* 4: 265-270.

De Ley, J. (1961) Comparative carbohydrate metabolism and a proposal for a phylogentic relationship in the acetic acid bacteria. *J Gen Microbiol* 24: 31-50.

De Ley, J., Schwings, J., (1984) Genus *Gluconobacter*. Kreig, N.R. Holt, J. G. (eds). *Bergey´s Manual of Systematic Bacteriology vol* 1: 267-278.

Deana, A., Belasco, J.G. (2005) Lost in translation: the influence of ribosomes on bacterial mRNA decay. *Gene Dev* 19: 2526-2533.

Denoeud, F., Kapranov, P., Ucla, C., Frankish, A., Castelo, R., Drenkow, J., … Reymond, A. (2007) Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Research* 17: 746-759.

Deppenmeier, U., Ehrenreich, A. (2009) Physiology of acetic acid bacteria in light of the genome sequence of *Gluconobacter oxydans*. *J Mol Microb Biotech* 16: 69-80.

Deppenmeier, U., Hoffmeister, M., Prust, C. (2002) Biochemistry and biotechnological applications of *Gluconobacter* strains. *Appl Microbiol Biot* 60: 233-242.

Dressaire, C., Redon, E., Gitton, C., Loubière, P., Monnet, V., Cocaign-Bousquet, M. (2011) Investigation of the adaptation of *Lactococcus lactis* to isoleucine starvation integrating dynamic transcriptome and proteome information. *Microbial cell factories* 10 Suppl 1: S18.

Dressman, D., Yan, H., Traverso, G., Kinzler, K.W., Vogelstein, B. (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *P Natl Acad Sci USA* 100: 8817-8822.

Ebright, R.H. (2000) RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *Journal of Molecular Biology* 304: 687-698.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., … Turner, S. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323: 133-138.

Emory, S.A., Bouvet, P., Belasco, J.G. (1992) A 5'-terminal stem-loop structure can stabilize mRNA in *Escherichia coli*. *Gene Dev* 6: 135-148.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., Turcatti, G. (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research* 34: e22.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., … Venter, J.C. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.

Glenn, T.C. (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* 11: 759-769.

Gosselé, F., Swings, J., De Ley, J. (1980) Growth factor requirements of *Gluconobacter*. *Zentralbl Bakterioll Parasiten Kd Infektionskr Hyg Abt. irig. Ser. C.* 1: 348-350.

Gregory, S.T., O'Connor, M., Dahlberg, A.E. (1996) Functional *Escherichia coli* 23S rRNAs containing processed and unprocessed intervening sequences from *Salmonella typhimurium*. *Nucleic Acids Research* 24: 4918-4923.

Güell, M., Yus, E., Lluch-Senar, M., Serrano, L. (2011) Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nature Reviews. Microbiology* 9: 658-669.

Gupta, A., Singh, V.K., Qazi, G.N., Kumar, A. (2001) *Gluconobacter oxydans*: its biotechnological applications. *J Mol Microb Biotech* 3: 445-456.

Hambraeus, G., von Wachenfeldt, C., Hederstedt, L. (2003) Genome-wide survey of mRNA half-lives in *Bacillus subtilis* identifies extremely stable mRNAs. *Molecular Genetics and Genomics : MGG* 269: 706-714.

Hanke, T., Nöh, K., Noack, S., Polen, T., Bringer, S., Sahm, H., Wiechert, W., Bott, M. (2013) Combined fluxomics and transcriptomics analysis of glucose catabolism *via* a partially cyclic pentose phosphate pathway in *Gluconobacter oxydans* 621H. *Appl Environ Microb* 79: 2336-2348.

Hanke, T., Richhardt, J., Polen, T., Sahm, H., Bringer, S., Bott, M. (2012) Influence of oxygen limitation, absence of the cytochrome bc(1) complex and low pH on global gene expression in *Gluconobacter oxydans* 621H using DNA microarray technology. *Journal of Biotechnology* 157: 359-372.

Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., … Xie, Z. (2008) Single-molecule DNA sequencing of a viral genome. *Science* 320: 106-109.

Hawley, D.K., McClure, W.R. (1983) Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic Acids Research* 11: 2237-2255.

Hekmat, D., Bauer, R., Fricke, J. (2003) Optimization of the microbial synthesis of dihydroxyacetone from glycerol with *Gluconobacter oxydans*. *Bioproc Biosyst Eng* 26: 109-116.

Herrmann, U., Merfort, M., Jeude, M., Bringer-Meyer, S., Sahm, H. (2004) Biotransformation of glucose to 5-keto-D-gluconic acid by recombinant *Gluconobacter oxydans* DSM 2343. *Appl Microbiol Biot* 64: 86-90.

Hölscher, T., Weinert-Sepalage, D., Görisch, H. (2007) Identification of membrane-bound quinoprotein inositol dehydrogenase in *Gluconobacter oxydans* ATCC 621H. *Microbiology* 153: 499-506.

Hrdlickova, R., Toloue, M., Tian, B. (2017) RNA-Seq methods for transcriptome analysis. *Wires RNA* 8.

Hunkapiller, T., Kaiser, R.J., Koop, B.F., Hood, L. (1991) Large-scale and automated DNA sequence determination. *Science* 254: 59-67.

Ishihama, A. (2000) Functional modulation of *Escherichia coli* RNA polymerase. *Annual Review of Microbiology* 54: 499-518.

Jain, M., Tyson, J.R., Loose, M., Ip, C.L.C., Eccles, D.A., O'Grady, J., … Reference, C. (2017) MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Research* 6: 760.

Kiefler, I., Bringer, S., Bott, M. (2017) Metabolic engineering of *Gluconobacter oxydans* 621H for increased biomass yield. *Appl Microbiol Biot* 101: 5453-5467.

Kostner, D., Luchterhand, B., Junker, A., Volland, S., Daniel, R., Buchs, J., Liebl, W., Ehrenreich, A. (2015) The consequence of an additional NADH dehydrogenase paralog on the growth of *Gluconobacter oxydans* DSM3504. *Appl Microbiol Biot* 99: 375-386.

Kröger, C., Dillon, S.C., Cameron, A.D., Papenfort, K., Sivasankaran, S.K., Hokamp, K., … Hinton, J.C. (2012) The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. *P Natl Acad Sci USA* 109: E1277-1286.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., … International Human Genome Sequencing, C. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., Studholme, D.J. (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification* 3: 1-8.

Lawrence, J.G., Roth, J.R. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 143: 1843-1860.

Li, Z., Deutscher, M.P. (2002) RNase E plays an essential role in the maturation of *Escherichia coli* tRNA precursors. *Rna* 8: 97-109.

Lin, P.H., Singh, D., Bernstein, J.A., Lin-Chao, S. (2008) Genomic analysis of mRNA decay in *E. coli* with DNA microarrays. *Methods in Enzymology* 447: 47-64.

Lodish, H., Berk, A., Zipursky, S.L., (2000) Bacterial Transcription Initiation. Freeman, W.H. (ed). *Molecular Cell Biology* 4th edition: Section 10.2.

Maeda, H., Fujita, N., Ishihama, A. (2000) Competition among seven *Escherichia coli* sigma subunits: relative binding affinities to the core RNA polymerase. *Nucleic Acids Research* 28: 3497-3503.

Makrides, S.C. (1996) Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiological Reviews* 60: 512-538.

Mamlouk, D., Gullo, M. (2013) Acetic Acid bacteria: physiology and carbon sources oxidation. *Indian Journal of Microbiology* 53: 377-384.

Mandal, M., Breaker, R.R. (2004) Gene regulation by riboswitches. *Nat Rev Mol Cell Bio* 5: 451-463.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., … Rothberg, J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.

Mathew, R., Chatterji, D. (2006) The evolving story of the omega subunit of bacterial RNA polymerase. *Trends in Microbiology* 14: 450-455.

Mathy, N., Benard, L., Pellegrini, O., Daou, R., Wen, T., Condon, C. (2007) 5'-to-3' exoribonuclease activity in bacteria: role of RNase J1 in rRNA maturation and 5' stability of mRNA. *Cell* 129: 681-692.

Mendoza-Vargas, A., Olvera, L., Olvera, M., Grande, R., Vega-Alvarado, L., Taboada, B., … Morett, E. (2009) Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. *PloS one* 4: e7526.

Mentz, A., Neshat, A., Pfeifer-Sancar, K., Pühler, A., Rückert, C., Kalinowski, J. (2013) Comprehensive discovery and characterization of small RNAs in *Corynebacterium glutamicum* ATCC 13032. *BMC Genomics* 14: 714.

Merfort, M., Herrmann, U., Bringer-Meyer, S., Sahm, H. (2006a) High-yield 5-keto-D-gluconic acid formation is mediated by soluble and membrane-bound gluconate-5-dehydrogenases of *Gluconobacter oxydans*. *Appl Microbiol Biot* 73: 443-451.

Merfort, M., Herrmann, U., Ha, S.W., Elfari, M., Bringer-Meyer, S., Görisch, H., Sahm, H. (2006b) Modification of the membrane-bound glucose oxidation system in *Gluconobacter oxydans* significantly increases gluconate and 5-keto-D-gluconic acid accumulation. *Biotechnology Journal* 1: 556-563.

Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nature Reviews. Genetics* 11: 31-46.

Mientus, M., Kostner, D., Peters, B., Liebl, W., Ehrenreich, A. (2017) Characterization of membrane-bound dehydrogenases of *Gluconobacter oxydans* 621H using a new system for their functional expression. *Appl Microbiol Biot* 101: 3189-3200.

Mitschke, J., Vioque, A., Haas, F., Hess, W.R., Muro-Pastor, A.M. (2011) Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in *Anabaena sp.* PCC7120. *P Natl Acad Sci USA* 108: 20130-20135.

Mohanty, B.K., Kushner, S.R. (2016) Regulation of mRNA Decay in Bacteria. *Annual Review of Microbiology* 70: 25-44.

Morey, J.S., Van Dolah, F.M. (2013) Global analysis of mRNA half-lives and *de novo* transcription in a dinoflagellate, *Karenia brevis*. *PloS One* 8: e66347.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., … Finn, R.D. (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Research* 43: D130-137.

Nicolas, P., Mäder, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., … Noirot, P. (2012) Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science* 335: 1103-1106.

Nossal, N.G., Singer, M.F. (1968) The processive degradation of individual polyribonucleotide chains. I. *Escherichia coli* ribonuclease II. *The Journal of Biological Chemistry* 243: 913-922.

Olijve, W., Kok, J.J. (1979) Analysis of *Gluconobacter oxydans* in glucose containing media. *Arcg Microbiol* 121: 283-290.

Osbourn, A.E., Field, B. (2009) Operons. *Cellular and molecular life sciences : CMLS* 66: 3755-3775.

Paget, M.S., Helmann, J.D. (2003) The sigma70 family of sigma factors. *Genome Biology* 4: 203.

Pappenberger, G., Hohmann, H.P. (2014) Industrial production of L-ascorbic Acid (vitamin C) and D-isoascorbic acid. *Adv Biochem Eng Biot* 143: 143-188.

Pfeifer-Sancar, K., Mentz, A., Rückert, C., Kalinowski, J. (2013) Comprehensive analysis of the *Corynebacterium glutamicum* transcriptome using an improved RNAseq technique. *BMC Genomics* 14: 888.

Pinto, A.C., Melo-Barbosa, H.P., Miyoshi, A., Silva, A., Azevedo, V. (2011) Application of RNA-seq to reveal the transcript profile in bacteria. *Genetics and Molecular Research : GMR* 10: 1707-1718.

Pribnow, D. (1975) Bacteriophage T7 early promoters: nucleotide sequences of two RNA polymerase binding sites. *Journal of Molecular Biology* 99: 419-443.

Prust, C., Hoffmeister, M., Liesegang, H., Wiezer, A., Fricke, W.F., Ehrenreich, A., Gottschalk, G., Deppenmeier, U. (2005) Complete genome sequence of the acetic acid bacterium *Gluconobacter oxydans*. *Nature Biotechnology* 23: 195-200.

Raghavan, R., Groisman, E.A., Ochman, H. (2011) Genome-wide detection of novel regulatory RNAs in *E. coli*. *Genome Research* 21: 1487-1497.

Redon, E., Loubière, P., Cocaign-Bousquet, M. (2005) Role of mRNA stability during genome-wide adaptation of *Lactococcus lactis* to carbon starvation. *The Journal of Biological Chemistry* 280: 36380-36385.

Richhardt, J., Bringer, S., Bott, M. (2012) Mutational analysis of the pentose phosphate and Entner-Doudoroff pathways in *Gluconobacter oxydans* reveals improved growth of a Δ*edd* Δ*eda* mutant on mannitol. *Appl Environ Microb* 78: 6975-6986.

Richhardt, J., Bringer, S., Bott, M. (2013) Role of the pentose phosphate pathway and the Entner-Doudoroff pathway in glucose metabolism of *Gluconobacter oxydans* 621H. *Appl Microbiol Biot* 97: 4315-4323.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., … Bustillo, J. (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475: 348-352.

Rothman-Denes, L.B. (2013) Structure of *Escherichia coli* RNA polymerase holoenzyme at last. *P Natl Acad Sci USA* 110: 19662-19663.

Saichana, N., Matsushita, K., Adachi, O., Frebort, I., Frebortova, J. (2015) Acetic acid bacteria: A group of bacteria with versatile biotechnological applications. *Biotechnology Advances* 33: 1260-1271.

Sanger, F., Nicklen, S., Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *P Natl Acad Sci USA* 74: 5463-5467.

Schlüter, J.P., Reinkensmeier, J., Barnett, M.J., Lang, C., Krol, E., Giegerich, R., Long, S.R., Becker, A. (2013) Global mapping of transcription start sites and promoter motifs in the symbiotic alpha-proteobacterium *Sinorhizobium meliloti* 1021. *BMC Genomics* 14: 156.

Schweiger, P., Gross, H., Deppenmeier, U. (2010) Characterization of two aldo-keto reductases from *Gluconobacter oxydans* 621H capable of regio- and stereoselective alpha-ketocarbonyl reduction. *Appl Microbiol Biot* 87: 1415-1426.

Shao, W., Price, M.N., Deutschbauer, A.M., Romine, M.F., Arkin, A.P. (2014) Conservation of transcription start sites within genes across a bacterial genus. *mBio* 5: e01398-01314.

Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., … Vogel, J. (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 464: 250-255.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., … Church, G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309: 1728-1732.

Soini, J., Ukkonen, K., Neubauer, P. (2008) High cell density media for *Escherichia coli* are generally designed for aerobic cultivations - consequences for large-scale bioprocesses and shake flask cultures. *Microbial Cell Factories* 7: 26.

Sorek, R., Cossart, P. (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Reviews. Genetics* 11: 9-16.

Srivastava, R.A., Srivastava, N., Apirion, D. (1992) Characterization of the RNA processing enzyme RNase III from wild type and overexpressing *Escherichia coli cells* in processing natural RNA substrates. *The International Journal of Biochemistry* 24: 737-749.

Stenstrom, C.M., Holmgren, E., Isaksson, L.A. (2001) Cooperative effects by the initiation codon and its flanking regions on translation initiation. *Gene* 273: 259-265.

Swerdlow, H., Gesteland, R. (1990) Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research* 18: 1415-1419.

Swerdlow, H., Wu, S.L., Harke, H., Dovichi, N.J. (1990) Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *Journal of chromatography* 516: 61-67.

Tatusova, T., Ciufo, S., Federhen, S., Fedorov, B., McVeigh, R., O'Neill, K., Tolstoy, I., Zaslavsky, L. (2015) Update on RefSeq microbial genomes resources. *Nucleic Acids Research* 43: D599-605.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., … Ostell, J. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research* 44: 6614-6624.

Thomason, M.K., Storz, G. (2010) Bacterial antisense RNAs: how many are there, and what are they doing? *Annual Review of Genetics* 44: 167-188.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., … Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* 28: 511-515.

Turcatti, G., Romieu, A., Fedurco, M., Tairi, A.P. (2008) A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids research* 36: e25.

Uspenskaia, S.N., Loitsianskaia, M.S. (1979) Efficiency of glucose utilization by *Gluconobacter oxydans*. *Mikrobiologiia* 48: 400-405.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., … Zhu, X. (2001) The sequence of the human genome. *Science* 291: 1304-1351.

Wang, L., Si, Y., Dedow, L.K., Shao, Y., Liu, P., Brutnell, T.P. (2011) A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq. *PloS One* 6: e26426.

Wang, Z., Gerstein, M., Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics* 10: 57-63.

Yamada, Y., Hoshino, K., Ishikawa, T. (1997) The phylogeny of acetic acid bacteria based on the partial sequences of 16S ribosomal RNA: the elevation of the subgenus *Gluconoacetobacter* to the generic level. *Bioscience, Biotechnology, and Biochemistry* 61: 1244-1251.

4. Appendix

4.1 Supplementary data: High precision genome sequencing of engineered *G. oxydans* 621H by combining long nanopore and short accurate Illumina reads

Angela Kranz[*,1,4], Alexander Vogel[*,2,3,4], Ursula Degner[1,4], Ines Kiefler[1,4], Michael Bott[1,4], Björn Usadel[2,3,4] and Tino Polen[1,4,@]

[1)] Institute of Bio- and Geosciences IBG-1: Biotechnology, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany.

[2)] IBMG: Institute for Biology I, RWTH Aachen University, Worringer Weg 2, 52074 Aachen, Germany.

[3)] IBG-2 Plant Sciences, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany.

[4)] The Bioeconomy Science Center (BioSC), c/o Forschungszentrum Jülich GmbH, 52425 Jülich, Germany.

SUPPLEMENTARY TABLES

Table S1 List of oligonucleotides used for PCR and Sanger sequencing to validate SNVs, MNVs, and InDels detected by genome sequencing using Illumina reads.

