

RESEARCH PAPER



Primary transcriptome analysis reveals importance of IS elements for the shaping of the transcriptional landscape of *Bordetella pertussis*

Fabian Amman^{a,†}, Alexandre D'Halluin^{b,†}, Rudy Antoine^b, Ludovic Huot^b, Ilona Bibova^c, Kristina Keidel^c, Stéphanie Slupek^b, Peggy Bouquet^b, Loïc Coutte^b, Ségolène Caboche^b, Camille Locht^{b,¥}, Branislav Vecerek^{c,¥} and David Hot^{b,¥}

^aUniversity of Vienna, Theoretical Biochemistry Group, Institute for Theoretical Chemistry, Vienna, Austria; ^bUniv. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019 - UMR8204 - CIL - Center for Infection and Immunity of Lille, Lille, France; ^cInstitute of Microbiology of the ASCR, Laboratory of post-transcriptional control of gene expression, Prague, Czech Republic

ABSTRACT

Bordetella pertussis is the causative agent of whooping cough, a respiratory disease still considered as a major public health threat and for which recent re-emergence has been observed. Constant reshuffling of *Bordetella pertussis* genome organization was observed during evolution. These rearrangements are essentially mediated by Insertion Sequences (IS), a mobile genetic elements present in more than 230 copies in the genome, which are supposed to be one of the driving forces enabling the pathogen to escape from vaccine-induced immunity.

Here we use high-throughput sequencing approaches (RNA-seq and differential RNA-seq), to decipher *Bordetella pertussis* transcriptome characteristics and to evaluate the impact of IS elements on transcriptome architecture. Transcriptional organization was determined by identification of transcription start sites and revealed also a large variety of non-coding RNAs including sRNAs, leaderless mRNAs or long 3' and 5'UTR including seven riboswitches. Unusual topological organizations, such as overlapping 5'- or 3'-extremities between oppositely orientated mRNA were also unveiled. The pivotal role of IS elements in the transcriptome architecture and their effect on the transcription of neighboring genes was examined. This effect is mediated by the introduction of IS harbored promoters or by emergence of hybrid promoters. This study revealed that in addition to their impact on genome rearrangements, most of the IS also impact on the expression of their flanking genes. Furthermore, the transcripts produced by IS are strain-specific due to the strain to strain variation in IS copy number and genomic context.

ARTICLE HISTORY

Received 30 March 2018
Accepted 3 April 2018

KEYWORDS

bordetella pertussis; insertion sequence; transcriptome

Introduction

Bordetella pertussis, the main causative agent of whooping cough, is a strictly human pathogen, responsible world-wide for an estimated 24.1 million pertussis cases, associated with 160,700 pertussis-linked death in 2014 [1]. The disease affects mainly children, but adolescents and adults are also susceptible [2]. Vaccination programs, first introduced in the 1950's, have resulted in a drastic decrease of pertussis incidence [3]. However, despite a high global vaccination coverage, in many countries a re-emergence of the disease has been observed over the past 10 years, which qualifies whooping cough today as the most prevalent vaccine-preventable childhood disease [4–6]. Pertussis resurgence has been hypothesized to result from a number of possible reasons, including waning of acellular pertussis vaccine efficiency and asymptomatic infections and transmission [7–11]. In addition, recently, many *B. pertussis* strains lacking pertactin, one of the major vaccine components of acellular vaccines, have been isolated in countries where acellular vaccines are used [12–15]. The lack of pertactin production in these strains is due to various mechanisms, including

frame-shift mutations, deletions, insertions and inversions, suggesting vaccine-induced pressure.



Massive gene loss, pseudogene formation and a reduction of the genome size of *B. pertussis* are among the main genomic features of this organism that became apparent when the genomes of several *Bordetella* species were compared [16]. Observed changes resulted mostly from major genetic rearrangements, including inversion and disruption of genomic islands [17–19]. These rearrangements are essentially mediated by the Insertion Sequence IS481, a mobile genetic element present in more than 230 copies in the *B. pertussis* genome [16]. Compared to other *Bordetella* species, this represents an exceptionally high potential for genomic rearrangements and could be one of the drivers of the pathogen ability to escape from vaccine-induced immunity [20].

Insertion Sequences (IS) encode a transposase and are flanked by inverted repeats involved in excision and insertion at a specific consensus target site [21]. These mobile genetic elements are tightly linked to the evolution and adaptation of pathogens to their hosts, to antibiotic

CONTACT David Hot  david.hot@pasteur-lille.fr; Camille Locht  Camille.Locht@pasteur-lille.fr  1 rue du professeur Calmette, 59000 Lille, France.

[†]Co first-authors.

[¥]These authors contributed equally to this work.

 Supplemental data for this article can be accessed on the  publisher's website.

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

resistance and to the response to an environmental stress [22–27]. The exceptional expansion of IS481 in the *B. pertussis* genome leads to a high structural genome diversity between strains, resulting from homologous recombination and site-directed insertion, despite an otherwise rather conserved genetic content [28]. IS have also the ability to activate expression of the neighboring genes, either by an extended transcription from an internal promoter “escaping the IS” (named P_{out} or P_{in}) or by the generation of a hybrid promoter (P_{hyb}) [29,30]. In the latter case, a strong promoter can be formed by insertion of an IS carrying –35 box that is in proper distance from a –10 box located in adjacent chromosomal sequence. It has been shown that the *katA* gene in the *B. pertussis* strain BP504 and the *bteA* gene in the vaccine-derived strain BP155 are transcribed by the outward promoter (P_{out}) of the neighboring IS481 [31,32]. Because of the numerous copies of IS481 in the *B. pertussis* genome, their impact on gene transcription is likely to occur at many sites. Thus, IS481 could generate strain-specific transcripts (messenger and regulatory RNA), which might contribute to *B. pertussis* strain adaptation.

To assess the global impact of IS481 on the transcriptome of *B. pertussis*, we used high through-put sequencing (HTS) combined with different bioinformatics analysis methods to establish the primary transcriptome of the Tohama I reference strain. This study revealed that most IS481 elements impact the transcriptional level of their flanking genes and can result in an uncharacterized regulation. Furthermore, as the distribution of IS elements is discriminative in each strain during the evolution, the transcripts driven by IS481 are strain-specific.

