Methods 86 (2015) 89–101

Contents lists available at ScienceDirect

Methods

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

Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in *Helicobacter pylori*

Thorsten Bischler [a], Hock Siew Tan [a], Kay Nieselt [b], Cynthia M. Sharma [a,*]

[a] *Research Center for Infectious Diseases (ZINF), University of Würzburg, Josef-Schneider-Str. 2/Bau D15, 97080 Würzburg, Germany*
[b] *Integrative Transcriptomics, ZBIT (Center for Bioinformatics Tübingen), University of Tübingen, Sand 14, D-72076 Tübingen, Germany*

ARTICLE INFO

Article history:
Received 8 April 2015
Received in revised form 7 June 2015
Accepted 9 June 2015
Available online 16 June 2015

Keywords:
Differential RNA-seq
Transcriptional start sites
Comparative transcriptomics
Small RNAs
Promoter motifs
Gene regulation
5′UTR

ABSTRACT

The global mapping of transcription boundaries is a key step in the elucidation of the full complement of transcriptional features of an organism. It facilitates the annotation of operons and untranslated regions as well as novel transcripts, including *cis*- and *trans*-encoded small RNAs (sRNAs). So called RNA sequencing (RNA-seq) based on deep sequencing of cDNAs has greatly facilitated transcript mapping with single nucleotide resolution. However, conventional RNA-seq approaches typically cannot distinguish between primary and processed transcripts. Here we describe the recently developed differential RNA-seq (dRNA-seq) approach, which facilitates the annotation of transcriptional start sites (TSS) based on deep sequencing of two differentially treated cDNA library pairs, with one library being enriched for primary transcripts. Using the human pathogen *Helicobacter pylori* as a model organism, we describe the application of dRNA-seq together with an automated TSS annotation approach for generation of a genome-wide TSS map in bacteria. Besides a description of transcriptome and regulatory features that can be identified by this approach, we discuss the impact of different library preparation protocols and sequencing platforms as well as manual and automated TSS annotation. Moreover, we have set up an easily accessible online browser for visualization of the *H. pylori* transcriptome data from this and our previous *H. pylori* dRNA-seq study.

1. Introduction

RNA-sequencing (RNA-seq) based on deep sequencing of cDNA libraries has been increasingly used as the method of choice for gene expression analysis and annotation of whole transcriptomes [1]. In comparison to hybridization-based techniques, such as microarrays or tiling arrays, RNA-seq has a higher dynamic range and requires less input material. Instead of applying previously designed probes that are prone to suffer from cross-hybridization issues, RNA-seq directly records the amount and boundaries of each transcript with single nucleotide (nt) resolution. The prior knowledge of the genomic sequence can facilitate the analysis and mapping of the sequenced cDNA reads, but is not necessarily required to detect and quantify a transcript. RNA-seq has greatly facilitated the annotation of transcript boundaries and the identification of novel transcripts in both pro- and eukaryotes [2–4]. While a major challenge for early bacterial RNA-seq experiments was the presence of highly abundant RNA species like rRNAs and tRNAs, which make up more than 95% of the RNA pool in a bacterial cell, this issue was overcome in eukaryotes by solely reverse-transcribing poly(A)-tailed mRNAs via oligo-d(T) priming during cDNA library preparation [4]. Since poly(A)-tails represent a degradation signal in bacteria, several strategies for rRNA removal including oligonucleotide-based removal of rRNAs with magnetic beads or size fractionation using gel electrophoresis (reviewed in [2,5]) were employed. The steadily dropping sequencing costs, bundled with a major increase in sequencing depth, nowadays provide sufficient coverage for the mRNA and non-abundant sRNA fractions without the necessity for additional depletion steps which were recently shown to introduce coverage bias [6].

In a typical RNA-seq experiment total RNA or a fraction thereof is first converted into cDNA in a reverse-transcription reaction, followed by PCR-based amplification of the library. Different library protocols are available, which are highly specific for the applied sequencing technique but can be subdivided into strand-specific and non-strand-specific protocols. Non-strand-specific protocols, for example, based on random hexamer priming and ligation of adapters to double-stranded cDNA have the drawback that they

* Corresponding author. Fax: +49 931 3182578.
E-mail address: cynthia.sharma@uni-wuerzburg.de (C.M. Sharma).

http://dx.doi.org/10.1016/j.ymeth.2015.06.012
1046-2023/© 2015 Elsevier Inc. All rights reserved.

lose the information whether sequencing reads originate from the sense or the antisense strand. To overcome this problem, strand-specific protocols have been developed including direct sequencing of first strand cDNA [7], template switching PCR [8], RNA C to U conversion using bisulfite [9] or second strand synthesis with dUTP followed by degradation after adapter ligation [10]. Our below listed protocol combines 5′ end RNA linker ligation with poly(A)-tailing using *Escherichia coli* poly(A) polymerase [11–13]. After cDNA library construction and different quality checks, the samples are sequenced on one of the available deep sequencing platforms, resulting in millions of cDNA reads. The most commonly used techniques are the Illumina (Solexa), the 454 Life Sciences system, and ABI SOLiD sequencing. More recently developed single-molecule sequencing technologies comprise SMRT sequencing (Pacific Biosciences) or nanopore sequencing (Oxford Nanopore Technologies). Depending on the applied protocol and sequencing method, the reads are subjected to different pre-processing steps such as quality filtering or adapter or poly(A)-tail trimming. Afterwards, cDNA reads are commonly aligned to a genomic sequence and can then be used for gene expression profiling based on existing annotations, the generation of nucleotide-wise coverage plots for visualization in a genome browser and the annotation of novel transcripts.

RNA-seq-based mapping of bacterial transcript boundaries enables a global elucidation of operon structures and facilitates annotation of untranslated regions (UTRs) of protein coding genes, which potentially contain gene regulatory elements. Additionally, it allows for detection of novel transcripts such as small regulatory RNAs (sRNAs) and facilitates the discovery of previously non- or misannotated ORFs. Primer extension [14] or 5′ RACE (rapid amplification of cDNA ends) [15–17] are established methods for the determination of transcript 5′ ends of single genes, but they are time-consuming and impractical for global analysis. Therefore, several RNA-seq-based protocols for sequencing of 5′ ends of RNAs including a modified 5′ RACE approach have been developed, but many of them cannot clearly distinguish transcriptional start sites (TSS) from processing sites [18–23].

Here we give a detailed description of the differential RNA-seq (dRNA-seq) method, which allows for global annotation of all expressed TSS under the examined growth condition in an organism of interest in one sequencing experiment [24]. While it was originally developed to study the primary transcriptome of the major human pathogen *Helicobacter pylori* [12] it has since been successfully applied for determination of TSS in a wide range of pro- and eukaryotic organisms [24]. With >1900 unique TSS and at least one antisense TSS to 50% of all genes, the dRNA-seq approach revealed a very complex and compact transcriptional output from the small *H. pylori* genome and an unexpected number of >60 sRNAs [12]. While our previous *H. pylori* dRNA-seq approach was based on 454 sequencing of dRNA-seq libraries from *H. pylori* strain 26695 grown under different growth conditions, we here exemplify the use of Illumina-based dRNA-seq for annotation of TSS under a representative growth condition. We compare the results from the different sequencing platforms and among different replicates. Furthermore, we perform an automated TSS annotation using TSSpredator (http://it.inf.uni-tuebingen.de/TSSpredator), which we had initially applied for a comparative TSS annotation in multiple *Campylobacter jejuni* strains [25] and the generation of a global TSS map of *Escherichia coli* K12 MG1655 [26], and compare the automated TSS annotation with manual TSS annotations from the previous *H. pylori* dRNA-seq study [12]. We provide the global TSS maps and cDNA coverage plots of the previous and newly generated *H. pylori* 26695 dRNA-seq data in an easily accessible online browser (http://hpylori-tss.imib-zinf.net/).

2. Materials and methods

2.1. Helicobacter pylori growth conditions

Helicobacter pylori wild type strain 26695 (CSS-0065, kindly provided by D. Scott Merrell, Bethesda, MD) was grown on GC-agar (Oxoid) plates supplemented with 10% (V/V) donor horse serum (Biochrom AG), 1% (V/V) vitamin mix, 10 μg/ml vancomycin, 5 μg/ml trimethoprim, and 1 μg/ml nystatin as described previously [44]. For liquid cultures, 15 or 50 ml Brain Heart Infusion medium (BHI, Becton, Dickinson and Company) supplemented with 10% (V/V) FBS (Biochrom AG) and 10 μg/ml vancomycin, 5 μg/ml trimethoprim, and 1 μg/ml nystatin was inoculated with *H. pylori* grown on plates to a final OD_{600} of 0.02–0.05 and grown under agitation at 140 rpm in 25 cm^3 or 75 cm^3 cell culture flasks (Corning). Bacteria were grown at 37 °C in a HERAcell 150i incubator (Thermo scientific) in a microaerophilic environment (10% CO_2, 5% O_2, and 85% N_2). When the cultures reached mid-log phase (OD_{600} ~0.6), culture volumes of cells corresponding to a total amount of 4 OD_{600} were mixed with 0.2 volumes of stop-mix (95% EtOH and 5% phenol, V/V), frozen in liquid N_2 and stored at −80 °C until RNA extraction. In total, three biological replicates of bacteria grown to mid-log phase (ML) were harvested: B1, which was grown separately from B2 and B3, which were grown on the same day.

2.2. RNA extraction and DNase I treatment

Frozen cell pellets were thawed on ice and resuspended in lysis solution containing 600 μl of 0.5 mg/ml lysozyme in TE buffer (pH 8.0) and 60 μl 10% SDS. Bacterial cells were lysed by incubating the samples for 1–2 min at 65 °C. Afterwards, total RNA was extracted using the hot-phenol method as described previously [12,27]. DNase I (Fermentas) treatment was performed on total RNA according to manufacturer's instruction. Removal of residual genomic DNA was subsequently verified by control PCR using the oligos CSO-0790: GTTTTTTCTAGACGTTTAAAACAAGCCTGGT and CSO-0791: GTTTTTGAATTCCATGATGACTCCTTTAATTGAAA which amplify a ~594 nt long product of the HP1432 gene.

