Mapping the complex transcriptional landscape of the phytopathogenic bacterium *Dickeya dadantii*

Raphaël Forquet #, Xuejiao Jiang, William Nasser, Florence Hommais, Sylvie Reverchon and Sam Meyer #

Université de Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, CNRS, UMR5240 MAP, F-69622, France

Corresponding authors:

Raphaël Forquet: Tel: +33 4 72 43 85 16; Email: raphael.forquet@insa-lyon.fr

Sam Meyer: Tel: +33 4 72 43 85 16; Email: sam.meyer@insa-lyon.fr

Abstract:

Dickeya dadantii is a phytopathogenic bacterium that causes soft rot in a wide range of plant hosts worldwide and a model organism for studying gene regulation during the pathogenic process. The present study provides a comprehensive and annotated transcriptomic map of *D. dadantii* obtained by a computational method combining three independent transcriptomic datasets covering a wide range of conditions which closely reproduce the variations of transcription occurring in the course of plant infection: (1) paired-end RNA-seq data for a precise reconstruction of the RNA landscape, (2) DNA microarray data reflecting gene response to sudden environmental shocks mimicking conditions encountered by bacteria in the plant, (3) dRNA-seq data for a specific high-resolution mapping of transcription start sites. We define transcription units throughout the genome, and map the associated transcription start and termination sites with a quantitative magnitude analysis. Our results show that transcription units sometimes coincide with predicted operons but are generally longer, most of them exhibiting internal promoters and terminators that generate alternative transcripts of variable gene composition. We characterize the occurrence of transcriptional read-through at terminators, which might play a basal regulation role and explain the extent of transcription beyond the scale of operons. We finally highlight the presence of noncontiguous operons and excludons in *D. dadantii* genome, novel genomic arrangements that might contribute to the basal coordination of transcription. The highlighted transcriptional organization may allow *D. dadantii* finely adjusting its gene expression program for a rapid adaptation to fast changing environment relevant to plant infection.

Importance:

This is the first transcriptomic map of a phytopathogen, characterized under physiological conditions encountered by the bacteria during the infection process. It might therefore significantly contribute to further progress in the field of phytopathogenicity. Our findings also provide insights into basal rules of coordination of transcription that might be valid for other bacteria, and may raise interest in the field of microbiology in general. In particular, we demonstrate that gene expression is coordinated at the scale of transcription units rather than operons, which are larger functional genomic units capable of generating transcripts with variable gene composition for a fine-tuning of gene expression in response to environmental changes. In line with recent studies, our findings indicate that the canonical operon model is insufficient to explain the complexity of bacterial transcriptomes.

Introduction

Bacteria of the *Dickeya* genus are Gram-negative phytopathogens that cause soft rot in a wide range of plant hosts worldwide (1–3). This severe disease leads to tissue maceration and eventually plant death *(4)*. *Dickeya* species have received considerable attention in recent years due to their emergence in European and Asian countries where they cause major economic losses by infecting agriculturally important crops such as apple, potato, tomato and carrot (5–8). Besides, the ability of *Dickeya* species to colonize various ecological niches such as soil, ground water, surfaces and interior of plant hosts and insects (2, 9, 10) makes them remarkably versatile models for the scientific community. *Dickeya dadantii* strain 3937 is the main model used for studying the regulation of gene expression during the plant pathogenic process since its genome was the first to be sequenced and curated in the *Dickeya* genus (11), yet the annotated map of its transcriptome is still lacking (as for all other phytopathogens). Obtaining this map is thus an important step for further progress in the field, and is the object of the present study.

Classically, bacterial transcription is described with the model of Jacob and Monod based on operons, defined as sets of contiguous and functionally-related genes co-transcribed from a single promoter up to a single terminator (12). In recent years however, accumulating studies in various bacterial species demonstrated that most operons actually comprise internal promoters and terminators, generating transcripts of variable gene composition, generally in a condition-dependent manner (13–22). This phenomenon, also known as suboperonic regulation (23), may be compared to alternative splicing in eukaryotes (24) and demonstrates a higher complexity of bacterial transcriptional landscapes than previously thought. Besides, transcription has been shown to extend beyond operons (18, 25), the latter being actually part of larger functional genomic units, referred to as transcription units (TUs) throughout the manuscript.

For studying the regulation of *D.dadantii* gene expression and pathogenicity program, its transcriptome should be characterized under physiological conditions encountered by the bacteria during the infection process. The latter consists of an asymptomatic phase where bacteria penetrate and colonize plant tissues, and a symptomatic phase where a set of plant cell wall-degrading enzymes are secreted by the bacteria to macerate plant tissues *(1)*. As for all pathogens, each stage of the infection process can be assimilated to an environmental change to which bacteria need to adapt for a successful infection of the host, by coordinating the virulence and global gene expression programs. *D. dadantii* presents the advantage that most of these successive stages can be closely reproduced through various shocks applied in culture conditions, generating a set of transcriptomic data which precisely mimic the dynamics of infection *(26, 27)*. This study is built upon a combination of such data, obtained by three different techniques providing complementary knowledge that take advantage of the possibility to reproduce many of the environmental conditions encountered by *D. dadantii* during pathogenesis: (1) high-resolution paired-end RNA-seq data; (2) DNA microarray data; (3) dRNA-seq data treated with or without

Terminator exonuclease (TEX). Together, these data provide a fine and exhaustive description of transcription dynamics and architecture directly relevant to plant infection. By means of an integrative computational method designed specifically for this study, we obtain a detailed and annotated map of the TUs defining the *Dickeya dadan*tii transcriptom*e, i.e*., the sets of contiguous co-expressed genes, by reconstructing the RNA landscape and validating the degree of co-expression among genes of the same TU throughout the pathogenic process. We then quantitatively map transcription start and termination sites in the investigated biologically relevant conditions, and analyze the associated predicted promoter and terminator motifs. We show that TUs sometimes coincide with predicted operons but are generally longer, most of them exhibiting internal promoters and terminators. We characterize the occurrence of transcriptional read-through at terminators, a mechanism that has been suggested as a basal coordinator and regulator of gene expression yet never explored in phytopathogens and still poorly understood across genomes in general. We finally detect putative noncontiguous operons and excludons in *D. dadantii* genome. This transcriptomic map might serve as a community resource to help elucidating the regulation *of D. dadantii* gene expression including virulence program in the context of plant pathogenicity. It also provides insights into basal rules of coordination of transcription that might be valid for other bacteria, specifically for other *Dickeya* species for which a core genome of 1300 genes was identified by comparative genomics (9).

Results and discussion

Characterization of *Dickeya dadantii* transcription units in conditions relevant to plant infection

The biological relevance of the target transcriptional map crucially relies on that of the datasets used as input information, in this case, with respect to the pathogenic process. In the course of infection, *D. dadantii* faces sequential and sudden environmental changes that result partly from plant defenses: acidic and oxidative stresses occur during plant penetration and colonization respectively (asymptomatic phase), when bacteria grow exponentially by consuming simple sugars and small soluble oligosaccharides available in plant apoplast *(1)*. When the latter are exhausted, the symptomatic phase initiates, bacteria enter into early stationary growth phase and start producing plant cell wall-degrading enzymes, mainly pectinases, that lead to the soft rot symptoms. The maceration of plant tissues releases vacuolar components in the apoplast, exposing the bacteria to osmotic stress. Such conditions have been closely reproduced experimentally, and transcriptome profiling was performed to provide a complete catalog of genes' responses and RNA landscapes along the infection process.

Dataset 1 was generated from high-resolution paired-end, strand-specific RNA-seq covering 6 growth conditions summarizing most of these environmental variations (Materials and Methods): minimal medium M63 supplemented with sucrose that mimics the asymptomatic phase (in exponential phase), addition of polygalacturonate (PGA), a pectic polymer present in plant cell wall, reproducing the symptomatic phase (early stationary phase) and the release of plant cell wall components that act as carbon sources for growth and inducers of pectinase synthesis (28–30). The fast response to aggression was considered by monitoring the effect of a global and transient chromosomal DNA relaxation that occurs *in vivo* in response to plant acidic and oxidative stresses (31) and plays a crucial role as a coordinator of gene expression including virulence program (26, 31–33).

Dataset 2 was generated from DNA microarray data covering a more exhaustive and detailed survey of conditions encountered during infection, with 32 growth conditions (Materials and Methods) extending those of dataset 1 with the presence of leaf extracts, but more importantly in each medium, by the separate exposure to acidic, oxidative or osmotic stresses closely reproducing the plant chemical defenses (27).

Finally, dataset 3 was generated from differential RNA sequencing (dRNA-seq) experiments carried out on four samples obtained by mixing RNAs from the large variety of environmental conditions of dataset 2 followed by treatment with Terminator exonuclease (TEX) prior to sequencing. TEX enzyme which degrades processed 5'-monophosphate RNAs and consequently enriches the samples in primary 5′-triphosphate end transcripts (34), thus precisely locating TSSs utilized during infection at single-nucleotide resolution.

This collection of diverse and complementary transcriptomic datasets provided a solid ground for precisely characterizing *D. dadantii* transcription units in conditions relevant to

plant infection, rather than basing our analysis on genomic data alone like most operon predictors (intergenic distances between genes, functional links among products).

In a first step, we reconstructed the RNA landscape throughout the whole genome at single nucleotide resolution based on dataset 1 (paired-end strand-specific RNA-Seq). Putative TUs were defined by fusing adjacent genes as long as RNA fragments were found in their intergenic region, a signature of co-transcription (Materials and Methods). As a second criterion, if genes within the same putative TU are indeed co-transcribed, they should be co-expressed during the successive steps of the infection process, which was analyzed based on the diversity of available DNA microarray data (dataset 2). This criterion provides an orthogonal cross-validation with respect to the first one, by quantitatively comparing the expression levels of genes (rather than intergenic signal) in a wider range of conditions. Briefly, we developed a hierarchical clustering framework (35) constrained to group adjacent genes as long as the correlation in their expression across all conditions was above a cutoff value indicative of similar expression profiles (see Materials and Methods). The resulting clusters were considered as separate TUs when the correlation dropped below the cutoff value; otherwise, the putative TU was validated on the basis of co-expression as well as intergenic RNA signal.