Pos. of variant	Forward Primer (5´ to 3´)	Reverse Primer (5´ to 3´)
7,185	GCAGATCGTTTTTGACGGG	GTGTGACGGTTTTGAAAGG
13,318	CTGACGGCTTTTGGTATTCC	AAAGATTTGGCGGGAGAGG
25,803	GAACAGTCCTTTTCCAGCC	TCACCTACAACCACATCCG
102,363 & 102,446^102,447	ATGTCGAAAGCCGATGGG	AAGCTCATGGAAGGCACC
117,915^117,916	CAAGACTTCTGTATCCCCC	CATTAAGACGACCGTGACC
127,329^127,330	GGGAAGTTCACGGATGAAG	CAGCGTTATGTCAGCCTCTC
203,740^203,741 & 203,756^203,757	CGAAGCAGCAAAAGACAGG	ATTTCCCGAACAATCCGACC
216,398^216,399	ACATAACCCTCTTCCTCGC	ATTACCTGTACCGCACGC
236,349 & 236,514^236,515	TACAATACCCCGCTGACGC	TTCACGCAGTTCCTTTTCCC
248,758^248,759	CAGTTCCATACGACGCTTAC	TGATCTGTCACTACCCTTC
273,806	TGCGTCTTCTGTCTGGTGG	GTTCGGAATCGGTCTTGGC
290,102^290,103	TCAGGAACTGGTGGAAATGG	TGCGACTGGAGATCGTCAC
364,311^364,312 & 364,456^364,457	ATCATTCTGGACGGCTTTCG	GTGGGAACCTGTCGCAGATG
404,262^404,263 & 404,274	TGGCGTCATCGTTCTCAAGG	GGTCGGTCGTGTGAAGATG
417,228^417,229 & 417,277^417,278	AAGGGCTTGGTCAGATAGTC	CAGGAGATGTTGCGGATGG
551,566	AGGGTTTTGACAGCTATGG	TGACGTTGAACAGAAGCGG
590,971^590,972	GCTCTGATACTGCCCCTG	GCTCTGATACTGCCCCTG
620,795^620,796	CTTCCTCTCCAAGATCCCC	TGTTTTCCCAGTCCTGCC

684,876^684,877	ACCTGTTCCGCCAGAGAC	GGTCTGTTTCTGTTTGTCGC
693,157, 693,198, 693,207 & 693,219	AAGATCGGAACGCAGAGG	GCGGTTTCAGGGTTATGC
776,179^776,180	CTTCCACAAAGGACTCGC	AATCTGAAATGGTCCGTCG
784,899	CAGTTCGGGCATTTCGTCC	AAATCAATGGTTTCGCCTGCG
833,443^833,444	ACTACAGAACTCGATCCCC	ATCTCGTCATCCTCCAACC
845,374	GCCTGACCGTAGAAAATGG	ACGCATGAAAACGAGACCC
858,340 & 858,607	CGATGAACGTGTTGAAGCCC	ACAGACCCGAAAAGCCACC
873,583^873,584	ACGATTAAAGCCAACTTCCC	GTCAGTCTTCATGCTCTTCC
882,467^882,468	TGATTGACGATGGTGACGG	ATATTTTCGGCAGCTCGC
908,865..908,866, 908,874 & 908,886	CGAATAAGGCAGGAACACC	GGCAGAAGAAAGACCATCG
992,802^992,803	CCATCATCTTCTGCCTTTCC	ATCTGGTCCGACTTCTCG
1,129,165	TGAATGGCAGAGAGAGAGG	GAATACTGGGCTTTTTCGGG
1,302,341^130,234,2	TGTATCTGAGAGGGCATTCC	CAGCGGAGAAAAAAGAACATCC
1,316,129	GTTTCGGCGTCTTTTGCG	AGTCCAACTTCCAGCACC
1,375,017	CCCACATGCTGGTCAAAGG	GGACGAAGAGGCGTAAAC
1,417,494	CGGAGTGCGTATAATGACAG	CATGCGGATTGCGGGTATTG
1,492,855	TCACGATGCCCAGAATGAC	CTCGAAGACGACATCAAAATC
1,706,402	CGCTTCTCCCAGCAGATTC	GCTAGAGTTTCGTCCATTATG
1,804,454	GGTTCCGTAAACGCTCTGTC	CTTTCCGCTGGAGATCGTC
1,825,341	GTTTCAGTCGTATCGGAGG	CATCTCGTTCTGTCTCATGG
1,879,659^1,879,660	ACCATCCCAGCCCATAAAGC	TGTCATGTTCCTTTCCCGCC
1,915,762^1,915,763	AAATTGGTATCCGCCTGC	ATCTTCTTGCCGAACTGC
1,925,626^1,925,627	GAGGCGTTCGGGGAAATC	CCTTTGGGGGCAATCAGAC
1,937,248^1,937,249	ATGTCTGGATCAACGAAGGC	GATTTCTACGGTGTCAGCC
1,957,230	GGTTGAGAGTTTGATCGTATGG	TGTGGTCGTGATACTCTGG
2,112,099^2,112,100	TGGAAGCACAAGAAGCAGG	TGAAGACGAAAACCAGACCC
2,112,557	TGCTGTTTTTCTTCCTGTTCC	CTGTATTTCGCACTGAGCC
2,133,844^2,133,845	TGGAGTTCTTCCTGTTCGG	AAATTCCAGATCGCTGCC

2,156,390^2,156,391	GGATGTGTGAAGCGTTTGC	TCACGGAAAGGGAAGATCGG
2,196,994^2,196,995	CTTGTTTTCCCAGCCTTCC	AAACCATTCCCGCCTTCC
2,235,276^2,235,277	CGAGGCTGTGGAGAACGG	CCAGGGCGGAATACTCACC
2,348,459^2,348,460	CAAAGCCCATCGTCGTGTG	ATGTGCTGTCGTGGTTCCTG
2,409,938^2,409,939	TCAGCAAATCGCATCAGTGG	CCGAATAGGGATCGTTGACC
2,491,561 & 2,491,596	TTTCGCCAGCCATTCCTGTG	GGTTGCTCCGTGACAGAATC
2,582,687^2,582,688 & 2,582,691^2,582,692	GCACTGACATATCTGCTAGTG	GCTGATCGTGAAGCCAATG
2,647,659 & 2,647,663	GGCTCCTACACGAATGTTG	CTACATACCCGCTCCGTTG
2,648,330	CCTACAGCCTTGACGACAG	CTGCCATCCATCATGCGTG
10,149^10,150 (P1*)	ACCCCAAGCTCTATTTCCC	TAATCTTGTCTGCTCGCCC
81,824^81,825 (P1)	CTATCCGAAAAACGACGCC	GTGACTTCCACTACAATGCC
112,715 (P1)	ACAATCACTCCGACAAGC	AATGACCATAAACGCCCC
155,095 & 155,109 (P1)	ACCTTGCCCTCTATGACCAG	CGCCTTGCCGTTCCATTGC

*) P1 = Variants detected in sequence of pGOX1 (NC_006672)

Table S2 List of oligonucleotides used for PCR and Sanger sequencing to validate structural variants detected by long nanopore reads.

Variant position	Forward Primer (5´ to 3´)	Reverse Primer (5´ to 3´)
1,233,221	AGGAGTGCCAGACATACC	ACCACTACTGCGTCAAGC
2,587,655	GCAGGAGGCGAATTTGAAGC CCAATCCTCCGCAGGTTATG AGCGTTCCTTCCCGTGATAC GCTGATTGAGAATGGCGCTATAGTTG	CTGCGGAGGATTGGATTTCG GAACGCTTTATCGCGCTGAC TCAGCATCATTATGAACGGCTGAATGG AAACGACCGGGACGCCTTTG
15,633 (P1)	TTTAGGAACTGTCGAGCC	CATGCGTCGTGTTTATGATT
13,804 (P2)	GGTTTATACTCTGTTGGG	GATGGATATGTATTGTGGG

P1 = SV detected in sequence of pGOX1 (NC_006672)
P2 = SV detected in sequence of pGOX2 (NC_006673)

Table S3 Frequencies of variants detected in intergenic regions of *G. oxydans* 621H wild type and engineered strains compared to the reference chromosome sequence (NC_006677) and plasmid pGOX1 to pGOX5 (NC_006672 to NC_006676).

					Frequency (%)						
nt Positions	**Type**	**Reference**	**Allele**	**Flanking genes**	**WT-DSMZ**	**WT-BM**	**WT-E**	**Δ*upp***	**IK001**	**IK002.1 IK002.1***	**IK003.1 IK003.1***
213,541	Deletion	C	-	GOX_RS02140 GOX_RS02145	-	-	-	-	-	11 -	- -
321,941	Deletion	CG	-	GOX_RS02670 GOX_RS02675	99	97	99	99	98	98 94	99 100
358,167	Insertion	-	G	GOX_RS02845 GOX_RS02850	94	96	97	93	97	97 99	98 91
417,286	Insertion	-	C	GOX_RS03130 GOX_RS03135	100	100	100	100	100	100 100	99 100
619,577	Insertion	-	T	GOX_RS04035 GOX_RS04040	98	95	96	98	96	95 97	96 93
700,226	Deletion	C	-	GOX_RS04350 GOX_RS04355	94	92	98	96	94	95 97	95 92
706,886	Insertion	-	GG	GOX_RS04375 GOX_RS04380	94	92	93	95	93	93 93	94 88
1,233,214	SNV	A	C	GOX_RS06790 GOX_RS06795	-	-	11	43	44	34 42	28 36
1,635,833	Insertion	-	C	GOX_RS08675 GOX_RS08680	-	-	85	-	-	- -	- -
1,734,742	Insertion	-	G	GOX_RS09155 GOX_RS09160	96	96	99	96	99	99 98	96 98
1,804,771	Insertion	-	G	GOX_RS09465 GOX_RS09470	93	95	96	97	95	94 94	93 95
1,975,439	Deletion	G	-	GOX_RS10245 GOX_RS10250	98	100	100	100	99	98 97	98 96

2,018,596	Insertion	-	A	GOX_RS10495 GOX_RS10500	98	94	94	97	96	95 97	97 94
2,257,175	Insertion	-	T	GOX_RS11580 GOX_RS11585	97	96	95	97	96	98 92	96 95
2,293,553	Insertion	-	A	GOX_RS11730 GOX_RS11735	100	94	99	99	95	95 95	99 97
2,360,930	Insertion	-	A	GOX_RS12020 GOX_RS12025	100	97	97	97	100	98 100	96 98
2,441,754	SNV	C	A	GOX_RS12410 GOX_RS12415	37	31	27	36	37	36 43	35 37
2,463,001	Insertion	-	T	GOX_RS12520 GOX_RS1525	96	97	98	99	99	94 91	97 98
2,570,397	SNV	C	T	GOX_RS13135 GOX_RS14075	100	100	100	100	100	100 100	100 100
2,570,495	SNV	G	A	GOX_RS13135 GOX_RS14075	100	100	100	100	100	100 100	100 100
2,570,736	SNV	T	C	GOX_RS13135 GOX_RS14075	96	100	100	100	98	100 96	100 100
2,648,181	Deletion	A	-	GOX_RS13575 GOX_RS13580	99	98	100	99	99	99 97	99 100
32 (P1)	SNV	T	C	GOX_RS00810 GOX_RS00005	53	55	54	52	63	52 58	55 56
53 (P1)	SNV	G	C	GOX_RS00810 GOX_RS00005	51	51	54	49	62	47 53	52 54
55 (P1)	SNV	G	A	GOX_RS00810 GOX_RS00005	51	51	54	49	62	47 53	52 53
57 (P1)	SNV	T	A	GOX_RS00810 GOX_RS00005	51	51	54	48	63	47 53	52 54
59 (P1)	SNV	C	T	GOX_RS00810 GOX_RS00005	51	51	55	49	63	48 53	52 55
61 (P1)	SNV	C	T	GOX_RS00810 GOX_RS00005	51	52	54	49	63	48 52	52 55

86 (P1)	SNV	C	T	GOX_RS00810 GOX_RS00005	53	54	55	53	55	46 54	51 56
99 (P1)	SNV	G	T	GOX_RS00810 GOX_RS00005	56	52	55	50	55	45 54	50 56
118 (P1)	SNV	G	A	GOX_RS00810 GOX_RS00005	60	52	60	55	51	45 57	51 52
149 (P1)	SNV	C	T	GOX_RS00810 GOX_RS00005	59	57	56	58	54	48 58	49 55
204 (P1)	SNV	C	T	GOX_RS00810 GOX_RS00005	56	52	55	60	42	48 60	50 51
216 (P1)	MNV	AT	TC	GOX_RS00810 GOX_RS00005	50	45	51	51	38	46 54	48 58
219 (P1)	SNV	T	C	GOX_RS00810 GOX_RS00005	50	44	51	49	38	46 53	49 47
221 (P1)	SNV	C	G	GOX_RS00810 GOX_RS00005	49	43	49	49	37	46 53	48 48
224 (P1)	SNV	T	A	GOX_RS00810 GOX_RS00005	53	44	51	52	37	48 55	53 50
244 (P1)	SNV	A	G	GOX_RS00810 GOX_RS00005	52	41	53	54	40	49 60	56 49
261 (P1)	SNV	C	T	GOX_RS00810 GOX_RS00005	51	43	46	51	42	50 59	55 47
281 (P1)	SNV	C	T	GOX_RS00810 GOX_RS00005	52	42	41	46	42	45 52	50 43
291 (P1)	Deletion	T	-	GOX_RS00810 GOX_RS00005	50	40	38	38	34	42 49	48 41
295 (P1)	MNV	TC	CT	GOX_RS00810 GOX_RS00005	48	40	37	37	34	41 48	47 40
302 (P1)	SNV	A	G	GOX_RS00810 GOX_RS00005	51	44	38	38	36	45 51	49 40
308 (P1)	Insertion	-	T	GOX_RS00810 GOX_RS00005	50	45	35	38	35	44 49	49 39

309 (P1)	SNV	-	T	GOX_RS00810 GOX_RS00005	50	45	35	38	35	44 50	49 39
316 (P1)	SNV	T	C	GOX_RS00810 GOX_RS00005	43	39	28	31	29	40 44	43 32
318 (P1)	MNV	GC	TT	GOX_RS00810 GOX_RS00005	43	39	28	29	30	40 45	43 30
324 (P1)	MNV	GA	AG	GOX_RS00810 GOX_RS00005	43	40	27	29	31	40 44	45 31
328 (P1)	MNV	TT	AC	GOX_RS00810 GOX_RS00005	44	38	27	28	30	40 46	46 31
349 (P1)	SNV	A	C	GOX_RS00810 GOX_RS00005	49	49	37	38	45	44 50	49 37
360 (P1)	SNV	A	T	GOX_RS00810 GOX_RS00005	50	49	38	39	44	45 51	46 38
363 (P1)	MNV	AG	GA	GOX_RS00810 GOX_RS00005	47	49	37	39	45	44 52	46 36
371 (P1)	SNV	C	T	GOX_RS00810 GOX_RS00005	46	48	38	38	47	46 52	45 36
374 (P1)	Insertion	-	C	GOX_RS00810 GOX_RS00005	44	48	39	36	47	45 52	43 36
377 (P1)	MNV	TC	GA	GOX_RS00810 GOX_RS00005	44	49	38	35	46	45 49	42 35
381 (P1)	Deletion	A	-	GOX_RS00810 GOX_RS00005	40	47	37	30	42	44 46	41 34
388 (P1)	SNV	C	T	GOX_RS00810 GOX_RS00005	46	49	39	32	44	49 50	45 38
410 (P1)	SNV	C	A	GOX_RS00810 GOX_RS00005	44	49	39	34	43	45 44	44 37
427 (P1)	SNV	G	T	GOX_RS00810 GOX_RS00005	43	49	38	37	44	43 45	42 37
510 (P1)	MNV	CC	AA	GOX_RS00810 GOX_RS00005	36	39	38	37	34	35 31	34 33

529 (P1)	SNV	G	A	GOX_RS00810 GOX_RS00005	34	35	38	36	30	26 25	27 30
534 (P1)	Deletion	T	-	GOX_RS00810 GOX_RS00005	30	33	36	35	28	25 23	26 26
541 (P1)	Insertion	-	G	GOX_RS00810 GOX_RS00005	29	30	35	32	24	23 21	23 25
60,729 (P1)	Insertion	-	C	GOX_RS00305 GOX_RS00310	95	95	95	99	97	95 98	98 99
145,803 (P1)	Insertion	-	C	GOX_RS00700 GOX_RS00705	98	97	96	92	99	96 94	97 96
163,156 (P1)	SNV	C	G	GOX_RS00810 GOX_RS00005	31	26	33	36	25	27 25	29 24
163,162 (P1)	SNV	G	T	GOX_RS00810 GOX_RS00005	34	28	34	35	25	29 26	31 26
163,166 (P1)	MNV	TT	CA	GOX_RS00810 GOX_RS00005	34	28	34	35	26	30 25	30 27
163,175 (P1)	MNV	CC	TT	GOX_RS00810 GOX_RS00005	37	31	36	36	28	34 28	32 31
163,179 (P1)	MNV	TC	AT	GOX_RS00810 GOX_RS00005	37	32	36	36	29	34 28	33 31S

P1 variant detected in the sequence of pGOX1 (NC_006672)

Table S4 Frequencies of 93 detected variants in 73 coding regions of *G. oxydans* 621H wild type and engineered strains compared to the reference chromosome sequence (NC_006677) and plasmid pGOX1 to pGOX5 (NC_006672 to NC_006676).

			Frequency (%)						
nt Positions	**Coding region change**	**Amino acid exchange**	**WT-DSMZ**	**WT-BM**	**WT-E**	**Δ*upp***	**IK001**	**IK002.1 IK002.1***	**IK003.1 IK003.1***
7,185	GOX_RS01165:c.740delA	Gln247fs	97	99	98	99	96	98 98	99 99
13,318	GOX_RS01180:c.2741delC	Ala914fs	97	96	94	94	99	97 96	98 97
25,803	GOX_RS01230: c.816A>T	Glu272Asp	100	100	100	100	100	100 100	100 100
102,363	GOX_RS01580:c.835delC	Leu279fs	97	100	99	100	100	99 94	99 98
102,446^102,447	GOX_RS01580:c.751_752insG	Val251fs	98	99	100	97	99	97 99	99 97
117,915^117,916	GOX_RS01655:c.1029_1030insC	Cys344fs	96	95	98	97	97	98 100	95 96
127,329^127,330	GOX_RS01685:c.1194_1195insC	Trp399fs	96	95	98	95	95	92 96	96 94
169,377^169,378	GOX_RS01915:c.250_251insC	Thr84fs	99	98	97	99	98	97 97	100 92
203,740^203,741	GOX_RS02090:c.989_990insG	Gly330fs	96	94	99	95	90	93 93	94 97
203,756^203,757	GOX_RS02090:c.973_974insT	Ser325fs	95	99	95	97	98	98 100	98 98
216,398^216,399	GOX_RS02160:c.410_411insC	Pro137fs	95	100	95	94	96	96 97	97 99
236,349	GOX_RS02235:c.980delC	Ala327fs	79	90	90	95	93	96 81	93 91

236,514^236,515	GOX_RS02235:c.1145_1146insC	Asp382fs	99	100	100	100	99	99 99	99 98
248,758^248,759	GOX_RS02315:c.597_598insG	Thr200fs	95	99	99	95	97	96 98	98 99
273,806	GOX_RS02445:c.284delC	Ser95fs	97	94	96	95	98	95 96	98 95
290,102^290,103	GOX_RS02520:c.109_110insC	Thr37fs	90	94	97	93	96	96 96	94 97
318,447	GOX_RS02660:c.30delG	Thr10fs	94	89	93	98	96	95 92	94 96
364,311^364,312	GOX_RS02865:c.1304_1305insC	Gly435fs	93	94	96	90	93	93 93	94 96
364,456^364,457	GOX_RS02865:c.1449_1450insG	Gly484fs	93	89	93	94	97	97 93	93 95
375,623	GOX_RS02905:c.2094C>T		100	100	100	100	100	100 100	100 100
375,626	GOX_RS02905:c.2097C>G		100	100	100	100	100	100 100	100 100
404,262^404,263	GOX_RS03085:c.1802_1803insC	Pro601fs	94	95	96	98	98	96 97	95 95
404,274	GOX_RS03085:c.1791delC	Thr597fs	98	99	99	100	98	99 95	99 97
417,228^417,229	GOX_RS03130:c.55_56insG	Ala19fs	97	98	99	99	98	97 98	97 100
417,277^417,278	GOX_RS03130:c.6_7insG	Ser3fs	95	94	98	97	97	94 98	96 93
450,256^450,257	GOX_RS03265:c.220_221insG	Ala74fs	96	96	98	97	97	97 100	98 97
551,566	GOX_RS03760:c.565delC	Pro189fs	97	98	97	95	99	97 96	96 100
590,971^590,972	GOX_RS03905:c.568_569insC	Gln190fs	98	94	97	96	98	97 95	95 95

620,795^620,796	GOX_RS04040:c.1212_1213insC	Pro405fs	94	93	95	95	93	98 92	98 95
684,876^684,877	GOX_RS04300:c.455_456insC	Val152fs	100	100	100	100	100	100 100	100 100
693,157	GOX_RS04330:c.1478delA	Lys493fs	92	96	100	98	96	98 96	99 97
693,198	GOX_RS04330:c.1437T>G		100	100	100	100	100	100 100	100 100
693,207	GOX_RS04330:c.1428delA	Glu476fs	99	99	100	100	100	99 97	99 100
693,219	GOX_RS04330:c.1416delT	Val472fs	99	99	99	100	99	99 100	100 99
776,179^776,180	GOX_RS04675:c.408_409insC	Cys137fs	97	96	100	100	97	95 95	96 95
784,899	GOX_RS04705:c.496delG	Gly166fs	98	98	97	99	98	97 100	98 94
833,443^833,444	GOX_RS04945:c.1341_1342insG	Thr448fs	97	99	98	97	100	96 97	95 96
845,374	GOX_RS05015:c.650delC	Pro217fs	98	94	98	99	98	99 98	95 99
858,340	GOX_RS05065:c.1270delG	Ala424fs	99	98	99	97	96	98 99	94 94
858,607	GOX_RS05065:c.1003delC	Leu335fs	96	99	95	96	95	93 99	97 99
873,507	GOX_RS05130:c.2201delT	Ile734fs	93	91	95	96	98	95 96	93 94
873,583^873,584	GOX_RS05130:c.2277_2278insG	Gly760fs	94	97	98	98	94	94 97	96 98
882,467^882,468	GOX_RS05185:c.395_396insC	Pro132fs	99	98	98	100	97	99 97	97 98
908,865..908,866	GOX_RS05325:c.48_49delGAinsAG	Arg16_Arg17delinsArgGly	100	100	99	100	100	100 100	100 100