2.3. dRNA-seq library preparation and sequencing

Terminator exonuclease (TEX) treatment of RNA samples was performed as previously described [12]. cDNA libraries for Illumina sequencing were constructed by Vertis Biotechnology AG, Germany (http://www.vertis-biotech.com/) in a strand-specific manner as previously described for eukaryotic microRNA [28] but omitting the RNA size-fractionation step prior to cDNA synthesis. In brief, ~200 ng of RNA sample were poly(A)-tailed using 2.5 U *E. coli* poly(A) polymerase (NEB) for 5 min at 37 °C. TEX treatment (+TEX) and mock treatment without the enzyme (−TEX) were carried out after poly(A)-tailing. To this end, poly(A)-tailed RNA was denatured for 2 min at 90 °C, cooled on ice for 5 min and treated with 1.5 U of Terminator Exonuclease (Epicentre) for 30 min at 30 °C. Then, the 5′-PPP structures were removed using tobacco acid pyrophosphatase (TAP). TAP treatment was performed by incubating +TEX and −TEX samples with 5 U TAP for 15 min at 37 °C.

Afterwards, an RNA adapter (5′ Illumina sequencing adapter, 5′-UUUCCCUACACGACGCUCUUCCGAUCU-3′) was ligated to the 5′-P of the TAP-treated, poly(A)-tailed RNA for 30 min at 25 °C. First strand cDNA was synthesized by using an oligo(dT)-adapter primer (see below) and the M-MLV reverse transcriptase (AffinityScript, Agilent) by incubation at 42 °C for 20 min, ramp to 55 °C followed by 55 °C for 5 min. In a PCR-based amplification step using a high fidelity DNA polymerase (Herculase II Fusion DNA Polymerases, Agilent) the cDNA concentration was increased to 10–20 ng/μl (initial denaturation at 95 °C for 2 min, 16–18 cycles 95 °C for 20 s and 68 °C for 2 min). A library-specific barcode for multiplex sequencing was included as part of a 3′ sequencing adapter. The TruSeq index primers for PCR amplification were used according to the instructions of Illumina. For all libraries the Agencourt AMPure XP kit (Beckman Coulter Genomics) was used to purify the DNA (1.8 × sample volume), and cDNA sizes were examined by capillary electrophoresis on a MultiNA microchip electrophoresis system (Shimadzu).

The following adapter sequences flank the cDNA inserts:

TruSeq_Sense_primer: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′, TruSeq_Antisense_NNNNNN_primer (NNNNNN = 6n barcode for multiplexing): 5′-CAAGCAGAAGACGGCATACGAGAT–NNNNNN-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC(dT25)-3′

The first biological replicate (B1) was sequenced on an Illumina HiSeq 2000 machine with 97 cycles while the second and third replicate (B2/B3) were sequenced on a HiSeq 2500 with 100 cycles. All sequencing was conducted in single-read mode.

2.4. Analysis of deep sequencing data

2.4.1. Read mapping and generation of coverage plots

To assure high sequence quality, the Illumina reads in FASTQ format were trimmed with a cutoff phred score of 20 by the program fastq_quality_trimmer from FASTX toolkit version 0.0.13. After trimming, poly(A)-tail sequences were removed and a size filtering step was applied in which sequences shorter than 12 nt were eliminated. The collections of remaining reads were mapped to the *H. pylori* 26695 (NCBI Acc.-No: NC_000915.1) genome using the RNA-seq pipeline READemption [29] and *segemehl* [30] with an accuracy cutoff of 95%. Coverage plots representing the numbers of mapped reads per nucleotide were generated. Reads that mapped to multiple locations with an equal score contributed a fraction to the coverage value. For example, reads mapping to three positions contributed only 1/3 to the coverage values. We chose this approach of including reads that map to multiple locations with relative scores rather than solely using uniquely mapped reads. It represents a tradeoff between introducing some uncertainty regarding the true origin of reads that map to multiple locations and not excluding all transcripts with true multiple copies in the genome like rRNAs, tRNAs and some of the sRNAs. Since only read mappings with equal scores were considered, most of the non-uniquely mapped reads likely corresponded to such duplicated or repetitive genes, rather than representing unspecifically mapped reads. Each resulting cDNA coverage graph was normalized by the number of reads that could be mapped from the respective library (typically several million reads when using Illumina sequencing) and afterwards multiplied by 1,000,000.

2.4.2. Coverage plot normalization by TSSpredator

Prior to the comparative analysis, the expression graphs with the cDNA coverages that resulted from the read mapping were further normalized using TSSpredator (http://it.inf.uni-tuebingen.de/TSSpredator). A percentile normalization step was applied to normalize the +TEX graphs. To this end, the 90th percentile of all data values was calculated for each +TEX graph. This value was then used to normalize the +TEX graph as well as the respective −TEX graph. Thus, the relative differences between each +TEX and −TEX graph were not changed in this normalization step. Again, afterwards all graphs were multiplied with the overall lowest value to restore the original data range. To account for different enrichment rates, a third normalization step was applied. During this step, prediction of TSS candidates was performed for each replicate. These candidates were then used to determine the median enrichment factor for each ±TEX library pair. Using these medians all −TEX libraries were then normalized against the library with the strongest enrichment. Besides annotation of TSS, the resulting graphs were also used for visualization in the Integrated Genome Browser [31].

2.4.3. Automated TSS annotation using TSSpredator

Based on the normalized expression graphs automated TSS prediction was performed similar to Thomason and Bischler et al. [26] and Dugar et al. [25] using TSSpredator. In brief, for each position *(i)* in the expression graph corresponding to the +TEX libraries, the algorithm calculates an expression height, $e(i)$, and compares that expression height to the preceding position by calculating $e(i)-e(i-1)$, which is termed the flank height. Additionally, the algorithm calculates a factor of height change $e(i)/e(i-1)$. To determine if a TSS is a real TSS and not a processed transcript end an enrichment factor is calculated as $e_{+\mathrm{TEX}}(i)/e_{-\mathrm{TEX}}(i)$, where $e_{+\mathrm{TEX}}(i)$ is the expression height for the TEX-treated sample and $e_{-\mathrm{TEX}}(i)$ is the expression height for the untreated sample. For all positions where these parameters (flank height, factor of height change, and enrichment) exceed the predefined thresholds a TSS is annotated.

We set the thresholds for the *minimum flank height* and the *minimum factor of height change*, which are used to determine if a TSS is detected to 0.3 and 2.0, respectively. Here, the value for the *minimum flank height* is a factor to the minimum 90th percentile over all libraries resulting in an absolute value of 2.94 (for predictions based on 4 replicates). If the TSS candidate reaches these thresholds in at least one replicate, the thresholds are decreased for the other replicates to 0.1 (0.98 absolute) and 1.5, respectively. Furthermore, we set the *matching replicates* parameter, which determines the number of replicates in which a TSS must exceed these thresholds in order to be marked as detected to 3. A TSS candidate is considered to be enriched, if the enrichment factor at the respective nucleotide position for at least one replicate is ⩾2.0. In order to take into account slight variations between TSS positions the respective parameter for clustering between replicates was set to a value of 1. In doing so, a consensus TSS position in a 3 nt window is determined based on the maximum flank height among the respective libraries.

Predicted TSS were assigned to five different classes based on their location with respect to predefined annotations: primary TSS (pTSS, main TSS within 300 nt upstream of a gene or operon), secondary TSS (sTSS, alternative TSS with lower flank height), internal TSS (iTSS, TSS within a gene), antisense TSS (asTSS, TSS antisense to a gene in a distance ⩽100 nt), and orphan TSS (oTSS, TSS not associated with annotation). Please note that compared to our previous manually annotated TSS used in [12], we reduced the maximal window for pTSS and sTSS classification from 500 nt to 300 nt to have a more strict TSS classification. This might affect some of the classifications of previously annotated TSS, i.e. TSS ⩽500 and >300 nt upstream of annotated genes. For example, some of the TSS that are also classified as iTSS or asTSS might have lost the primary or secondary classification whereas TSS solely classified as pTSS or sTSS would be annotated as oTSS. Moreover, our automated TSS prediction and classification employs an updated annotation file, which now also contains the annotations for validated sRNAs from *H. pylori*. Thus, these are now also listed with their primary TSS in Table S1.

2.4.4. Availability of sequencing data

Raw sequencing reads in FASTQ format and coverage files normalized by TSSpredator in wiggle (WIG) format are available via Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/-geo) under accession number GSE67564.

Additionally, we used previous 454 sequencing data of a TEX-treated (+TEX) and an untreated library (−TEX) based on a sample collected at mid-log growth (B0) from the previous *H. pylori* dRNA-seq study [12] for which raw data were previously uploaded to the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra) under accession number SRA010186.

3. Results and discussion

3.1. The dRNA-seq approach for global mapping of TSS

The dRNA-seq approach allows for the precise mapping of TSS on a genome-wide scale via selective sequencing of primary transcripts [24]. For each biological sample, a cDNA library pair consisting of one library (+TEX) generated from RNA treated with terminator 5′ phosphate dependent exonuclease (TEX) and a second library (−TEX) generated from untreated total RNA is sequenced. TEX selectively digests processed transcripts with a 5′-P which results in an enrichment for primary transcripts that still carry a 5′-PPP in the +TEX library [12]. Fig. 1A depicts how TEX treatment of total RNA eliminates most of the processed RNAs including the abundant 16S and 23S rRNA. Another method that relies on initial TEX-treatment for depletion of processed transcripts employs a modified 5′ RACE approach [21,22]. However, compared to the dRNA-seq approach this approach does not include a direct comparison to an untreated library based on the same sample, which facilitates the discrimination of primary and processed transcripts.