With this approach, we mapped the first layer of transcription organization in *D.dadantii.* According to our findings, the 4211 protein-coding genes are organized into 2028 transcription units (provided in Tab. S1A), among which 1118 are monocistronic and 910 are polycistronic, ranging from 2 to 28 genes (Fig. 1A, 1B and Tab. S4). At the genomic scale, we compared our results with those of Rockhopper, a popular operon predictor that uses expression data as well as genomic information as input (36) (Materials and Methods). Remarkably, TUs are generally longer than predicted operons: the average TU, including monocistronic ones, has 2.1 genes whereas the average polycistronic TU has 3.4 genes against 1.6 and 3.1 respectively for predicted operons (Fig. 1C, Tab. S4). Almost three quarters (73.5%) of all genes are co-transcribed in TUs against 56.9% for predicted operons (Fig. 1A, Tab. S4). Our results indicate that TUs are indeed larger functional genomic units as 45% of predicted operons are extended by at least one gene (Fig. 1D), in agreement with recent findings in *E. coli* based on long-read sequencing (18).

As an example, the *sapABCDF* operon encoding a transporter involved in antimicrobial peptide resistance and virulence in numerous bacteria including *D. dadantii* (37) is extended to include the enoyl-acyl carrier protein reductase *fabI* that catalyzes an essential step in the biosynthesis of fatty acids of the membrane (38) (Fig. 3B). It might be noted that *fabI* has a different genomic location in *E. coli* and is consequently not co-transcribed with *sapABCDF* in that species (39) although this synteny is conserved in other *Dickeya* genomes, showing that TUs can merge and/or vary over time at the evolutionary scale. Since these genes are functionally unrelated (except for a general relation with the membrane), the biological relevance and role of this event requires further investigation.

The *glg* genes involved in glycogen metabolism constitute another instructive example. They were initially classified in two separate operons in *E. coli* (40), and later identified as

a single TU involving alternative transcripts of variable gene composition depending on growth conditions (41). The latter is also true in *D. dadantii* according to our findings (Fig. 3C), illustrating how transcription extends beyond the scale of the operon. Nevertheless, 45% of predicted operons exactly coincide with a TU in our analysis (Fig. 1D), including known examples such as *smtAmukFEB* involved in chromosome partitioning (Fig. 2A) (42). Besides, many identified TUs are likely operons of unknown functions and features (Fig. 2B), which represent interesting starting points to discover new transcriptional functional units.

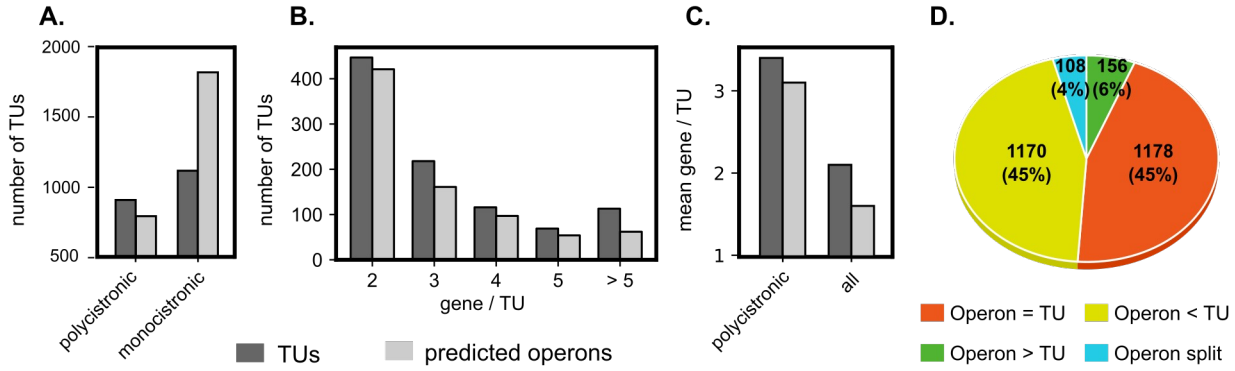

Figure 1. Transcription units comprise operons predicted by Rockhopper but are usually longer. (A) Repartition of monocistronic and polycistronic TUs identified by our analysis and comparison to predicted operons. (B) Size distributions. (C) Average number of genes. (D) Fate of predicted operons that are mostly found as or within TUs in our algorithm.

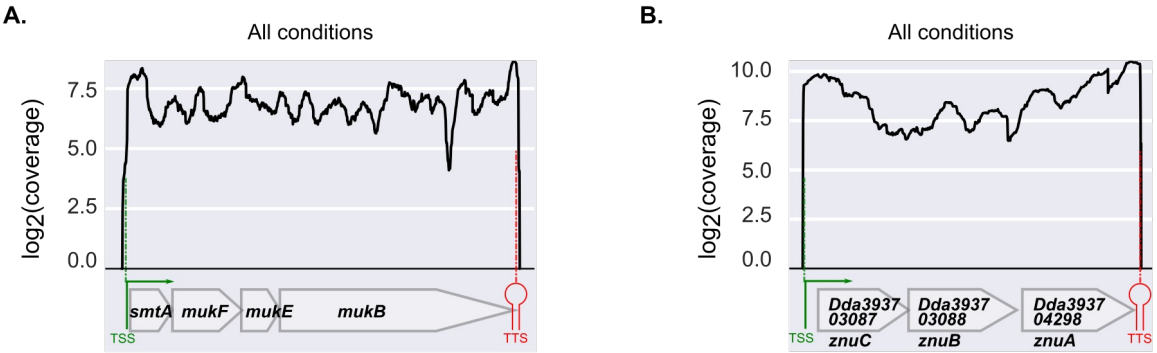

Figure 2. Many transcription units identified by our approach coincide with operons. (A) Example of a known operon (*smtAmukFEB*). (B) Identification of a new TU exhibiting uniform read coverage and strong internal cross-correlations (Fig. S1A), clearly indicative of an operon. Its function was unknown but a homology analysis revealed that it corresponds to the cluster of genes *znuCBA,* a Zn^{2+} uptake system (43).

Genome-wide identification of *D. dadantii* transcription start and termination sites reveals a complex transcriptional landscape

Once *D. dadantii* transcription units were defined, the next step was to elaborate a map of transcription start sites (TSSs) and transcription termination sites (TTSs) for each TU along the genome. First, as mentioned above, dRNA-seq experiments were carried out to build a large library of 9313 putative TSSs of high-resolution (33) covering a wide range of physiological conditions relevant to plant infection (dataset 3, Tab. S2). These were obtained by treating the RNA samples with TEX prior to sequencing, then the TSSer workflow was applied for a precise determination of TSS positions (44), followed by a visual curation (Materials and Methods). Putative TTS positions, on the other hand, were predicted based on the two main mechanisms of transcription termination in bacteria: Rho-independent (intrinsic) termination and Rho-dependent (regulated) termination (45) using ARNold (46) and RhoTermPredict (47) programs, respectively (Materials and Methods), resulting in 3564 intrinsic TTSs (Tab. S3A) and 5851 Rho-dependent TTSs (Tab. S3B).

A quantitative mapping of the transcription landscape was then performed in order to estimate the contribution of each TSS/TTS to the TU and filter out weak TSSs and TTSs. While most comparable maps define TSSs/TTSs by their position only, we exploited the complementarity of the input data to also systematically analyze their magnitude (or strength) in the investigated conditions. The +TEX library and TTS predictions are not suitable for the latter purpose, which required building a second list of TSSs and TTSs of poorer resolution but quantitative magnitude from the non-treated paired-end RNA-seq data (dataset 1). Briefly, TSSs and TTSs were defined based on the enrichment in RNA fragment starts and stops upstream of gene starts and downstream of gene stops respectively, and the number of fragments associated to these sites across all samples was considered as the global strength (Materials and Methods). The lists obtained with the two methods (from dataset 1 and dataset 3) were then merged into a unified list of TSSs/TTSs of optimal spatial resolution, quantitative magnitude, and with an estimated level of confidence dependent on the level of agreement between the two datasets (see Materials and Methods). These TSSs and TTSs were then assigned to TUs and internal TSSs/TTSs were retained only if they yielded at least 15% of the total start/stop magnitude of the TU and were thus biologically relevant in the investigated conditions, resulting in a total of 2595 TSSs and 1699 TTSs (including internal ones) over all TUs (Tab. S1A to S1C). Finally, a scan for promoter motifs upstream of TSSs was conducted with bTSSfinder (48): we found promoters upstream of 1848 (71%) TSSs in total (Tab. S1B, Fig. 3).

Altogether, this algorithm allowed us to quantitatively map TSSs and TTSs across all TUs based on a cross-validation of different types of data and predictions. This analysis allowed us to refine the comparison of TUs and operons presented above. According to our findings, only 20% of polycistronic TUs (181) exhibit a single promoter and terminator (Fig. 2 and 3A) and thus fit into the classical definition of operons, only 47% of which (85) are predicted as such by Rockhopper. The 80% remaining TUs (729) are complex: 32%

(287) have at least one internal TSS without any internal TTS (Fig. 3A) such as *sapABCDFfabI* (Fig. 3B). 37% (339) have both internal TSS(s) and TTS(s) (Fig. 3A), such as *glgBXCAP* (Fig. 3B) and *pelCZ* (Fig. 3C). Finally, 11% (103) have at least one internal TTS without any internal TSS (Fig. 3A) such as *rhlB-gppA-pehV* (Fig. 4A), *pelD-paeY-pemA* (Fig. 4B) and *gcvTHP* (Fig. 6). Most *D. dadantii* TUs can consequently generate alternative transcripts of variable gene composition, resulting in a dense and complex transcriptional landscape.