908,874	GOX_RS05325:c.57delG	Ala19fs	95	98	97	97	93	96 97	95 97
908,886	GOX_RS05325:c.69delC	Gly23fs	98	94	99	97	98	98 98	95 95
992,802^992,803	GOX_RS05700:c.1414_1415insC	Thr472fs	95	92	98	96	97	97 99	98 94
1,129,165	GOX_RS06280:c.226T>G	Cys76Gly	100	100	100	100	100	100 100	100 100
1,201,294	GOX_RS06635:c.867T>G		100	100	100	100	100	100 100	100 100
1,302,341^ 1,302,342	GOX_RS07120:c.16_17insG	Arg6fs	96	89	96	97	95	95 97	95 97
1,316,129	GOX_RS07200:c.196delG	Gly66fs	95	99	95	100	95	95 96	94 97
1,336,284^ 1,336,285	GOX_RS07310:c.1133_1134insG	Gln378fs	100	96	100	99	97	98 98	96 98
1,375,017	GOX_RS07480:c.168delC	Pro56fs	93	94	100	93	95	94 96	100 95
1,384,378	GOX_RS07515:c.78G>C		100	100	100	100	100	100 100	100 100
1,417,494	GOX_RS07655:c.192delG	Gln64fs	99	96	94	98	99	97 91	98 98
1,492,855	GOX_RS08000:c.578delC	Pro193fs	97	94	99	98	94	94 99	96 98
1,706,402	GOX_RS09035:c.252delC	Pro84fs	94	94	93	95	92	93 95	93 92
1,784,492	GOX_RS09385:c.610delG	Gly204fs	97	99	97	100	97	97 99	99 98
1,785,478	GOX_RS09385:c.1596C>G		-	17	13	22	12	14 18	26 15
1,786,017	GOX_RS09390:c.755C>A	Ala252Glu	-	-	20	-	21	24 17	19 20

1,804,454	GOX_RS09465:c.14delG	Gly5fs	96	98	99	99	98	98 95	98 95
1,825,341	GOX_RS09575:c.373delA	Lys125fs	99	100	95	98	97	99 97	97 99
1,879,659^ 1,879,660	GOX_RS09830:c.96_97insG	Cys33fs	98	100	96	96	98	97 97	97 97
1,894,981	GOX_RS09905:c.2382delA	Glu794fs	99	99	99	99	99	99 100	98 96
1,915,762^ 1,915,763	GOX_RS09980:c.1312_1313insC	Ala438fs	100	97	99	100	97	100 100	98 96
1,925,626^ 1,925,627	GOX_RS10030:c.253_254insC	Ser85fs	98	94	96	96	96	95 98	98 99
1,937,248^ 1,937,249	GOX_RS10075:c.1900_1901insC	Thr634fs	98	99	98	100	100	97 98	99 99
1,957,230	GOX_RS10175:c.401delC	Pro134fs	98	97	98	99	97	97 95	95 98
2,015,781^ 2,015,782	GOX_RS10475:c.96_97insC	Thr33fs	98	98	93	96	93	97 90	95 86
2,087,993^ 2,087,994	GOX_RS10790:c.2766_2767insA	Glu923fs	100	100	100	100	100	100 100	100 100
2,112,099^ 2,112,100	GOX_RS10885:c.427_428insA	Stop (TAG) 143 Stop (TAA)	97	96	100	99	97	99 94	98 99
2,112,557	GOX_RS10890:c.770A>T	Glu257Val	100	100	100	100	100	100 100	100 100
2,133,844^ 2,133,845	GOX_RS10980:c.1247_1248insC	Ala416fs	96	100	93	100	98	98 96	97 95
2,156,390^ 2,156,391	GOX_RS11115:c.446_447insG	Gln149fs	86	95	94	97	94	95 93	94 96
2,196,994^ 2,196,995	GOX_RS11305:c.570_571insG	Ile191fs	97	96	98	97	96	97 95	97 95

2,235,276^ 2,235,277	GOX_RS11470:c.738_739insA	Lys247fs	88	90	95	93	93	93 88	92 96
2,348,459^ 2,348,460	GOX_RS11980:c.548_549insC	Pro183fs	99	95	99	100	96	91 91	95 96
2,409,938^ 2,409,939	GOX_RS12290:c.797_798insC	Val266fs	98	96	98	99	95	94 97	95 95
2,491,561	GOX_RS12660:c.970T>C	Phe324Leu	100	100	100	100	100	100 100	100 100
2,491,596	GOX_RS12660:c.1005T>C		100	100	100	100	100	100 99	100 100
2,572,385.. 2,572,386	GOX_RS13155:c.265_266delAAinsGG	Asn89Gly	100	97	100	100	100	100 100	100 100
2,572,407	GOX_RS13155:c.287C>G	Ala96Gly	98	97	100	100	100	100 100	100 100
2,582,687^ 2,582,688	GOX_RS13205:c.616_617insG	Arg206fs	95	95	99	96	96	92 91	96 93
2,582,691^ 2,582,692	GOX_RS13205:c.620_621insG	Gln207fs	96	96	94	92	95	94 98	98 95
2,647,659	GOX_RS13575:c.513delC	Ala171fs	99	97	97	94	99	99 97	98 96
2,647,663	GOX_RS13575:c.509delA	Asn170fs	98	99	100	98	99	98 99	96 95
2,648,330	GOX_RS13580:c.1299delG	Gly433fs	95	97	98	98	93	94 96	98 98
10,149^10,150 (P1)	GOX_RS00055:c.911_912insG	Leu304fs	98	95	98	93	99	99 100	97 98
81,824^81,825 (P1)	GOX_RS13955:c.466_467insG	Arg156fs	96	96	97	98	98	97 99	96 98
97,917 (P1)	GOX_RS13970:c.423delC	Ala141fs	97	97	100	94	94	99 100	99 99
112,715 (P1)	GOX_RS00555:c.19delA	Lys7fs	100	99	99	99	98	95 98	98 97

155,095 (P1)	GOX_RS00760:c.742delG	Gly248fs	99	93	98	97	95	98 97	97 99
155,109 (P1)	GOX_RS00760:c.756delC	Asn252fs	98	99	95	97	97	98 98	98 97

*) after 24 h or 28 h of growth under controlled conditions in a DASGIP bioreactor (Kiefler et al., 2017)
P1 variant detected in the sequence of pGOX1 (NC_006672)

Table S5 Results of BLASTP search with re-annotated open reading frames (ORFs). The relevant change on the nucleotide level for affected locus tags can be found in Table S4.

Locus tag	Product	Protein ID	Old start	Old stop	Old AS length	New start	New stop	New AS length	Best BLAST hit	AS length of BLAST hit
GOX_RS01165	Hypothetical protein	- (pseudo)	6,446	7,238	-	6,446	7,237	263	Hypothetical protein WP_024717314.1 (MULTISPECIES)	263
GOX_RS01180	Oxidoreductase	- (pseudo)	10,578	13,530	-	10,577	13,528	983	Oxidoreductase WP_034954733.1 (*Gluconobacter oxydans* DSM 2003)	983
GOX_RS01580	Clp protease	- (pseudo)	103,197	101,932	-	103,195	101,930	421	ATP-dependent Clp protease ATP-binding subunit ClpX WP_024717294.1 (MULTISPECIES)	421
GOX_RS01655	Hypothetical protein	- (pseudo)	118,944	117,656	-	118,943	117,654	429	Membrane protein WP_024717288.1 (MULTISPECIES)	429

GOX_RS01685	ABC transporter ATP-binding protein	- (pseudo)	128,523	126,686	-	128,523	126,685	612	ABC transporter ATP-binding protein WP_024717286.1 (MULTISPECIES)	612
GOX_RS01915	UDP-N-acetylmuramyl-tripeptide--D-alanyl-D-alanine ligase	- (pseudo)	169,128	170,500	-	169,128	170,501	457	UDP-N-acetylmuramyl-tripeptide—D-alanyl-D-alanine ligase WP_024717277.1 (MULTISPECIES)	457
GOX_RS02090	Aspartate aminotransferase	- (pseudo)	204,729	203,520	-	204,732	203,521	403	Aspartate aminotransferase WP_024717131.1 (*Gluconobacter oxydans* DSM 2003)	403
GOX_RS02160	Methionine biosynthesis protein	- (pseudo)	216,808	216,204	-	216,812	216,207	201	Methionine biosynthesis protein WP_024716992.1 (MULTISPECIES)	201
GOX_RS02235	5-oxoprolinase	- (pseudo)	235,370	236,938	-	235,374	236,942	522	5-oxoprolinase WP_024717186.1 (MULTISPECIES)	522

GOX_RS02315	Lysyl-tRNA synthetase	- (pseudo)	249,355	248,361	-	249,360	248,365	331	EF-P lysine aminoacylase GenX WP_034954554.1 (MULTISPECIES)	331
GOX_RS02445	Tryptophanyl-tRNA synthetase	- (pseudo)	273,523	274,513	-	273,528	274,517	329	Tryptophan-tRNA ligase WP_024717209.1 (MULTISPECIES)	329
GOX_RS02520	Cell division protein FtsX	- (pseudo)	289,994	290,901	-	289,998	290,906	302	Cell division protein FtsX WP_024717213.1 (MULTISPECIES)	302
GOX_RS02660	Peptidylprolyl isomerase	- (pseudo)	318,418	319,792	-	318,423	319,796	457	Peptidylprolyl isomerase WP_024717224.1 (MULTISPECIES)	457
GOX_RS02865	Hypothetical protein	- (pseudo)	363,008	365,045	-	363,011	365,050	679	Hypothetical protein WP_024717236.1 (*Gluconobacter oxydans* DSM 2003)	679

GOX_RS03085	DNA-directed RNA polymerase subunit beta	WP_041242769.1	406,064	401,892	1,390	406,069	401,897	1,390	DNA-directed RNA polymerase subunit beta WP_062269565.1 (*Gluconobacter oxydans* LMG 1406)	1390
GOX_RS03130	DNA-binding response regulator	WP_011251979.1	417,283	416,528	251	417,264	416,533	243	DNA-binding response regulator WP_034954579.1 (MULTISPECIES)	243
GOX_RS03265	Flagellar MS-ring protein	- (pseudo)	450,476	448,798	-	450,485	448,806	559	Flagellar MS-ring protein WP_024716534.1 (MULTISPECIES)	559
GOX_RS03760	Sugar transporter	- (pseudo)	551,002	552,228	-	550,975	552,072	365	Sugar transporter WP_062448510.1 (*Gluconobacter oxydans* LMG 1399)	408
GOX_RS03905	Cobalt transporter	- (pseudo)	591,539	590,368	-	591,548	590,376	390	Cobalt transporter WP_051391295.1 (MULTISPECIES)	390
GOX_RS04040	Aldehyde dehydrogenase	- (pseudo)	619,584	620,875	-	619,552	620,886	444	Aldehyde dehydrogenase WP_024716853.1 (MULTISPECIES)	444

GOX_RS04300	Hypothetical protein	- (pseudo)	685,331	683,744	-	685,343	683,934	469	Hypothetical protein WP_062458192.1 (*Gluconobacter oxydans* LMG 27012)	529
GOX_RS04330	Histidine kinase	WP_011252221.1	694,634	691,002	1,209	694,643	691,014	1,209	Histidine kinase WP_051391302.1 (MULTISPECIES)	1209
GOX_RS04675	Glycosyl transferase	- (pseudo)	776,587	773,634	-	776,598	773,644	984	Glycosyl transferase WP_024716641.1 (*Gluconobacter oxydans* DSM 2003)	984
GOX_RS04705	Hypothetical protein	- (pseudo)	784,404	785,859	-	784,415	785,869	484	Hypothetical protein WP_024716643.1 (*Gluconobacter oxydans* DSM 2003)	484
GOX_RS04945	Chemotaxis protein	- (pseudo)	834,784	833,223	-	834,795	833,233	520	Chemotaxis protein WP_024716656.1 (MULTISPECIES)	520
GOX_RS05015	PhoB family transcriptional regulator	- (pseudo)	846,023	845,255	-	846,033	845,266	255	DNA-binding response regulator WP_024716661.1 (MULTISPECIES)	255

GOX_RS05065	Peptidase M1	- (pseudo)	859,609	857,607	-	859,617	857,617	666	Peptidase M1 WP_024716664.1 (MULTISPECIES)	666
GOX_RS05130	Cell division protein FtsK	- (pseudo)	871,307	873,988	-	871,315	873,996	893	Cell division protein FtsK WP_051391300.1 (*Gluconobacter oxydans* DSM 2003)	893
GOX_RS05185	PTS system transporter subunit IIA	- (pseudo)	882,862	882,423	-	882,871	882,431	146	PTS sugar transporter subunit IIB WP_024716673.1 (MULTISPECIES)	146
GOX_RS05325	Hypothetical protein	WP_011252410.1	908,818	909,204	128	908,834	909,211	125	Hypothetical protein WP_024716682.1 (MULTISPECIES)	125
GOX_RS05700	Mechanosensitive ion channel protein MscS	- (pseudo)	991,389	993,817	-	991,396	993,825	809	Mechanosensitive ion channel protein MscS WP_051391306.1 (*Gluconobacter oxydans* DSM 2003)	809

GOX_RS07115 (no Frameshift)/ GOX_RS07120 (Frameshift)	Hypothetical protein/ Hypothetical protein	- (pseudo) / WP_041242737.1	1,301,088 1,302,326	1,302,191 1,303,345	- 339	1,301,096	1,303,354	752	Putative phosphatase AHK71176.1 (*Gluconobacter oxydans* DSM 3504)	752
GOX_RS07200	Folylpolyglutamate synthase	- (pseudo)	1,315,934	1,317,257	-	1,315,943	1,317,265	440	Folylpolyglutamate synthase WP_024717023.1 (MULTISPECIES)	440
GOX_RS07310	Hypothetical protein	- (pseudo)	1,335,152	1,336,285		1,335,160	1,336,299	379	Hypothetical protein WP_024717030.1 (*Gluconobacter oxydans* DSM 2003)	379
GOX_RS07480	Hypothetical protein	- (pseudo)	1,375,184	1,374,035	-	1,375,192	1,374,044	382	Membrane protein WP_024717038.1 (MULTISPECIES)	382
GOX_RS07655	Amidase	- (pseudo)	1,417,303	1,418,638	-	1,417,311	1,418,645	444	Amidase WP_024717045.1 (MULTISPECIES)	444
GOX_RS08000	Transcriptional regulator	- (pseudo)	1,493,432	1,492,826	-	1,493,438	1,492,833	201	DNA-binding response regulator WP_024716615.1 (MULTISPECIES)	201

GOX_RS09035	Hypothetical protein	- (pseudo)	1,706,653	1,704,877	-	1,706,659	1,704,884	591	Hypothetical protein WP_024716932.1 (*Gluconobacter oxydans* DSM 2003)	591
GOX_RS09385	FAD-dependent oxidoreductase	- (pseudo)	1,783,883	1,785,491	-	1,783,890	1,785,497	535	FAD-dependent oxidoreductase WP_024717189.1 (MULTISPECIES)	535
GOX_RS09465	Hypothetical protein	WP_011253168.1	1,804,467	1,803,838	209	1,804,842	1,803,844	332	Hypothetical protein WP_024716934.1 (MULTISPECIES)	332
GOX_RS09575	ATP12 chaperone protein	- (pseudo)	1,824,969	1,825,665	-	1,824,975	1,825,670	231	ATP12 chaperone protein WP_024716939.1 (*Gluconobacter oxydans* DSM 2003)	231
GOX_RS09830	NAD (FAD)-utilizing dehydrogenase	- (pseudo)	1,879,755	1,878,533	-	1,879,761	1,878,538	407	NAD (FAD)-utilizing dehydrogenase WP_024716951.1 (MULTISPECIES)	407
GOX_RS09905	Ribonuclease E	- (pseudo)	1,897,362	1,894,569	-	1,897,367	1,894,575	930	Ribonuclease E WP_024716814.1 (*Gluconobacter oxydans* DSM 2003)	930

GOX_RS09980	Rubredoxin-NAD(+) reductase	- (pseudo)	1,916,021	1,914,451	-	1,914,456	1,916,027	523	Rubredoxin-NAD(+) reductase WP_024716817.1 (MULTISPECIES)	523
GOX_RS10030	Transcriptional regulator	- (pseudo)	1,925,374	1,925,984	-	1,925,380	1,925,991	203	SMC-Scp complex subunit ScpB WP_024716820.1 (MULTISPECIES)	203
GOX_RS10075	Hypothetical protein	WP_011253285.1	1,935,349	1,937,310	653	1,935,356	1,937,395	679	Hypothetical protein WP_024716822.1 (MULTISPECIES)	679
GOX_RS10175	Carbonic anhydrase	- (pseudo)	1,957,630	1,956,946	-	1,957,637	1,956,954	227	Carbonic anhydrase WP_024716826.1 (MULTISPECIES)	227
GOX_RS10475	Hypothetical protein	- (pseudo)	2,015,877	2,014,919	-	2,015,884	2,014,925	319	Hypothetical protein WP_024716846.1 (*Gluconobacter oxydans* DSM 2003)	319
GOX_RS10790	DNA helicase	- (pseudo)	2,090,759	2,087,839	-	2,090,768	2,087,847	973	DNA helicase WP_024717077.1 (*Gluconobacter oxydans* DSM 2003)	973

GOX_RS10885	Hypothetical protein	WP_011253441.1	2,111,673	2,112,101	142	2,111,682	2,112,110	142	Hypothetical protein WP_011253441.1 (MULTISPECIES)	142
GOX_RS10980	Hypothetical protein	- (pseudo)	2,135,091	2,132,420	-	2,135,102	2,132,430	890	Hypothetical protein WP_024716757.1 (*Gluconobacter oxydans* DSM 2003)	890
GOX_RS11115	Hypothetical protein	- (pseudo)	2,155,945	2,156,705	-	2,155,956	2,156,717	253	Hypothetical protein WP_024716767.1 (MULTISPECIES)	253
GOX_RS11305	Hemolysin C	- (pseudo)	2,197,564	2,196,618	-	2,197,577	2,196,630	315	Hemolysin C WP_024716779.1 (MULTISPECIES)	315
GOX_RS11470	Hypothetical protein	WP_011253551.1	2,234,539	2,235,333	264	2,234,552	2,235,298	248	Hypothetical protein WP_024716793.1 (MULTISPECIES)	248
GOX_RS11980	MFS transporter	- (pseudo)	2,347,716	2,349,007	-	2,349,024	2,347,732	430	MFS transporter WP_024716957.1 (MULTISPECIES)	430
GOX_RS12290	Oleate hydratase	- (pseudo)	2,409,142	2,411,114	-	2,409,160	2,411,133	657	Oleate hydratase WP_034954644.1 (MULTISPECIES)	657

GOX_RS13205	Sodium:proton antiporter	- (pseudo)	2,582,072	2,583,281	-	2,582,092	2583303	403	Sodium:proton antiporter WP_024716906.1 (*Gluconobacter oxydans* DSM 2003)	403
GOX_RS13575 (Frameshift)/ GOX_RS13580 (Frameshift)	Hypothetical protein	WP_011253936.1 / WP_011253937.1	2,648,171 2,649,628	2,647,380 2,648,303	263 441	2,652,080	2,650,038	680	Hypothetical protein WP_024716717.1 (Gluconobacter oxydans DSM 2003)	680
GOX_RS00055 (P1; Frameshift)/ GOX_RS13940 (P1; no Frameshift)	Hypothetical protein/ Hypothetical protein	WP_011251397.1 / WP_049750700.1	9,239 10,279	10,267 11,145	342 288	9,239	11,146	635	Hypothetical protein WP_024716805.1 (*Gluconobacter oxydans* DSM 2003)	635
GOX_RS00555 (P1)	DotG	WP_011251497.1	112,697	113,290	197	112,162	113,292	376	Hypothetical protein WP_061510690.1 (*Gluconobacter thailandicus*)	417
GOX_RS00760 (P1)	Transcriptional initiation protein Tat	- (pseudo)	154,354	155,153	-	154,357	155,154	265	Transcriptional initiation protein Tat WP_010516738.1 (MULTISPECIES)	265