An alternative strategy to identify TSS on a global scale is based on treatment of RNA with tobacco acid pyrophosphatase (TAP) and has been used for global identification of sRNAs and their TSS in *Clostridium difficile* [32] or the generation of a transcriptome map and analysis of pervasive transcription in *Propionibacterium acnes*

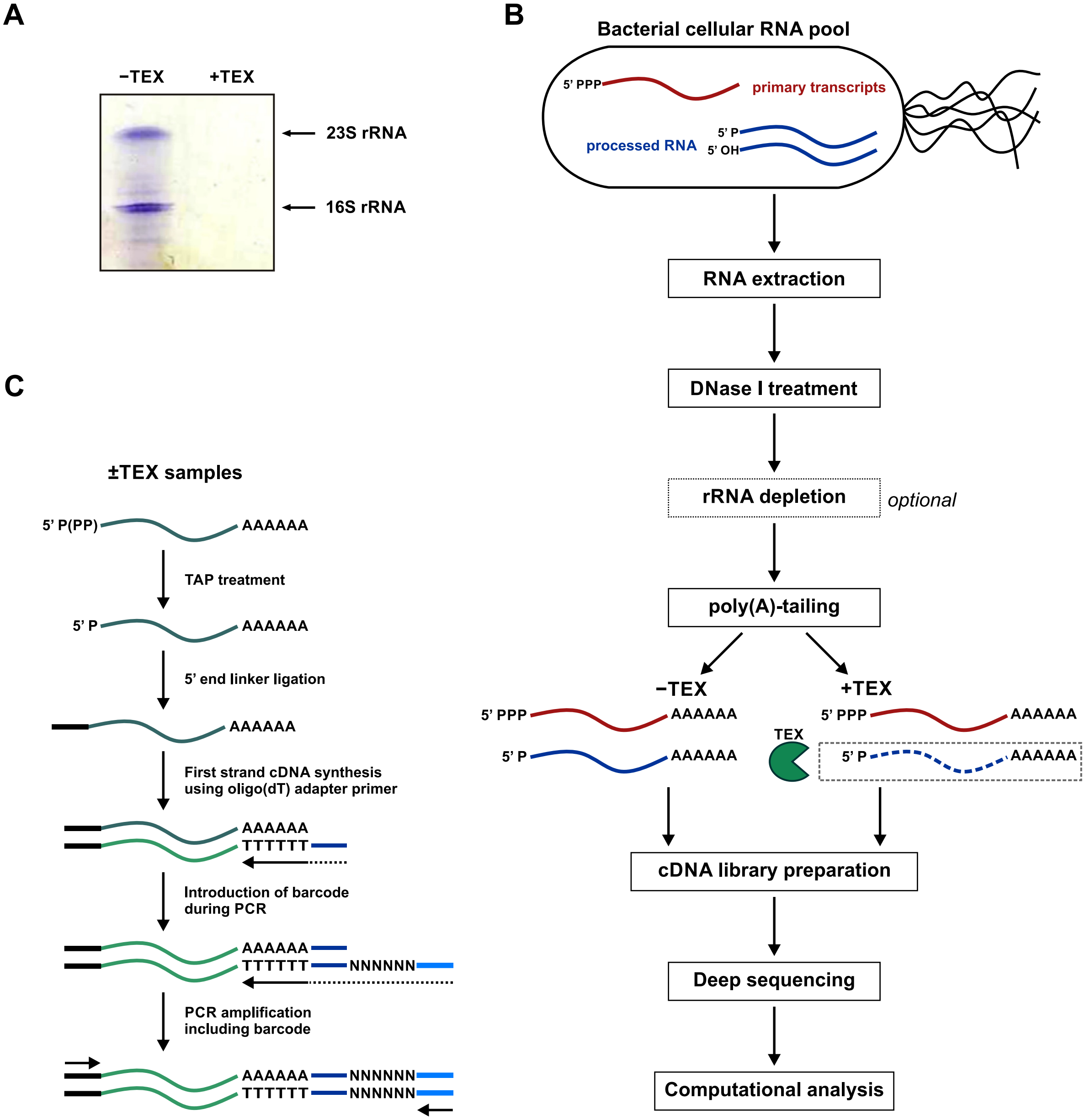

Fig. 1. Workflow for dRNA-seq-based primary transcriptome analysis. (A) *H. pylori* 26695 total RNA harvested at OD_{600} 0.6 with (+) and without (−) TEX treatment was separated on a 4% 7 M Urea polyacrylamide gel and stained with Stains-All (Sigma–Aldrich). Positions of bands for 16S and 23S rRNA are indicated on the right. (B) Representative workflow of a dRNA-seq experiment. (C) Illumina sequencing-specific cDNA library preparation protocol applied to both, +TEX and −TEX samples.

[33] as well as *Streptomyces coelicolor* and *Escherichia coli* [47]. TAP removes pyrophosphates from the 5′-PPP group of primary transcripts leaving a 5′-P end and making them accessible for 5′ end linker ligation. Comparison of library pairs generated from RNA with (+TAP) and without (−TAP) TAP treatment enables determination of TSS based on enrichment of primary transcripts in the +TAP versus the −TAP library. However, in contrast to the dRNA-seq approach this approach does specifically enrich for primary transcripts and does not deplete the abundant rRNAs and tRNAs so that deeper sequencing coverage might be required. A similar strategy was applied for TSS mapping in *E. coli* using 5′ polyphosphatase instead of TAP [23].

3.2. dRNA-seq of *Helicobacter pylori* strain 26695

Here, we describe the application of dRNA-seq using the gastric pathogen *H. pylori* 26695 as an example bacterium. *H. pylori* thrives in the acidic environment of the human stomach where it can cause gastritis, ulcers, gastric cancer and lead to lifelong, persistent infections [34,35]. *H. pylori* has a relatively small genome of 1.67 Mbp and encodes only for a small number of transcriptional regulators. The strain 26695 was originally isolated from a gastritis patient in the United Kingdom and was one of the first bacteria with a sequenced genome [36]. Strain 26695 is one of the most widely used strains in *H. pylori* research and a genome-wide map of TSS and operons based on dRNA-seq data was previously generated for this strain grown under five different biological conditions [12]. The conditions comprised bacteria grown to mid-logarithmic phase (ML), which represented the reference growth condition, or under acid stress (AS), grown in contact with responsive gastric epithelial AGS cells (AG) or non-responsive liver cells (HU), or in cell culture medium alone (PL). For each of these conditions, a single library was constructed and between 200,000 and 500,000 cDNA reads for each sample were generated by 454 pyrosequencing.

For every RNA-seq experiment, one important decision to make is the selection of an appropriate sequencing technology. Several platforms with differences in read length, reads per run, accuracy, price and time per run [37] are available on the market. Here, we will focus on protocols and data analyses that apply to the Illumina sequencing technology, which is currently the most widely-used platform for RNA-seq. To illustrate the necessary steps for generation of a TSS map and to assess the effects of the deeper Illumina sequencing coverage and the use of several replicates in comparison to the previous TSS annotations, we collected three biological replicate RNA samples (B1–B3) from *H. pylori* 26695 wild-type cells grown to mid-logarithmic phase in rich BHI medium +10% FCS. The protocol used to generate dRNA-seq data from these samples is shown in Fig. 1B and details are listed in the Materials and Methods section. After collection of cell samples, total RNA was isolated using the hot-phenol extraction [12,25,27] (see Section 2.2). It is crucial to obtain high-quality RNA in this step to avoid extensive sequencing of rRNA degradation fragments. Thus, an RNA quality check on agarose gels or using Bioanalyzer chips is recommended after removal of residual genomic DNA via DNase I treatment (see Section 2.2). An additional rRNA depletion is optional and is not necessary in most cases since sequencing coverage is no longer limiting. Due to the removal of processed RNAs, TEX treatment also decreases the fraction of rRNAs and tRNAs. Thus, together with the additional depletory effects due to lower preference of poly(A) addition by *E. coli* poly(A) polymerase (PAP I) described below and lower efficiency in reverse transcription for structured rRNAs during library construction, no additional rRNA depletion steps are required in a typical dRNA-seq experiment.

For the preparation of dRNA-seq libraries, either the ±TEX treatment can be the first step or each RNA sample can be first polyadenlyated using PAP I, followed by differential TEX treatment. Here we describe the latter order (Fig. 1B and C), which has the advantage that it ensures equal poly(A)-tailing for the corresponding ±TEX library pairs. The cDNA libraries for Illumina sequencing were generated in the same way for +TEX and −TEX samples and experimental details are given in Section 2.3. Strand-specificity of the sequencing is crucial to distinguish sense from antisense transcripts. In our method this is achieved by attaching a 5′ RNA adapter and a poly(A)-tail to each fragment prior to cDNA synthesis. First a poly(A)-tail was attached to the RNA molecules. It was shown that PAP I has a preference of polyadenylating mRNAs over rRNAs resulting in an inherent rRNA depletion in the resulting cDNA library [38]. Afterwards, each of the poly(A)-tailed biological replicates B1-B3 was split into two halves which were then differentially treated with TEX, resulting in −TEX samples covering RNAs with a 5′-P and a 5′-PPP and +TEX samples that are enriched for 5′-PPP RNAs. Next, the ±TEX samples were treated with TAP to cleave the 5′-PPP groups of primary transcripts leaving a 5′-P. This step is necessary to enable subsequent ligation of the 5′ end linkers that cannot be ligated to a 5′-PPP end. Please note that processed transcripts with a 5′-OH are not covered in the final cDNA libraries, although they are resistant to TEX removal, since they are not accessible for 5′ end RNA linker ligation. In case one is interested in capturing this class of transcripts, an additional treatment with polynucleotide kinase and ATP is required to generate 5′-P ends (for a protocol see [39]). After TAP treatment, an RNA linker was ligated to the transcripts in the ±TEX samples. Next, first strand cDNA was generated using an oligo(dT)-adapter primer and library-specific barcodes were introduced during PCR amplification of each library. All libraries were sequenced on either an Illumina 2000 (B1-HS1) or 2500 machine (B2-HS2 and B3-HS2). In total we sequenced between 4.1 and 8.1 Mio cDNA reads per library (Table 1). This represents a more than 10-fold higher coverage compared to the previous 454 libraries [12].

It was shown that the construction of cDNA libraries can be a major source of variation among RNA-seq experiments based on the same organism in both pro- and eukaryotes [26,40]. Especially, additional bias might be introduced by distinct library preparation protocols for different sequencing platforms due to differences in ligation efficiency and RNA structure or G/C-content-dependent differences in reverse transcription or PCR amplification efficacy [41]. The resulting variation in amplification of certain transcripts could be an explanation for observed differences among distinct studies of the same organism [40]. When comparing biological and technical replicates in a dRNA-seq analysis of *E. coli*, we observed larger variation for distinct library preparations from the same biological sample than among biological replicates for which the libraries were generated in parallel [26]. We therefore recommend, if possible, conducting cDNA library preparation for all samples simultaneously. This is even more important for quantitative gene-expression profiling experiments compared to qualitative transcriptome annotation approaches such as dRNA-seq.

3.3. dRNA-seq data analysis

After Illumina sequencing of the six B1-B3 ±TEX libraries, we assessed read quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). This software provides for each library a summary, which includes different quality metrics as, for example, sequence length distribution, GC content distribution, presence of duplicated or overrepresented sequences, per-base N content and most importantly base call quality scores. Read quality for Illumina sequencing typically decreases at the 3′ end of longer

Table 1
Mapping statistics for the *H. pylori* 26695 Illumina dRNA-seq libraries. This table summarizes the total number of sequenced cDNA reads after quality trimming, as well as the number of mapped and uniquely mapped reads for each library. Percentage values are relative to the number of cDNA reads that are >11 nt after poly(A) trimming.