A notable feature of complex TUs is the heterogeneity of transcription levels along the genes due to internal TSSs / TTSs, usually in a condition-dependent manner, resulting in a moderate correlation in the expression of genes within the TU (13). As an example, in *sapABCDFfabI* TU (Fig. 3B), *fabI* is expressed both as part of the entire transcript and as an independent transcript generated from a strong internal TSS (37% of total magnitude), explaining the lower correlation between *fabI* and the remaining genes (Fig. S1B). In *glgBXCAP* TU (Fig. 3C), alternative transcripts of variable gene composition can be generated depending on TSS and TTS usage. Another example relevant to plant infection is the *pelCZ* cluster (Fig. 3D) encoding two endopectate lyases secreted by *D. dadantii* which degrade pectin contained in plant cell walls (49). The substrates of Pel enzymes are pectic oligomers, *e.g.* PGA, that act as inducers of *pel* expression. By Northern blotting, the *pelCZ* genes were shown to be co-transcribed into a single polycistronic transcript under inducing conditions by PGA, in addition to the two monocistronic mRNAs encoded by *pelC* or *pelZ* under non-inducing conditions (50). Our present findings are in full agreement with these observations, as *pelCZ* is detected as a single TU harboring one internal TSS and TTS, each giving rise to monocistronic transcripts. In our data, *pelCZ* expression profiles are similar in presence or absence of PGA in spite of a drastically different global expression level (Fig. S2), suggesting that in absence of inducer, this very low level previously prevented a reliable detection of the entire transcript.

Altogether, our findings clearly indicate that the canonical operon model is insufficient to explain the complexity of the *D. dadantii* transcriptional landscape, in line with results in many other organisms (13–22). The existence of alternative entry and exit points for RNA Polymerase inside TUs allows the cells to adjust the relative expression level of adjacent genes within a global coordination of expression of the entire TU (Fig. 3) that may allow, in the case of *D. dadantii* during plant infection, a rapid adaptation to fast changing environment.

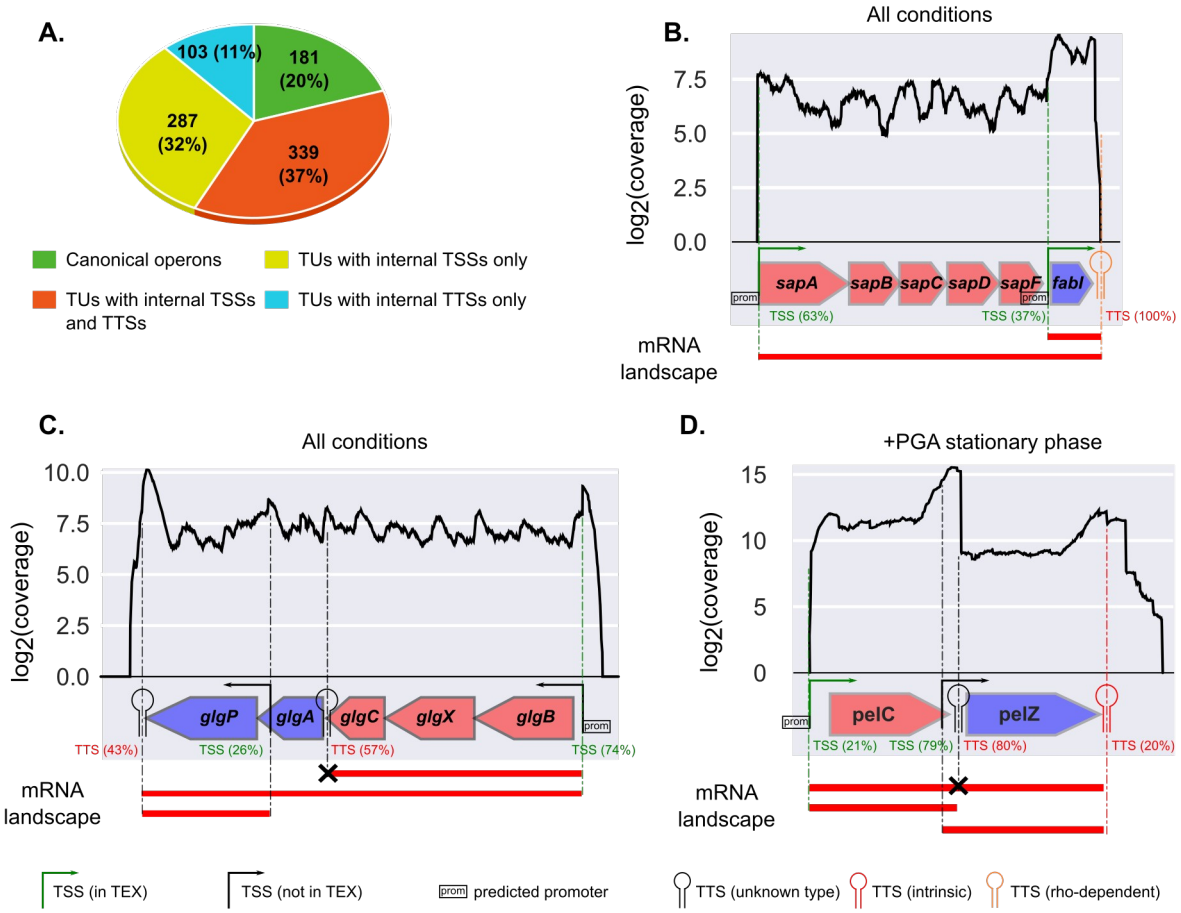

Figure 3. Most transcription units are complex with internal TSSs and/or TTSs that generate alternative transcripts. (A) Characteristics of TUs. (B) The *sapABCDF* and *fabI* genes, predicted by Rockhopper (36) as two separate red and blue operons respectively, were identified as a single TU, with a strong internal TSS (37% total magnitude) expressing *fabI* alone. (C) The *glg* genes were identified as a single TU (involving several isoforms) containing two separate predicted operons (blue and red genes), as suggested by the uniform read coverage and in line with results in *E. coli* (18, 20, 41) although the identified transcripts differ among studies. (D) Identification of the *pelCZ* TU with different isoforms depending on the condition, as previously determined (50). The two genes are split into different operons by Rockhopper. A strong internal TSS, followed by a strong TTS, contributes to the complexity of its expression (see text).

Transcriptional read-through, the root of transcription extension ?

We showed that transcription units comprise predicted operons, yet are generally longer. This extension of transcription might, in part, result from the ability of RNA Polymerase to stochastically override an imperfect terminator by a mechanism referred to as transcriptional read-through (18, 25). The latter has long been identified in specific operons (51–53) and was shown more recently to be widespread in bacterial genomes (17, 18, 25), where it may in fact play a basal coordination and regulation role (21).

A condition-independent rate of stochastic termination might result in the co-expression of the genes located before and after the TTS (as in a classical operon), but with a reduced transcriptional level of the latter, a mechanism possibly relevant to functionally related genes that must be expressed at different strengths while keeping a constant ratio (53). However, the termination efficiency can also be subject to regulation, depending on environmental conditions and metabolic needs, resulting in a variable degree of read-through and thus of relative expression levels (51, 52). Such conditional read-through can involve Rho and other proteins assisting termination (54–57) as well as other conditional premature termination mechanisms such as attenuation (58, 59), T-box conditional termination (60, 61) and riboswitches (62, 63).

An example of condition-independent read-through occurs at the *rhlB-gppA-pehV* TU (Fig. 4A). The *rhlB* gene encodes a component of the RNA degradosome (64, 65) whereas *gppA* encodes guanosine 5'-triphosphate 3'-diphosphate (pppGpp) pyrophosphatase involved in bacterial stringent response (66) and the *pehV* gene encodes a polygalacturonase involved in pectin degradation (67). These genes are functionally unrelated (except for a distant link to nutritional stress) yet appear co-transcribed, which is in fact quite frequent among operons (42, 68). This TU exhibits a variable expression level (by up to 50%) across the sampled conditions, but the internal (relative) expression pattern is condition-independent: *rhlB* and *gppA* are expressed at a similar level, whereas *pehV* is systematically less transcribed (Fig. 4A). This observation is correlated with the presence of an intrinsic internal TTS downstream of *gppA*. By computing the expression ratio of *pehV* compared to *rhlB/gppA*, we inferred the associated termination probability (or terminator strength) and found a constant value $P(TTS_{gppA}) = 92 \pm 1\%$ (95% confidence interval) characteristic of a non-conditional transcriptional read-through. Thus, the three genes are co-transcribed from a single promoter of condition-dependent activity, with a reduced transcriptional level of *pehV* exhibiting a constant ratio (8%) compared to the other genes (Fig. S1C). The biological relevance of this mechanism remains to be clarified. In *E. coli*, *rhlB* and *gppA* were also recently shown to be co-transcribed (18, 20). Another example of condition-independent read-through occurs at the *gcvTHP* TU involved in glycine cleavage (69) (Fig. 6). We detected an internal TTS downstream of *gcvH* in accordance with studies in *E. coli* (18, 20) and inferred its termination probability $P(TTS_{gcvH}) = 71 \pm 22\%$ (95% confidence interval), based on the expression ratio of *gcvP*

compared to *gcvT* and *gcvH* across RNA-seq conditions. It is unclear whether the variability is due to RNA-seq signal variations or a weak regulation of the termination rate. The GcvT, H, and P proteins are part of the glycine cleavage system with GcvL (70), and GcvP activity might be required at lower rate in the investigated conditions.

By definition, all identified internal TTSs (549) experience transcriptional read-through. As a rough estimate, condition-independent read-through was detected for 77 (14%) of all identified internal TTSs, based on the constant expression ratio of the genes located downstream vs upstream of each internal TSS across RNA-seq conditions (Fig. 4, Materials and Methods). The remaining internal TTSs rather experience condition-dependent read-through; however, the systematic estimation of stochastic termination rates at internal TTSs is delicate based on our data only, due to the limited number of RNA-seq conditions and the presence of nearby TSSs that contribute to the heterogeneous expression levels along the TU, as illustrated by *pelCZ* (Fig. 3D).