Results

Bordetella pertussis transcriptome analysis

To characterize the primary transcriptome of *Bordetella pertussis* we used HTS of strand-specific cDNA obtained from transcripts of the Tohama I derivative BPSM, grown in standard conditions at the exponential growth phase [33]. We combined two independent sequencing approaches to maximize information about the transcriptome architecture, a classical RNA-sequencing method (RNA-seq) on a library of strand-orientated cDNA prepared from total RNA depleted for rRNA and differential RNA-seq method (dRNA-seq) detecting the transcriptional start sites (TSS) by using terminator 5′-phosphatedependent Exonuclease (TEX)-treated and -untreated RNA samples. Libraries from four independent biological replicates were created. A total of 36.7 million reads could be mapped to the four samples (19.9 million for TEX-treated and 16.8 million for the untreated libraries).

The data obtained by the two approaches were combined to quantify the transcripts abundance and to determine the primary transcriptome architecture (BioProject ID: PRJNA430365). All raw and processed data are compiled in an Assembly Hub (Fig. 1) and can be publicly accessed and incorporated into any UCSC Genome Browser instance by the following address: <http://bit.ly/2Euo5HZ>.

TSS localization and transcript extent of annotated coding genes

We first used the above described data to determine the transcriptome organization for the annotated genes to determine the TSS and abundance of mRNA transcripts and their relative arrangement. The analysis of the dRNA-seq data using the software TSSAR revealed 2,722 potential TSSs (Fig. S1) [34]. The TSSs were classified according to their genomic context using the annotation file NC_002929.2. A TSS within ≤ 250 nt upstream of annotated genes was classified either as primary TSS for the furthest upstream, and as secondary TSS for the others. TSS localized in the forward orientation within open reading frame (ORF) of annotated genes was classified as internal TSS. The position of 614 primary TSSs regarded as the potential +1 position of transcription could be identified (Table S1). Among these 614 genes, 99 also display at least one secondary TSS. We validated the TSS of several transcripts by 5′RACE. The TSS of several coding transcripts (*ptxA*, *bipA*, *fim2*) and of candidate transcripts (see below) were confirmed or determined using this approach (Fig. S1). The TSSs were then used to determine putative promoter sequence motifs in their upstream region using MEME [35]. Beside a convincing TATA box motif (sequence TANAAT) in the –10 regions, no significant motif was shared by more than 50 TSS (Fig. S2).

Among the 614 primary TSS, 35 were localized at the position of the translational start codon of the annotated gene (Fig. S2). More than a third of them are of unknown function or are annotated to code for hypothetical proteins. However, within the functionally annotated genes a large fraction ($\sim 26\%$) consists of genes encoding regulators, including *ompR* (BP3222) and *basR* (BP3534). A hypergeometric test for these 35 genes confirmed that regulators are significantly enriched in the set of leaderless transcripts (*p*-value 0.0014).

On the other side, 120 genes were preceded by a long 5′UTR (> 100 nt). Long 5′UTRs often contain regulatory structures such as riboswitches or thermosensors. Among the 120 long 5′UTRs, 7 riboswitches were predicted *in silico* (Fig. S3). Using RT-PCR spanning the region between the predicted riboswitch and adjacent gene, 4 of them were confirmed to be co-transcribed with the adjacent gene (Fig. S3). The three remaining predicted riboswitches (type *yybPykoY* riboswitch, Cobalamin riboswitch and FMN riboswitch in front of *BP3410*, *BP3595* and *ribB*, respectively) could not be confirmed by RT-PCR to be part of the same transcript. However, all three correspond to riboswitches, which are expected to be switched off (leading to aborted transcripts) under the growth conditions used here (*i.e.* no manganese ion for *yybP-ykoY* riboswitch, presence of adenosylcobalamin for cobalamin riboswitch and presence of FMN for FMN-riboswitch) [36]. The abortive transcription was confirmed by the absence of RT-PCR products corresponding to the adjacent gene, whereas RT-PCR products were detected using the 5′UTR-specific (Fig. S3). 5′RACE also confirmed the TSSs of the SAH, SAM-alpha and Glycine riboswitches (Fig. S1).

To characterize the transcriptional organization of the annotated genes we used the data obtained from RNA-seq. The transcript organization was deduced for 2,262 out of the 3,871 annotated ORF (including IS transposase genes) allowing us to determine the mono- and poly-cistronic genes (NC_002929.2)

(see Fig. 1). We could determine 1,315 transcript structures, of which 838 were mono- and 477 were polycistronic transcripts (table S2). A histogram of number of cistrons/operon structure is shown in supplementary data Fig. S4.

Furthermore, 5' overlapping, divergent head-to-head and 3' overlapping, convergent tail-to-tail transcript organizations were identified. Occasionally the 3'UTR of a transcript extends well beyond the annotated stop codon, sometimes yielding a 3'UTR of up to several hundreds of nt reaching into the antisense region of the neighboring gene. A striking example of this structure is the transcriptional organization of *bvgR* and *bvgAS* genes, a major virulence gene regulator system in *Bordetella pertussis*. The transcript of *bvgR* overlaps with the *bvgS* ORF by more than 410 nt and the *bvgS* transcript overlaps with the *bvgR* ORF by more than 170 nt (Fig. 2). Thus, including intergenic regions (IGR) these two transcripts share an overlapping region composed of 630 nt. This overlap was confirmed by a two-step RT-PCR described in supplementary data part 6. Within the entire genome, overlapping transcripts were detected 37 times for 5' divergent tail-to-tail overlaps and 89 times for 3' convergent head-to-head overlaps (table S2).