Library	Total number of reads after quality trimming	Number of reads long enough after poly(A) trimming (>11 nt)	Mapped reads	% Mapped reads	Uniquely mapped reads	% Uniquely mapped reads
ML B1-HS1 + TEX	4,105,444	2,904,136	2,855,756	98.3	1,776,256	61.2
ML B1-HS1 – TEX	4,709,180	4,393,218	4,303,527	98.0	2,191,371	49.9
ML B2-HS2 + TEX	6,979,343	6,747,193	6,694,900	99.2	4,121,103	61.1
ML B2-HS2 – TEX	8,128,096	8,051,963	8,008,388	99.5	4,011,524	49.8
ML B3-HS2 + TEX	6,700,169	6,402,059	6,351,658	99.2	3,952,168	61.7
ML B3-HS2 – TEX	7,435,053	7,344,125	7,302,057	99.4	4,160,876	56.7

reads. Therefore, preprocessing of reads is important to facilitate alignment to the reference genome. In our pipeline (Fig. 2A) we conducted quality trimming from the 3′ end, poly(A) trimming, and size filtering in order to generate a set of high quality reads which were afterwards mapped to the *H. pylori* 26695 reference genome (NC_000915.1). For all steps starting from poly(A) trimming until coverage plot generation we used the RNA-seq analysis tool READemption [29].

To examine the percentage of reads mapped to individual RNA classes, we calculated the number of reads that overlapped for at least 10 nt in either sense or antisense direction, annotations for 5′UTRs, mRNAs, sRNAs, rRNAs, tRNAs and housekeeping RNAs (RNase P RNA, SRP RNA, tmRNA and 6S RNA) based on the previously generated *H. pylori* transcriptome map (Table 2) [12]. The amount of reads mapping to rRNA ranged between 48 and 55% for the −TEX libraries and between 35 and 38% for the +TEX libraries, indicating that TEX depletes these processed transcripts. Moreover, even in the −TEX libraries the observed rRNA fraction is lower than the expected 90–95% for abundant rRNAs which might be caused by multiple factors: (i) the poly(A)-tailing with lower preference for rRNAs during library construction mentioned above, (ii) the fact that no RNA fragmentation was conducted prior to cDNA synthesis, which would result in a large amount of rRNA fragments and (iii) the lower efficiency of reverse transcription of structured RNA. This further shows that an additional rRNA depletion step is not necessarily required in our protocol. Moreover, a clear enrichment (at least 2-fold) of the fractions of reads mapping to sRNAs as well as 5′UTRs is observed in the +TEX libraries compared to the respective −TEX libraries, showing a successful enrichment of primary transcripts and the 5′ ends of transcripts. Please note that sequencing initiates from the 5′ adapter, which further enriches for the 5′ ends of transcripts.

Based on the read mappings we computed per-strand coverage plots for each library (for details see Section 2.4) that indicate the number of mapped reads per nucleotide. In case a read mapped with the same score to multiple regions in the genome, only a corresponding fraction, e.g. a score of 0.5 reads in case of two equal mappings, was counted for the respective positions. The resulting cDNA coverage plots allow examination of the transcriptome in a genome browser, e.g., the Integrated Genome Browser (IGB) [31], with single nucleotide resolution. The visualized RNA-seq data can then be used for annotation of transcript boundaries or novel transcripts such as sRNAs.

3.4. Identification of TSS based on dRNA-seq

The differential RNA-seq approach leads to a characteristic cDNA coverage pattern of dRNA-seq library pairs at TSS. The

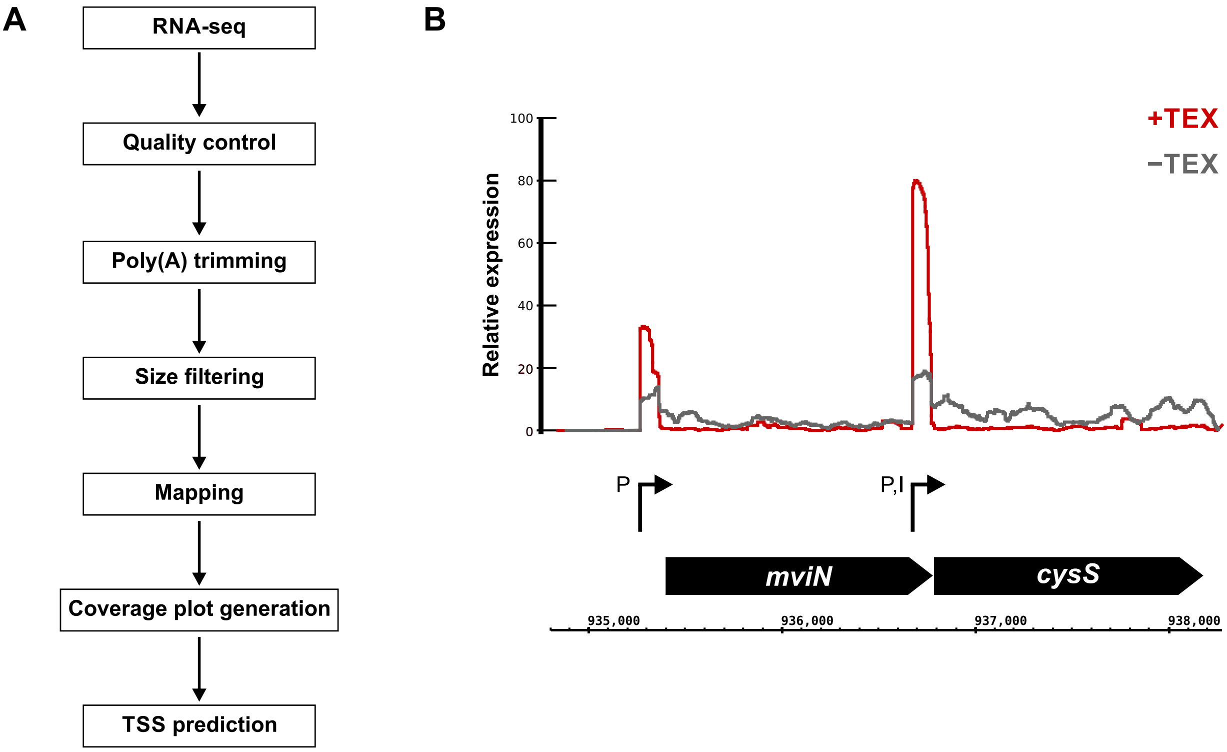

Fig. 2. Computational dRNA-seq analysis and TSS enrichment. (A) Workflow of the dRNA-seq data analysis pipeline. (B) Illustration of a representative cDNA enrichment pattern in the +TEX versus −TEX library at a TSS located upstream of the *mviN* gene and at a TSS internal to *mviN* and upstream of the *cysS* gene.

Table 2
Mapping statistics of cDNA reads based on strand and type of RNA class. This table indicates the number of cDNA reads that were mapped to the different RNA classes (5′UTR, mRNA, sRNA, rRNA, tRNA and housekeeping RNA) for each library. The numbers for the mapped reads per RNA class are shown with percentage values calculated from the total number of mapped reads regardless of mapped location (taken from Table 1) for the respective library. Housekeeping RNAs are RNase P RNA, SRP RNA, tmRNA and 6S RNA.

		Illumina library											
		ML B1 HS1 + TEX		ML B1 HS1 – TEX		ML B2 HS2 + TEX		ML B2-HS2 − TEX		ML B3 HS2 + TEX		ML B3 HS2 − TEX	
	Total*	2,855,756		4,303,527		6,694,900		8,008,388		6,351,658		7,302,057	
Sense	5′UTR	163,535	(6%)	119,712	(3%)	843,181	(13%)	336,795	(4%)	863,048	(14%)	349,806	(5%)
	mRNA	297,647	(10%)	895,532	(21%)	1,109,869	(17%)	2,467,397	(31%)	1,072,325	(17%)	2,641,220	(36%)
	sRNA	596,772	(21%)	255,731	(6%)	1,386,141	(21%)	210,665	(3%)	1,293,252	(20%)	196,253	(3%)
	rRNA	1,087,360	(38%)	2,205,744	(51%)	2,426,360	(36%)	4,382,063	(55%)	2,246,962	(35%)	3,497,798	(48%)
	tRNA	351,622	(12%)	428,298	(10%)	155,451	(2%)	76,466	(1%)	144,869	(2%)	75,009	(1%)
	Housekeeping RNA	113,695	(4%)	125,315	(3%)	201,105	(3%)	94,001	(1%)	183,972	(3%)	93,288	(1%)
Antisense	5′UTR	6,553	(0%)	4,650	(0%)	27,975	(0%)	4,052	(0%)	24,387	(0%)	4,551	(0%)
	mRNA	120,994	(4%)	125,865	(3%)	246,790	(4%)	119,946	(1%)	243,755	(4%)	130,290	(2%)
	sRNA	81,927	(3%)	39,202	(1%)	169,463	(3%)	49,897	(1%)	152,405	(2%)	43,984	(1%)
	rRNA	181	(0%)	204	(0%)	312	(0%)	54	(0%)	399	(0%)	57	(0%)
	tRNA	202	(0%)	401	(0%)	376	(0%)	693	(0%)	432	(0%)	765	(0%)
	Housekeeping RNA	114	(0%)	191	(0%)	200	(0%)	80	(0%)	308	(0%)	103	(0%)

* Total reads for each library also include reads that mapped to locations other than the listed RNA classes.

specific enrichment pattern (Fig. 2B) of the +TEX library compared to the corresponding −TEX library [12] indicates the start of a primary transcript and can thus be used to annotate TSS. Based on global examination of these enrichment patterns it is possible either to conduct manual TSS annotation based on visual inspection of the coverage plots in a genome browser, such as the IGB, or to use a tool for automated TSS annotation based on dRNA-seq data. In the previous dRNA-seq analysis *of H. pylori* 26695, we manually annotated 1907 unique TSS based on visual inspection of the enrichment patterns among the five examined growth conditions. Manual TSS annotation is laborious and time-consuming, especially when applied to large genomes or the comparative analysis of multiple strains or conditions including biological replicates. In comparison to automated TSS prediction, which follows defined rules, it is also likely to introduce human bias, based on individual perception of the data and thus might lead to different results for unclear cases. However, manual inspection can be useful in single cases to either confirm automated predictions or check for TSS patterns that were misinterpreted based on predefined parameter thresholds.