An example of condition-dependent read-through occurs at the *pelD-paeY-pemA* TU (Fig. 4B), which is identified by our approach but was also characterized by Northern blotting *(*71*)*. It encodes three genes involved in pectin degradation: *pelD*, *paeY* and *pemA*. In the initial step of pectinolysis occuring in plants, *paeY* (acetylesterase) and *pemA* (methylesterase) remove acetyl and methyl groups from pectin, which can then be efficiently degraded by the pectate lyase *pelD (*3*)*. The *pelD* gene is essentially transcribed as a monocistronic RNA, although its terminator (predicted as intrinsic) can be overstepped to generate a polycistronic transcript comprising the three genes *(*71*)*. In exponential phase, the three genes are homogeneously (but weakly) transcribed as a unique polycistronic RNA, suggesting that the internal TTS is not efficient ($P(TTS_{pelD})$ = 0 %). In stationary phase in presence of PGA, the whole TU is up-regulated, and the internal TTS becomes more efficient ($P(TTS_{pelD})$ = 86±3 %, 95% confidence interval), resulting in the extensive synthesis of the *pelD* monocistronic RNA and a lower expression level of the two further genes. The regulation events occurring at this TTS remain to be characterized, but may adjust the relative expression levels of the genes following metabolic needs, since *pelD* has a predominant role in pectin degradation and virulence *(*72*,* 73*)* and must likely be expressed at much higher rates than the two other enzymes. Another example occurs at the *cytABCD* TU (Fig. S3) which encodes four insecticidal toxins and was previously shown to produce a polycistronic mRNA comprising the four genes, besides the possible existence of alternative isoforms *(*74*)*. The sequencing coverage together with the putative internal intrinsic TTS detected after *cytA* are clearly indicative of a condition-dependent read-through (Fig. S3), with termination occurring less efficiently at *cytA* in stationary phase in presence of PGA compared to exponential phase. This variation in termination efficiency at *cytA* associated to an environmental change may again allow tuning the relative amounts of the corresponding toxins, especially if a precise and condition-dependent balance between them is required for optimal activity during the insect infection process *(*74*)*. Interestingly, this cluster of four genes was acquired by horizontal transfer.

Since transcriptional read-through partly relies on basal RNA Polymerase / TTS interactions, it might be conserved during horizontal transfer among bacterial species without requiring an independent acquisition and integration of regulatory signals in the transcriptional regulatory network of the recipient cell.

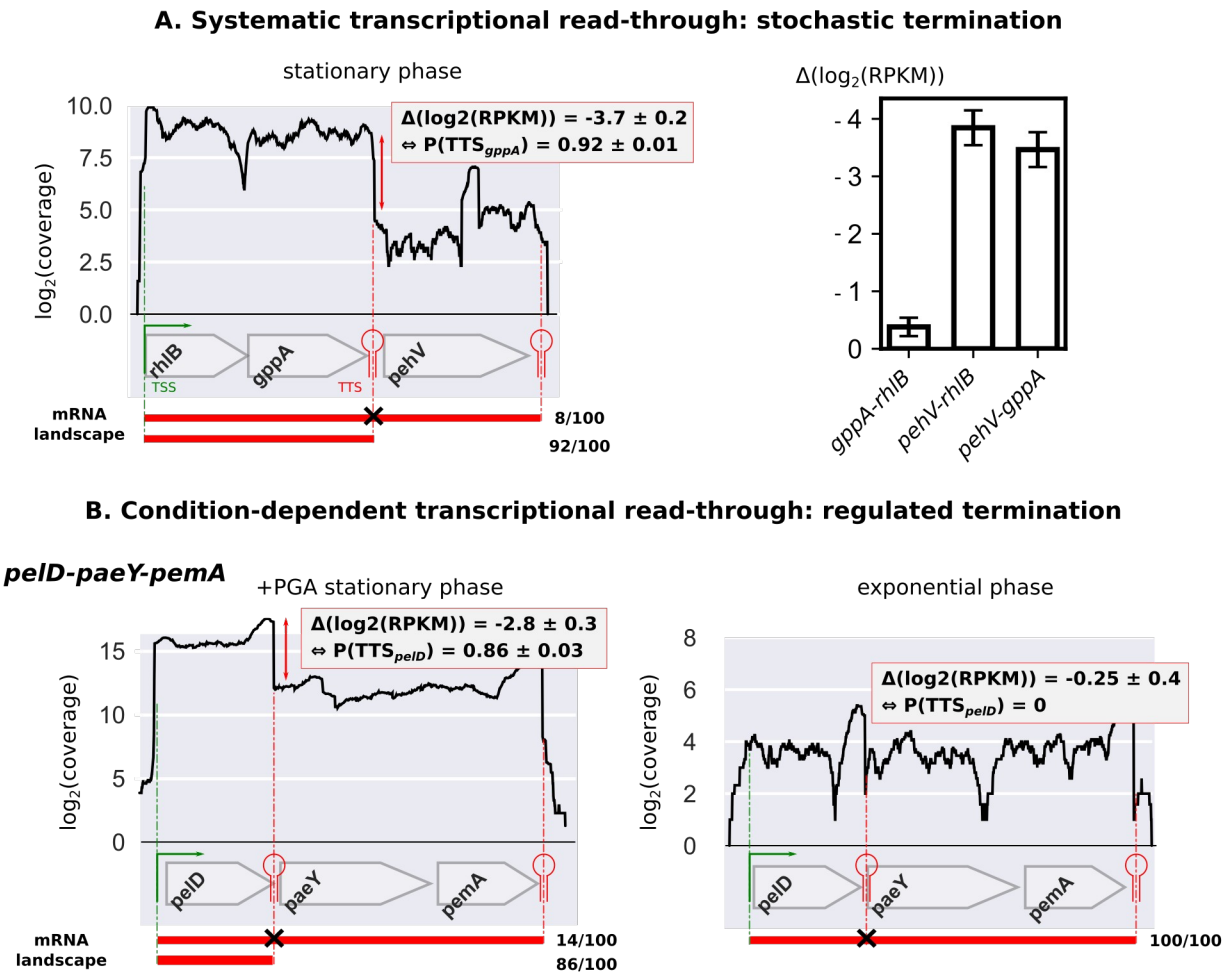

Figure 4. Quantification of transcriptional read-through. (A) Non-conditional read-through: example of the *rhlB-gppA-pehV* TU. The first two genes are homogeneously transcribed among conditions, resulting in an expression variation Δ($\log_2$(RPKM)) close to 0 (95% confidence intervals are shown). The internal TTS downstream of *gppA* (*TSS_{gppA})* is occasionally overstepped, resulting in two transcript isoforms. This partial read-through exhibits a constant rate across investigated conditions and results in a systematic and regular decrease of *pehV* expression compared to the other genes while their expression level changes by up to 50%. The probability of stochastic termination (p(TSS_{gppA}) = 0.92 ± 0.01) is inferred from this expression variation. (B) Condition-dependent read-through: example of the *pelD-paeY-pemA* TU. A TTS is identified downstream of *pelD* in agreement with previous studies *(71)*. Its termination probability is regulated and depends on growth phase and presence of PGA (0.86 ± 0.03 vs 0), besides a global up- or down-regulation of the whole TU.

Detection of putative excludons and noncontiguous transcriptions units

All previous examples involved genes located on the same DNA strand; yet recent studies also describe interactions between overlapping antisense coding transcripts, involved in a mutual regulation. In particular, noncontiguous operons refer to operons that contain a gene or group of genes that is transcribed in the opposite direction (75). 83 TUs with such features were found in *D. dadantii* genome (provided in Tab. S5A). Among them, an example is the *indCvfmAZBCDFG* TU encoding a component of the *vfm* quorum sensing system required for plant cell wall-degrading enzymes production and pathogenesis (Fig. 5) *(76)*. The *vfmE* gene, located on the opposite strand and within this TU, is also part of this system and known to encode a transcriptional activator (of the AraC family) of the *vfm* locus. Since all genes of the TU are co-transcribed within a single mRNA, it is likely that these two overlapping antisense transcripts could negatively regulate each other, e.g. by transcriptional interference (RNA Polymerases collision) or RnaseIII-mediated double-stranded RNA processing *(77)*. An expression increase of the *vfm* locus would then reduce the expression of *vfmE,* and in turn its own expression, forming a genome-embedded negative feedback loop for controlling the production of quorum sensing signal and plant cell-wall degrading enzymes.

Finally, “excludons” refer to genomic regions in which convergent or divergent genes display overlapping transcription (78). By exploiting the map of transcription start and termination sites , we found 160 putative convergent excludons (overlapping 3’ UTRs) and 63 putative divergent excludons (overlapping 5’ UTRs) (provided in Tab. S5B). An example is the divergent excludon between *greB* and *ompRenvZ* transcription units, encoding a transcript cleavage factor required for effective transcription elongation (79) and a two-component signal transduction system involved in osmotic stress response *(80)*, respectively (Fig. S5). Both TUs comprise long 5’UTRs, forming a region of overlapping transcription that was previously identified in *E. coli* (81) and may underpin a mutual post-transcriptional regulation.