GOX_RS13955 (P1; Frameshift)/ GOX_RS13960 (P1; no Frameshift)	Hypothetical protein	WP_011251462.1 / WP_049750704	81,359 81,832	81,835 82,449	158 205	81,361	82,452	363	Hypothetical protein WP_064275438.1 (*Gluconobacter cerinus*)	363
GOX_RS13965 (P1; no Frameshift) GOX_RS13970 (P1; Frameshift)	Hypothetical protein	WP_049750705.1 / WP_049750706.1	97,729 98,339	96,980 97,902	249 145	98,342	96,984	452	Hypothetical protein WP_024716505.1 (*Gluconobacter oxydans* DSM 2003)	452

P1 variant detected in the sequence of pGOX1 (NC_006672)

SUPPLEMENTARY FIGURES

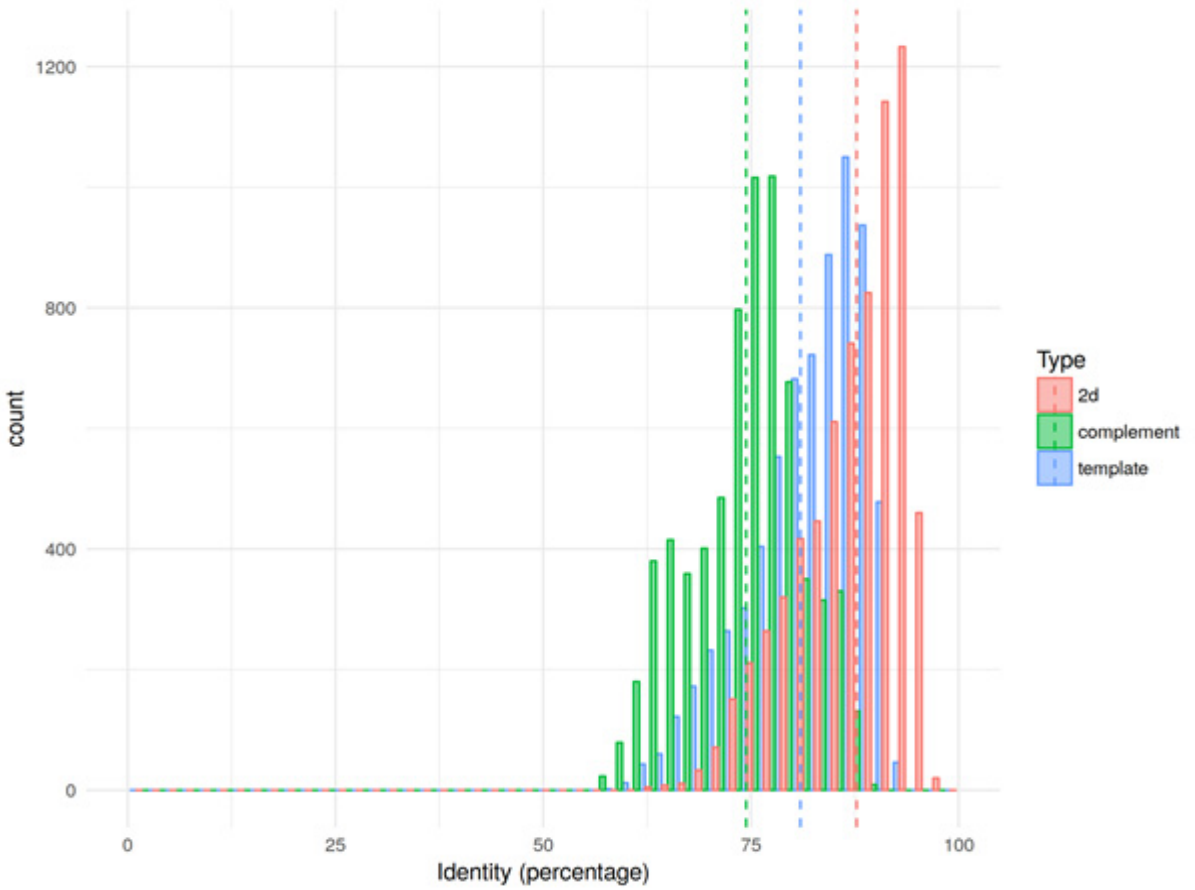

Fig. S1. Identity for R9 chemistry nanopore reads
Read identity (per aligned bases including InDels) for nanopore reads mapped back to the updated genome reference of *Gluconobacter oxydans* 621H using BWA-MEM (v0.7.15-r1140, arXiv:1303.3997). Values are shown for pass-filter 2D reads as well as their consisting template and complement reads. Dotted lines indicate mean identity of the corresponding read type (2D: 87.74, template: 80.97, complement: 74.44).

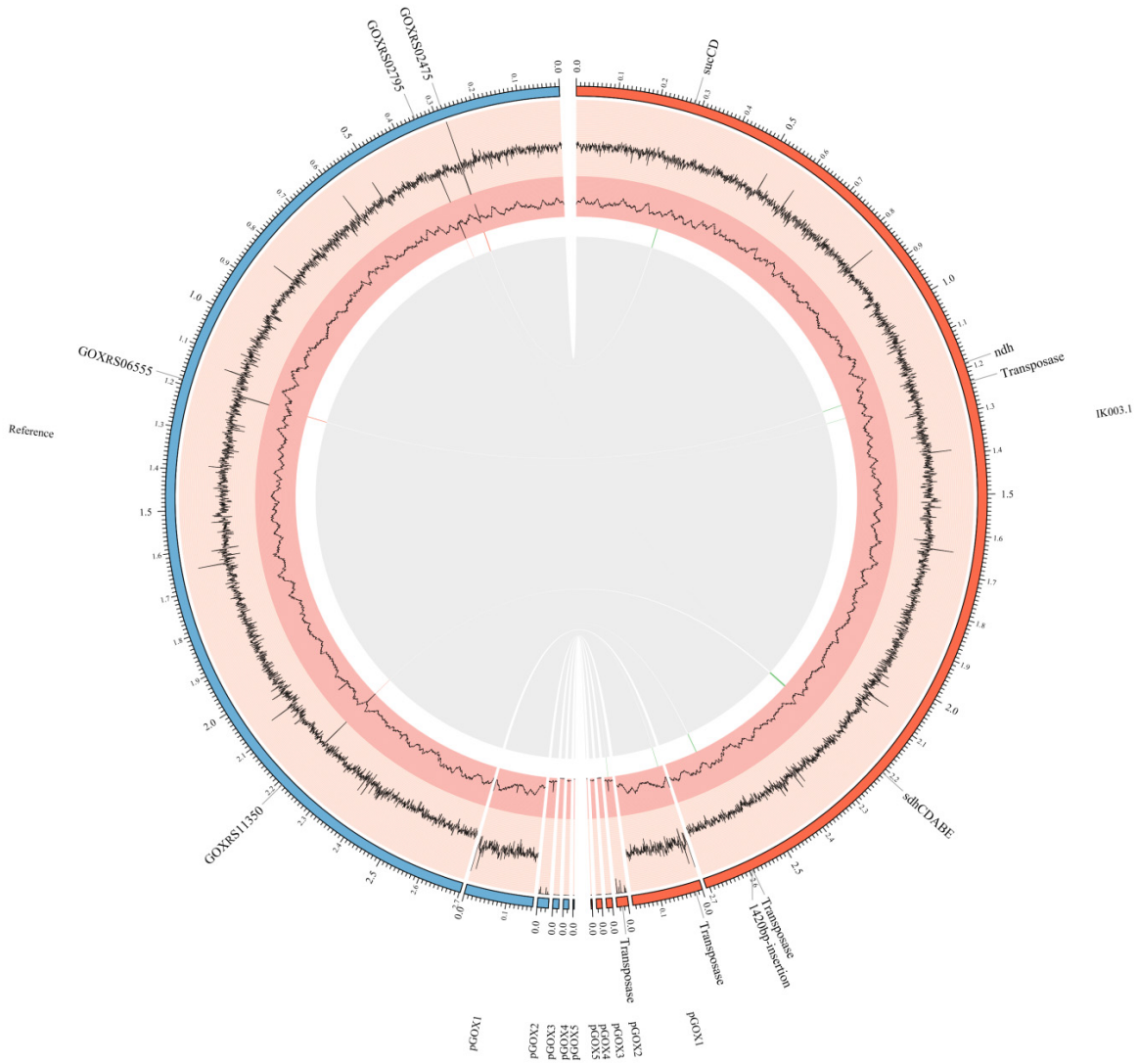

Fig. S2. Circos visualization of genome comparison.
Genome-scale comparison between the mappings to the *G. oxydans* wild type reference (left) and the engineered IK003.1 strain (right) by LAST alignment revealed the four known deletions of *gdhM* (GOX_RS02475), *upp* (GOX_RS02795), *pdc* (GOX_RS06555) and *gdhS* (GOX_RS11350) marked in red and the three known heterologous gene insertions of *sucCD*, *ndh*, and *sdhCDABE* marked in green. The four transposase labels indicate the additional insertions and the novel 1420 bp sequence according to Table 5. Illumina reads (bright red) and long nanopore reads (dark red) were aligned back to the assembly and the reference and are visualized as average coverage over a 500 bp window.

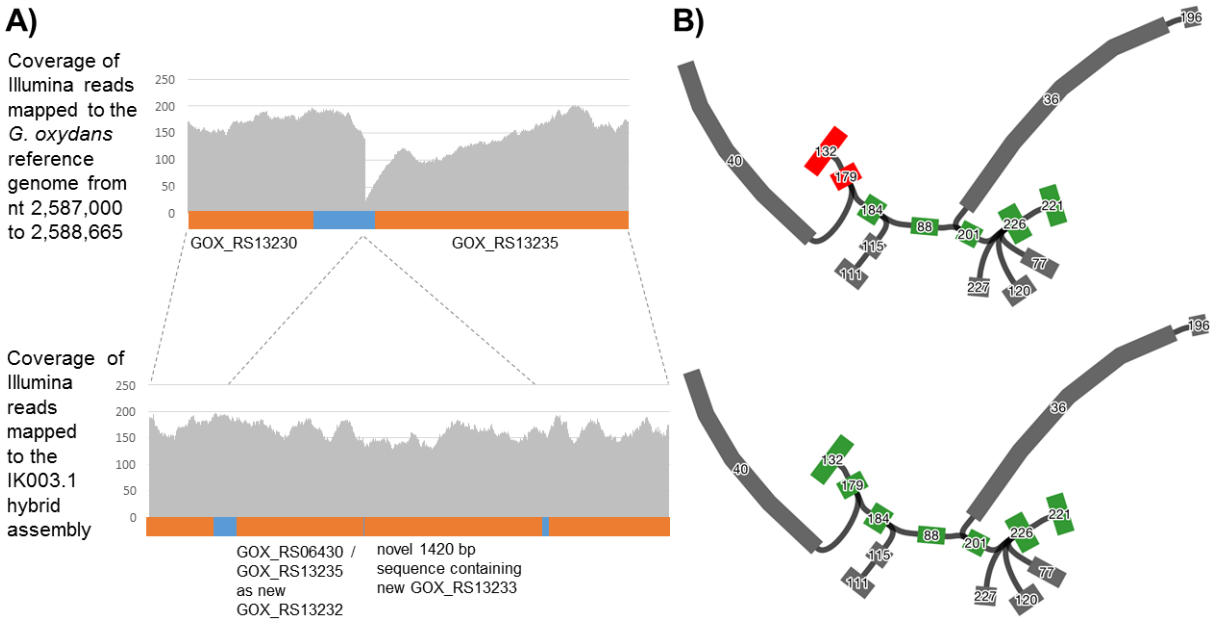

Fig. S3. Resolving assembly ambiguity of the novel 1420 bp sequence insertion at nt 2,587,655 framed by an transposon repeat using nanopore long reads.
A) Alignment of short Illumina reads to the *G. oxydans* 621H reference is unable to resolve the repeat structure and novel sequence insertion (upper panel). Short read de-novo assembly provided incomplete assembly of the region (not shown). The hybrid assembly using the short reads and the long nanopore reads completely resolved the full length structure of the repeats with the novel 1420 bp sequence insertion (lower panel). 25 long reads anchor the whole region with an overlap of least 400 bp to the flanking regions of the *G. oxydans* chromosome. The two CDS ( ) in the sequence present in the intergenic region ( ) of GOX_RS13230 and GOX_RS13235 were assigned new locus tags (www.gluconobacterfactory.de).
B) A selected part of the SPAdes assembly graph using the short Illumina reads schematically represents the ambiguity of the region at nt 2,587,655 due to transposase sequences present at several positions in the genome (upper panel). This assembly was unable to bridge node 184 and node 179 (green and red indicate two different contigs in the assembly). The hybrid assembly approach enabled the resolution of the transposon and the 1420 bp sequence insertion (lower panel). Visualization of assembly graphs by the Bandage tool (Wick et al., 2015).

REFERENCES

Kiefler, I., Bringer, S., Bott, M., (2017) Metabolic engineering of *Gluconobacter oxydans* 621H for increased biomass yield.

Wick, R.R., Schultz, M.B., Zobel, J., Holt, K.E., (2015) Bandage: interactive visualization of de novo genome assemblies. Bioinformatics 31, 3350-3352.

4.2 Supplementary data: Global RNA decay and 23S rRNA fragmentation in *G. oxydans* 621H

Angela Kranz[1,2], Andrea Steinmann[1,2], Ursula Degner[1], Aliye Mengus-Kaya[1], Susana Matamouros[1], Michael Bott[1,2] and Tino Polen[1,2,*]

[1)] IBG-1: Biotechnology, Institute of Bio- and Geosciences, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

[2)] The Bioeconomy Science Center (BioSC), c/o Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