TSS classification and identification of sRNA

In addition to determining the transcript organization of the coding genes, we also used the dRNAseq data to define the TSS of internal (I) and antisense (A) transcripts, corresponding to TSS within an annotated ORF on the sense and the anti-sense strand, respectively. Furthermore, we identified orphan (O) TSS, localized in IGR at a ≥ 250 nt distance from the closest gene in the same orientation. The "A" and "O" TSSs can be considered as the transcription starts of potentially new transcripts not detected so far and will be referred in the rest of the text as candidate transcripts. Among the total 362 candidate transcripts deduced from A and O TSS, 118 were localized within IGR (O) and 244 were antisense to annotated genes (Fig. 3). Moreover, the RNA-seq data revealed 268 novel candidate transcripts not overlapping with any of the annotated genes. 218 of them are in the antisense orientation to annotated

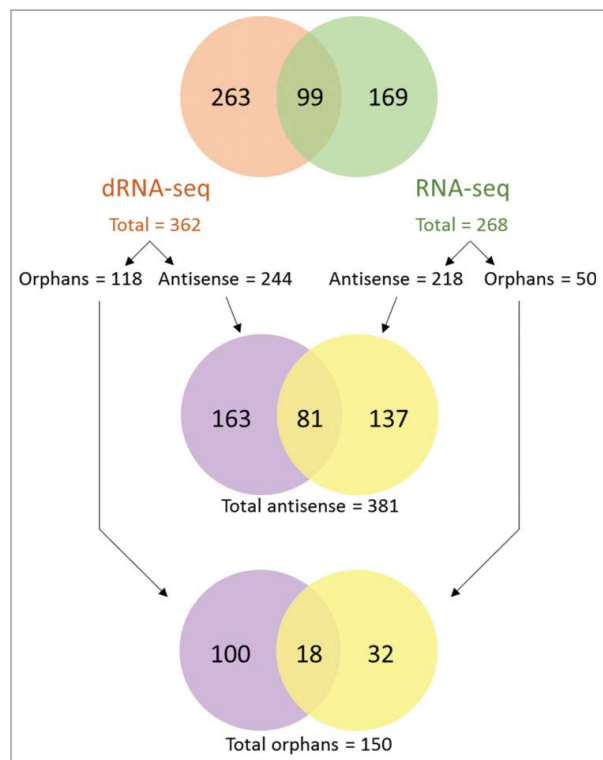


Figure 3. Candidate transcripts classification. Venn diagrams of candidate transcripts numbers deduced from dRNA-seq and RNA-seq. Number of antisense and orphan candidate transcripts are counted as the function of the detection methods.

genes, either within the ORF or overlapping with the 5' or 3' ends, and the 50 remaining candidate transcripts are within IGR. The distribution of antisense and orphan candidate transcripts identified by these two approaches is detailed in Fig. 3.

Some of these transcripts are localized in the IGR in the vicinity and in the same orientation as the adjacent annotated gene, either upstream, close to the translational start codon, or downstream, next to a stop codon. However, they may potentially correspond to long 5' or 3' UTR, respectively. To

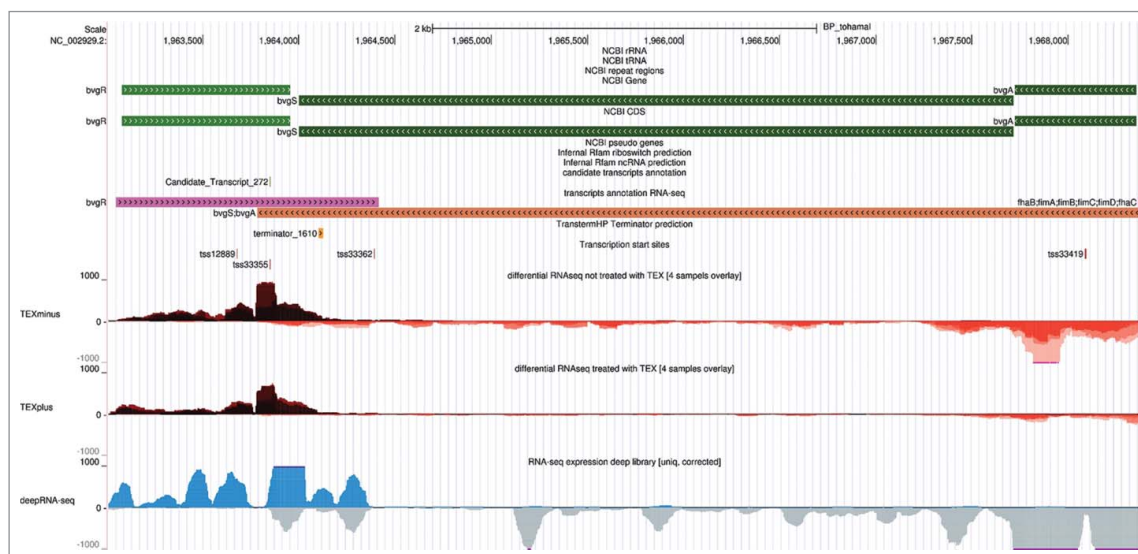


Figure 2. *bvgR* and *bvgAS* transcription organization. Detailed view of genome browser for the *bvgAS* and *bvgR* sequence region. Prediction of transcripts organization is indicated.

investigate this possibility further we analyzed three candidate transcripts (numbers 229, 233 and 235). As for the riboswitches, the independent transcription of each candidate with respect to the adjacent gene was tested by a two-step RT-PCR spanning a region between the candidate transcript and the neighboring ORF (Fig. S3). Candidate transcript 233 was detected as an independent transcript by RT-PCR. However, it is upstream of the threonine-tRNA ligase gene and could therefore result from an abortive transcript of the *thrS* transcript [37,38]. The two remaining candidate transcripts (229 and 235) were confirmed as independent transcripts. The predicted TSSs of 3 other candidate transcripts were validated by 5'RACE: transcript 110 (antisense sRNA to *cyaA*), 273 (antisense to *fhaB*) and 521 (Fig. S1 and table S3).

All these novel predicted transcripts can be seen as a potential reservoir of regulatory RNAs. Therefore, we plan to analyze the gene expression profiles of identified non-coding RNAs under different growth conditions.