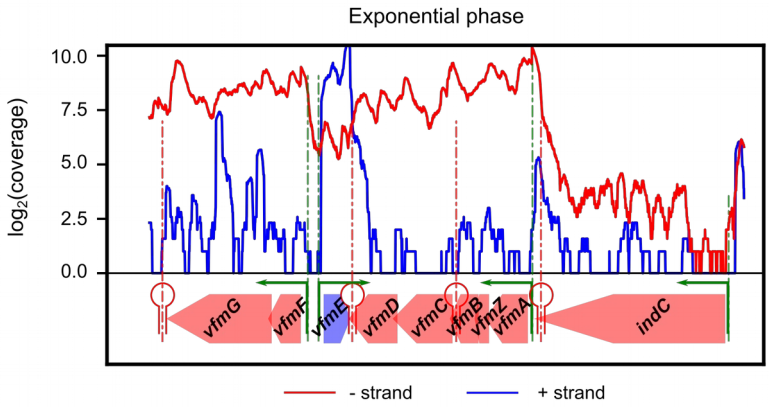

Figure 5. The *vfm* locus forms a potential noncontiguous transcription unit. The *vfmE* gene is transcribed in the opposite direction of the *indCvfmAZBCDFG* TU, generating two overlapping mRNAs that might be involved in a mutual regulation (see text).

Concluding statement

By combining three transcriptomic datasets designed to reproduce the variety of the transcriptional landscape of the infection process and yielding complementary information, together with a computational method designed for this study, we precisely determined and annotated the transcriptomic map of *D. dadantii* in conditions relevant to plant infection, the first of its kind in phytopathogens. Beyond its practical aspect as a community resource to help the scientific community unraveling gene regulation including virulence program in this and related species, our findings clearly indicate, after others, that the canonical operon model is insufficient to explain the complexity of bacterial transcriptional landscapes, in line with many studies. The ability of the cell to differentially express genes of the same operon depending on metabolic needs and environmental conditions was first described with suboperonic regulation years ago. Later, with the emergence of next-generation sequencing, transcriptomic analyses confirmed at the genomic scale that most operons were able to generate alternative transcripts of variable gene composition. Recent findings include non contiguous operons and excludons, where the expression of an operon transcript can be mutually regulated with that of a gene located on the opposite strand at the same locus. Transcriptional read-through at terminators is another mechanism that might play a basal coordination and regulation role, and explain the extent of transcription beyond the scale of operons. It might also contribute to transcription overlapping between adjacent genes. In the present study, we have demonstrated and quantified, for the first time in a phytopathogen, transcriptional read-through at terminators. Altogether, ours findings provide insights into the mechanisms of basal coordination of transcription and might contribute to the revision of the canonical view of operon structure and transcription organization.

Materials and Methods

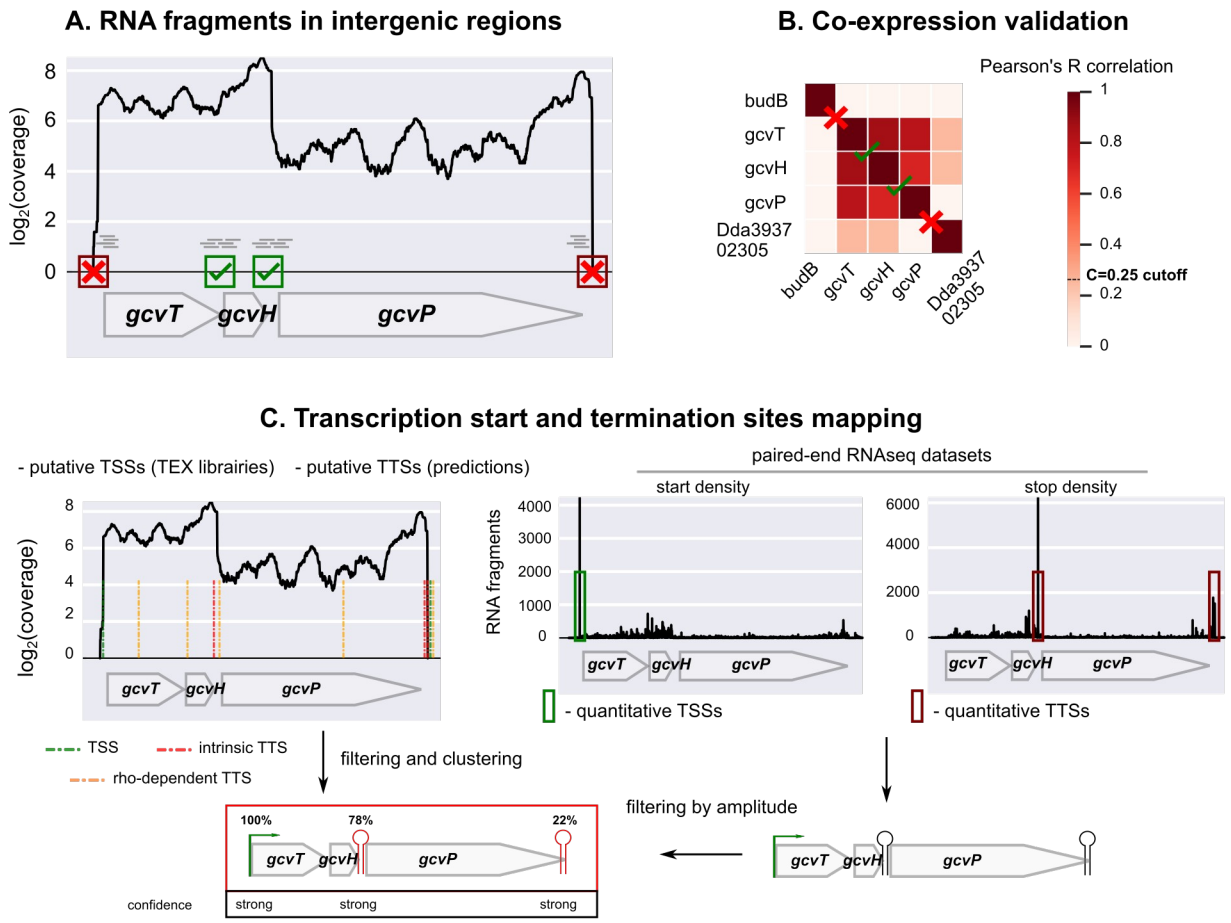

Figure 6. Algorithm for the characterization of *D. dadantii* transcription units. (A) Definition of putative TUs based on RNA-seq coverage (dataset 1) in intergenic regions, with isodirectional genes being split when the coverage drops to zero (here before *gcvT* and after *gcvP*). (B) Validation based on correlation of expression across 32 conditions (dataset 2). The genes of identified putative TUs are correlated, in contrast to surrounding isodirectional genes (*budB* and *Dda3937_02305*). (C) TSS and TTS mapping based on dRNA-seq (TEX libraries, dataset 3), TTS predictions, and paired-end RNA-seq data (dataset 1). First, putative TSSs of high resolution but qualitative strength were defined from an analysis of TEX libraries, and putative TTSs were predicted. Second, a list of TSSs and TTSs of quantitative strength (for the investigated experimental conditions) but poorer resolution was defined from the enrichment of RNA-seq paired-end fragment starts (start density) and stops (stop density) upstream of gene starts and downstream of gene stops, respectively. Third, only TSSs and TTSs with sufficient strength were retained, and compared to the closest (independently obtained) TEX TSS / predicted hairpin loop, in order to define their exact position and level of confidence. As a result of the analysis, this TU included the *gcvTHP* genes, the first two genes being expressed both as part of the

entire transcript and as an independent transcript generated from a strong internal TTS (76% of total magnitude), explaining the lower correlation between *gcvP* and the remaining genes.

Bacterial strain, genome annotation and genome-wide predictions of operons. The genome sequence and annotation files from *Dickeya dadantii* strain *3937* were obtained from NCBI under accession NC_014500.1 (11). This work focused on coding genes only (CDS, representing 4211 genes over 4411 in total). *D. dadantii* operons were predicted using Rockhopper, a recent computational tool for operon prediction based on RNA-seq expression data as well as genomic and functional information (36), by providing dataset 1 as input.

RNA-sequencing data (dataset 1) and definition of putative transcription units based on intergenic signals. Strand-specific, paired-end RNA-seq processed data used in this study are described elsewhere (32). Raw data can be found in the EBI Gene Expression (ArrayExpress) database under accession E-MTAB-7650. Transcriptomes were obtained in 6 conditions (with two biological replicates each) including various growth (exponential or stationary phase, in presence or absence of PGA) and DNA supercoiling conditions (novobiocin shock). For each sample, RNA fragments were reconstructed from paired-end reads information, and genome-wide coverage was computed from resulting RNA fragments coordinates using a Python home-made script. To define putative transcription units, separately for each strand, adjacent genes were fused in the same putative TU as long as the coverage was greater than 0 at each position of their intergenic region (independently of its size) for at least half of the samples (Fig. 6A).

Microarray data (dataset 2) and co-expression validation of the putative transcription units using hierarchical clustering. Microarray processed data used in this study are described elsewhere (26). Data are available in the EBI Gene Expression (ArrayExpress) database under accession *E-MTAB-541*. They comprise 32 conditions (with two biological replicates each) including various growth conditions encountered by *D. dadantii* during plant infection: cells were harvested in exponential or stationary phase, in presence or in absence of PGA and leaf extract, and exposed or not to environmental perturbations (acidic, osmotic and oxidative stress). For all microarray datasets, Pearson's correlation coefficient (R) was computed among all gene pairs over all conditions on the logarithm of the normalised expression level (derived from probes intensity). For each putative TU, adjacent genes were grouped into clusters based on this correlation, using a hierarchical clustering framework constrained to group adjacent genes only, with a custom Python script. At each iteration of the algorithm, the median of cross-correlations among all clusters (or genes) was computed, and the adjacent clusters with maximal median were fused. The hierarchical clustering ends when a cutoff value *C* for the correlation is reached (Fig. 6B). If the agglomeration of all genes of the TU is achieved without reaching *C*, the TU is validated. Otherwise, the final clusters are considered as separate TUs. A high *C* value results in short highly correlated TUs, whereas a short *C* value yields longer moderately correlated TUs (Tab. S4). We defined the value *C=0.25* such that 20% of

operon predictions were discarded (Fig. S4), since it is the number of false predictions (i.e. specificity) evaluated for Rockhopper in *E. coli*, a *D. dadantii enterobacterium* relative. Varying the precise value of C did not qualitatively change the main results (Tab. S4). The identified TUs exhibit a similar length distribution as those reported in *E. coli* (13, 18).