Multiple groups have in the meanwhile developed diverse TSS prediction tools, which make use of the information provided by dRNA-seq [42,43]. Here, we computationally annotated TSS in the three B1-B3 dRNA-seq data sets of *H. pylori* 26695 grown to mid-log phase utilizing the software TSSpredator (http://it.inf.uni-tuebingen.de/TSSpredator). We had originally applied this tool for a comparative TSS annotation in a dRNA-seq analysis of multiple *C. jejuni* strains [25], and also successfully applied it for TSS predictions among different growth conditions in *E. coli* K12 MG1655 [26]. To our knowledge, it is the most flexible of all currently available programs, with implemented support for comparative analysis and varying numbers of replicates. While other tools incorporate elaborate statistics to decide which genomic positions represent a TSS, TSSpredator applies specific heuristics to imitate manual TSS annotation with a set of tunable parameters. To ensure comparability between replicates, TSSpredator conducts additional normalization steps on the expression graphs of both, +TEX and −TEX libraries. Afterwards, each genomic position is checked for the presence of a potential TSS by assessing flank height and factor of height change in the +TEX libraries as well as enrichment between +TEX and −TEX libraries. When run comparatively, a TSS is annotated if it is detected and enriched in at least one strain or condition and in case multiple replicates are available the *matching replicates* parameter can be adjusted to determine the number of replicates in which a TSS must be detected but enrichment is only required in one of them. Here, we used the default settings of TSSpredator, which were established based on our manual annotation in *H. pylori* 26695 [12] and already applied in our previous studies [25,26].

In order to compare the TSS prediction based solely on the new Illumina mid-log dRNA-seq libraries to the manual annotations (1907 TSS) based on 454 data from the initial study [12], we ran TSSpredator using the three Illumina data sets as replicates. Requiring detection of a TSS in all replicates (*matching replicates* = 3), we predicted 1949 TSS. A comparison of these TSS positions with the 1907 manual TSS annotations requiring a precise match (cutoff 0 nt) resulted in 971 matching positions. The same comparison allowing for a maximum distance of three nt revealed an overlap of 1208 positions. This difference might be due to slight fluctuations in the actual TSS position for some promoters where transcription initiation is wobbly and the coverage shows a staircase-like pattern. In these cases, annotation of the major TSS is not always straightforward and slight variations in the libraries can lead to the annotation of neighboring positions. For this reason, we decided to tolerate such slight variations for this as well as subsequent comparisons reported below. The 1208 matching positions represent ~62% of our current predictions and ~63% of the previously annotated TSS (Fig. 3A). The additional 741 TSS predicted based on our current data are in most cases a result of the deeper coverage gained by Illumina sequencing and the support by several replicates for the mid-log growth condition. Previous TSS positions that are not detected in the Illumina dRNA-seq libraries are mainly caused by absence of or very low expression in at least one of the three Illumina replicates. In these cases, the respective TSS commonly shows a signal in one or more of the four other conditions assessed in the previous 454 study and was thus annotated. For example, in Fig. 3B the two TSS upstream of the HP0531 gene were annotated with matching positions in both, the previous manual annotation and the current TSSpredator prediction. The TSS within the HP0531 gene was only annotated by TSSpredator because there was no clear enrichment in the 454 data. In contrast, the TSS internal to HP0532 was only annotated in the 454 data as it was mainly expressed in the AS and HU conditions, but only very lowly expressed in the ML condition.

Furthermore, we noted some overall cDNA coverage variations, even within the same growth conditions, as observed for example between the B1-HS1 replicate and the B2/B3-HS2 replicates in Fig. 3B, which might be due to variations during library preparation. While such variations could be problematic for monitoring gene expression, especially of lowly expressed genes, when using

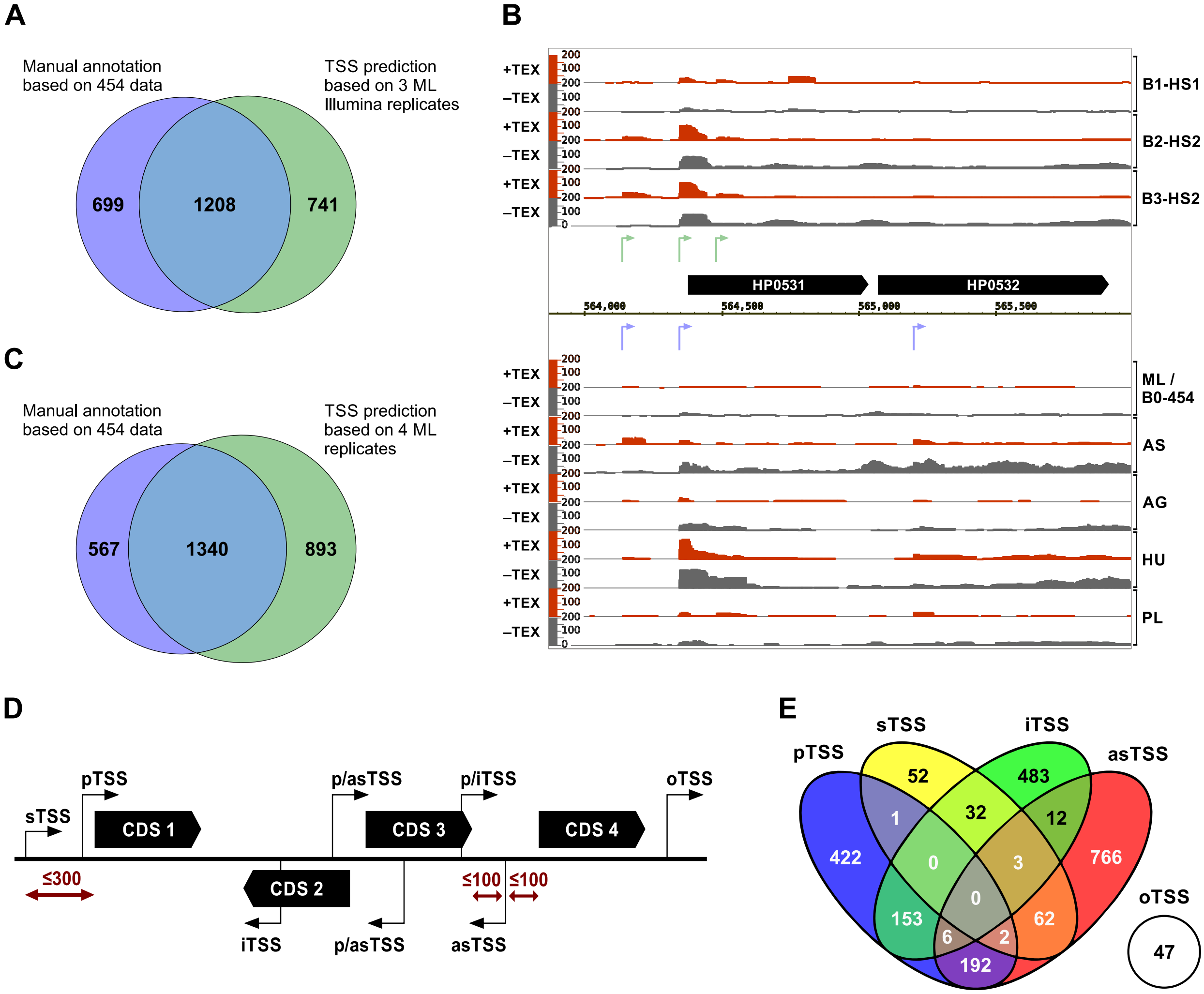

Fig. 3. TSS predictions in *H. pylori* 26695. (A) Comparison of manual TSS annotations from a previous study based on 454 sequencing of five different growth conditions to current TSSpredator predictions based on three biological replicates for the ML condition. (B) Example region encompassing the HP0531 and HP0532 genes, encoding the *cag* pathogenicity island proteins Cag11 and Cag12, respectively. cDNA coverage plots for the three Illumina ML replicates of our current data set are shown at the top with predicted TSS colored in green, while coverage plots for the five growth conditions from the previous 454 study [12] are shown at the bottom with manually annotated TSS colored in blue. Conditions or replicate identifiers are shown on the right (ML: mid-log growth; AS: acid stress; AG: *H. pylori* grown in the presence of AGS gastric cells; HU: *H. pylori* grown in the presence of Huh7 liver cells; PL: *H. pylori* in cell culture medium), while presence or absence of TEX treatment is indicated on the left. The *x*-axis reflects genomic positions while the *y*-axis indicates relative expression based on normalized read coverage. (C) Comparison of manual TSS annotations from the previous study based on 454 sequencing of five different growth conditions to current TSSpredator predictions based on four biological replicates (454 and Illumina) for the ML condition. (D) The location relative to annotated genes is depicted for the five different TSS classes (primary, secondary, internal, antisense, and orphan). The height of the black arrows indicates differences in expression strength while the distance cutoffs for flanking genes are shown in red. (E) The distribution according to TSS classes is depicted for the 2233 TSS predicted based on four ML replicates.

these data sets as replicates, this variability should not impede a qualitative dRNA-seq-based 5′ end mapping based on one or more conditions, since it does not affect the position of a TSS to be annotated. For measuring gene expression changes between different conditions, we recommend an approach that includes RNA fragmentation to cover full-length transcripts in combination with sample and library preparation in one experiment to reduce biological and technical variation among samples.

3.5. Comparison of H. pylori 454 and Illumina dRNA-seq data

To further examine the overlap between the old and new mid-log libraries and to investigate potential variation due to different sequencing platforms and library preparations, we complemented our three newly sequenced replicates with coverage plots for the ML condition based on 454 sequencing from [12] as an additional replicate (B0) and performed comparative TSSpredator predictions treating the four ML replicates as conditions. Using this setup, the resulting set of TSS encompassed 3240 distinct positions that were detected and enriched in at least one replicate and 1122 TSS which were found in all four (Fig. S1). A very good overlap (2211 TSS) with very few unique TSS positions for each library was observed between the B2-HS2 and B3-HS2 replicates, which were grown on the same day and for which library preparation was performed together, indicating that a careful and similar sample treatment is important to minimize variations. The second best overlap was observed between these two and the B1-HS1 replicate (1767 TSS), suggesting that slight differences in cultivation and potential biases introduced during library preparation can lead to differences in TSS expression and detection. The B1-HS1 and B0-454 replicates both introduced a similar amount of uniquely detected positions (392 and 329, respectively). This indicates again that not only differences in applied sequencing technologies and depth can play a role but also other experimental or technical variation like differences in treatment and library preparation. In order to generate a comprehensive and reliable TSS map for the ML condition, we repeated the TSS prediction, now treating the 4 ML libraries as replicates. As a tradeoff between reliability and tolerance of data variation, we set the *matching replicates* parameter to a value of 3 (for details see Section 2.4). In total, we predicted 2233 TSS that were detected in at least three and enriched in at least one of the 4 ML replicates (Table S1). Out of these 2233 TSS found in ML growth, 1340 TSS were found in the set of our 1907 manually annotated TSS [12] (Fig. 3C).