A large proportion of antisense RNA results from transcriptional activity of IS

The *B. pertussis* IS481 element harbors a promoter, called P_{in} , which drives the transcription of the transposase gene and a second, outward facing promoter, P_{out} , localized downstream of P_{in} in the opposite direction [39]. The P_{out} activity yields a small transcript which blocks the translation of the transposase mRNA by hybridizing to the ribosomal binding site and destabilizing the mRNA [40,41]. The transcript issued from P_{out} can also serve as a transcriptional activator for flanking genes in the same orientation.

We screened the expression profiles of all IS elements in the BP Tohama I genome (supplementary table S4). For that purpose a sequencing read mapping procedure was performed without filtering of multi-mapping reads in order to detect transcription signal in the repeated sequence regions of the IS. We therefore obtained a profile of expression within all IS which corresponds to the mean IS expression. The promoters' expression levels were assessed by specific reads overlapping IS extremities and genomic flanking regions. The P_{in} and P_{out}

transcription activities were clearly detected (see example of the IS BP0704 on Fig. 4A). We observed 70 IS (66 IS481, 2 IS1002 and 2 IS1663) for which the P_{out} transcript extend into the flanking gene which is in the same orientation, resulting in a potential activation of its transcription as described for *kataA* and *bteA*. Such a configuration is illustrated by Fig. 4B. Interestingly, for 150 other IS (139 IS481, 2 IS1002 and 9 IS1663) the transcript from P_{out} extends into the flanking gene which is in opposite orientation resulting in the generation of a transcript antisense to these genes (Fig. 4C). Finally, for 23 IS (18 IS481 and 5 IS1663) the P_{out} transcription extends into an IGR generating potential non-coding RNAs. The transcription starts for the P_{out} of 10 randomly chosen IS were checked by 5'RACE (Fig. S1) and were all confirmed at the predicted positions as shown in Fig. 4.

In addition, in 101 cases (97 IS481, 1 IS1002 and 3 IS1663) the transcription of the transposase mRNA driven by P_{in} extends beyond the IS element, into the downstream gene, either in the same orientation (46 times) or in the opposite orientation (38 times) of the downstream gene, or into an IGR (17 times). Some transcripts (6 IS481) appear to result from a hybrid promoter, where the -10 box is provided by the gene lying close to insertion site and the -35 box is located in the inverted repeat of the IS. The transcription activity of one hybrid promoter (from candidate transcript 256) was confirmed by 5'RACE (Fig. S1). Globally, almost all ISs show an impact on transcriptional activity in the neighboring regions as only 6 IS do not display any additional transcriptional activity under the studied condition. Some IS (94) use simultaneously a P_{in} and P_{out} causing intense transcription activity, and two of them show in addition also a P_{hybrid} activity.

Table S4 summarizes the detected transcripts resulting from P_{out} , P_{in} and P_{hybrid} of all IS in the Tohama I genome. Most of them were annotated by our analysis pipeline as new transcripts. This pool of transcripts may have specific regulatory function which is yet to be characterized. Alternatively, they could also account for functionless byproducts of random IS displacement and therefore, these specific candidates are labeled as such in the 'Comments' column of the table S3.

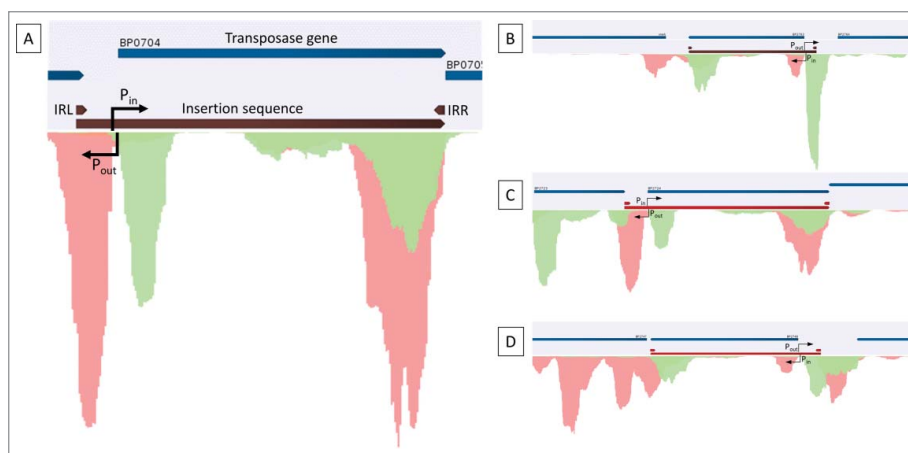


Figure 4. Promoters and transcription in insertion sequences. Gene ORF are indicated as plain blue arrows. IS and inverted repeat left and right (IRL and IRR) are indicated as plain brown arrows. The sequencing depth graphs are shown underneath in green for positive (; same orientation as published sequence) orientated reads and in red for negatively orientated reads. Promoter positions and orientation are indicated as black arrows. (A) General organization of transcription in IS region. (B) Representative example of flanking gene transcription activation due to P_{out} activity. (C) Representative example of antisense transcription in regard to flanking gene due to P_{out} activity. (D) P_{in} polar transcription activity in the gene downstream of the transposase.

Impact of IS number and localization on the *B. pertussis* transcriptome landscape

In order to assess the potential impact on the transcription generated from the P_{out} on the *B. pertussis* transcriptome in general, we examined the numbers and genomic contexts of all the P_{out} of 21 fully sequenced and annotated *B. pertussis* genome sequences (Fig. S5). Unique P_{out} structures identified only in one given strain are listed in the table of Fig. S5. As an example, when we compared Tohama I strain, representing the reference strain of *B. pertussis* to the strain D420, representing the currently circulating strains, we observed that 9 out of the 246 P_{out} identified in Tohama I are localized in a unique genomic context (one of them was verified by RT-PCR, Fig. S5) [42]. On the other side, 15 out of 252 D420 P_{out} promoters are in a unique context (one of them was verified by RT-PCR, Fig. S5). More globally, almost all strains present at least one difference compared to all the other strains. Some strains, such as CS, B203, 137 and, at the extreme, 18323, have a high number of unique P_{out} contexts. These differences result from differences in IS insertion and rearrangement events. Each difference in the P_{out} context between strains can thus result in a different levels of gene expression.