Differential RNA-sequencing experiments and genome-wide identification of putative transcription start sites (dataset 3). RNAs from the samples used for microarray analysis (dataset 2, (26)) were pooled into four samples S1 to S4. Each of them resulted in a combination of stress conditions (pH, NaCl, H_2O_2) and growth conditions: exponential phase with stress (S1), exponential phase without stress (S2), stationary phase with stress (S3), stationary phase without stress (S4). Those samples were then supplied to Vertis Biotechnologie AG for TEX treatment and Illumina sequencing (dataset 3). Briefly, ribosomal RNA molecules were depleted from the total RNA samples using the Ribo-Zero rRNA Removal Kit for bacteria (Epicentre). From the rRNA depleted RNA samples, the small RNA fractions <200 nt were separated from the large fractions using the RNeasy MinElute Cleanup Kit (Qiagen). TSS cDNA libraries were generated from the rRNA depleted large RNA fractions. First, the samples were fragmented using RNase III. Samples were then poly(A)-tailed using poly(A) polymerase, and split into two halves. One half was treated with Terminator exonuclease (+TEX, Epicentre), while the other one was left untreated (-TEX). 5'PPP structures were then converted into 5'P ends using RNA 5' Polyphosphatase (5'PP, Epicentre). Afterwards, an RNA adapter was ligated to the 5'P of the RNA. First-strand cDNA synthesis was performed using an oligo(dT)-adapter primer and the M-MLV reverse transcriptase. The resulting cDNAs were PCR-amplified using a high fidelity DNA polymerase, and purified using the Agencourt AMPure XP kit (Beckman Coulter Genomics). The cDNA pools were then sequenced on an Illumina NextSeq 500 system using 60 bp read length and a single-end, strand-specific protocol. Sequencing reads were trimmed, and both polyA and adapters removed: the resulting sequence files (fastq format) are available in the EBI Gene Expression (ArrayExpress) database under accession *E-MTAB-9075*. Putative TSS positions were then determined based on the enrichment of sequencing reads in TEX-treated samples (+TEX) compared to non-treated ones (-TEX) using TSSer, an automated annotation program from dRNA-seq data with default parameters: TSS positions within 5 bases on the same strand were clustered together and the position with the highest amount of read increase in the +TEX library was retained. TSSs obtained with such approach are provided in Tab. S2.

Genome-wide detection of transcription start and termination sites from RNA-seq data, mapping to the transcription units. For each position of the genome, we computed the total number of RNA fragments starting at this particular position, across all samples of dataset 1, using a Python home-made script, and similarly for RNA fragments ending at this particular position. From these start and stop densities (Fig. 6C), we defined TSSs/TTSs as positions upstream of gene start codons (up to 250 bp, based on 5'UTR lengths in *E. coli*) and downstream of gene stop codons (up to 100bp, based on 3'UTR lengths in *E. coli*) enriched for RNA fragments starts and stops, respectively. In each of

these regions, we started from site *i* with highest start/stop signal k_i. We imposed three successive conditions to filter start/stop sites important for the expression of the TU: (1) k_i is significantly different from 0 (with 99% confidence); (2) k_i is greater or equal than a density cutoff value *D*, set as ten times the median of the density values of the region investigated for TSSs, and five times the median density values for TTSs, showing that the recorded transcripts indeed start at that position, rather than along a poorly defined starting region. In that case, the position *i* was considered as a putative TSS/TTS, of strength k_i. Setting a low density cutoff *D* would tend to identify all peaks, including weak TSSs / TTSs but also peaks resulting from RNA-seq signal variations, whereas a high value of *D* would tend to identify only the strongest peaks, resulting in an absence of detection of weak / secondary TSSs / TTSs. We selected the value of *D* by (1) visually curating the density graphs and distinguishing k_i resulting from RNA-seq variations from potential *bona fide* TSSs / TTSs positions and (2) such that TSSs and TTSs were detected for known operons and experimentally characterized TUs (described in the manuscript).

Then, for each TSS identified with such approach, if a putative TSS obtained from dataset 3 (TEX libraries) was close enough (+/- 20 bp), its position (supposedly more precise) was retained, and the level of confidence of such TSS, validated by both approaches, was set as strong (Fig. 1, Tab. S1A and S1B, quoted “strong”). For TTSs, the same method was applied by using the position of the closest predicted hairpin loop (+/- 50 bp) (Tab. S1A and S1C, quoted “strong”), in which case the corresponding type of termination (intrinsic/Rho-dependent) was indicated (Tab. S1C). Remaining putative TSSs from dataset 3 and predicted TTSs that were not assigned to a TSS/TTS identified from dataset 1 were discarded due to a weak strength in the investigated conditions. Remaining TSSs and TTSs obtained from dataset 1 that were not validated were still retained (Tab. S1A to S1C, confidence quoted “medium” for both TSSs and TTSs, type of TTS quoted “unknown”). Finally, for TUs with internal TSSs and/or TTSs, only internal TSSs and TTSs with 15% relative amplitude (*i.e.* $k_i / \Sigma(k_i)$ quoted “relative strength”) were retained, resulting in a total of 2595 TSSs and 1699 TTSs over all TUs (Tab. S1A to S1C). The primary TSS refers to the TU’s first gene TSS whereas the primary TTS refers to the TU’s last gene TTS. For TUs with no primary TSS/TTS, we indicated the closest putative TSS/TTS (up to 250 bp upstream TU start/up to 100bp downstream TU stop such as before) (Tab. S1A to S1C, confidence quoted “low”). Otherwise, the primary TSS/TTS was arbitrarily set as the TU’s first gene start codon position/TU’s last gene stop codon (Tab. S1A, confidence quoted “not found”).

Detection of condition-independent readthrough at internal TTSs. For each internal TTS, the expression ratio $\Delta(\log_2(\text{RPKM}))$ of the gene located downstream compared to the gene located upstream was computed across RNA-seq conditions. We imposed two successive conditions to consider the transcriptional read-through rate of the former internal TTS as condition-independent: (1) $\Delta(\log_2(\text{RPKM})) \leq -0.5$ for at least 8 samples over 12 corresponding at least to a termination probability P(TTS) = 71%, (2) standard

error of the mean $\sigma_{P(TTS)} \leq 12.5\%$ corresponding to a relatively constant mean expression ratio and subsequent termination probability.

Data availability

- *Dickeya dadantii* strain 3937 genome sequence and annotation files: NCBI accession number NC_014500.1 (https://www.ncbi.nlm.nih.gov/nuccore/NC_014500) (11)
- RNA-seq data (dataset 1): EBI Gene Expression (ArrayExpress) accession number E-MTAB-7650 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-7650/) (32)
- Microarray data (dataset 2): EBI Gene Expression (ArrayExpress) accession number E-MTAB-541 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-541/) (26)
- Differential RNA-seq data (dataset 3): EBI Gene Expression (ArrayExpress) accession number E-MTAB-9075 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9075/). *Note: not publicly available until the manuscript is accepted. Please use login details Username = Reviewer_E-MTAB-9075, Password = jenuxmon.*

Acknowledgements

We thank the whole CRP team for useful discussions as well as Ivan Junier.

Funding. R.F was funded by a research allocation from the French Ministry. This work also benefited from INSA Lyon grants [BQR 2016 to S.M.]; IXXI; Agence Nationale de la Recherche [ANR-18-CE45-0006-01 to S.M.], Breakthrough Phytobiome IDEX LYON project, Université de Lyon Programme d'investissements d'Avenir [ANR16-IDEX-0005]; Centre National de la Recherche Scientifique [to S.R., F.H. W.N. and S.M.]; Université Claude Bernard Lyon 1 [to S.R., F.H. W.N. and S.M.].

Conflict of interest statement. None declared.

613 **References**

1. Reverchon S, Muskhelisvili G, Nasser W. 2016. Virulence Program of a Bacterial Plant Pathogen: The Dickeya Model. Prog Mol Biol Transl Sci 142:51–92.

2. Reverchon S, Nasser W. 2013. Dickeya ecology, environment sensing and regulation of virulence programme. Environ Microbiol Rep 5:622–636.

3. Hugouvieux Cotte Pattat N, Condemine G, Gueguen E, Shevchik VE. 2020. Dickeya Plant Pathogens, p. 1–10. *In* eLS. American Cancer Society.

4. Czajkowski R, Pérombelon MCM, Veen JA van, Wolf JM van der. 2011. Control of blackleg and tuber soft rot of potato caused by Pectobacterium and Dickeya species: a review. Plant Pathology 60:999–1013.

5. Hossain A, Masum MMI, Wu X, Abdallah Y, Ogunyemi SO, Wang Y, Sun G, Li B, An Q. 2020. Screening of Bacillus strains in biocontrol of pathogen Dickeya dadantii causing stem and root rot disease of sweet potato. Biocontrol Science and Technology 0:1–19.

6. Fujikawa T, Ota N, Sasaki M, Nakamura T, Iwanami T. 2019. Emergence of apple bacterial quick decline caused by Dickeya dadantii in Japan. J Gen Plant Pathol 85:314–319.

7. Charkowski AO. 2018. The Changing Face of Bacterial Soft-Rot Diseases. Annual Review of Phytopathology 56:269–288.

8. Toth IK, Wolf JM van der, Saddler G, Lojkowska E, Hélias V, Pirhonen M, Tsror (Lahkim) L, Elphinstone JG. 2011. Dickeya species: an emerging problem for potato production in Europe. Plant Pathology 60:385–399.

9. Duprey A, Taib N, Leonard S, Garin T, Flandrois J-P, Nasser W, Brochier Armanet C, Reverchon S. 2019. The phytopathogenic nature of Dickeya aquatica 174/2 and the dynamic early evolution of Dickeya pathogenicity. Environmental Microbiology 21:2809–2835.