TSSpredator automatically assigns TSS to five different classes according to their location in relation to annotated genes: primary TSS (pTSS), secondary TSS (sTSS), internal TSS (iTSS), antisense TSS (asTSS), and orphan TSS (oTSS) (for details see Section 2.4 and Fig. 3D). Notably, one TSS can independently be assigned to more than one category as, for example, in the presence of alternative suboperons the pTSS of the downstream gene can also be internal to the upstream gene (Figs. 2B and 3D). Similarly, in the case of overlapping 5′UTRs, the associated TSS can be both, a pTSS and an asTSS. Among the 2233 mid-log TSS, we identified 776 pTSS (422 classified only as pTSS), 152 sTSS (52 classified only as sTSS), 689 iTSS (483 classified only as iTSS), 1043 asTSS (766 classified only as asTSS) and 47 oTSS (Fig. 3E). This classification is based on the same gene annotations for the 26695 strain from NCBI (NC_000915.1) which we already used in our previous study [12], supplemented with annotations for validated sRNAs that we discovered at this time.

3.6. Detection of regulatory elements

Knowledge of genome-wide TSS positions facilitates the discovery of diverse transcriptome features including regulatory elements. Global inference of promoter motifs upstream of TSS can help to understand which sequence elements are important for transcription initiation and elucidate gene regulation pathways. In the previous *H. pylori* dRNA-seq study, an extended −10 box downstream of periodic AT-rich stretches was identified as the canonical promoter motif for the housekeeping sigma factor σ^{80} [12]. The same motif was later confirmed in our comparative dRNA-seq analysis as the consensus for the housekeeping σ^{08} in *Campylobacter jejuni* [25] reinforcing that this is a common feature of ε-proteobacterial promoters.

Annotated pTSS and sTSS of mRNA genes can be used to generate transcriptome-wide 5′UTR maps that can subsequently be utilized to search for *cis*-regulatory elements such as riboswitches and RNA thermometers. Additionally, they can contain sRNA binding sites, for example, the 5′UTR of the chemotaxis receptor TlpB which contains a poly(G) stretch far upstream of the start codon which is targeted by the sRNA RepG [44]. Our TSS map includes 925 pTSS and sTSS of which 790 are associated with mRNA genes. 20 of these TSS give rise to leaderless transcripts while the remaining 770 5′UTRs show an average and median length of ~81 and 45 nt, respectively, and a clear peak in the distribution in a range between 20 and 40 nt (data not shown). This is consistent with earlier findings [12] and while leaderless mRNAs used to be considered rare in prokaryotes, unexpectedly high numbers have also been discovered in other bacteria [45–47]. On the other hand, in archaea, where leaderless mRNAs seem to represent the standard translational template, a dRNA-seq-based study in *Methanosarcina mazei* revealed that most mRNAs carry long 5′UTRs [48]. These findings underline the importance of 5′UTRs for translational control and the usefulness of dRNA-seq for their annotation.

3.6.1. Identification of cis- and trans-encoded sRNAs

The TSS map does not only provide information on transcription starts and regulatory mechanisms associated with already annotated genes or operons but also facilitates the discovery of novel regulatory elements, including sRNAs expressed from intergenic regions or antisense to ORFs. In the previous 454 dRNA-seq study we identified >60 sRNAs in *H. pylori* and an extensive antisense transcriptome. Fig. 4A shows an oTSS located in the intergenic region between the HP1399 and HP1400 genes annotated as arginase and iron(III) dicitrate transport protein FecA, respectively. This TSS was annotated as pTSS for HP1400 in our previous study as the 454 data did not provide any evidence for the existence of the newly predicted pTSS 294 nucleotides further downstream. The downstream TSS is clearly visible in the Illumina data and was also already mapped before by Ernst et al. [49] 2 nt further upstream via primer extension. Such oTSS could either belong to separate standing sRNA genes (e.g. an unannotated sRNA of ~220 nt in the case of the TSS upstream of HP1400) or represent alternative promoters leading to transcription of longer 5′UTRs. The example of the TSS upstream of HP1400 indicates that even more transcriptome features can still be discovered when sequencing at higher coverage or more conditions are included.

Another prominent class of transcripts that is getting more and more attention are antisense RNAs (asRNAs) [50,51]. In the 454 data, >900 asTSS were detected and at least one asTSS expressed opposite of >50% of all genes. Based on our new mid-log data we detected 766 TSS solely classified as asTSS, indicating again a large set of asRNA candidates. 52% of these asTSS overlap with the 684 TSS solely classified as asTSS in the 454 datasets. Fig. 4B shows an example for an asTSS located internal and antisense to the *ispDF* gene annotated as bifunctional 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase/2-C-methyl-D-erythritol 2,4-cyclodiphosphate. This TSS was also previously annotated based on the 454 data. We do not know how many of our predicted asTSS represent functional RNAs or the amount corresponding only to spurious transcripts, as the number of reported antisense RNAs strongly varies and the function of most of these transcripts is still unclear [52]. However, we think that a global TSS map is the optimal starting point to find an answer to this question by conducting additional experiments like, for example, detection on Northern blots and discovery of associated phenotypes. Moreover, regulation of the asTSS expression under different growth or stress conditions as well as conservation in multiple strains could be further indications that they indeed have regulatory functions [26].

3.6.2. Overlapping 5′UTRs

Association of a TSS to more than one class, as mentioned above, can be used to select for regulatory elements. Divergently transcribed gene pairs with overlapping regions in the 5′UTR or even coding sequence (CDS) can result in asRNA-mediated gene regulation (reviewed in [53]) or affect promoter occupancy [54]. We found 200 pTSS and 67 sTSS that were additionally classified as asTSS. Requiring a minimum overlap of 10 nt and considering only TSS for mRNA genes, we identified 40 distinct overlapping 5′UTRs associated with 28 divergently transcribed gene pairs (Table S2). One example is shown in Fig. 4C, which depicts two hypothetical proteins (HP1162 and HP1163) with their associated pTSS. The 5′UTR of HP1162 almost completely overlaps CDS and 5′UTR of HP1163, possibly resulting in an asRNA-mediated regulation.

3.7. Accessibility of the H. pylori 26695 TSS map in an online browser

In the previous 454 dRNA-seq study, we provided the TSS map in a table that indicated the TSS positions. While such a table format is very useful for downstream analysis such as promoter motif predictions or 5′UTR calculations, sometimes it is also helpful to look at the cDNA coverage plots to see the overall read distribution for a gene of interest. Thus, we here used GenomeView [55] to set up an easily accessible online browser that directly includes the

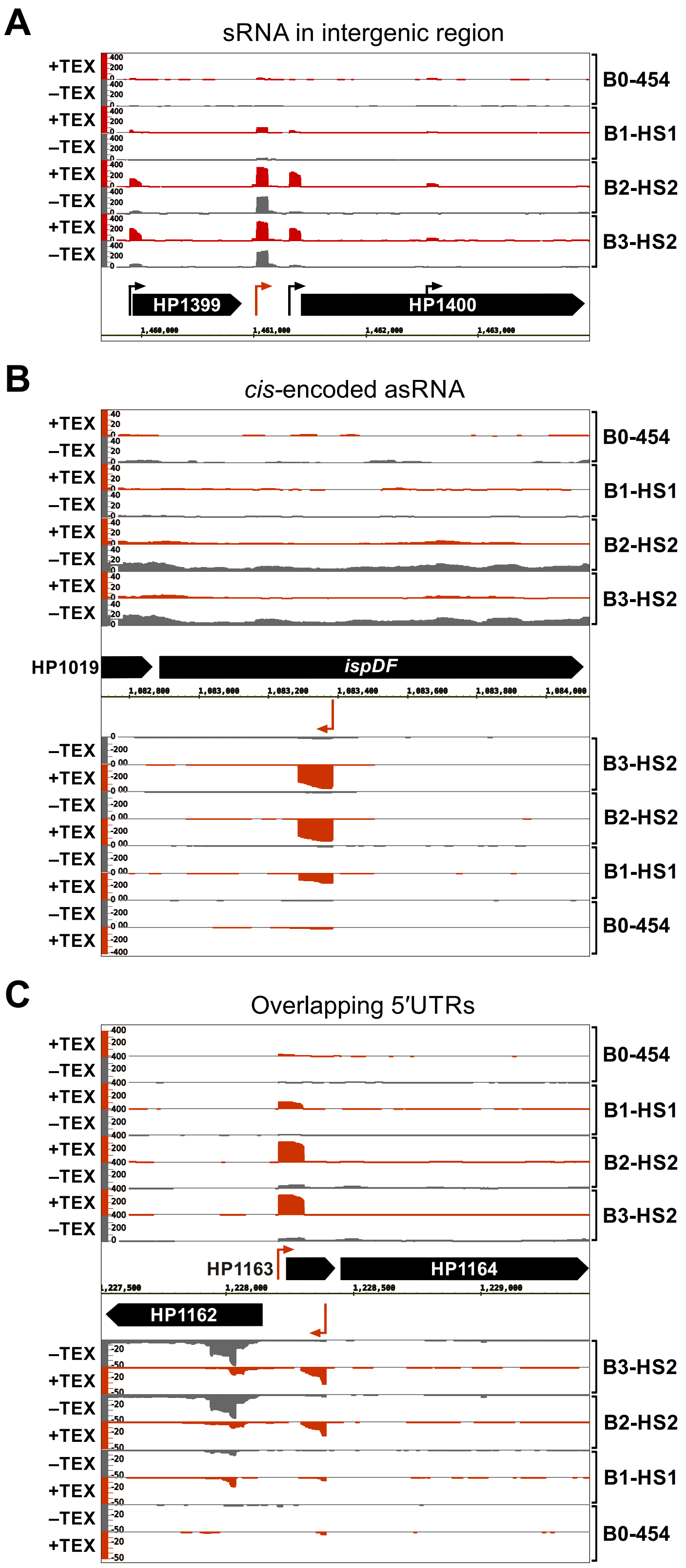

Fig. 4. Examples for transcripts and regulatory elements. Screenshots from IGB showing the relative cDNA coverage plots for ±TEX libraries of the four ML replicates. Red arrows indicate the genomic position of (A) a putative sRNA in the intergenic region between the HP1399 and HP1400 genes, (B) a putative asRNA transcribed from the opposite strand of the *ispDF* gene, and (C) two predicted p/asTSS for the divergently transcribed HP1162 and HP1163 genes, which indicate the presence of overlapping 5′UTRs.

complete set of predicted TSS from this study together with the respective coverage plots and gene annotations that were used for the prediction (Fig. 5). For comparison, we also added the previous manual TSS annotations from [12] for which coverage plots of all five biological conditions are loaded but not displayed by default. The browser allows for manual inspection of the data