The levels of transcription driven by P_{out} were measured by calculating RPKM in the 200-nt region downstream of the left inverted repeat of each IS481 element. A wide range of transcriptional levels was observed. Some P_{out} had no detectable activity, while some P_{out} induced high transcriptional rates. A deep RNA-seq experiment was performed on D420 grown in the same condition as BPSM and the resulting P_{out} RPKM were compared. A high correlation ($R^2 = 0.8$) between $\log_2(\text{RPKM})$ of the P_{out} of these two strains was observed indicating the same level of P_{out} transcriptional activity between the two strains (Fig. S5).

Discussion

In this study we combined two approaches, RNA-seq and differential RNA-seq, to elucidate the detailed organization of the transcriptional landscape of *B. pertussis* Tohama I, including the operon structures, 5' UTR, 3' UTR and novel transcripts. A total of 2,722 potential TSS were identified, 614 of which defined the +1 position of transcription, and 8 were confirmed by 5'RACE. A total of 1,315 transcript structures could be determined, of which 838 were mono- and 477 polycistronic operons.

The primary transcriptome of *B. pertussis* shows a complex regulatory network, defined by transcriptomic structures seen in other pathogens already [43,44]. Among the 614 TSS, 120 yielded long 5'UTR (> 100 nt). Seven out of these 120 long 5'UTR are predicted to be riboswitches, which are involved in functions as diverse as enzyme cofactor sensing (SAM, SAH, TPP, FMN and cobalamin), amino acid sensing (glycine) and metal ion sensing (*yybP-ykoY*). The presence of riboswitch elements, of which four could be confirmed by RT-PCR to be part of the same operon as the downstream ORF, provides yet another level of possible regulation mechanism of *B. pertussis* to adequately and rapidly adapt to changing conditions during infection for example in order to coordinate control of crucial processes along the respiratory track of the host. In addition to

these transcript structures containing a long 5'-UTR, 35 leaderless transcripts (0.90% of annotated genes) were also detected. Leaderless transcripts defy the well-accepted paradigm by which translation of bacterial mRNA is initiated by binding of the ribosome via the complementary region of 16S rRNA on a Shine-Dalgarno sequence in the 5' UTR [45]. The number of leaderless transcripts found in *B. pertussis* is in the range of those found in other bacteria for which global transcriptome landscape was determined, with 18 (0.46%), 23 (0.48%), and 83 (1.41%) leaderless mRNAs identified in *Helicobacter pylori*, *Salmonella typhimurium*, and *Klebsiella pneumoniae*, respectively, but is far lower than the over 500 (>12.4%) leaderless transcripts found in *Mycobacterium tuberculosis* [43,46–48]. Intriguingly, many of the 35 *B. pertussis* leaderless transcripts code for putative regulatory proteins, which suggests a highly specific regulatory role of these genes in *B. pertussis*.

The transcriptomic landscape of *Bordetella pertussis* show multiple antisense transcript shaping potential regulatory effect between close related genes. A total of 37 5'-overlapping divergent head-to-head and, among these unusual transcript structures, the example of a TetR-family transcriptional regulator (BP1202) stands out in particular, since the entire gene overlaps with the 5' UTR of the convergent transcript coding for the two uncharacterized proteins BP1203 and BP1204 [49]. This transcriptional organization corresponds to the definition of an excludon as proposed by Wurtzel *et al.* and might therefore indicate that BP1203 and BP1204 encode products which functions might be opposite to the function of the transcriptional regulator TetR [50]. 89 3'-overlapping convergent tail-to-tail transcripts were also found, sometimes extending well beyond the stop codon up to several hundred nucleotides into the antisense region of the neighboring gene. The overlapping of *bvgAS* and *bvgR* transcripts arises a potential new level of complexity in the regulation of the master two-component system of *Bordetella pertussis* [51,52]. Indeed, this configuration implies a potential cross-regulation of *bvgAS* and *bvgR* transcripts. The 3'UTR overlapping of two flanking genes has already been seen in different bacteria, and has been suggested to induce the decay of one of the mRNA depending on sigma factor expression [53–56].

One of the major findings of this study was that a large number of transcripts arise from the invading transcription driven by IS elements. Transcription issued from internal promoters, P_{in} and P_{out} , or newly formed P_{hyb} promoters can extend into the neighboring genomic regions and affect the gene expression as they can initiate the transcription of the downstream gene, as shown in *Bordetella pertussis* and other prokaryotes, including pathogens [26,29,32]. However, our primary transcriptome analysis reveals new types of IS-driven RNAs either transcribed within intergenic regions or antisense to their flanking gene(s). Interestingly, the new candidate RNAs transcribed antisense to neighboring gene(s) are highly prevalent over those affecting sense genes. Antisense RNA originating from IS elements could have a regulatory function for those genes, probably by inducing degradation of the mRNA as already shown with antisense small RNA transcribed at the 3' of a gene [57]. Genome rearrangements mediated by IS481 are considered as one of the driving forces of the *B. pertussis* adaptation to

human respiratory tract and vaccine-induced immunity. The different localization and number of those elements in *Bordetella pertussis* strains show strain-specific transcripts with a potential regulatory function validated by RT-PCR. Thus, this evidence could explain some gene expression alterations associated with IS481 even without insertion in the gene sequence. This could be the case of some antigens used in acellular vaccine such as *fim3* and *fhaB* gene, which are altered in Bpe280 strain after 12 *in vitro* passages on agar plate and located near an IS481 element [58]. More importantly, some of the resulting gene reshufflings appeared to be conserved through positive selection indicating that some of the IS481-induced changes in gene expression are beneficial to these isolates. We also found that transcription levels driven by IS481 P_{out} promoters can be variable depending on the nature of the locus the IS element is inserted into, but that conserved loci drive similar P_{out} transcriptional levels in different strains. Thus, the exceptionally high number of IS481 elements contributes to *B. pertussis* evolution not only by gene rearrangements but also by alteration of the transcriptional landscape. As many different IS show some outgoing transcription from those promoters, this global transcriptional impact can extend to other prokaryotic pathogen containing high numerous IS, as seen in *Shigella flexneri* strain 2457T and *Yersinia pestis* strain CO92 [59,60]. The regulatory functions of the antisense RNAs originating from IS elements will be further investigated.