10. Grenier A-M, Duport G, Pagès S, Condemine G, Rahbé Y. 2006. The Phytopathogen Dickeya dadantii (Erwinia chrysanthemi 3937) Is a Pathogen of the Pea Aphid. Appl Environ Microbiol 72:1956–1965.

11. Glasner JD, Yang C-H, Reverchon S, Hugouvieux-Cotte-Pattat N, Condemine G, Bohin J-P, Van Gijsegem F, Yang S, Franza T, Expert D, Plunkett G, San Francisco MJ, Charkowski AO, Py B, Bell K, Rauscher L, Rodriguez-Palenzuela P, Toussaint A, Holeva MC, He SY, Douet V, Boccara M, Blanco C, Toth I, Anderson BD, Biehl BS, Mau B, Flynn SM, Barras F, Lindeberg M, Birch PRJ, Tsuyumu S, Shi X, Hibbing M, Yap M-N, Carpentier M, Dassa E, Umehara M, Kim JF, Rusch M, Soni P, Mayhew GF, Fouts DE, Gill SR, Blattner FR, Keen NT, Perna NT. 2011. Genome Sequence of the Plant-Pathogenic Bacterium Dickeya dadantii 3937▿. J Bacteriol 193:2076–2077.

12. Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–356.

13. Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J, San Miguel P, Shimada T, Ishihama A, Mori H, Wanner BL. 2014. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. mBio 5:e01442-01414.

14. Junier I, Unal EB, Yus E, Lloréns-Rico V, Serrano L. 2016. Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium. Cell Syst 2:391–401.

15. Li S, Dong X, Su Z. 2013. Directional RNA-seq reveals highly complex condition-dependent transcriptomes in E. coli K12 through accurate full-length transcripts assembling. BMC Genomics 14:520.

16. Mao X, Ma Q, Liu B, Chen X, Zhang H, Xu Y. 2015. Revisiting operons: an analysis of the landscape of transcriptional units in E. coli. BMC Bioinformatics 16.

17. Nicolas P, Mäder U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S, Becher D, Bisicchia P, Botella E, Delumeau O, Doherty G, Denham EL, Fogg MJ, Fromion V, Goelzer A, Hansen A, Härtig E, Harwood CR, Homuth G, Jarmer H, Jules M, Klipp E, Le Chat L, Lecointe F, Lewis P, Liebermeister W, March A, Mars RAT, Nannapaneni P, Noone D, Pohl S, Rinn B, Rügheimer F, Sappa PK, Samson F, Schaffer M, Schwikowski B, Steil L, Stülke J, Wiegert T, Devine KM, Wilkinson AJ, van Dijl JM, Hecker M, Völker U, Bessières P, Noirot P. 2012. Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science 335:1103–1106.

18. Yan B, Boitano M, Clark TA, Ettwiller L. 2018. SMRT-Cappable-seq reveals complex operon variants in bacteria. 1. Nat Commun 9:1–11.

19. Warrier I, Ram-Mohan N, Zhu Z, Hazery A, Echlin H, Rosch J, Meyer MM, Opijnen T van. 2018. The Transcriptional landscape of Streptococcus pneumoniae TIGR4 reveals a complex operon architecture and abundant riboregulation critical for growth and virulence. PLOS Pathogens 14:e1007461.

20. Ju X, Li D, Liu S. 2019. Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. 11. Nat Microbiol 4:1907–1918.

21. Mejía-Almonte C, Busby SJW, Wade JT, van Helden J, Arkin AP, Stormo GD, Eilbeck K, Palsson BO, Galagan JE, Collado-Vides J. 2020. Redefining fundamental concepts of transcription initiation in bacteria. Nature Reviews Genetics 1–16.

22. Rosinski-Chupin I, Sauvage E, Fouet A, Poyart C, Glaser P. 2019. Conserved and specific features of Streptococcus pyogenes and Streptococcus agalactiae transcriptional landscapes. BMC Genomics 20.

23. Adhya S. 2003. Suboperonic regulatory signals. Sci STKE 2003:pe22.

24. Kornblihtt AR, Schor IE, Alló M, Dujardin G, Petrillo E, Muñoz MJ. 2013. Alternative splicing: a pivotal step between eukaryotic transcription and translation. 3. Nat Rev Mol Cell Biol 14:153–165.

25. Junier I, Rivoire O. 2016. Conserved Units of Co-Expression in Bacterial Genomes: An Evolutionary Insight into Transcriptional Regulation. PLoS ONE 11:e0155740.

26. Jiang X, Sobetzko P, Nasser W, Reverchon S, Muskhelishvili G. 2015. Chromosomal “Stress-Response” Domains Govern the Spatiotemporal Expression of the Bacterial Virulence Program. mBio 6.

27. Jiang X, Zghidi Abouzid O, Oger Desfeux C, Hommais F, Greliche N, Muskhelishvili G, Nasser W, Reverchon S. 2016. Global transcriptional response of Dickeya dadantii to environmental stimuli relevant to the plant infection. Environmental Microbiology 18:3651–3672.

28. Nasser W, Condemine G, Plantier R, Anker D, Robert-Baudouy J. 1991. Inducing properties of analogs of 2-keto-3-deoxygluconate on the expression of pectinase genes of Erwinia chrysanthemi. FEMS Microbiol Lett 81:73–78.

29. Nasser W, Reverchon S, Condemine G, Robert-Baudouy J. 1994. Specific interactions of Erwinia chrysanthemi KdgR repressor with different operators of genes involved in pectinolysis. J Mol Biol 236:427–440.

30. Reverchon S, Robert-Baudouy J. 1987. Molecular cloning of an Erwinia chrysanthemi oligogalacturonate lyase gene involved in pectin degradation. Gene 55:125–133.

31. Ouafa Z-A, Reverchon S, Lautier T, Muskhelishvili G, Nasser W. 2012. The nucleoid-associated proteins H-NS and FIS modulate the DNA supercoiling response of the pel genes, the major virulence factors in the plant pathogen bacterium Dickeya dadantii. Nucleic Acids Res 40:4306–4319.

32. El Houdaigui B, Forquet R, Hindré T, Schneider D, Nasser W, Reverchon S, Meyer S. 2019. Bacterial genome architecture shapes global transcriptional regulation by DNA supercoiling. Nucleic Acids Res 47:5648–5657.

33. Martis B. S, Forquet R, Reverchon S, Nasser W, Meyer S. 2019. DNA Supercoiling: an Ancestral Regulator of Gene Expression in Pathogenic Bacteria? Computational and Structural Biotechnology Journal 17:1047–1055.

34. Sharma CM, Vogel J. 2014. Differential RNA-seq: the approach behind and the biological insight gained. Curr Opin Microbiol 19:97–105.

35. Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, Achas M, Adebiyi E. 2016. Clustering Algorithms: Their Application to Gene Expression Data. Bioinform Biol Insights 10:237–253.

36. Tjaden B. 2019. A computational system for identifying operons based on RNA-seq data. Methods https://doi.org/10.1016/j.ymeth.2019.03.026.

37. López-Solanilla E, García-Olmedo F, Rodríguez-Palenzuela P. 1998. Inactivation of the sapA to sapF locus of Erwinia chrysanthemi reveals common features in plant and animal bacterial pathogenesis. Plant Cell 10:917–924.

38. Cronan JE, Thomas J. 2009. Bacterial Fatty Acid Synthesis and its Relationships with Polyketide Synthetic Pathways. Methods Enzymol 459:395–433.

39. Bergler H, Fuchsbichler S, Högenauer G, Turnowsky F. 1996. The enoyl-[acyl-carrier-protein] reductase (FabI) of Escherichia coli, which catalyzes a key regulatory step in fatty acid biosynthesis, accepts NADH and NADPH as cofactors and is inhibited by palmitoyl-CoA. Eur J Biochem 242:689–694.

40. Preiss J. 2009. Glycogen: Biosynthesis and Regulation. EcoSal Plus 3.

41. Montero M, Almagro G, Eydallin G, Viale AM, Muñoz FJ, Bahaji A, Li J, Rahimpour M, Baroja-Fernández E, Pozueta-Romero J. 2011. Escherichia coli glycogen genes are organized in a single glgBXCAP transcriptional unit possessing an alternative suboperonic promoter within glgC that directs glgAP expression. Biochem J 433:107–117.

42. Yamanaka K, Ogura T, Niki H, Hiraga S. 1995. Characterization of the smtA gene encoding an S-adenosylmethionine-dependent methyltransferase of Escherichia coli. FEMS Microbiol Lett 133:59–63.

43. Patzer SI, Hantke K. 2000. The zinc-responsive regulator Zur and its control of the znu gene cluster encoding the ZnuABC zinc uptake system in Escherichia coli. J Biol Chem 275:24321–24332.

44. Jorjani H, Zavolan M. 2014. TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data. Bioinformatics 30:971–974.

45. Ray-Soni A, Bellecourt MJ, Landick R. 2016. Mechanisms of Bacterial Transcription Termination: All Good Things Must End. Annual Review of Biochemistry 85:319–347.

46. Naville M, Ghuillot-Gaudeffroy A, Marchais A, Gautheret D. 2011. ARNold: a web tool for the prediction of Rho-independent transcription terminators. RNA Biol 8:11–13.

47. Di Salvo M, Puccio S, Peano C, Lacour S, Alifano P. 2019. RhoTermPredict: an algorithm for predicting Rho-dependent transcription terminators based on Escherichia coli, Bacillus subtilis and Salmonella enterica databases. BMC Bioinformatics 20:117.

48. Shahmuradov IA, Mohamad Razali R, Bougouffa S, Radovanovic A, Bajic VB. 2017. bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and Escherichia coli. Bioinformatics 33:334–340.

49. Garibaldi A, Bateman DF. 1971. Pectic enzymes produced by Erwinia chrysanthemi and their effects on plant tissue. Physiological Plant Pathology 1:25–40.