SUPPLEMENTARY DATA

FIGURES

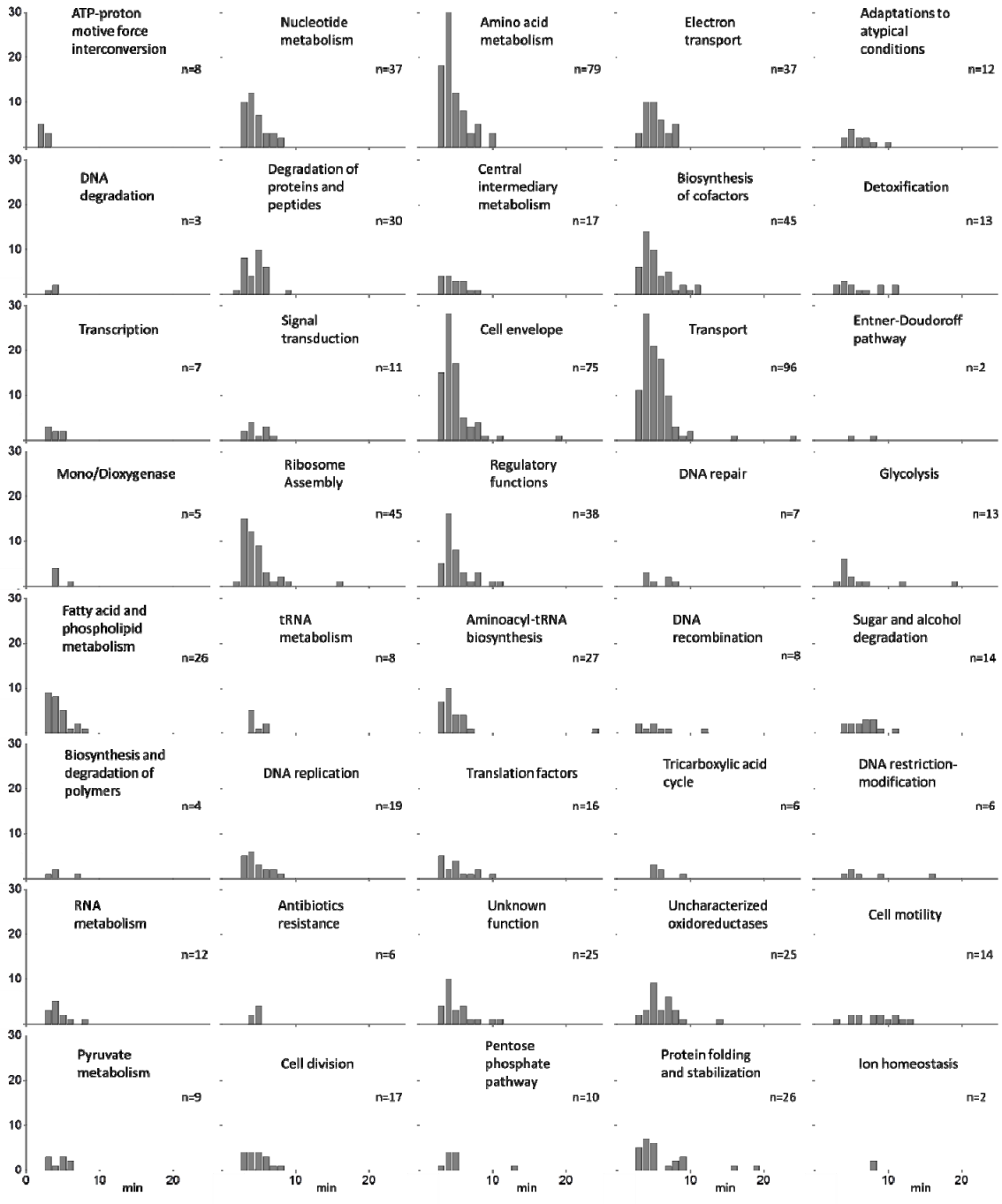

Figure S1. Distribution of mRNA half-lives according to gene product functions [1].

```
Ec_rrlA                 GGTTAAGCGACT------AA-GCGTACACGGTGGATGCCCTGGCAGTCAGAGGCGATGAA   53
St_rrnH                 -CTAAGCGTACACGGTGGAT-GCCCTGGCAGTCAGAGGCGATGAAGGGCGTGCTAATCTG   58
Rs_RSP_4295             --------TCAAGCGCGAAAAGGGCGTTTGGTGGATGCCTAGGCAGCAAGAGGCGATGAA   52
Go_GOX1319              ---------------GAGAAGGGCGTTCGGTGGATGCCTTGGCACTAAGAGGCGATGAA   44
Go_GOX1159              ---------------GAGAAGGGCGTTCGGTGGATGCCTTGGCACTAAGAGGCGATGAA   44
Ar_ARAD_RS00940         ------GATTAAGTGTTATAAGGGCATTTAGTGGATGCCTAGGCATGCACAGGCGATGAA   54
Rl_RLEG3_RS21775        ------GATTAAGTGTCGTAAGGGCATTTGGTGGATGCCTTGGCATGCACAGGCGATGAA   54
Rp_RpaI_R0046           ------AATCAAGTGCCTTAAGGGTGTTCGACGGATGCCTTGGCGCTGAGAGGCGATGAA   54
Bj_BJ6T_RS07380         -----CAATCAAGTGCCTTAAGGGTGTTCGGTGGATGCCTTGGCGCTGAGAGGCGATGAA   55
                                          *             * *   *       *   **

Ec_rrlA                 GGACGTGCTAATCTGCGATAAGCGTCGGTAAGGTGATATGAACCGTTATAACCGGCGATT   113
St_rrnH                 CGATAA---------------GCGCCGGTAAGGTGATATGAACCGTTATAACCGGCGATA   103
Rs_RSP_4295             GGACGTGATACCCTGCGTTAAGCCATGGGGAGCCGGGAATGGGCTTTGATCCA-TGGATG   111
Go_GOX1319              GGACGTGGTACGCTGCGAAAAGCCATGGGGAGCCGCGAACAGGCTTTGATCCG-TGGATG   103
Go_GOX1159              GGACGTGGTACGCTGCGAAAAGCCATGGGGAGCCGCGAACAGGCTTTGATCCG-TGGATG   103
Ar_ARAD_RS00940         GGACGTGATACGCTGCGAAAAGCCGTGGGGAGCTGCGAATAAGCTTTGATCCA-TGGATC   113
Rl_RLEG3_RS21775        GGACGTGATACGCTGCGAAAAGCCGTGGGGAGCTGCGAATGAGCTTTGATCCA-TGGATC   113
Rp_RpaI_R0046           GGACGTGCTACGCTGCGATAAGCCATGGGGAGCTGCGAAGAAGCTTTGATCCG-TGGATT   113
Bj_BJ6T_RS07380         GGACGTGCTACGCTGCGATAAGCCGTGGGGAGCTGCGAAGAAGCTTTGATCCA-TGGATT   114
                         **                  **   **  **  *  *     * **    *    ***

Ec_rrlA                 TCCGAATGGGGAAACCCA-------------------------------------------   131
St_rrnH                 CCCGAATGGGGAAACCCA-------------------------------------------   121
Rs_RSP_4295             TCCGAATGGGGAAACCCACCTGACATTCTGCTATTGTTATCCAACGGATATCG-------   164
Go_GOX1319              TCCGAATGGGGCAACCCCCTCGCAAGA---------------------------------   130
Go_GOX1159              TCCGAATGGGGCAACCCCCTCGCAAGA---------------------------------   130
Ar_ARAD_RS00940         TCCGAATGGGGCAACCCACCTTAAATGCTTAGAAAATCCAAACTGCTGCATAGCA-----   168
Rl_RLEG3_RS21775        TCCGAATGGGGCAACCCACCTTAAATGCTTGGAAAATCATTCTGGTTGTCTGGCCTT---   170
Rp_RpaI_R0046           TCCGAATGGGGAAACCCACCTTCGATAGCCGGAACTCCAAGGCCTTGCGTTTCGA-----   168
Bj_BJ6T_RS07380         TCCGAATGGGGAAACCCACCTTCGATAGCCGGAACTCCAAGACCTTTGTCGAAAGACATC   174
                         ********** *****

Ec_rrlA                 -------------------------------------------------------------   131
St_rrnH                 -------------------------------------------------------------   121
Rs_RSP_4295             -------------------------------------------------------------   164
Go_GOX1319              -------------------------------------------------------------   130
Go_GOX1159              -------------------------------------------------------------   130
Ar_ARAD_RS00940         ------------------------------------------------------------G   169
Rl_RLEG3_RS21775        ----------------------------------------------TGGCCCGGCAACCC   184
Rp_RpaI_R0046           ----------------------------------------------AAGAGACGTGAGGG   182
Bj_BJ6T_RS07380         GGTGTGGGGTTCGATCAGATGATGTGAGAAGCCAGGCCTTTAGATTTCGATCGAAGAGGT   234

Ec_rrlA                 -------------GTGTGTTTCGACACACTATCATTAACTGAATCCATAGGTTAATGAGG   178
St_rrnH                 -------------GTGTGATTCGTCACACTATCATTAACTGAATCCATAGGTTAATGAGG   168
Rs_RSP_4295             ---------ATAGTGGGGTGAGACAGGTATCTTAACCCTGAATACATAGGGGTTTTGAAG   215
Go_GOX1319              ----------------------------GGATCATGCACTGAATACATAGGTGTATGAGG   162
Go_GOX1159              ----------------------------GGATCATGCACTGAATACATAGGTGTATGAGG   162
Ar_ARAD_RS00940         CTTGGGTTTCTAAGCATTGTGATAAGGTATCTACACCTGAATAAAATAGGGTGTAAGAAG   229
Rl_RLEG3_RS21775        GAGTGGTTTCCAAGCATTGTGATAAGGTATCTACACCTGAAT-ACATAGGGTGTAAGAAG   243
Rp_RpaI_R0046           TTTGGATTTCCGGTTATCAAGAGAAGGTATGAGATCTCTGAATACATAGGAGGTTTCAAG   242
Bj_BJ6T_RS07380         TTTGGATTTCCGGTTATCAAGAGAAGGTATGAGACTTCTGAATACATAGGAGGTTTCAAG   294
                                                                          *      * *

Ec_rrlA                 CGAACCGGGGGAACTGAAACATCTAAGTACCCCGAGGAAAAGAAATCAACCGAGATTCCC   238
St_rrnH                 CGAACCGGGGGAACTGAAACATCTAAGTACCCCGAGGAAAAGAAATCAACCGAGATTCCC   228
Rs_RSP_4295             CGAACCCGGGGAACTGAAACATCTAAGTACCCGGAGGAAAGGAAATCAACAGAGACTCCG   275
Go_GOX1319              CAAACCCGGGGAACTGAAACATCTCAGTACCTGGAGGAAAAGACATCAACAGAGATTCCG   222
Go_GOX1159              CAAACCCGGGGAACTGAAACATCTCAGTACCTGGAGGAAAAGACATCAACAGAGATTCCG   222
Ar_ARAD_RS00940         CGAACGCAGGGAACTGAAACATCTAAGTACCTGCAGGAAAGGACATCAACCGAGACTCCG   289
Rl_RLEG3_RS21775        CGAACGCAGGGAACTGAAACATCTAAGTACCTGCAGGAAAGGACATCAACCGAGACTCCG   303
Rp_RpaI_R0046           CGAACCCAGGGAACTGAAACATCTAAGTACCTGGAGGAAAGGACATCAACCGAGACTCCG   302
Bj_BJ6T_RS07380         CAAACCCAGGGAACTGAAACATCTAAGTACCTGGAGGAAAGGACATCAACAGAGACTCCG   354
                        * ***   *************** ******   ****** ** ****** **** ***

Ec_rrlA                 CCAGTAGCGGCGAGCGAACGGGGAGCAGCCCAGAGCCTGAATCAGTGTGTGTGTTAGTGG   298
St_rrnH                 CCAGTAGCGGCGAGCGAACGGGGAGGAGCCCAGAGCCTGAATCAGCATGTGTGTTAGTGG   288
```

```
Rs_RSP_4295          CTAGTAGTGGCGAGCGAACGCGGACCAGCCGATCTCCGAAGAGTG----------ACTGG   325
Go_GOX1319           CTAGTAGTGGCGAGCGAACGCGGAGCAGGCCAATGCCTGATCAGGAAGA-----AGCTGA   277
Go_GOX1159           CTAGTAGTGGCGAGCGAACGCGGAGCAGGCCAATGCCTGATCAGGAAGA-----AGCTGA   277
Ar_ARAD_RS00940      CAAGTAGTGGCGAGCGAACGCGGACCAGGCCAGTGGCAATGAGTGTTAA-----AGTGGA   344
Rl_RLEG3_RS21775     CAAGTAGTGGCGAGCGAACGCGGACCAGGCCAGTGGCAATTGTGATTAA-----AGTGGA   358
Rp_RpaI_R0046        CTAGTAGTGGCGAGCGAACGCGGACCAGGCCAGTCATCATTGGAAGACA-----ATTGGA   357
Bj_BJ6T_RS07380      TTAGTAGTGGCGAGCGAACGCGGACCAGGCCAGTGATACATCAAAGACA-----ATCGGA   409
                       ***** ************ ***  ** * *                          *

Ec_rrlA              AAGCGTCTGGAAAGGCGTGCGATACAGGGTGACAGCCCCGTACACAAAAATGCACATGCT   358
St_rrnH              AAGCGTCTGGAAAGGCGCGCGATACAGGGTGACAGCCCCGTACACAAAAGCGCATGTGCT   348
Rs_RSP_4295          AATGGCCTGGAAAGGCCAGCCACAGCGGGTGACAGCCCCGTACAGGAAG------CTCCA   379
Go_GOX1319           -ACGGTCTGGAAAGTCCGGCAAGAATGGGTGATAGCCCCGTAAGCGTAG-----TGTTCT   331
Go_GOX1159           -ACGGTCTGGAAAGTCCGGCAAGAATGGGTGATAGCCCCGTAAGCGTAG-----TGTTCT   331
Ar_ARAD_RS00940      -AGAACCTGGAAAGGTTTGCCGTAGAGGGTGATAGCCCCGTACGCGTAG-----ATACAC   398
Rl_RLEG3_RS21775     -ACGCTCTGGAAAGTGCGGCCGTAGTGGGTGACAGCCCCGTACGCGTAG-----ATATCA   412
Rp_RpaI_R0046        ATCTGTCAGGAAAGCAGAGCCTTAGAGGGTGATAGCCCCGTACAAGTAA-----TTCAAC   412
Bj_BJ6T_RS07380      ACCGGTCAGGAAAGCCGGGCCTCAGAGGGTGATAGCCCCGTACGAGTAA-----TGCGAT   464
                           * ******    **   *  ****** *********     *

Ec_rrlA              GTGAGCTCGATGAGTAGGGCGGGACACGTGGTATCCTGTCTGAATATGGGGGGACCATCC   418
St_rrnH              GTGAGCTCGATGAGTAGGGCGGGACACGTGGTATCCTGTCTGAATATGGGGGGACCATCC   408
Rs_RSP_4295          GGAGACATATCAAGTAGGGCGGGACACGTGAAATCCTGTCTGAAGATCGGGGGACCACCC   439
Go_GOX1319           GATGAGGATTCGAGTAGGGCGGGGCACGTGAAACCCTGTCTGAACATGGGGGGACCACCC   391
Go_GOX1159           GATGAGGATTCGAGTAGGGCGGGGCACGTGAAACCCTGTCTGAACATGGGGGGACCACCC   391
Ar_ARAD_RS00940      TTATTGTCCTAGAGTAGGGCGGGACACGTGAAATCCTGTCTGAACATGGGGCGACCACGC   458
Rl_RLEG3_RS21775     TGATTGTCCTAGAGTAGGGCGGGACACGAGAAATCCTGTCTGAACATGGGGAGACCACTC   472
Rp_RpaI_R0046        CATTGATGCTCGAGTAAGGCGGGACACGTGAAATCCTGTCTGAACATGGGGGGACCACCC   472
Bj_BJ6T_RS07380      GATGTATCCACGAGTAAGGCGGGACACGTGAAATCCTGTCTGAACGCGG-GGGACCACCC   523
                               **** ****** **** *  * **********    * * *****  *

Ec_rrlA              TCCAAGGCTAAATACTCCTGACTGACCGATAGTGAACCAGTACCGTGAGGGAAAGGCGAA   478
St_rrnH              TCCAAGGCTAAATACTCCTGACTGACCGATAGTGAACCAGTACCGTGAGGGAAAGGCGAA   468
Rs_RSP_4295          CCGAAGGCTAAGTACTCCTTGCTGACCGATAGCGAACCAGTACCGTGAGGGAAAGGTGAA   499
Go_GOX1319           TCCAAGCCTAAATACTCCTTAGTGACCGATAGCGAACAAGTACCGTGAGGGAAAGGTGAA   451
Go_GOX1159           TCCAAGCCTAAATACTCCTTAGTGACCGATAGCGAACAAGTACCGTGAGGGAAAGGTGAA   451
Ar_ARAD_RS00940      TCCAAGCCTAAGTACTCGTGCATGACCGATAGCGAACCAGTACCGTGAGGGAAAGGTGAA   518
Rl_RLEG3_RS21775     TCCAAGCCTAAGTACTCGTGCATGACCGATAGCGAACAAGTACCGTGAGGGAAAGGTGAA   532
Rp_RpaI_R0046        TCCAAGCCTAAGTACTCCTCAGCGACCGATAGTGAACCAGTACCGTGAGGGAAAGGTGAA   532
Bj_BJ6T_RS07380      TCCAAGCCTAAGTACTCCTCAGCGACCGATAGTGAACCAGTACCGTGAGGGAAAGGTGAA   583
                      * *** **** ***** *    ********* **** ****************** ***

Ec_rrlA              AAGAACCCCGGCGAGGGGAGTGAAAAAGAACCTGAAACCGTGTACGTACAAGCAGTGGGA   538
St_rrnH              AAGAACCCCGGCGAGGGGAGTGAAAAAGAACCTGAAACCGTGTACGTACAAGCAGTGGGA   528
Rs_RSP_4295          AAGCACCCCGACGAGGGGAGTGAAACAGTACCTGAAACCGGACGCCTACAAGCAGTCGGA   559
Go_GOX1319           AAGCACCCCGATGAGGGGAGTGAAAGAGAC-CTGAAACCGGACGCCTACAAGCAGTCG--   508
Go_GOX1159           AAGCACCCCGATGAGGGGAGTGAAAGAGAC-CTGAAACCGGACGCCTACAAGCAGTCG--   508
Ar_ARAD_RS00940      AAGCACCCCGACAAGGGGAGTGAAATAGAACCTGAAACTGGATGCCTACAAACAGTCGGA   578
Rl_RLEG3_RS21775     AAGCACCCCGACAAGGGGAGTGAAATAGAACCTGAAACCGGATGCCTACAAACAGTCGGA   592
Rp_RpaI_R0046        AAGCACCCCGACGAGGGGAGTGAAATAGTTCCTGAAATCGGACACCTACAAACAGACGGA   592
Bj_BJ6T_RS07380      AAGCACCCCGACGAGGGGAGTGAAATAGA-CCTGAAACCGGACACCTACAAACAGATGGA   642
                     *** ******   ************ **   ******  *    * ***** ***  *

Ec_rrlA              GCACG-------------------------------------------------------   543
St_rrnH              GCCCCACCACTAAGCCAGTGGTGAACTCCACATCCGCATCCTTTGCTGAGGATACGGTTA   588
Rs_RSP_4295          GGG-----T---------------------------------------------------   563
Go_GOX1319           ----GAGCC---------------------------------------------------   513
Go_GOX1159           ----GAGCC---------------------------------------------------   513
Ar_ARAD_RS00940      GCCCG-------------------------------------------------------   583
Rl_RLEG3_RS21775     GCCCG-------------------------------------------------------   597
Rp_RpaI_R0046        GCCCAAGAT---------------------------------------------------   601
Bj_BJ6T_RS07380      GCCCAAGAT---------------------------------------------------   651

Ec_rrlA              -----------------------------------------------CTTAGGCGTGTGAC   557
St_rrnH              ACGGAGCGAAAGCGACGTTCAACCGCAAAACAAGCAGAGGGGGCTTAGTGGTGGGGTGAC   648
Rs_RSP_4295          -----------------------------------------------CCATGAGACCTGAC   577
Go_GOX1319           -----------------------------------------------TCTTATGGGGTGAC   527
Go_GOX1159           -----------------------------------------------TCTTATGGGGTGAC   527
```

```
Ar_ARAD_RS00940      ----------------------------------------------------CAAGGGTGAC   593
Rl_RLEG3_RS21775     ----------------------------------------------------CAAGGGTGAC   607
Rp_RpaI_R0046        ------------------------------------------------ACGTTCTGGGTGAC   615
Bj_BJ6T_RS07380      ------------------------------------------------ACGTTCTGGGTGAC   665
                                                                        ****

Ec_rrlA              TGCGTACCTTTTGTATAATGGGTCAGCGACTTATATTCTGTAGCAAGGTTAACCGAATAG   617
St_rrnH              TGCGTACCTTTTGTATAATGGGTCAGCGACTTATATTCTGTAGCAAGGTTAACCGAATAG   708
Rs_RSP_4295          GGCGTACCTTTTGTATAATGGGTCAACGACTTGGTCTCACGAGCAAGCTTAAGCCGGTAG   637
Go_GOX1319           GGCGTACCTTTTGTATAATGGGTCAGCGAGTTTCTGTTTGCAGCGAGCTTAAGCCGTTAG   587
Go_GOX1159           GGCGTACCTTTTGTATAATGGGTCAGCGAGTTTCTGTTTGCAGCGAGCTTAAGCCGTTAG   587
Ar_ARAD_RS00940      GGCGTACCTTTTGTATAATGGGTCAACGACTTAGTGTAACAAGCAAGCTTAAGCCGGTAG   653
Rl_RLEG3_RS21775     GGCGTACCTTTTGTATAATGGGTCAACGACTTAGTGTAACAAGCAAGCTTAAGCCGGTAG   667
Rp_RpaI_R0046        GTCGTACCTTTTGTATTATGGGCCAGCGACTTAATTTAACGAGCAAGCTTAAGCCGGTAG   675
Bj_BJ6T_RS07380      ATCGTACCTTTTGTATTATGGGCCAGCGACTTAATTTAACGAGCAAGCTTAAGCCGATAG   725
                       ************** ***** ** *** **    *    *** ** **** *   ***

Ec_rrlA              -GGGAGCCGAAGGGAAACCGAGTCTTAACTGGGCGT-TAAGTTGCAGGGTATAGACCCGA   675
St_rrnH              -GGGAGCCGGAGGGAAACCGAGTCTTAATTGGGCGT-TAAGTTGCAGGGTATAGACCCGA   766
Rs_RSP_4295          GTGGAGGCGCAGCGAAAGCGAGTCTTAAAAGGGCGT-CGAGTTCGTGGGATCAGACCCGA   696
Go_GOX1319           GTGTAGGCGTAGCGAAAGCGAGTCTGAATAGGGCGAATGAGTTGCTGGCAGAAGACCCGA   647
Go_GOX1159           GTGTAGGCGTAGCGAAAGCGAGTCTGAATAGGGCGAATGAGTTGCTGGCAGAAGACCCGA   647
Ar_ARAD_RS00940      GTGTAGGCGAAGCGAAAGCGAGTCTGAATAGGGCGATTTAGTTTGTTGCATTAGACCCGA   713
Rl_RLEG3_RS21775     GTGTAGGCGAAGCGAAAGCGAGTCTGAATAGGGCGATATAGTTTGTTGCATTAGACCCGA   727
Rp_RpaI_R0046        GTGTAGGCGCAGCGAAAGCGAGTCTGAATAGGGCGT-CAAGTTCGTTGGATTAGACCCGA   734
Bj_BJ6T_RS07380      GCGAAGGCGCAGCGAAAGCGAGTCTGAATAGGGCGT-CAAGTTCGTTGTATTAGACCCGA   784
                       * ** ** ** **** ******* **  *****    ****    *    ********

Ec_rrlA              AACCCGGTGATCTAGCCATGGGCAGGTTGAAGGTTGGGTAACACTAACTGGAGGACCGAA   735
St_rrnH              AACCCGGTGATCTAGCCATGGGCAGGTTGAAGGTTGGGTAACACTAACTGGAGGACCGAA   826
Rs_RSP_4295          AACCAGGTGATCTAGCCATGAGCAGGATGAAGTCAGGGTAACACCTGATGGAGGTCCGAA   756
Go_GOX1319           AACCGAGTGATCTAGCCATGGCCAGGCTGAAGGTGCGGTAACACGCACTGGAGGGCCGAA   707
Go_GOX1159           AACCGAGTGATCTAGCCATGGCCAGGCTGAAGGTGCGGTAACACGCACTGGAGGGCCGAA   707
Ar_ARAD_RS00940      AACCGAGTGATCTAGCCATGAGCAGGTTGAAGGTTGGGTAACACCAACTGGAGGACCGAA   773
Rl_RLEG3_RS21775     AACCGAGTGATCTAGCCATGAGCAGGTTGAAGGTTGGGTAACACCAACTGGAGGACCGAA   787
Rp_RpaI_R0046        AACCTAGTGATCTAGCCATGAGCAGGTTGAAGGTGAGGTAACACTCACTGGAGGACCGAA   794
Bj_BJ6T_RS07380      AACCTAGTGATCTAGCCATGAGCAGGTTGAAGGTGAGGTAACACTCACTGGAGGACCGAA   844
                     ****  **************  **** *****    ********    ****** *****

Ec_rrlA              CCGACTAATGTTGAAAAATTAGCGGATGACTTGTGGCTGGGGGTGAAAGGCCAATCAAAC   795
St_rrnH              CCGACTAATGTTGAAAAATTAGCGGATGACCTGTGGCTGGGGGTGAAAGGCCAATCAAAC   886
Rs_RSP_4295          CCAACACCCGTTGAAAAGGGTCTGGATGACTTGTGGCTAGGGGTGAAAGGCCAATCAAAC   816
Go_GOX1319           CCCACGCCTGTTGAAAAAGTCGGGGATGAGCTGTGGCTAGGGGTGAAAGGCCAATCAAAC   767
Go_GOX1159           CCCACGCCTGTTGAAAAAGTCGGGGATGAGCTGTGGCTAGGGGTGAAAGGCCAATCAAAC   767
Ar_ARAD_RS00940      CCCGCATCTGTTGCAATAGATTGGGATGACTTGTGGCTAGGGGTGAAAGGCCAATCAAAC   833
Rl_RLEG3_RS21775     CCCGCATCTGTTGCAATAGATTGGGATGACTTGTGGCTAGGGGTGAAAGGCCAATCAAAC   847
Rp_RpaI_R0046        CGGGTGCCTGTTGAAAAAGGCTCCGATGACTTGTGGTTAGGGGTGAAAGGCCAATCAAAC   854
Bj_BJ6T_RS07380      CGGGTGCCTGTTGAAAAAGGCTCCGATGACTTGTGGTTAGGGGTGAAAGGCCAATCAAAC   904
                     *        **** **        *****  ***** * ********************

Ec_rrlA              CGGGAGATAGCTGGTTCTCCCCGAAAGCTATTTAGGTAGCGCCTCGTGAATTCATCTCCG   855
St_rrnH              CGGGAGATAGCTGGTTCTCCCCGAAAGCTATTTAGGTAGCGCCTCGTGAATTCATCTCCG   946
Rs_RSP_4295          CTGGAGATAGCTGGTTCTCCGCGAAAGCTATTTAGGTAGCGCCTCGGACGAATACCTCGG   876
Go_GOX1319           TCGGAGATAGCTGGTTCTCCGCGAAATCTATTGAGGTAGATCGTCGGGTGTTTACCCTGG   827
Go_GOX1159           TCGGAGATAGCTGGTTCTCCGCGAAATCTATTGAGGTAGATCGTCGGGTGTTTACCCTGG   827
Ar_ARAD_RS00940      TCGGAAATAGCTGGTTCTCCGCGAAATCTATTTAGGTAGAGCGTCGAGCGAATACTCCCG   893
Rl_RLEG3_RS21775     TCGGAAATAGCTGGTTCTCCGCGAAATCTATTTAGGTAGAGCGTCGAGCGAATACCCCCG   907
Rp_RpaI_R0046        TGGGAAATAGCTGGTTCTCCGCGAAAGATATTTAGGTATCGCCTCGGACGAATACCTCAG   914
Bj_BJ6T_RS07380      TGGGAAATAGCTGGTTCTCCGCGAAAGATATTTAGGTATCGCCTCGGATGAATACCTCAG   964
                       *** ************** *****  **** *****   * ***       *     *