Material and methods

Bacterial strain and oligonucleotides

Bordetella pertussis BPSM, a streptomycin-resistant derivative of Tohama I was grown as previously described at 37°C in modified Stainer-Scholte medium, containing 100 µg/ml streptomycin (Sigma Chemicals) [33,52]. Cells were harvested at exponential phase ($OD_{600\text{ nm}} = 2.1$) by adding 2 ml of 5:95 (v:v) phenol/ethanol to 8 ml of culture medium. After centrifugation for 8 min at 2800 × g the pellets were stored at -80°C until further use.

Primers (IDT DNA Technologies) were designed using FastPCR and sequence specificity were checked using Blast [61,62].

RNA-seq library preparation and Illumina sequencing

Bacterial pellets were resuspended in 400 µl of 1 mg/ml Lysozyme (Sigma Aldrich). Total RNA was extracted with the TRI Reagent Kit (Ambion) according to the manufacturer's protocols. Ten µg of total RNA were first DNase treated using DNase I (Sigma Aldrich) for 10 min at room temperature and the ribosomal RNA was then depleted using Ribo-Zero rRNA removal kit (Epicentre) according to manufacturer instructions. Quantity and quality assessment were checked using the Nanodrop 2000 and the Bioanalyzer 2100 (Agilent Technologies), respectively. Sequencing libraries were prepared using TruSeq stranded total RNA library prep kit (Illumina) and the sequencing

run was performed on a HiSeq® 2500 sequencing system (Illumina).

Differential RNA-seq libraries preparation and Ion Torrent PGM sequencing

Five µg of DNase treated total RNA were treated with 1U of Terminator™ 5'phosphate-dependant exonuclease (TEX) (Epicentre) for 60 min at 30°C. Reaction was stopped by adding 1 µl of EDTA 100 mM pH 8 and purified by an organic extraction. An equal amount of treated (TEX plus) and untreated (TEX minus) RNA with the Terminator™ 5'phosphate-dependant exonuclease were incubated with 1U of Tobacco Acid Pyrophosphatase (Epicentre) for 60 min at 37°C.

cDNA libraries were constructed with 500 ng of TEX plus or TEX minus RNA using the Ion Total RNA-seq Kit v2 (Life Technologies), following the manufacturer's recommendations and purified twice with 1.8 volume of Agencourt AMPure (Beckman Coulter). Emulsion PCR, enrichment and sequencing were made using the Ion PGM™ template OT2 400 Kit and the Ion PGM™ sequencing 400 Kit with 12 pM of each library. Enriched beads were sequenced on an Ion Torrent PGM machine using a 318 v2 chip.

Bioinformatics analysis of dRNA-seq data

The four independent replicates of TEX-treated and untreated library pairs were quality controlled using fastqc, reads were quality trimmed using fastq_quality_trimmer from the fastx-toolkit, with a phred cutoff of 22, reads with length below 16 bases after trimming were discarded. The libraries were mapped on the BP reference genome (NC_002929.2) using the short read aligner Segemehl with default settings [63]. For all libraries only the uniquely mapped reads were further considered. All four TEX treated-untreated library pairs were analyzed with TSSAR demanding a minimal peak size of 2. For each genomic position the relative enrichment of signal intensity in the TEX treated library in the single analysis were merged applying Fisher's combined probability test and corrected for multiple testing with the Benjamini & Hochberg method [34,64,65]. Positions with an enrichment in the TEX treated library in comparison to the untreated library based on statistical significance level of 0.1 were denominated as transcription start sites (TSS). Finally, consecutive TSS positions separated by less than 10 nt were joined and represented by the most prominent peak.

5'race using PGM Ion Torrent sequencing technology

Starting sites were determined by 5'RACE for NGS sequencing as described by Beauregard *et al* [66]. PCR was performed with a first primer specific for the adaptor sequence and a second primer specific for the target cDNA (see Table in supplementary data part 1). A and P1 sequences for PGM sequencing were ligated using the Ion Plus Fragment Library Kit, following manufacturer's recommendations (Life Technology). PCR products were purified with 1.8 volume of Agencourt AMPure (Beckman Coulter) and sequenced on a 314 chip using Hi-Q kits for PGM Ion Torrent machine (Life Technology). Reads

trimming, filtering and counting to determine transcript start was as described in supplementary data part 1.

Operonic structure determination and novel transcripts annotation

For the annotation of transcription units, mapped reads from all four TEX minus libraries from the dRNA-seq experiments were merged. The position wise read coverage were calculated using bedtools genomecov [67]. Since strand-specificity imposes a problem in current NGS protocols, we developed a simple yet effective approach to detect and correct for the fraction of anti-sense shadow and predicted transcripts architectures by calculation taking care of TSS, transcription terminators, annotated genes and observed transcription (detailed in supplemental material part 6) [68].

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.


Acknowledgment


Funding: FA was supported by the Austrian Science Fund (FWF) project SFB F43. This work was supported by the Czech Science Foundation (www.gacr.cz) (grant 16-34825L to B.V.), by the Czech Health Research Council (www.azvcr.cz/) (grant 16-30782A to B.V.) and by funding from RVO61388971. This work was also supported by the Ministry of Education, Youth and Sports of the Czech Republic projects (CZ.1.07/2.3.00/20.0055 to B.V. and I.B., CZ.1.07/2.3.00/30.0003 to B.V. and K.K.).

Funding

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic (CZ.1.07/2.3.00/20.0055). This work was supported by the Czech Science Foundation (www.gacr.cz) (16-34825L). This work was supported by the Austrian Science Fund (FWF) (SFB F43). This work was supported by the Agentura Pro Zdravotnický Výzkum České Republiky. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic projects. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic projects (CZ.1.07/2.3.00/20.0055). This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic (CZ.1.07/2.3.00/30.0003). Czech Health Research Council (16-30782A).