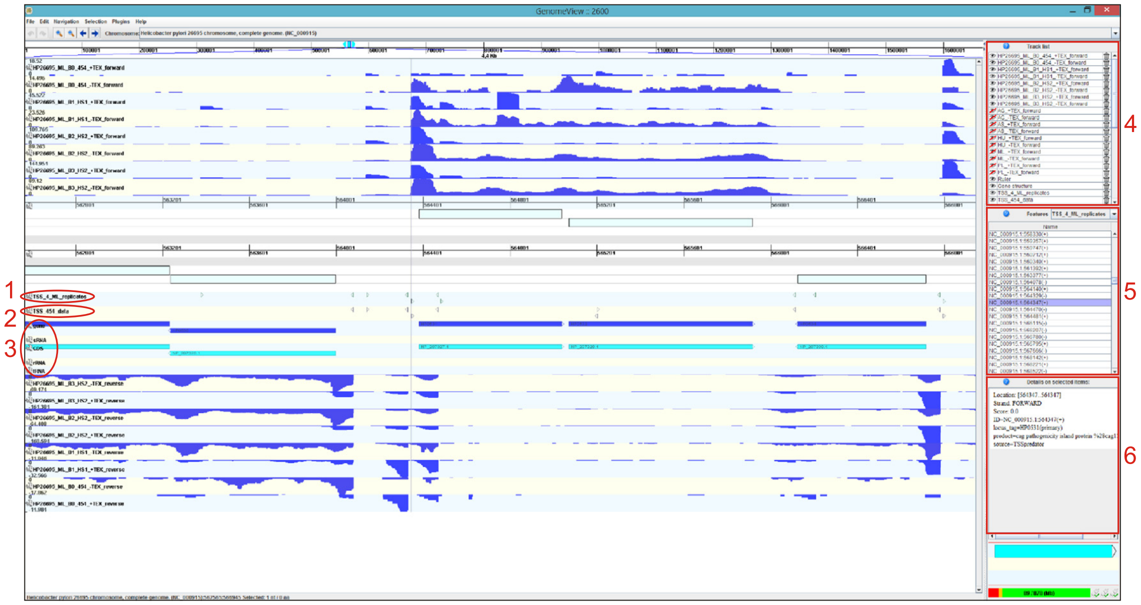

Fig. 5. Example screenshot of the online browser. cDNA coverage plots for the forward and reverse strand are displayed above and below the genomic axis, respectively. The browser depicts our new TSS annotations based on all 4 ML replicates (1), TSS annotations from the previous study based on 454 sequencing [12] (2), and annotations for genes, coding sequences (CDS), sRNAs, rRNAs and tRNAs (3). On the right the display and order of tracks can be altered in the track list (4), specific features can be selected (5) and details for selected items are displayed (6).

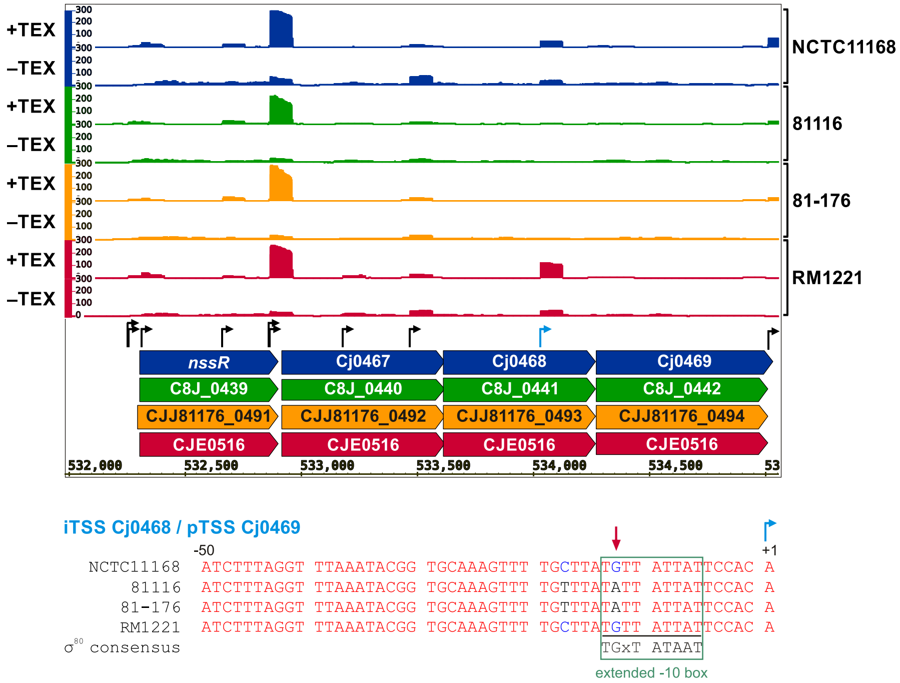

Fig. 6. Comparative TSS annotation in *C. jejuni* strains reveals strain-specific promoter usage. (Top) cDNA coverage for an example region of the SuperGenome of four *C. jejuni* strains which encompasses the *nssR*, Cj0467, Cj0468, Cj0469 operon. Black arrows indicate annotated TSS and the blue arrow a p/iTSS internal to Cj0468 and upstream of Cj0469 which is only detected in two strains (NCTC11168 and RM1221) and shows no expression in the other strains (81116 and 81-176). (Bottom) Multiple alignment of the promoter region −50 to +1 upstream of the blue p/iTSS based on the four *C. jejuni* strains. Differential expression of this TSS is likely caused by a G to A single nucleotide polymorphism (SNP) (red arrow) in the extended −10 box of strains 81116 and 81-176

including expression and enrichment at annotated TSS. Please note that there is no option for a consistent scaling of all coverage plots, which makes it necessary to compare the numbers representing relative expression values at the left end of each track. This online browser, which is available under http://www.imib-wuerzburg.de/research/hpylori/, greatly facilitates the data accessibility and allows researchers to examine the cDNA coverage plots and TSS for their genes of interest.

4. Conclusions

Genome-wide annotation of transcriptional features is crucial to understand the full complement of transcriptional regulation in an organism. Knowledge of precise positions of transcript 5′ ends gained by dRNA-seq is fundamental for a variety of downstream analyses like global prediction of promoter motifs or automated annotation of *cis*-regulatory features in 5′UTRs. In addition, it provides a basis for the annotation of a plethora of novel transcripts including sRNAs and asRNAs as well as specific regulatory features like antisense-mediated regulation via overlapping 5′UTRs. Here, we provided a detailed description of the application of the dRNA-seq method to generate a transcriptome-wide TSS map using *H. pylori* 26695 as an example organism. We utilized an automated TSS prediction approach implemented in the tool TSSpredator, which greatly facilitates TSS annotation on a global scale.

We compared the predicted TSS positions based on four replicates of the ML condition to previous manual annotations from our initial study [12] and detected 1340 matching positions but in addition 893 novel TSS, which were previously missed due to low coverage or insufficient support by several growth conditions. Other TSS positions might have been missed as they were not expressed in the ML condition or due to a lack of enrichment in the +TEX library. This could for example be caused by processing of primary transcripts by the RNA pyrophosphohydrolase RppH which was shown to initiate degradation via cleavage of the 5′-PPP [56]. Moreover, some of the differences in TSS could also be due to slight differences in growth conditions or mutations in the 26695 clones upon sequential passages in different labs.

The TSSpredator tool is also capable of comparative analysis based on different bacterial strains or biological conditions. In a previous study, we used dRNA-seq together with TSSpredator to annotate TSS in four *Campylobacter jejuni* strains [25] in a comparative manner. Using a whole-genome alignment of multiple strains calculated by Mauve [57], TSSpredator computes a common coordinate system for all strains referred to as SuperGenome and TSS are then annotated by directly applying the above-mentioned detection and enrichment criteria to corresponding genomic positions. An example for a TSS that is only present in two of the four strains is shown in Fig. 6. The difference is likely caused by a single base mutation in the extended −10 box of the promoter region for the p/iTSS displayed in blue. The G at the second position of the consensus motif (TGxTATAAT) is replaced by an A in strains 81116 and 81-176 abolishing transcription in these strains. In strains NCTC11168 and RM1221 the TSS within Cj0468 uncouples transcription of the Cj0469 gene encoding an amino-acid ABC transporter ATP-binding protein from the *nssR*, Cj0467, Cj0468, Cj0469 operon. This indicates that while most comparative genomics studies consider SNPs in open reading frames that can lead to frameshift mutations or change protein function, also SNPs in non-coding parts can contribute to strain-specific gene expression and regulation and thereby add yet another layer of complexity. Such a comparative transcriptome analysis of multiple isolates might also help to examine the conservation and potential functions of the increasing number of *cis*-encoded antisense RNAs and helps to reveal conserved and strain-specific or species-specific sRNAs [25,58].

Overall, a comparative TSS analysis of multiple *H. pylori* strains and or *H. pylori* grown under different stress or growth conditions will provide further insight into conserved and strain-specific transcriptional features of this widespread human pathogen, which might underlie phenotypic differences among closely related strains. Together with variable host factors, these might contribute to the different clinical outcomes observed for *H. pylori* infections and to establish life-long persistent infections and adaptation to changing conditions in the human stomach.

Acknowledgements

We thank Anika Lins for technical assistance, Richard Reinhardt for help with deep sequencing, Alexander Herbig for discussions on TSSpredator, and Jörg Vogel for sharing previous 454 *H. pylori* RNA-seq data. The Sharma lab received financial support from the ZINF Young Investigator program at the Research Center for Infectious Diseases (ZINF) in Würzburg, Germany, the Bavarian Research Network for Molecular Biosystems (BioSysNet), DFG project Sh580/1-1, the Daimler-Benz-Foundation, and the Young Academy program of the Bavarian Academy of Sciences.

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.ymeth.2015.06.012.