Ec_rrlA              GGGGTAGAGCACTGTTTCGGCAAGGGGGTCATCCCGACTTACCAACCCGATGCAAACTGC   915
St_rrnH              GGGGTAGAGCACTGTTTCGGCTAGGGGGCCATCCCGGCTTACCAACCCGATGCAAACTGC   1006
Rs_RSP_4295          GGGGTAGAGCACTGCATGGATGATGGGGGCCCACAGCCTTACTGAGTCTAAGCAAACTCC   936
Go_GOX1319           GGGGTAGAGCACTGGATGGGCTAGGGGGGCCCAAAGCCTTACCAAACCTAACCAAACTCC   887
Go_GOX1159           GGGGTAGAGCACTGGATGGGCTAGGGGGGCCCAAAGCCTTACCAAACCTAACCAAACTCC   887
Ar_ARAD_RS00940      GGGGTAGAGCACTGGATGGGCTATGGGGACTCACCGTCTTACTGATCCTAACCAAACTCC   953
Rl_RLEG3_RS21775     GGGGTAGAGCACTGGATGGGCTATGGGGACTCACCGTCTTACTGATCCTAACCAAACTCC   967
Rp_RpaI_R0046        GGGGTAGAGCACTGGATGGGCTAGGGGGACTTACCGTCTTACCAAACCCAACCAAACTCC   974
```

```
Bj_BJ6T_RS07380      GGGGTAGAGCACTGGATGGGCTAGGGGGACTTACCGTCTTACCAAACCCAACCAAACTCC   1024
                     **************  * *   * **** *     * *****  *  * *  ****** *

Ec_rrlA              GAATACCGGAGAATGTTATCACGGGAGACACACGGCGGGTGCTAACGTCCGTCGTGAAGA   975
St_rrnH              GAATACCGGAGAATGTTATCACGGGAGACACACGGCGGGTGCTAACGTCCGTCGTGAAGA   1066
Rs_RSP_4295          GAATACCCGAGAGTACTA-TCCGGGAGACACACGGCGGGTGCTAACGTCCGTCGTGAAGA   995
Go_GOX1319           GAATACCCAGGAGTATAG-CCCGGCAGACAGACAGTGGGTGCTAAGGTCCATTGTCGAGA   946
Go_GOX1159           GAATACCCAGGAGTATAG-CCCGGCAGACAGACAGTGGGTGCTAAGGTCTATTGTCGAGA   946
Ar_ARAD_RS00940      GAATACCGGGAAGTACTA-CTCGGCAGACACACGGCGGGTGCTAACGTCCGTCGTGAAAA   1012
Rl_RLEG3_RS21775     GAATACCGGGGAGTACTA-CTCGGCAGACACACGGCGGGTGCTAACGTCCGTCGTGAAAA   1026
Rp_RpaI_R0046        GAATACCTGAGAGTACTA-TCCGGGAGTCACACGGCGGGTGCTAACGTCCGTCGTGGAGA   1033
Bj_BJ6T_RS07380      GAATACCTGAGAGTACTA-TCCGGGAGTCACACGGCGGGTGCTAACGTCCGTCGTGGAGA   1083
                     *******    * *       *** ** ** ** * ********* ***  * **  * *

Ec_rrlA              GGGAAACAACCCAGACCGCCAGCTAAGGTCCCAAAGTCATGGTTAAGTGGGAAACGATGT   1035
St_rrnH              GGGAAACAACCCAGACCGCCAGCTAAGGTCCCAAAGTCATGGTTAAGTGGGAAACGATGT   1126
Rs_RSP_4295          GGGAAACAACCCTGACCTGCAGCTAAGGCCCCCAATTCGTGGCTAAGTGGGAAAGCATGT   1055
Go_GOX1319           GGGAAACAGCCCAGACCACCAGCTAAGGCCCCTAAATCGTGGCTAAGTGGGAAAGGATGT   1006
Go_GOX1159           GGGAAACAGCCCAGACCACCAGCTAAGGCCCCTAAATCGTGGCTAAGTGGGAAAGGATGT   1006
Ar_ARAD_RS00940      GGGCAACAACCCTAACCTCCAGCTAAGGTCCCCAAGTCATGGCTAAGTGGGAAAGGATGT   1072
Rl_RLEG3_RS21775     GGGCAACAACCCTAACCTCCAGCTAAGGTCCCCAAGTCATGGCTAAGTGGGAAAGGATGT   1086
Rp_RpaI_R0046        GGGAAACAACCCTGACCTACAGCTAAGGCCCCCAATTCGTGGCTAAGTGGGAAAGGATGT   1093
Bj_BJ6T_RS07380      GGGAAACAACCCGGACCTACAGCTAAGGCCCCTAATTCGTGGCTAAGTGGGAAAGGATGT   1143
                     *** **** ***  ***  ********* *** ** ** *** **********  ****

Ec_rrlA              GGGAAGGCCCAGACAGCCAGGATGTTGGCTTAGAAGCAGCCATCATTTAAAGAAAGCGTA   1095
St_rrnH              GGGAAGGCCCAGACAGCCAGGATGTTGGCTTAGAAGCAGCCATCATTTAAAGAAAGCGTA   1186
Rs_RSP_4295          GGGACGGCCAAAACAACCAGGAGGTTGGCTTAGAAGCAGCCATCCTTTAAAGATAGCGTA   1115
Go_GOX1319           GGGGATTCCAAAACAACCAGGAGGTTGGCTTAGAAGCAGCCATCCTTTAAAGAAAGCGTA   1066
Go_GOX1159           GGGGATTCCAAAACAACCAGGAGGTTGGCTTAGAAGCAGCCATCCTTTAAAGAAAGCGTA   1066
Ar_ARAD_RS00940      GAGGATCCCAAAACAACCAGGATGTTGGCTTAGAAGCAGCCATCATTTAAAGAAAGCGTA   1132
Rl_RLEG3_RS21775     GAGGATCCCAAAACAACCAGGATGTTGGCTTAGAAGCAGCCATCATTTAAAGAAAGCGTA   1146
Rp_RpaI_R0046        GGAAATCCCAAAACAACCAGGAGGTTGGCTTAGAAGCAGCCATCCTTTAAAGAAAGCGTA   1153
Bj_BJ6T_RS07380      GGAAATCCCAAAACAACCAGGAGGTTGGCTTAGAAGCAGCCATCCTTTAAAGAAAGCGTA   1203
                     *      ** * *** ****** ********************** ******** ******

Ec_rrlA              ATAGCTCACTGGTCGAGTCGGCCTGCGC-----GGAAGATGTAACGGGGCTAAACCATGC   1150
St_rrnH              ATAGCTCACTGGTCGAGTCGGCCTGCGC-----GGAAGATGTAACGGGGCTAAACCATGC   1241
Rs_RSP_4295          ACAGCTCACTGGTCTAGATAAGCTGTCCTGCGGCGAAGATGTAACGGGGCTCAAGCCACG   1175
Go_GOX1319           ATAGCTCACTGGTCTATTAGAAAC--CCTGCGCCGAAAATGTAACGGGGCTCAAGCCACG   1124
Go_GOX1159           ATAGCTCACTGGTCTATTAGAAAC--CCTGCGCCGAAAATGTAACGGGGCTCAAGCCACG   1124
Ar_ARAD_RS00940      ACAGCTCACTGGTCTAGTCAAGGGTCTTTGCGCCGAAAATGTAACGGGGCTAAAGCCATG   1192
Rl_RLEG3_RS21775     ACAGCTCACTGGTCTAAATAAGGGTCTTTGCGCCGAAAATGTAACGGGGCTGAAGCCATG   1206
Rp_RpaI_R0046        ACAGCTCACTGGTCTAAATAAGGGTTTCTGCGCCGAAGATGTAACGGGGCTCAAGCCACG   1213
Bj_BJ6T_RS07380      ACAGCTCACTGGTCTAAATAAGGGTTTCTGCGCCGAAGATGTAACGGGGCTCAAGCCACG   1263
                     * ************ *                 *** ************* ** *

Ec_rrlA              A-CCGAAGCTGCGGCAGCGACACTATG---------------------------------   1176
St_rrnH              A-CCGAAGCTGCGGCAGCGAACGTATCACCCAAGACAA--------CTTTACGGAGTT--   1290
Rs_RSP_4295          AGCCGAAGCTCAGGATGCACAGCAATGTGCGTGGT----AGCGGAGCGTTCCGTGATATA   1231
Go_GOX1319           TGCCGAAGCTGTGGGTGCATAC---TATGTATGCGCGGTAGCGGAGCGTTCCGTAAGT--   1179
Go_GOX1159           TGCCGAAGCTGTGGGTGCATAC---TATGTATGCGCGGTAGCGGAGCGTTCCGTAAGT--   1179
Ar_ARAD_RS00940      CACCGAAGCTGAGGATTTGCGAG--CAATCGCAAGTGGTAGCGGAGCGTTCCGTAAGC--   1248
Rl_RLEG3_RS21775     CACCGAAGCTGAGGATGTGTAGC----AATACACGTGGTAGCGGAGCGTTCCGTAAGC--   1260
Rp_RpaI_R0046        AGCCGAAGCTTAGGGTGTGCGCA------AGCACGCGGTAGCGGAGCGTTCTGTAAGC--   1265
Bj_BJ6T_RS07380      AGCCGAAGCTTAGGGTGTGATCCGCAAGGGTCACGCGGTAGCGGAGCGTTCTGTAAGC--   1321
                       ********  **

Ec_rrlA              ---------------------------------------------------------   1176
St_rrnH              -------GACGATTGACGGAGCGAAGCGA---CGTCAAAGCGTTCATTAAAGTCGAGTTG   1340
Rs_RSP_4295          GCTCATTGTGTGTTTATCGAACGCACCACCGGTCCGAACGAGGGCACTGCCCTCAAGTAG   1291
Go_GOX1319           ---------------------------------------------------------   1179
Go_GOX1159           ---------------------------------------------------------   1179
Ar_ARAD_RS00940      ---------------------------------------------------------   1248
Rl_RLEG3_RS21775     ---------------------------------------------------------   1260
Rp_RpaI_R0046        ---------------------------------------------------------   1265
Bj_BJ6T_RS07380      ---------------------------------------------------------   1321
```

```
Ec_rrlA              ------------TGTTGTTGGGTAGGGGAGCGTTCTGTAAGCCTGTGAAGGTGTGCTGTG   1224
St_rrnH              GCTTAGGGATACGTTCGTTGGGTAGGGGAGCGTTCTGTAAGCCTGTGAAGGTGTGTCGTG   1400
Rs_RSP_4295          CGAAGCGGTAGGGCGCGCGGTAGCGCACACAAAGAGCTTTCTGTGAAGCCGGGCCGTAAG   1351
Go_GOX1319           ----------------------------------------CTGCGAAGGAGACGGGGTG   1198
Go_GOX1159           ----------------------------------------CTGCGAAGGAGACGGGGTG   1198
Ar_ARAD_RS00940      ----------------------------------------CTGTGAAGGGGTACCTGTG   1267
Rl_RLEG3_RS21775     ----------------------------------------CGATGAAGGGAGACCCGTG   1279
Rp_RpaI_R0046        ----------------------------------------CTGCGAAGGGCGACCCGTG   1284
Bj_BJ6T_RS07380      ----------------------------------------CTGCGAAGGGCGACTCGTG   1340
                                                                             *

Ec_rrlA              AGGCATGCTGGAGGTATCAGAAGTGCGAATGCTGACATAAGTAACGATAAAGCGGGTGAA   1284
St_rrnH              AGGCATGCTGGAGGTATCAGAAGTGCGAATGCTGACATAAGTAACGATAAAGCGGGTGAA   1460
Rs_RSP_4295          GCATCCGGTGGAGAGATCGGAAGCGAGAATGTTGACATGAGTAGCGACAAACAGGGTGAG   1411
Go_GOX1319           ACCCTCTCTGGAGATATCGGAAGTGCGAATGCTGACATGAGTAGCGACAAACAGTGCGAG   1258
Go_GOX1159           ACCCTCTCTGGAGATATCGGAAGTGCGAATGCTGACATGAGTAGCGACAAACAGTGCGAG   1258
Ar_ARAD_RS00940      AGGGGCCCTGGAGGTATCGGAAGTGCGAATGTTGACATGAGTAACGATAAAGAGGGTGAG   1327
Rl_RLEG3_RS21775     AGGGCTCCTGGAGGTATCGGAAGTGCGAATGTTGACATGAGTAACGATAAAGAGGGTGAG   1339
Rp_RpaI_R0046        AGGGCGCCTGGAGGTATCAGAAGTGCGAATGCTGGCATGAGTAACGACAAACACTGTGAA   1344
Bj_BJ6T_RS07380      AGAGCGCCTGGAGGTATCAGAAGTGCGAATGCTGGCATGAGTAACGACAAACACTGTGAA   1400
                             *****  *** **** * ***** ** *** **** *** ***    * **

Ec_rrlA              AAGCCCGCTCGCCGGAAGACCAAGGGTTCCTGTCCAACGTTAATCGGGGCAGGGTGAGTC   1344
St_rrnH              AAGCCCGCTCGCCGGAAGACCAAGGGTTCCTGTCCAACGTTAATCGGGGCAGGGTGAGTC   1520
Rs_RSP_4295          AGACCCTGTCGCCGAAAGTCCAAGGGTTCCTGCTTAAAGCTAATCTGAGCAGGGTAAGCC   1471
Go_GOX1319           AAACACTGTCGCCGAAAGTCCAAGGGTTCCTGCGCAAGGTTAATCCACGCAGGGTGAGCC   1318
Go_GOX1159           AAACACTGTCGCCGAAAGTCCAAGGGTTCCTGCGCAAGGTTAATCCACGCAGGGTGAGCC   1318
Ar_ARAD_RS00940      AGACCCTCTCGCCGAAAGACCAAGGGTTCCTGCTTAAAGTTAATCTGAGCAGGGTTAGCC   1387
Rl_RLEG3_RS21775     AGACCCTCTCGCCGAAAGACCAAGGGTTCCTGCTTAAAGTTAATCTGAGCAGGGTTAGCC   1399
Rp_RpaI_R0046        AGACAGTGTCGCCGAAAGTCCAAGGGTTCCTGCGTAAAGTTAATCTTCGCAGGGTTAGCC   1404
Bj_BJ6T_RS07380      AGACAGTGTCGCCGAAAGTCCAAGGGTTCCTGCGTAAAGTTAATCTTCGCAGGGTTAGCC   1460
                     *  *    ****** *** *************   ** * *****   ******* ** *

Ec_rrlA              GACCCCTAAGGCGAGGCCGAAAGGCGTAGTCGATGGGAAACAGGTTAATATTCCTGTACT   1404
St_rrnH              GACCCCTAAGGCGAGGCCGAAAGGCGTAGTCGATGGGAAACGGGTTAATATTCCCGTACT   1580
Rs_RSP_4295          GGCCCCTAAGGCGAGGCCGAAAGGCGTAGTCGATGGGAACCAGGTTAATATTCCTGGGCC   1531
Go_GOX1319           GGCCCCTAAGGCGAGGGCGAGAGCCGTAGTCGATGGGAACCAGTTCAATATTACTGGGCC   1378
Go_GOX1159           GGCCCCTAAGGCGAGGGCGAGAGCCGTAGTCGATGGGAACCAGTTCAATATTACTGGGCC   1378
Ar_ARAD_RS00940      GGCCCCTAAGACGAGGCGGACACGCGTAGTCGATGGGAACCACGTTAATATTCGTGGGCC   1447
Rl_RLEG3_RS21775     GGCCCCTAAGGCGAGGCAGAAATGCGTAGTCGATGGGAACCACGTTAATATTCGTGGGCC   1459
Rp_RpaI_R0046        GGTCCCTAAGGCGAGGCCGAAAGGCGTAGTCGATGGGAATCACGTGAATATTCGTGAGCC   1464
Bj_BJ6T_RS07380      GGTCCCTAAGGCGAGGCCGAAAGGCGTAGTCGATGGGAATGCAGTGAATATTCTGCAGCC   1520
                     *  ******* *****  ** *  ***************     * ******      *

Ec_rrlA              TGGTGTTACTGCGAAGGGGGGACGGAGAAGGCTATGTTGGCCGGGCGACGGTTGTCCCGG   1464
St_rrnH              TGGTGTTACTGCGAAGGGGGGACGGAGAAGGCTATGTTGGCCGGGCGACGGTTGTCCCGG   1640
Rs_RSP_4295          AGGAGGATGTGACGGATCGCAGGTGTAGT--------------------------TCGGT   1565
Go_GOX1319           TGCCAGAAGTGACGAATGAGAGATGTTGT-------------------------CTGTC   1412
Go_GOX1159           TGCCAGAAGTGACGAATGAGAGATGTTGT-------------------------CTGTC   1412
Ar_ARAD_RS00940      TGGTGGTAGTGACGGATTGCTTAACTTGT-------------------------TCACA   1481
Rl_RLEG3_RS21775     TGGTGGTAGTGACGGATTGCACAAGTTGT-------------------------TCATT   1493
Rp_RpaI_R0046        AGTGGATGGTGACGAATCCCTTATGTTGT-------------------------TCGAC   1498
Bj_BJ6T_RS07380      AGTGGATGGTGACGAATCCCGTGTGTTGT-------------------------CCGAC   1554
                      *       **

Ec_rrlA              TTTAAGCGTGTAGGCTGGTTTTCCAGGCAAATCCGGAA--AATCAAGGCTGAGGCGTGAT   1522
St_rrnH              TTTAAGCGTGTAGGTGTGTGTTCCAGGTAAATCCGGTTCACTTTAACACTGAGGCGTGAC   1700
Rs_RSP_4295          CTTAT-----------------------------------------------------   1570
Go_GOX1319           CTTAA-----------------------------------------------------   1417
Go_GOX1159           CTTAA-----------------------------------------------------   1417
Ar_ARAD_RS00940      CTTAT-----------------------------------------------------   1486
Rl_RLEG3_RS21775     CTAAT-----------------------------------------------------   1498
Rp_RpaI_R0046        CTTAC-----------------------------------------------------   1503
Bj_BJ6T_RS07380      CTTAC-----------------------------------------------------   1559
                      * *

Ec_rrlA              GACGAGGCACTACGGTGCTGAAGCAACAAATGCCCTGCTTCCAGGAAAAGCCTCTAAGCA   1582
St_rrnH              GACGAGGCACTACGGTGCTGAAGCAACAAATGCCCTGCTTCCAGGAAAAGCCTCTAAGCA   1760
```

```
Rs_RSP_4295          ------------------------CGGATTGACCGGGCTGCTGAGCGGTCCCTGGAAATA   1606
Go_GOX1319           ------------------------CGGATTGAACAGGCTTTTCAATCATTCCAGGAAATA   1453
Go_GOX1159           ------------------------CGGATTGAACAGGCTTTTCAATCATTCCAGGAAATA   1453
Ar_ARAD_RS00940      ------------------------TGGATTGTGTGGGCGGGGACGCGGTTCCAGGAAATA   1522
Rl_RLEG3_RS21775     ------------------------TGGATTGGGTGGGCAGCGGAGCGGTTCCAGGAAATA   1534
Rp_RpaI_R0046        ------------------------TGGATTGGTCGGGCCTCGACGGGGTTCCAGGAAATA   1539
Bj_BJ6T_RS07380      ------------------------TGGATTGGTTGGGCTTCGAAGGGGTTCCAGGAAATA   1595
                                             *       **           **   **  *

Ec_rrlA              TCAGGTAACATCAAATCGTACCCCAAACCGACACAGGTGGTCAGGTAGAGAATACCAAGG   1642
St_rrnH              TCAGGTAACACGAAATCGTACCCCAAACCGACACAGGTGGTCAGGTAGAGAATACCAAGG   1820
Rs_RSP_4295          GCCCT--CCATCAGACCGTACCCCAAACCGACACAGGTGGACTGGTAGAGAATACCAAGG   1664
Go_GOX1319           GCTCTGGCGTATAGACCGTACCCGAAACCGACACAGGTGGACTGGTAGAGAATACCAAGG   1513
Go_GOX1159           GCTCTGGCGTATAGACCGTACCCGAAACCGACACAGGTGGACTGGTAGAGAATACCAAGG   1513
Ar_ARAD_RS00940      GCTCCACCGTATAGACCGTACCCGAAACCGACACAGGTGGTCAGGTAGAGTATACCAAGG   1582
Rl_RLEG3_RS21775     GCTCCACCGTATAGACCGTACCCGAAACCGACACAGGTGGTCAGGTAGAGTATACCAAGG   1594
Rp_RpaI_R0046        GCCTCC-ACATCAGACCGTACCCGAAACCGACACAGGTGGACTGGTAGAGTATACCAAGG   1598
Bj_BJ6T_RS07380      GCCTCC-ACATCAGACCGTACCCGAAACCGACACAGGTGGACTGGTAGAGTATACCAAGG   1654
                      *          * * ******* ***************** * ******* *********

Ec_rrlA              CGCTTGAGAGAACTCGGGTGAAGGAACTAGGCAAAATGGTGCCGTAACTTCGGGAGAAGG   1702
St_rrnH              CGCTTGAGAGAACTCGGGTGAAGGAACTAGGCAAAATGGTGCCGTAACTTCGGGAGAAGG   1880
Rs_RSP_4295          CGCTTGAGAGAACCACATCAAAGGAACTCGGCAAAATGCCTCCGTAAGTTCGCGAGAAGG   1724
Go_GOX1319           CGCTTGAGAGAACGATGCTGAAGGAACTAGGCAAATTGCTCGTGTAACTTCGGGATAAAC   1573
Go_GOX1159           CGCTTGAGAGAACGATGCTGAAGGAACTAGGCAAATTGCTCGTGTAACTTCGGGATAAAC   1573
Ar_ARAD_RS00940      CGCTTGAGAGAACTATGTTGAAGGAACTCGGCAAATTGCACGCGTAACTTCGGAAGAAGC   1642
Rl_RLEG3_RS21775     CGCTTGAGAGAACTATGTTGAAGGAACTCGGCAAATTGCACGCGTAACTTCGGAAGAAGC   1654
Rp_RpaI_R0046        CGCTTGAGAGAACTATGTTGAAGGAACTCGGCAATTTACCTCCGTAACTTCGGGATAAGG   1658
Bj_BJ6T_RS07380      CGCTTGAGAGAACTATGTTGAAGGAACTCGGCAATTTACCTCCGTAACTTCGGGATAAGG   1714
                     *************       ******** *****  *      **** ****  * **

Ec_rrlA              CACGCTGATATGTAGGTGAAGCGACTTGCTCGTGGAGCTGAAATCAGTCGAAGATACCAG   1762
St_rrnH              CACGCTGACACGTAGGTGAAGTGATTTACTCATGGAGCTGAAGTCAGTCGAAGATACCAG   1940
Rs_RSP_4295          AGGCCCCGTCTGTAGGCAA-----------------CTATGGGCGGGGGGCACAAACCAG   1767
Go_GOX1319           GAGACCCGCTCGTGGGCAA-----------------CCATGGACGGGTGGCACAGACCAG   1616
Go_GOX1159           GAGACCCGCTCGTGGGCAA-----------------CCATGGACGGGTGGCACAGACCAG   1616
Ar_ARAD_RS00940      GTGACCCTTATCTACGCAA-----------------GTATGTGAGGGTGGCACAGACCAG   1685
Rl_RLEG3_RS21775     GTGACCCCAATCTACGCAA-----------------GTATTTTGGGGTGGCACAGACCAG   1697
Rp_RpaI_R0046        AGGCCTTCTGTTTGCGCAA-----------------GCAGGCAGGAGGGGCACAGACCAG   1701
Bj_BJ6T_RS07380      AGGCCCATTGCTCGCGCAA-----------------GCGGGCAGTGGGGGCACAGACCAG   1757
                         *          *  *                           *  * * * *****

Ec_rrlA              CTGGCTGCAACTGTTTATTAAAAACACAGCACTGTGCAAACACGAAAGTGGACGTATACG   1822
St_rrnH              CTGGCTGCAACTGTTTATTAAAAACACAGCACTGTGCAAACACGAAAGTGGACGTATACG   2000
Rs_RSP_4295          GGGGTGGCGACTGTTTACTTAAAACACAGGGCTGTGCGAAGCCGCAAGGCGACGTATACA   1827
Go_GOX1319           GGGGTAGCGACTGTTTAGTAAAAACACAGGGCTCTGCGAAATCGTGAGATGACGTATAGG   1676
Go_GOX1159           GGGGTAGCGACTGTTTAGTAAAAACACAGGGCTCTGCGAAATCGTGAGATGACGTATAGG   1676
Ar_ARAD_RS00940      GGGGTAGCGACTGTTTACCAAAAACACAGGGCTCTGCGAAGTCGCAAGACGACGTATAGG   1745
Rl_RLEG3_RS21775     GGGGTAGCGACTGTTTATCAAAAACACAGGGCTCTGCGAAGTCGCAAGACGACGTATAGG   1757
Rp_RpaI_R0046        GGGGTGGCAACTGTTTAACAAAAACACAGGGCTCTGCGAAATCGCAAGATGACGTATAGG   1761
Bj_BJ6T_RS07380      GGGGTGGCAACTGTTTAACAAAAACACAGGGCTCTGCGAAATCGCAAGATGACGTATAGG   1817
                       **  ** ********   *********  ** *** **  **  **  ********

Ec_rrlA              GTGTGACGCCTGCCCGGTGCCGGAAGGTTAATTGATGGGGTTAGCCGCAAGGCGAAGCTC   1882
St_rrnH              GTGTGACGCCTGCCCGGTGCCGGAAGGTTAATTGATGGGGTCAGCGC--AAGCGAAGCTC   2058
Rs_RSP_4295          GTCTGACGCCTGCCCGGTGCTGGAAGGTTAAAAGGAGGAGTGCA-----------AGCTC   1876