ORCID

Branislav Vecerek  <http://orcid.org/0000-0001-9520-6324>

David Hot  <http://orcid.org/0000-0002-2361-0616>

References

- [1] Crowcroft NS, Stein C, Duclos P, et al. How best to estimate the global burden of pertussis? *Lancet Infect Dis.* 2003;3:413–418.
- [2] Paddock CD, Sanden GN, Cherry JD, et al. Pathology and Pathogenesis of Fatal *Bordetella pertussis* Infection in Infants. *Clin Infect Dis.* 2008;47:328–338.
- [3] World Health Organization. Pertussis vaccines: WHO position paper - August 2015. *Weely Epidemiol Rec.* 2015;90:433–460.
- [4] Sealey KL, Belcher T, Preston A. *Bordetella pertussis* epidemiology and evolution in the light of pertussis resurgence. *Infect Genet Evol.* 2016;40:136–143.
- [5] Kline JM, Lewis WD, Smith EA, et al. Pertussis: a reemerging infection. *Am Fam Physician.* 2013;88:507–514.
- [6] Wadman M, You J. The vaccine wars. *Science.* (80-.). 2017;356:364–365.
- [7] Witt MA, Katz PH, Witt DJ. Unexpectedly limited durability of immunity following acellular pertussis vaccination in preadolescents in a north American outbreak. *Clin Infect Dis.* 2012;54:1730–1735.
- [8] Smits K, Pottier G, Smet J, et al. Different T cell memory in preadolescents after whole-cell or acellular pertussis vaccination. *Vaccine.* 2013;32:111–118.
- [9] Warfel JM, Edwards KM. Pertussis vaccines and the challenge of inducing durable immunity. *Curr Opin Immunol.* 2015;35:48–54.
- [10] Edwards KM, Berbers GAM. Immune responses to pertussis vaccines and disease. *J Infect Dis.* 2014;209:S10–S15.
- [11] Althouse BM, Scarpino S V. Asymptomatic transmission and the resurgence of *Bordetella pertussis*. *BMC Med.* 2015;13:146.
- [12] Martin SW, Pawloski L, Williams M, et al. Pertactin-negative *Bordetella pertussis* strains: evidence for a possible selective advantage. *Clin Infect Dis.* 2015;60:223–227.
- [13] Queenan AM, Cassiday PK, Evangelista A. Pertactin-Negative Variants of *Bordetella pertussis* in the United States. *N Engl J Med.* 2013;368:583–584.
- [14] Otsuka N, Han H-J, Toyozumi-Ajisaka H, et al. Prevalence and genetic characterization of pertactin-deficient *bordetella pertussis* in Japan. Miyaji EN, editor. *PLoS One.* 2012;7:e31985.
- [15] Williams MM, Sen K, Weigand MR, et al. *Bordetella pertussis* strain lacking pertactin and Pertussis Toxin. *Emerg Infect Dis.* 2016;22:319–322.
- [16] Parkhill J, Sebahia M, Preston A, et al. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet.* 2003;35:32–40.
- [17] Park J, Zhang Y, Buboltz AM, et al. Comparative genomics of the classical *Bordetella* subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. *BMC Genomics.* 2012;13:545.
- [18] Cummings CA, Brinig MM, Lepp PW, et al. *Bordetella* species are distinguished by patterns of substantial gene loss and host adaptation. *J Bacteriol.* 2004;186:1484–1492.
- [19] Diavatopoulos DA, Cummings CA, Schouls LM, et al. *Bordetella pertussis*, the causative agent of whooping cough, evolved from a distinct, human-associated lineage of *B. bronchiseptica*. *PLoS Pathog.* 2005;1:e45.
- [20] Xu Y, Liu B, Gröndahl-Yli-Hannuksila K, et al. Whole-genome sequencing reveals the effect of vaccination on the evolution of *Bordetella pertussis*. *Sci Rep.* 2015;5:12888.
- [21] Siguier P, Goubeyre E, Varani A, et al. Everyman's guide to bacterial insertion sequences. *Microbiol Spectr.* 2015;3:MDNA3-0030-2014.
- [22] Chain PSG, Hu P, Malfatti SA, et al. Complete genome sequence of *Yersinia pestis* strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen. *J Bacteriol.* 2006;188:4453–4463.
- [23] Siguier P, Goubeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol Rev.* 2014;38:865–891.
- [24] Depardieu F, Podglajen I, Leclercq R, et al. Modes and modulations of antibiotic resistance gene expression. *Clin Microbiol Rev.* 2007;20:79–114.
- [25] Humayun MZ, Zhang Z, Butcher AM, et al. Hopping into a hot seat: Role of DNA structural features on IS5-mediated gene activation and inactivation under stress. *Kalendar R, editor. PLoS One.* 2017;12:e0180156.
- [26] Safi H, Barnes PF, Lakey DL, et al. IS6110 functions as a mobile, monocyte-activated promoter in *Mycobacterium tuberculosis*. *Mol Microbiol.* 2004;52:999–1012.
- [27] Schneider D, Lenski RE. Dynamics of insertion sequence elements during experimental evolution of bacteria. *Res Microbiol.* 2004;155:319–327.
- [28] Weigand MR, Peng Y, Loparev V, et al. The history of *Bordetella pertussis* genome evolution includes structural rearrangement. *Becker A, editor. J Bacteriol.* 2017;199:e00806–16.