References

[1] K.O. Mutz, A. Heilkenbrinker, M. Lonne, J.G. Walter, F. Stahl, Curr. Opin. Biotechnol. 24 (2013) 22–30.
[2] N.J. Croucher, N.R. Thomson, Curr. Opin. Microbiol. 13 (2010) 619–624.
[3] A.H. van Vliet, FEMS Microbiol. Lett. 302 (2010) 1–7.
[4] Z. Wang, M. Gerstein, M. Snyder, Nat. Rev. Genet. 10 (2009) 57–63.
[5] R. Sorek, P. Cossart, Nat. Rev. Genet. 11 (2010) 9–16.
[6] N.F. Lahens, I.H. Kavakli, R. Zhang, K. Hayer, M.B. Black, H. Dueck, A. Pizarro, J. Kim, R. Irizarry, R.S. Thomas, G.R. Grant, J.B. Hogenesch, Genome Biol. 15 (2014) R86.
[7] N.J. Croucher, M.C. Fookes, T.T. Perkins, D.J. Turner, S.B. Marguerat, T. Keane, M.A. Quail, M. He, S. Assefa, J. Bahler, R.A. Kingsley, J. Parkhill, S.D. Bentley, G. Dougan, N.R. Thomson, Nucl. Acids Res. 37 (2009) e148.
[8] N. Cloonan, A.R. Forrest, G. Kolle, B.B. Gardiner, G.J. Faulkner, M.K. Brown, D.F. Taylor, A.L. Steptoe, S. Wani, G. Bethel, A.J. Robertson, A.C. Perkins, S.J. Bruce, C.C. Lee, S.S. Ranade, H.E. Peckham, J.M. Manning, K.J. McKernan, S.M. Grimmond, Nat. Methods 5 (2008) 613–619.
[9] Y. He, B. Vogelstein, V.E. Velculescu, N. Papadopoulos, K.W. Kinzler, Science 322 (2008) 1855–1857.
[10] D. Parkhomchuk, T. Borodina, V. Amstislavskiy, M. Banaru, L. Hallen, S. Krobitsch, H. Lehrach, A. Soldatov, Nucl. Acids Res. 37 (2009) e123.
[11] A. Sittka, S. Lucchini, K. Papenfort, C.M. Sharma, K. Rolle, T.T. Binnewies, J.C. Hinton, J. Vogel, PLoS Genet. 4 (2008) e1000163.
[12] C.M. Sharma, S. Hoffmann, F. Darfeuille, J. Reignier, S. Findeiß, A. Sittka, S. Chabas, K. Reiche, J. Hackermüller, R. Reinhardt, P.F. Stadler, J. Vogel, Nature 464 (2010) 250–255.
[13] P.T. McGrath, H. Lee, L. Zhang, A.A. Iniesta, A.K. Hottes, M.H. Tan, N.J. Hillson, P. Hu, L. Shapiro, H.H. McAdams, Nat. Biotechnol. 25 (2007) 584–592.
[14] J.A. Thompson, M.F. Radonovich, N.P. Salzman, J. Virol. 31 (1979) 437–446.
[15] L. Argaman, R. Hershberg, J. Vogel, G. Bejerano, E.G. Wagner, H. Margalit, S. Altuvia, Curr. Biol. 11 (2001) 941–950.
[16] J. Vogel, V. Bartels, T.H. Tang, G. Churakov, J.G. Slagter-Jager, A. Huttenhofer, E.G. Wagner, Nucl. Acids Res. 31 (2003) 6435–6443.
[17] B.A. Bensing, B.J. Meyer, G.M. Dunny, Proc. Natl. Acad. Sci. USA 93 (1996) 7794–7799.
[18] O. Wurtzel, R. Sapra, F. Chen, Y. Zhu, B.A. Simmons, R. Sorek, Genome Res. 20 (2010) 133–141.
[19] A. Mendoza-Vargas, L. Olvera, M. Olvera, R. Grande, L. Vega-Alvarado, B. Taboada, V. Jimenez-Jacinto, H. Salgado, K. Juarez, B. Contreras-Moreira, A.M. Huerta, J. Collado-Vides, E. Morett, PLoS One 4 (2009) e7526.
[20] B.K. Cho, K. Zengler, Y. Qiu, Y.S. Park, E.M. Knight, C.L. Barrett, Y. Gao, B.O. Palsson, Nat. Biotechnol. 27 (2009) 1043–1049.
[21] B.K. Cho, D. Kim, E.M. Knight, K. Zengler, B.O. Palsson, BMC Biol. 12 (2014) 4.
[22] D. Kim, J.S. Hong, Y. Qiu, H. Nagarajan, J.H. Seo, B.K. Cho, S.F. Tsai, B.O. Palsson, PLoS Genet. 8 (2012) e1002867.
[23] N. Singh, J.T. Wade, Methods Mol. Biol. 1103 (2014) 1–10.
[24] C.M. Sharma, J. Vogel, Curr. Opin. Microbiol. 19C (2014) 97–105.
[25] G. Dugar, A. Herbig, K.U. Förstner, N. Heidrich, R. Reinhardt, K. Nieselt, C.M. Sharma, PLoS Genet. 9 (2013) e1003495.
[26] M.K. Thomason, T. Bischler, S.K. Eisenbart, K.U. Förstner, A. Zhang, A. Herbig, K. Nieselt, C.M. Sharma, G. Storz, J. Bacteriol. 197 (2015) 18–28.
[27] P. Blomberg, E.G. Wagner, K. Nordstrom, EMBO J. 9 (1990) 2331–2340.
[28] E. Berezikov, F. Thuemmler, L.W. van Laake, I. Kondova, R. Bontrop, E. Cuppen, R.H. Plasterk, Nat. Genet. 38 (2006) 1375–1377.
[29] K.U. Förstner, J. Vogel, C.M. Sharma, Bioinformatics 30 (2014) 3421–3423.
[30] S. Hoffmann, C. Otto, S. Kurtz, C.M. Sharma, P. Khaitovich, J. Vogel, P.F. Stadler, J. Hackermüller, PLoS Comput. Biol. 5 (2009) e1000502.
[31] J.W. Nicol, G.A. Helt, S.G. Blanchard Jr., A. Raja, A.E. Loraine, Bioinformatics 25 (2009) 2730–2731.
[32] O.A. Soutourina, M. Monot, P. Boudry, L. Saujet, C. Pichon, O. Sismeiro, E. Semenova, K. Severinov, C. Le Bouguenec, J.Y. Coppee, B. Dupuy, I. Martin-Verstraete, PLoS Genet. 9 (2013) e1003493.
[33] Y.F. Lin, D.A. Romero, S. Guan, L. Mamanova, K.J. McDowall, BMC Genomics 14 (2013) 620.
[34] T.L. Cover, M.J. Blaser, Gastroenterology 136 (2009) 1863–1873.
[35] S. Suerbaum, P. Michetti, N. Engl. J. Med. 347 (2002) 1175–1186.
[36] J.F. Tomb, O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, B.A. Dougherty, K. Nelson, J. Quackenbush, L. Zhou, E.F. Kirkness, S. Peterson, B. Loftus, D. Richardson, R. Dodson, H.G.

Khalak, A. Glodek, K. McKenney, L.M. Fitzegerald, N. Lee, M.D. Adams, E.K. Hickey, D.E. Berg, J.D. Gocayne, T.R. Utterback, J.D. Peterson, J.M. Kelley, M.D. Cotton, J.M. Weidman, C. Fujii, C. Bowman, L. Watthey, E. Wallin, W.S. Hayes, M. Borodovsky, P.D. Karp, H.O. Smith, C.M. Fraser, J.C. Venter, Nature 388 (1997) 539–547.

[37] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, M. Law, J. Biomed. Biotechnol. 2012 (2012) 251364.

[38] J. Frias-Lopez, Y. Shi, G.W. Tyson, M.L. Coleman, S.C. Schuster, S.W. Chisholm, E.F. Delong, Proc. Natl. Acad. Sci. USA 105 (2008) 3805–3810.

[39] N. Heidrich, G. Dugar, J. Vogel, C.M. Sharma, Methods Mol. Biol. 1311 (2015) 1–21.

[40] P.A. t Hoen, M.R. Friedlander, J. Almlof, M. Sammeth, I. Pulyakhina, S.Y. Anvar, J.F. Laros, H.P. Buermans, O. Karlberg, M. Brannvall, J.T. den Dunnen, G.J. van Ommen, I.G. Gut, R. Guigo, X. Estivill, A.C. Syvanen, E.T. Dermitzakis, T. Lappalainen, Nat. Biotechnol. 31 (2013) 1015–1022.

[41] C.A. Raabe, T.H. Tang, J. Brosius, T.S. Rozhdestvensky, Nucl. Acids Res. 42 (2014) 1414–1426.

[42] F. Amman, M.T. Wolfinger, R. Lorenz, I.L. Hofacker, P.F. Stadler, S. Findeiß, BMC Bioinf. 15 (2014) 89.

[43] H. Jorjani, M. Zavolan, Bioinformatics 30 (2014) 971–974.

[44] S.R. Pernitzsch, S.M. Tirier, D. Beier, C.M. Sharma, Proc. Natl. Acad. Sci. USA 111 (2014) E501–E510.

[45] T. Cortes, O.T. Schubert, G. Rose, K.B. Arnvig, I. Comas, R. Aebersold, D.B. Young, Cell Rep. 5 (2013) 1121–1131.

[46] A. de Groot, D. Roche, B. Fernandez, M. Ludanyi, S. Cruveiller, D. Pignol, D. Vallenet, J. Armengaud, L. Blanchard, Genome Biol. Evol. 6 (2014) 932–948.

[47] D.A. Romero, A.H. Hasan, Y.F. Lin, L. Kime, O. Ruiz-Larrabeiti, M. Urem, G. Bucca, L. Mamanova, E.E. Laing, G.P. van Wezel, C.P. Smith, V.R. Kaberdin, K.J. McDowall, Mol. Microbiol. 94 (5) (2014) 963–987.

[48] D. Jager, C.M. Sharma, J. Thomsen, C. Ehlers, J. Vogel, R.A. Schmitz, Proc. Natl. Acad. Sci. USA 106 (2009) (1882) 21878–21882.

[49] F.D. Ernst, J. Stoof, W.M. Horrevoets, E.J. Kuipers, J.G. Kusters, A.H. van Vliet, Infection Immunity 74 (2006) 6821–6828.

[50] J. Georg, W.R. Hess, Microbiol. Mol. Biol. Rev. 75 (2011) 286–300.

[51] M.K. Thomason, G. Storz, Annu. Rev. Genet. 44 (2010) 167–188.

[52] J.T. Wade, D.C. Grainger, Nat. Rev. Microbiol. 12 (2014) 647–653.

[53] N. Sesto, O. Wurtzel, C. Archambaud, R. Sorek, P. Cossart, Nat. Rev. Microbiol. 11 (2013) 75–82.

[54] K.M. Bendtsen, J. Erdossy, Z. Csiszovszki, S.L. Svenningsen, K. Sneppen, S. Krishna, S. Semsey, Nucl. Acids Res. 39 (2011) 6879–6885.

[55] T. Abeel, T. Van Parys, Y. Saeys, J. Galagan, Y. Van de Peer, Nucl. Acids Res. 40 (2012) e12.

[56] A. Deana, H. Celesnik, J.G. Belasco, Nature 451 (2008) 355–358.

[57] A.C. Darling, B. Mau, F.R. Blattner, N.T. Perna, Genome Res. 14 (2004) 1394–1403.

[58] O. Wurtzel, N. Sesto, J.R. Mellin, I. Karunker, S. Edelheit, C. Becavin, C. Archambaud, P. Cossart, R. Sorek, Mol. Syst. Biol. 8 (2012) 583.