Go_GOX1319           GCCTGACGCCTGCCCGGTGCCGGAAGGTTAAGAGGAGGTGTGCA-----------AGCAC   1725
Go_GOX1159           GCCTGACGCCTGCCCGGTGCCGGAAGGTTAAGAGGAGGTGTGCA-----------AGCAC   1725
Ar_ARAD_RS00940      GTCTGACGCCTGCCCGGTGCTGGAAGGTTAAGAGGAGGGGTGCA-----------AGCTC   1794
Rl_RLEG3_RS21775     GTCTGACGCCTGCCCGGTGCTGGAAGGTTAAGAGGAGAGGTGCA-----------AGCTT   1806
Rp_RpaI_R0046        GTCTGACGCCTGCCCGGTGCCGGAAGGTTAAGAGGAGGAGTGCA-----------AGCTC   1810
Bj_BJ6T_RS07380      GTCTGACGCCTGCCCGGTGCCGGAAGGTTAAGAGGAGAGGTGCA-----------AGCCT   1866
                     *  ***************** **********  *  *  **             ***

Ec_rrlA              TTGATCGAAGCCC-CGGTAAACGGCGGCCGTAACTATAACGGTCCTAAGGTAGCGAAATT   1941
St_rrnH              CTGATCGAAGCCC-CGGTAAACGGCGGCCGTAACTATAACGGTCCTAAGGTAGCGAAATT   2117
Rs_RSP_4295          CGAATTGAAGCCCCAGTAA-ACGGCGGCCGTAACTATAACGGTCCTAAGGTAGCGAAATT   1935
Go_GOX1319           TGAATTGAAGCCCCGGTAAACCGGCGGCCGTAACTATAACGGTCCTAAGGTAGCGAAATT   1785
```

```
Go_GOX1159          TGAATTGAAGCCCCGGTAAA-CGGCGGCCGTAACTATAACGGTCCTAAGGTAGCGAAATT  1784
Ar_ARAD_RS00940     TGAATCGAAGCCCCAGTAA-ACGGCGGCCGTAACTATAACGGTCCTAAGGTAGCGAAATT  1853
Rl_RLEG3_RS21775    TGAATCGAAGCCCCAGTAA-ACGGCGGCCGTAACTATAACGGTCCTAAGGTAGCGAAATT  1865
Rp_RpaI_R0046       TGAATTGAAGCCCCGGTAA-ACGGCGGCCGTAACTATAACGGTCCTAAGGTAGCGAAATT  1869
Bj_BJ6T_RS07380     TGAATCGAAGCCCCGGTAA-ACGGCGGCCGTAACTATAACGGTCCTAAGGTAGCGAAATT  1925
                       ** *******  *  *  ************************************

Ec_rrlA             CCTTGTCGGGTAAGTTCCGACCTGCACGAATGGCGTAATGATGGCCAGGCTGTCTCCACC  2001
St_rrnH             CCTTGTCGGGTAAGTTCCGACCTGCACGAATGGCGTAATGATGGCCAGGCTGTCTCCACC  2177
Rs_RSP_4295         CCTTGTCGGGTAAGTTCCGACCTGCACGAATGGCGTAACGATCTCCCCGCTGTCTCTGAT  1995
Go_GOX1319          CCTTGTCGGGTAAGTTCCGACCTGCACGAATGGCGTAACGACTTCCCCACTGTCTCCAGC  1845
Go_GOX1159          CCTTGTCGGGTAAGTTCCGACCTGCACGAATGGCGTAACGACTTCCCCACTGTCTCCAGC  1844
Ar_ARAD_RS00940     CCTTGTCGGGTAAGTTCCGACCTGCACGAATGGCGTAACGACTTCCCCGCTGTCTCCAAC  1913
Rl_RLEG3_RS21775    CCTTGTCGGGTAAGTTCCGACCTGCACGAATGGCGTAACGACTTCCCCGCTGTCTCCAAC  1925
Rp_RpaI_R0046       CCTTGTCGGGTAAGTTCCGACCTGCACGAATGGCGTAATGACTTCCCCGCTGTCTCCAAC  1929
Bj_BJ6T_RS07380     CCTTGTCGGGTAAGTTCCGACCTGCACGAATGGCGTAATGACTTCCCCGCTGTCTCCAAC  1985
                    ************************************ **   **   *******

Ec_rrlA             CGAGACTCAGTGAAATTGAACTCGCTGTGAAGATGCAGTGTACCCGCGGCAAGACGGAAA  2061
St_rrnH             CGAGACTCAGTGAAATTGAACTCGCTGTGAAGATGCAGTGTACCCGCGGCAAGACGGAAA  2237
Rs_RSP_4295         GTGGACTCAGCGAAATTGAACTGTGTGTCAAGATGCACACTTCCCGCGGTTAGACGGAAA  2055
Go_GOX1319          ATCGACTCAGCGAAATTGAATTCCCCGTGAAGATGCGGGGTACCCGCGGTCAGACGGAAA  1905
Go_GOX1159          ATCGACTCAGCGAAATTGAATTCCCCGTGAAGATGCGGGGTACCCGCGGTCAGACGGAAA  1904
Ar_ARAD_RS00940     ATAGACTCAGTGAAATTGAATTCCCCGTGAAGATGCGGGGTTCCTGCGGTCAGACGGAAA  1973
Rl_RLEG3_RS21775    ATAGACTCAGTGAAATTGAATTCCCCGTGAAGATGCGGGGTTCCTGCGGTCAGACGGAAA  1985
Rp_RpaI_R0046       ATAGACTCAGTGAAATTGAATTCCCCGTGAAGATGCGGGGTTCCTGCGGTCAGACGGAAA  1989
Bj_BJ6T_RS07380     ATAGACTCAGTGAAATTGAATTCCCCGTGAAGATGCGGGGTTCCTGCGGTCAGACGGAAA  2045
                       ******* ********* *    ** *******    * ** ****  *********

Ec_rrlA             GACCCCGTGAACCTTTACTATAGCTTGACACTGAACATTGAGCCTTGATGTGTAGGATAG  2121
St_rrnH             GACCCCGTGAACCTTTACTATAGCTTGACACTGAACATTGAGCCTTGATGTGTAGGATAG  2297
Rs_RSP_4295         GACCCCATGAACCTTTACTATAGCTTCGCACTGGCATCAGGATTGTGATGTGCAGGATAG  2115
Go_GOX1319          GACCCTATGAACCTTTACTGCAGCTTTGCAGTGGCATCAGAGACATTCTGTGTAGGATAG  1965
Go_GOX1159          GACCCTATGAACCTTTACTGCAGCTTTGCAGTGGCATCAGAGACATTCTGTGTAGGATAG  1964
Ar_ARAD_RS00940     GACCCCGTGCACCTTTACTATAGCTTTACACTGGCATTCGTGTCGGCATGTGTAGGATAG  2033
Rl_RLEG3_RS21775    GACCCCGTGCACCTTTACTATAGCTTTACACTGGCATTCGTGTCGGCATGTGTAGGATAG  2045
Rp_RpaI_R0046       GACCCCGTGCACCTTTACTGTAGCTTTGCGCTGGTATTCGTGACTGTTTGTGTAGAATAG  2049
Bj_BJ6T_RS07380     GACCCCGTGCACCTTTACTGTAGCTTTGCGCTGGTATTCGTGACTGTTTGTGTAGAATAG  2105
                    *****  ** *********  *****  *  **      *        **** ** ****

Ec_rrlA             GTGGGAGGCTTTGAAGTGTGGACGCCAGTCTGCATGGAGCCGA-CCTTGAAATACCACCC  2180
St_rrnH             GTGGGAGGCTTTGAAGTGTGGACGCCAGTCTGCATGGAGCCGA-CCTTGAAATACCACCC  2356
Rs_RSP_4295         GTGGTAGGCATCGAAGCGGGGACGCCAGTTCCCGTGGAGCCAA-CCTTGAGATACCACCC  2174
Go_GOX1319          GTCGGAGGCTTTGAAACCCAGGCGCCAGCTTGGGTGGAGCCAT-CCTTGAAATACGACCC  2024
Go_GOX1159          GTCGGAGGCTTTGAAACCCAGGCGCCAGCTTGGGTGGAGCCAT-CCTTGAAATACGACCC  2023
Ar_ARAD_RS00940     GTGGTAGGCTTTGAAGCAGGGACGCCAGTTCTTGTGGAGCCAT-CCTTGAAATACCACCC  2092
Rl_RLEG3_RS21775    GTGGTAGGCTTTGAAGCGGGGACGCCAGTTTCCGTGGAGCCAT-CCTTGAAATACCACCC  2104
Rp_RpaI_R0046       GTGGTAGGCTTTGAAGCTCGGGCGCCAGCTCGGGTGGAGCCGCAATGTGAAATACCACCC  2109
Bj_BJ6T_RS07380     GTGGTAGGCTTTGAAGCCGTGGCGCCAGCCATGGTGGAGCCGAAATGTGAAATACCACCC  2165
                    ** * **** * ***     * ******      *******      *** **** ****

Ec_rrlA             TTTAATGTTTGATGTTCTAACGTTGACCCGTAATCCGGGTTGCGGACAGTGTCTGGTGGG  2240
St_rrnH             TTTAATGTTTGATGTTCTAACGTGGACCCGTAATCCGGGTTGCGGACAGTGTCTGGTGGG  2416
Rs_RSP_4295         TTCGCCATCTTGATGTCTAACCGCGGCCCGTTATCCGGGTCCGGGACCCTGCGTGGTGGG  2234
Go_GOX1319          TGAATTTTTCTTGATGTCTAACCGAGACCAGTAAGCCTGGTCCGGGACCCTGCATGGTGGG  2084
Go_GOX1159          TGAATTTTTCCGATGTCTAACCGAGACCAGTAAGCCTGGTCCGGGACCCTGCATGGTGGG  2083
Ar_ARAD_RS00940     TTATCGTCATGGATGTCTAACCGCGGTCCGTTATCCGGATCCGGGACAGTGTATGGTGGG  2152
Rl_RLEG3_RS21775    TTATCGTCATGGATGTCTAACCGCGGCCCGTTATCCGGGTCCGGGACAGTGTATGGTGGG  2164
Rp_RpaI_R0046       TAATGGTTATGGATATCTAACCGCGATCCCTTATCGGGTTCCGGGACAGCGCATGGTGGG  2169
Bj_BJ6T_RS07380     TAATGGTTATGGATATCTAACCGCGTCCCCTCAGCGGGGACCGGGACAGCGCATGGTGGG  2225
                    *             ******   *  *  * * *  *     ****   *  *******

Ec_rrlA             TAGTTTGACTGGGGCGGTCTCCTCCTAAAGAGTAACGGAGGAGCACGAAGGTTGGCTAAT  2300
St_rrnH             TAGTTTGACTGGGGCGGTCTCCTCCTAAAGAGTAACGGAGGAGCACGAAGGTTGGCTAAT  2476
Rs_RSP_4295         TAGTTTGACTGGGGCGGTCGCCTCCCAAACAGTAACGGAGGCGCGCGATGGTGGGCTCAG  2294
Go_GOX1319          CAGTTTGACTGGGGCGGTCGCCTCCCAAAGTGTAACGGAGGCGCGCGATGGTGGGCTCAG  2144
Go_GOX1159          CAGTTTGACTGGGGCGGTCGCCTCCCAAAGTGTAA-GGAGGCGCGCGATGGTGGGCTCAG  2142
Ar_ARAD_RS00940     TAGTTTGACTGGGGCGGTCGCCTCCGAAAGAGTAACGGAGGCGCGCGATGGTGGGCTCAG  2212
```

```
Rl_RLEG3_RS21775      TAGTTTGACTGGGGCGGTCGCCTCCGAAAGAGTAACGGAGGCGCGCGATGGTGGGCTCAG   2224
Rp_RpaI_R0046         CAGTTTGACTGGGGCGGTCGCCTCCCAAAGAGTAACGGAGGCGTGCGACGGTAGGCTCAG   2229
Bj_BJ6T_RS07380       CAGTTTGACTGGGGCGGTCGCCTCCCAAAGAGTAACGGAGGCGTGCGAAGGTAGGCTCAG   2285
                       ****************** ***** ***  **** ***** *  *** *** **** *

Ec_rrlA               CCTGGTCGGACATCAGGAGGTTAGTGCAATGGCATAAGCCAGCTTGACTGCGAGCGTGAC   2360
St_rrnH               CCTGGTCGGACATCAGGAGGTTAGTGCAATGGCATAAGCCAGCTTGACTGCGAGCGTGAC   2536
Rs_RSP_4295           ACCGGTCGGAAATCGGTCGTCGAGTGCAATGGCAGAAGCCCGCCTGACTGCAAGACTGAC   2354
Go_GOX1319            GCCGGTCGGAAACCGGCTGTCGAGTGCAATGGCATAAGCCCGCCTGACTGTGAGAGTGAC   2204
Go_GOX1159            GCCGGTCGGAAACCGGCTGTCGAGTGCAATGGCATAAGCCCGCCTGACTGTGAGAGTGAC   2202
Ar_ARAD_RS00940       ACCGGTCGGAAATCGGTCGTCGAGTGCAATGGCATAAGCCCGCCTGACTGCGAGACTGAC   2272
Rl_RLEG3_RS21775      ACCGGTCGGAAATCGGTCGTCGAGTGCAATGGCATAAGCCCGCCTGACTGCGAGACTGAC   2284
Rp_RpaI_R0046         AACGGTCGGAAATCGTTCGTCGAGTACAATGGCATAAGCCTGCCTGACTGCGAGACCAAC   2289
Bj_BJ6T_RS07380       AACGGTCGGAAATCGTTCGTCGAGTATAATGGCATAAGCCTGCCTGACTGCGAGATCTAC   2345
                         ******* * *    *   ***  ******* ***** ** ******  **    **

Ec_rrlA               GGCGCGAGCAGGTGCGAAAGCAGGTCATAGTGATCC-GGTGGTTCTGAATGGAAGGGCCA   2419
St_rrnH               GGCGCGAGCAGGTGCGAAAGCAGGTCATAGTGATCC-GGTGGTTCTGAATGGAAGGGCCA   2595
Rs_RSP_4295           AAGTCGAGCAGAGACGAAAGTCGGCCATAGTGATCC-GGTGGTCCCGAGTGGAAGGGCCA   2413
Go_GOX1319            AGCTCGATCAGAGACGAAAGTCGGCCATAGTGATCC-GGTGGTCCCATGTGGACGGGCCA   2263
Go_GOX1159            AGCTCGATCAGAGACGAAAGTCGGCCATAGTGATCC-GGTGGTCCCATGTGGACGGGCCA   2261
Ar_ARAD_RS00940       AAGTCGAGCAGAGACGAAAGTCGGTCATAGTGATCC-GGTGGTCCCGCGTGGAAGGGCCA   2331
Rl_RLEG3_RS21775      AAGTCGAGCAGAGACGAAAGTCGGTCATAGTGATCC-GGTGGTCCCGCGTGGAAGGGCCA   2343
Rp_RpaI_R0046         AAGTCGAGCAGAGACGAAAGTCGGTCATAGTGATCCCGGTGGTCCCGCGTGGATGGGCCA   2349
Bj_BJ6T_RS07380       GAATCGAGCAGAGACGAAAGTCGGTCATAGTGATCC-GGTGGTCCCGCGTGGATGGGCCA   2404
                          *** ***   ******  ** ************ ****** *    **** ******

Ec_rrlA               TCGCTCAACGGATAAAAGGTACTCCGGGGATAACAGGCTGATACCGCCCAAGAGTTCATA   2479
St_rrnH               TCGCTCAACGGATAAAAGGTACTCCGGGGATAACAGGCTGATACCGCCCAAGAGTTCATA   2655
Rs_RSP_4295           TCGCTCAACGGATAAAAGGTACTCTGGGGATAACAGGCTGATGATGCCCAAGAGTCCATA   2473
Go_GOX1319            TCGCTCAACGGATAAAAGGTACTCTAGGGATAACAGGCTGATCTCCCCCAAGAGTCCACA   2323
Go_GOX1159            TCGCTCAACGGATAAAAGGTACTCTAGGGATAACAGGCTGATCTCCCCCAAGAGTCCACA   2321
Ar_ARAD_RS00940       TCGCTCAACGGATAAAAGGTACGCCGGGGATAACAGGCTGATGACCCCCAAGAGTCCATA   2391
Rl_RLEG3_RS21775      TCGCTCAACGGATAAAAGGTACGCCGGGGATAACAGGCTGATGACCCCCAAGAGTCCATA   2403
Rp_RpaI_R0046         TCGCTCAACGGATAAAAGGTACGCCGGGGATAACAGGCTGATGACGCCCAAGAGTCCATA   2409
Bj_BJ6T_RS07380       TCGCTCAACGGATAAAAGGTACGCCGGGGATAACAGGCTGATGACGCCCAAGAGTCCATA   2464
                      ********************** *  ****************    ********* ** *

Ec_rrlA               TCGACGGCGGTGTTTGGCACCTCGATGTCGGCTCATCACATCCTGGGGCTGAAGTAGGTC   2539
St_rrnH               TCGACGGCGGTGTTTGGCACCTCGATGTCGGCTCATCACATCCTGGGGCTGAAGTAGGTC   2715
Rs_RSP_4295           TCGACGGCATCGTTTGGCACCTCGATGTCGGCTCATCTCATCCTGGGGCTGGAGCAGGTC   2533
Go_GOX1319            TCGACGGGGAGGTTTGGCACCTCGATGTCGGCTCATCACATCCTGGGGCTGGAGCAGGTC   2383
Go_GOX1159            TCGACGGGGAGGTTTGGCACCTCGATGTCGGCTCATCACATCCTGGGGCTGGAGCAGGTC   2381
Ar_ARAD_RS00940       TCGACGGGGTTGTTTGGCACCTCGATGTCGGCTCATCGCATCCTGGGGCTGGAGCAGGTC   2451
Rl_RLEG3_RS21775      TCGACGGGGTTGTTTGGCACCTCGATGTCGGCTCATCGCATCCTGGGGCTGGAGCAGGTC   2463
Rp_RpaI_R0046         TCGACGGCGTCGTTTGGCACCTCGATGTCGGCTCATCACATCCTGGGGCTGGAGAAGGTC   2469
Bj_BJ6T_RS07380       TCGACGGCGTCGTTTGGCACCTCGATGTCGGCTCATCACATCCTGGGGCTGGAGAAGGTC   2524
                      *******    ************************* ************* ** *****

Ec_rrlA               CCAAGGGTATGGCTGTTCGCCATTTAAAGTGGTACGCGAGCTGGGTTTAGAACGTCGTGA   2599
St_rrnH               CCAAGGGTATGGCTGTTCGCCATTTAAAGTGGTACGCGAGCTGGGTTTAGAACGTCGTGA   2775
Rs_RSP_4295           CCAAGGGTATGGCTGTTCGCCATTTAAAGAGGTACGTGAGCTGGGTTTAGAACGTCGTGA   2593
Go_GOX1319            CCAAGGGTTCGGCTGTTCGCCGATTAAAGTGGTACGTGAGCTGGGTTTAGAACGTCGTGA   2443
Go_GOX1159            CCAAGGGTTCGGCTGTTCGCCGATTAAAGTGGTACGTGAGCTGGGTTTAGAACGTCGTGA   2441
Ar_ARAD_RS00940       CCAAGGGTTTGGCTGTTCGCCAATTAAAGCGGTACGTGAGCTGGGTTCAGAACGTCGTGA   2511
Rl_RLEG3_RS21775      CCAAGGGTTTGGCTGTTCGCCAATTAAAGCGGTACGTGAGCTGGGTTCAGAACGTCGTGA   2523
Rp_RpaI_R0046         CCAAGGGTTCGGCTGTTCGCCGATTAAAGTGGTACGTGAGCTGGGTTCAGAACGTCGTGA   2529
Bj_BJ6T_RS07380       CCAAGGGTTCGGCTGTTCGCCGATTAAAGTGGTACGTGAGCTGGGTTCAGAACGTCGTGA   2584
                      ********  ***********  ****** ****** ********** ************

Ec_rrlA               GACAGTTCGGTCCCTATCTGCCGTGGGCGCTGGAGAACTGAGGGGGGCTGCTCCTAGTAC   2659
St_rrnH               GACAGTTCGGTCCCTATCTGCCGTGGGCGCTGGAGAACTGAGGGGGGCTGCTCCTAGTAC   2835
Rs_RSP_4295           GACAGTTCGGTCCCTATCTGCCGTGGGTGTAGGAGACTTGAGAAGAGTTGCCCCTAGTAC   2653
Go_GOX1319            GACAGTTCGGTCCCTATCTGCCGTGGGTGTAAGAGACTTGAGAGGATTTGTCCCTAGTAC   2503
Go_GOX1159            GACAGTTCGGTCCCTATCTGCCGTGGGTGTAAGAGACTTGAGAGGATTTGTCCCTAGTAC   2501
Ar_ARAD_RS00940       GACAGTTCGGTCCCTATCTGCCGTGGGTGTAGGAATATTGACAGGATCTGTCCCTAGTAC   2571
Rl_RLEG3_RS21775      GACAGTTCGGTCCCTATCTGCCGTGGGTGTAGGAATATTGACAGGATCTGTCCCTAGTAC   2583
Rp_RpaI_R0046         GACAGTTCGGTCCCTATCTGCCGTGGGTGTTGGAATGTTGAGAGGATTTGCCCCTAGTAC   2589
Bj_BJ6T_RS07380       GACAGTTCGGTCCCTATCTGCCGTGGGTGTTGGAATGTTGAGAGGATTTGCCCCTAGTAC   2644
```

```
                    ************************** *   **    ***   *   **  ********

Ec_rrlA             GAGAGGACCGGAGTGGACGCATCACTGGTGTTCGGGTTGTCATGCCAATGGCACTGCCCG   2719
St_rrnH             GAGAGGACCGGAGTGGACGCATCACTGGTGTTCGGGTTGTCATGCCAATGGCACTGCCCG   2895
Rs_RSP_4295         GAGAGGACCGGGGTGAACGATCCACTGGTGGACCAGTTGTCGTGCCAACGGCAGTGCTGG   2713
Go_GOX1319          GAGAGGACCGGGATGAACATACCTCTGGTGCACCGGTTGTCACGCCAGTGGCACAGCCGG   2563
Go_GOX1159          GAGAGGACCGGGATGAACATACCTCTGGTGCACCGGTTGTCACGCCAGTGGCACAGCCGG   2561
Ar_ARAD_RS00940     GAGAGGACCGGGATGGACATATCTCTGGTGGACCTGTTGTCCTGCCAAGGGCATAGCAGG   2631
Rl_RLEG3_RS21775    GAGAGGACCGGGATGGACATATCTCTGGTGGACCTGTTGTCCTGCCAAGGGCATAGCAGG   2643
Rp_RpaI_R0046       GAGAGGACCGGGGTGAACGTACCTCTGGTGGAGCTGTTGTCGCGCCAGCGGCAGTGCAGC   2649
Bj_BJ6T_RS07380     GAGAGGACCGGGGTGAACGTACCTCTGGTGGAGCTGTTGTCGCGCCAGCGGCAGTGCAGC   2704
                    ***********  ** **    * ******     ******  ****  ****  **

Ec_rrlA             GTAGCTAAATGCGGAAGAGATAAGTGCTGAAAGCATCTAAGCACGAAACTTGCCCCGAGA   2779
St_rrnH             GTAGCTAAATGCGGAAGAGATAAGTGCTGAAAGCATCTAAGCACGAAACTTGCCCCGAGA   2955
Rs_RSP_4295         GTAGCTATGATCGGACAGGATAACCGCTGAAGGCATCTAAGCGGGAAGCCCCCTTCAAAA   2773
Go_GOX1319          GTAGCTAAGTATGGACGGGATAACCGCTGAAAGCATCTAAGCGGGAAACCCACCTCAAAA   2623
Go_GOX1159          GTAGCTAAGTATGGACGGGATAACCGCTGAAAGCATCTAAGCGGGAAACCCACCTCAAAA   2621
Ar_ARAD_RS00940     GTAGCTACATATGGAATGGATAACCGCTGAAGGCATCTAAGCGGGAAACCAACCTGAAAA   2691
Rl_RLEG3_RS21775    GTAGCTACATATGGACGGGATAACCGCTGAAGGCATCTAAGCGGGAAACCCACCTGAAAA   2703
Rp_RpaI_R0046       ATAGCTATGTACGGACGGGATAACCGCTGAAAGCATCTAAGCGGGAAACCCACCTCAAAA   2709
Bj_BJ6T_RS07380     ATAGCTATGTACGGACGGGATAACCGCTGAAAGCATCTAAGCGGGAAACCCACCTCAAAA   2764
                     ******     ***   *****  ****** **********  *** *   *    * *

Ec_rrlA             TGAGTTCTCCCTGACTCCTTGAGAGTCCTGAAGGAACGTTGAAGACGACGACGTTGATAG   2839
St_rrnH             TGAGTTCTCCCTGAGACTTAGAGTCTCCTGAAGGAACGTTGAAGACGACGACGTTGATAG   3015
Rs_RSP_4295         CAAGGTCTCCCT-----------------TGAGGGCCGTGGAAGACCACCACGTCGATAG   2816
Go_GOX1319          CGAGGTCTC--------------------ATAGAGCCGTGACAGACCATCACGTTGATAG   2663
Go_GOX1159          CGAGGTCTC--------------------ATAGAGCCGTGACAGACCATCACGTTGATAG   2661
Ar_ARAD_RS00940     CGAGTGTTCCCTA----------------TCAGAGCCGTGGTAGACGACCACGTTGATAG   2735
Rl_RLEG3_RS21775    CGAGTATTCCCTA----------------TCAGAGCCGTGGAAGACGACCACGTTGATAG   2747
Rp_RpaI_R0046       CGAGCATTCCCT-----------------TGAGAACCGTGGAAGACGACCACGTTGATAG   2752
Bj_BJ6T_RS07380     CGAGCATTCCCT-----------------TGAGAACCGTGGAAGACCACCACGTTGATAG   2807
                      **   **                     **   ***   **** *  **** *****

Ec_rrlA             GCCGGGTGTGTAAGCGCAGCGATGCGTTGAGCTAACCGGTACTAATGAACCGTGAGGCTT   2899
St_rrnH             GCCGGGTGTGTAAGCGCAGCGATGCGTTGAGCTAACCGGTACTAATGAACCGTGAGGCTT   3075
Rs_RSP_4295         GCCAGAGGTGTAAGCGCAGCAATGCGTTCAGCTGACTGGTACTAATTGCCCGATAGGCTT   2876
Go_GOX1319          GCCGGGTGTGAAAGTGCAGTAATGCATGCAGCTAACCGGTCCTAATCG------------   2711
Go_GOX1159          GCCGGGTGTGAAAGTGCAGTAATGCATGCAGCTAACCGGTCCTAATCG------------   2709
Ar_ARAD_RS00940     GCCGGGTGTGGAAGTGCGGCAACGCATGAAGCTTACCGGTACTAATAGCTCGATTGGCTT   2795
Rl_RLEG3_RS21775    GCCGGGTGTGGAAGTGCGGCAACGCATGAAGCTTACCGGTACTAATAGCTCGATCGGCTT   2807
Rp_RpaI_R0046       GCCGGATGTGGAAGTGCAGCAATGCATGTAGCTTACCGGTACTAATCGTTCGATTGGCTT   2812
Bj_BJ6T_RS07380     GCCGGGTGTGGAAGTGCAGTAATGCATGCAGCTTACCGGTACTAATCGTTCGATTGGCTT   2867
                    *** *  *** *** ** *  * ** *  **** ** *** *****

Ec_rrlA             AACCTT------------  2905
St_rrnH             AACCTTACAACGCCGAAG  3093
Rs_RSP_4295         GATCT-------------  2881
Go_GOX1319          ----------------    2711
Go_GOX1159          ----------------    2709
Ar_ARAD_RS00940     GATTGTTCTCATT-----  2808
Rl_RLEG3_RS21775    GATCGTTCTCATT-----  2820
Rp_RpaI_R0046       GATTGCTCTCATTAT---  2827
Bj_BJ6T_RS07380     GATTGCTCTCATTTT---  2882
```