- [29] Prentki P, Teter B, Chandler M, et al. Functional promoters created by the insertion of transposable element IS1. *J Mol Biol.* 1986;191:383–393.
- [30] Wang A, Roth JR. Activation of silent genes by transposons Tn5 and Tn10. *Genetics.* 1988;120:875–885.
- [31] DeShazer D, Wood GE, Friedman RL. Molecular characterization of catalase from *Bordetella pertussis*: identification of the *katA* promoter in an upstream insertion sequence. *Mol Microbiol.* 1994;14:123–130.
- [32] Han H-J, Kuwae A, Abe A, et al. Differential expression of type III effector BteA protein due to IS481 insertion in *Bordetella pertussis*. Neyrolles O, editor. *PLoS One.* 2011;6:e17797.
- [33] Antoine R, Loch C. Roles of the disulfide bond and the carboxy-terminal region of the S1 subunit in the assembly and biosynthesis of pertussis toxin. *Infect Immun.* 1990;58:1518–1526.
- [34] Amman F, Wolfinger MT, Lorenz R, et al. TSSAR: TSS annotation regime for dRNA-seq data. *BMC Bioinformatics.* 2014;15:89.
- [35] Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings Int Conf Intell Syst Mol Biol.* 1994;2:28–36.
- [36] Dambach M, Sandoval M, Updegrave TB, et al. The Ubiquitous *yybP-ykoY* riboswitch is a manganese-responsive regulatory element. *Mol Cell.* 2015;57:1099–1109.
- [37] Grundy FJ, Henkin TM. tRNA as a positive regulator of transcription antitermination in *B. Subtilis*. *Cell.* 1993;74:475–482.
- [38] Condon C, Putzer H, Grunberg-Manago M. Processing of the leader mRNA plays a major role in the induction of *thrS* expression following threonine starvation in *Bacillus subtilis*. *Proc Natl Acad Sci U S A.* 1996;93:6992–6997.
- [39] McLafferty MA, Marcus DR, Hewlett EL. Nucleotide sequence and characterization of a repetitive DNA element from the genome of *bordetella pertussis* with characteristics of an insertion sequence. *Microbiology.* 1988;134:2297–2306.
- [40] Ma C, Simons RW. The IS10 antisense RNA blocks ribosome binding at the transposase translation initiation site. *EMBO J.* 1990;9:1267–1274.
- [41] Simons RW, Kleckner N. Translational control of IS10 transposition. *Cell.* 1983;34:683–691.
- [42] Boinett CJ, Harris SR, Langridge GC, et al. Complete Genome Sequence of *Bordetella pertussis* D420. *Genome Announc.* 2015;3:e00657–15.
- [43] Sharma CM, Hoffmann S, Darfeuille F, et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature.* 2010;464:250–255.
- [44] Stazic D, Voß B. The complexity of bacterial transcriptomes. *J Biotechnol.* 2016;232:69–78.
- [45] Shine J, Dalgarno L. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A.* 1974;71:1342–1346.
- [46] Seo J-H, Hong JS-J, Kim D, et al. Multiple-omic data analysis of *Klebsiella pneumoniae* MGH 78578 reveals its transcriptional architecture and regulatory features. *BMC Genomics.* 2012;13:679.
- [47] Kroger C, Dillon SC, Cameron ADS, et al. The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. *Proc Natl Acad Sci.* 2012;109:E1277–E1286.
- [48] Cortes T, Schubert OT, Rose G, et al. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell Rep.* 2013;5:1121–1131.
- [49] Georg J, Hess WR. cis-Antisense RNA, another level of gene regulation in bacteria. *Microbiol Mol Biol Rev.* 2011;75:286–300.
- [50] Wurtzel O, Sesto N, Mellin JR, et al. Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species. *Mol Syst Biol.* 2012;8:583.
- [51] Loch C, Antoine R, Jacob-Dubuisson F. *Bordetella pertussis*, molecular pathogenesis under multiple aspects. *Curr Opin Microbiol.* 2001;4:82–89.
- [52] Hot D, Antoine R, Renaud-Mongénie G, et al. Differential modulation of *Bordetella pertussis* virulence genes as evidenced by DNA microarray analysis. *Mol Genet Genomics.* 2003;269:475–486.
- [53] Toledo-Arana A, Dussurget O, Nikitas G, et al. The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature.* 2009;459:950–956.
- [54] Ren G-X, Guo X-P, Sun Y-C. Regulatory 3' Untranslated Regions of Bacterial mRNAs. *Front Microbiol.* 2017;8:1276.
- [55] Lasa I, Toledo-Arana A, Gingeras TR. An effort to make sense of antisense transcription in bacteria. *RNA Biol.* 2012;9:1039–1044.
- [56] Lasa I, Toledo-Arana A, Dobin A, et al. Genome-wide antisense transcription drives mRNA processing in bacteria. *Proc Natl Acad Sci U S A.* 2011;108:20172–20177.
- [57] Svensson SL, Sharma CM. Small RNAs in bacterial virulence and communication. *Virulence Mech Bact Pathog Fifth Ed American Society of Microbiology.* 2016;4:169–212.
- [58] Brinig MM, Cummings CA, Sanden GN, et al. Significant gene order and expression differences in *Bordetella pertussis* despite limited gene content variation. *J Bacteriol.* 2006;188:2375–2382.
- [59] Zaghoul L, Tang C, Chin HY, et al. The distribution of insertion sequences in the genome of *Shigella flexneri* strain 2457T. *FEMS Microbiol Lett.* 2007;277:197–204.
- [60] Parkhill J, Wren BW, Thomson NR, et al. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature.* 2001;413:523–527.
- [61] Kalendar R, Lee D, Schulman AH. FastPCR software for PCR, in silico PCR, and oligonucleotide assembly and analysis. *Methods Mol Biol.* 2014;1116:271–302.
- [62] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–410.
- [63] Hoffmann S, Otto C, Kurtz S, et al. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. Searls DB, editor. *PLoS Comput Biol.* 2009;5:e1000502.
- [64] Fisher RA. *Statistical Methods for Research Workers.* Springer, New York, NY; 1992. p. 66–70. DOI:10.1007/978-1-4612-4380-9_6
- [65] Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B.* 1995;57:289–300.
- [66] Beauregard A, Smith E, Petrone B, et al. Identification and characterization of small RNAs in *Yersinia pestis*. *RNA Biol.* 2013;10:397–405.
- [67] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841–842.
- [68] Levin JZ, Yassour M, Adiconis X, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods.* 2010;7:709–715.