Figure S2. Sequence alignment of 23S rRNA gene sequences from *E coli* K-12 substrain MG1655 (Ec_rrlA), *S. typhimurium* LT2 (St_rrnH), *R. sphaeroides* 2.4.1 (Rs_RSP_4295), *G. oxydans* 621H (Go_GOX1319 and Go_GOX1159), *A. radiobacter* K84 (Ar_ARAD_RS00940), *R. leguminosarum bv. trifolii* WSM1689, *R. palustris* TIE-1 (Rp_rpaI_R0046), *B. japonicum* USDA 6 (Bj_BJ6T_RS07380). For the rRNA genes of *G. oxydans*, the three regions with a read coverage <5% of the average coverage, which were found for all four gene copies, are highlighted in light grey. The additional region, which was only identified in the rRNA sequence of GOX1159/GOX_RS06970, is shown in darker grey. Differences between the rRNA sequences GOX1319/GOX_RS07780 and

GOX1159/GOX_RS06970 are underlined and bolded. Boxes at these positions show the base identity in GOX0221/GOX_RS02255 and GOX1467/GOX_RS08565, which are identical. Known or assumed fragmentation regions found in other bacteria are underlined [2-5].

REFERENCES

1. Prust C, Hoffmeister M, Liesegang H, Wiezer A, Fricke WF, Ehrenreich A, Gottschalk G, Deppenmeier U: Complete genome sequence of the acetic acid bacterium *Gluconobacter oxydans*. Nat Biotechnol. 2005, 23(2):195-200.
2. Burgin AB, Parodos K, Lane DJ, Pace NR: The excision of intervening sequences from *Salmonella* 23S ribosomal RNA. Cell. 1990, 60(3):405-414.
3. Selenska-Pobell S, Evguenieva-Hackenberg E: Fragmentations of the large-subunit rRNA in the family *Rhizobiaceae*. J Bacteriol. 1995, 177(23):6993-6998.
4. Zahn K, Inui M, Yukawa H: Characterization of a separate small domain derived from the 5' end of 23S rRNA of an α-proteobacterium. Nucleic Acids Res. 1999, 27(21):4241-425.
5. Zahn K, Inui M, Yukawa H: Divergent mechanisms of 5' 23S rRNA IVS processing in the α-proteobacteria. Nucleic Acids Res. 2000, 28(23):4623-4633.

Danksagung

Prof. Dr. Michael Bott möchte ich für die Überlassung des herausfordernden und sehr interessanten Themas danken sowie für die Durchsicht der Manuskripte und dieser Arbeit. Außerdem danke ich ihm sehr für die Möglichkeit an einem neuen spannenden Projekt arbeiten zu dürfen.

Bei Jun.-Prof. Dr. Ilka Axmann möchte ich mich für die Übernahme des Zweitgutachtens bedanken. Außerdem danke ich ihr für das große Interesse am Fortschritt meiner Doktorarbeit sowie die gute Betreuung als Mentorin.

Dr. Tino Polen danke ich sehr für die engagierte Betreuung meiner Arbeit, die vielen Diskussionen bezüglich der Interpretation der Ergebnisse sowie für die Korrektur der Publikationen und der Doktorarbeit.

Allen aktuellen und ehemaligen Mitgliedern der AG Polen möchte ich für die sehr gute Arbeitsatmosphäre, die Unterstützung und die netten gemeinsamen Freizeitaktivitäten danken. Doris Rittmann und Ulli Degner danke ich sehr für die Unterstützung im Labor sowie für die Hilfe bei Problemen jeglicher Art. Andrea Steinmann gilt ein besonderer Dank für die engagierte Mitarbeit am Projekt und für die lustigen Aktionen neben der Arbeit. Ohne Dr. Ines Kiefler wäre die Arbeit mit Gox sehr einsam gewesen, daher möchte ich mich bei ihr herzlich für die sehr nette Zusammenarbeit, die immer viel Spaß gemacht hat, bedanken.

Außerdem möchte ich mich bei allen Mitarbeitern des Instituts sowie bei der Infrastruktur für die angenehme Arbeitsatmosphäre sowie für die Hilfe bei kleineren und größeren Problemen bedanken.

Allen wichtigen Menschen in meinem Leben, die mir in den vergangenen Jahren bei allen Problemen geholfen haben und immer für mich da waren, gilt ein ganz besonderer Dank!

Der wichtigste Dank von ganzem Herzen geht an meine Eltern für ihre uneingeschränkte Unterstützung!

Erklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine Universität Düsseldorf“ erstellt worden ist. Ich habe bisher keine erfolglosen Promotionsversuche unternommen. Diese Dissertation wurde bisher an keiner anderen Fakultät vorgelegt.

Jülich,

Angela Kranz