50. Pissavin C, Robert-Baudouy J, Hugouvieux-Cotte-Pattat N. 1996. Regulation of pelZ, a gene of the pelB-pelC cluster encoding a new pectate lyase of Erwinia chrysanthemi 3937. J Bacteriol 178:7187–7196.

51. Gusarov I, Nudler E. 2001. Control of Intrinsic Transcription Termination by N and NusA: The Basic Mechanisms. Cell 107:437–449.

52. Morita T, Ueda M, Kubo K, Aiba H. 2015. Insights into transcription termination of Hfq-binding sRNAs of Escherichia coli and characterization of readthrough products. RNA 21:1490–1501.

53. Stringer AM, Currenti S, Bonocora RP, Baranowski C, Petrone BL, Palumbo MJ, Reilly AA, Zhang Z, Erill I, Wade JT. 2014. Genome-Scale Analyses of Escherichia coli and Salmonella enterica AraC Reveal Noncanonical Targets and an Expanded Core Regulon. J Bacteriol 196:660–671.

54. Boudvillain M, Figueroa-Bossi N, Bossi L. 2013. Terminator still moving forward: expanding roles for Rho factor. Current Opinion in Microbiology 16:118–124.

55. Burns CM, Richardson LV, Richardson JP. 1998. Combinatorial effects of NusA and NusG on transcription elongation and rho-dependent termination in Escherichia coli11Edited by M. Gottesman. Journal of Molecular Biology 278:307–316.

56. Mondal S, Yakhnin AV, Babitzke P. 2017. Modular Organization of the NusA- and NusG-Stimulated RNA Polymerase Pause Signal That Participates in the Bacillus subtilis trp Operon Attenuation Mechanism. J Bacteriol 199.

57. Lawson MR, Berger JM. 2019. Tuning the sequence specificity of a transcription terminator. Curr Genet 65:729–733.

58. Merino E, Yanofsky C. 2005. Transcription attenuation: a highly conserved regulatory strategy used by bacteria. Trends Genet 21:260–264.

59. Turnbough CL. 2019. Regulation of Bacterial Gene Expression by Transcription Attenuation. Microbiol Mol Biol Rev 83.

60. Green NJ, Grundy FJ, Henkin TM. 2010. The T box mechanism: tRNA as a regulatory molecule. FEBS Lett 584:318–324.

61. Zhang J, Chetnani B, Cormack ED, Alonso D, Liu W, Mondragón A, Fei J. 2018. Specific structural elements of the T-box riboswitch drive the two-step binding of the tRNA ligand. eLife 7:e39518.

62. Millman A, Dar D, Shamir M, Sorek R. 2017. Computational prediction of regulatory, premature transcription termination in bacteria. Nucleic Acids Res 45:886–893.

63. Proshkin S, Mironov A, Nudler E. 2014. Riboswitches in regulation of Rho-dependent transcription termination. Biochim Biophys Acta 1839:974–977.

64. Coburn GA, Miao X, Briant DJ, Mackie GA. 1999. Reconstitution of a minimal RNA degradosome demonstrates functional coordination between a 3' exonuclease and a DEAD-box RNA helicase. Genes Dev 13:2594–2603.

65. Py B, Higgins CF, Krisch HM, Carpousis AJ. 1996. A DEAD-box RNA helicase in the Escherichia coli RNA degradosome. Nature 381:169–172.

66. Hauryliuk V, Atkinson GC, Murakami KS, Tenson T, Gerdes K. 2015. Recent functional insights into the role of (p)ppGpp in bacterial physiology. 5. Nat Rev Microbiol 13:298–309.

67. Nasser W, Shevchik VE, Hugouvieux-Cotte-Pattat N. 1999. Analysis of three clustered polygalacturonase genes in Erwinia chrysanthemi 3937 revealed an anti-repressor function for the PecS regulator. Mol Microbiol 34:641–650.

68. de Lorenzo V, Danchin A. 2008. Synthetic biology: discovering new worlds and new words. The new and not so new aspects of this emerging research field. EMBO Rep 9:822–827.

69. Stauffer LT, Fogarty SJ, Stauffer GV. 1994. Characterization of the Escherichia coli gcv operon. Gene 142:17–22.

70. Kikuchi G, Motokawa Y, Yoshida T, Hiraga K. 2008. Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycinemia. Proc Jpn Acad Ser B Phys Biol Sci 84:246–263.

71. Shevchik VE, Hugouvieux-Cotte-Pattat N. 1997. Identification of a bacterial pectin acetyl esterase in Erwinia chrysanthemi 3937. Mol Microbiol 24:1285–1301.

72. Boccara M, Diolez A, Rouve M, Kotoujansky A. 1988. The role of individual pectate lyases of Erwinia chrysanthemi strain 3937 in pathogenicity on saintpaulia plants. Physiological and Molecular Plant Pathology 33:95–104.

73. Beaulieu C. 1993. Pathogenic Behavior of Pectinase-Defective *Erwinia chrysanthemi* Mutants on Different Plants. MPMI 6:197.

74. Costechareyre D, Dridi B, Rahbé Y, Condemine G. 2010. Cyt toxin expression reveals an inverse regulation of insect and plant virulence factors of Dickeya dadantii. Environ Microbiol 12:3290–3301.

75. Sáenz-Lahoya S, Bitarte N, García B, Burgui S, Vergara-Irigaray M, Valle J, Solano C, Toledo-Arana A, Lasa I. 2019. Noncontiguous operon is a genetic organization for coordinating bacterial gene expression. PNAS https://doi.org/10.1073/pnas.1812746116.

76. Nasser W, Dorel C, Wawrzyniak J, Van Gijsegem F, Groleau M-C, Déziel E, Reverchon S. 2013. Vfm a new quorum sensing system controls the virulence of Dickeya dadantii. Environ Microbiol 15:865–880.

77. Toledo-Arana A, Lasa I. 2020. Advances in bacterial transcriptome understanding: From overlapping transcription to the excludon concept. Mol Microbiol 113:593–602.

78. Sesto N, Wurtzel O, Archambaud C, Sorek R, Cossart P. 2013. The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. 2. Nature Reviews Microbiology 11:75–82.

79. Hsu LM, Vo NV, Chamberlin MJ. 1995. Escherichia coli transcript cleavage factors GreA and GreB stimulate promoter escape and gene expression in vivo and in vitro. Proc Natl Acad Sci U S A 92:11588–11592.

80. Cai SJ, Inouye M. 2002. EnvZ-OmpR interaction and osmoregulation in Escherichia coli. J Biol Chem 277:24155–24161.

81. Lybecker M, Zimmermann B, Bilusic I, Tukhtubaeva N, Schroeder R. 2014. The double-stranded transcriptome of Escherichia coli. Proc Natl Acad Sci USA 111:3134–3139.

Supplemental material

Supplemental figures

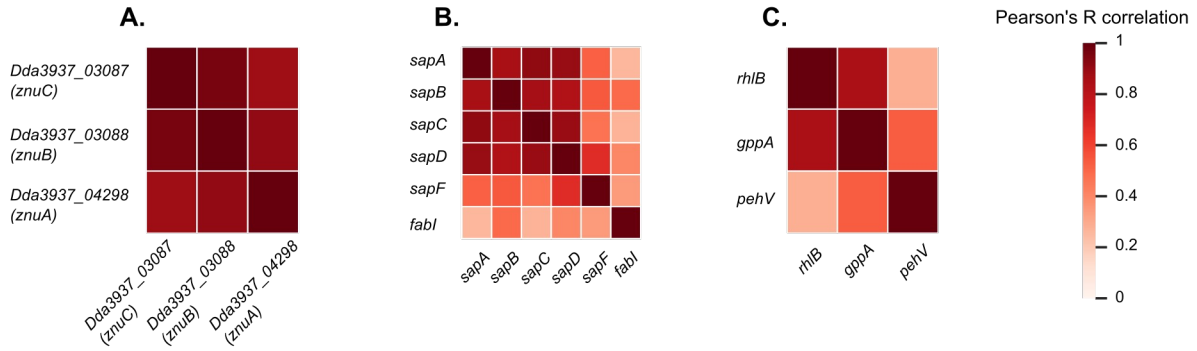

Figure S1. Co-expression validations. (A) *znuCBA* TU: the three genes exhibit strong internal cross-correlations clearly indicative of an operon. (B) *sapABCDFfabI* TU: the six genes are co-expressed despite a reduced correlation of *fabI* with the other genes due to the presence of a strong internal TSS upstream of *fabI* (Fig. 3B). (C) *rhlb-gppA-pehV* TU: the three genes are co-expressed despite a reduced transcriptional level of *pehV* (Fig. 4A) and a reduced correlation due to condition-independent read-through at the *gppA* intrinsic terminator.

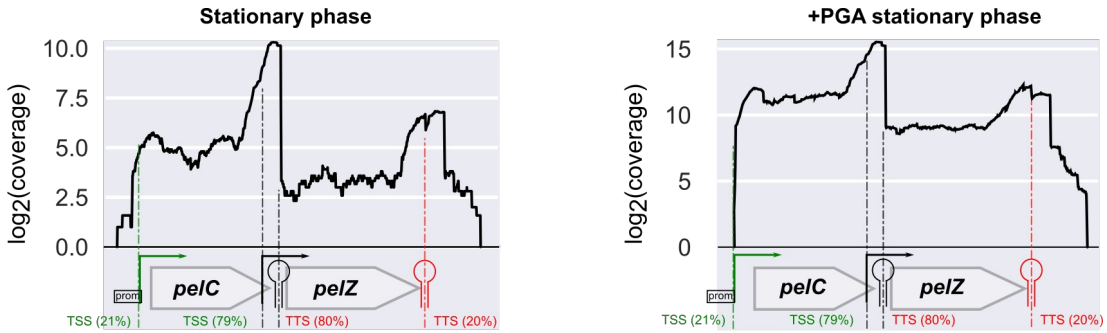

Figure S2. Effect of PGA on *pelCZ* TU expression. *pelC* and *pelZ* expression profiles are similar in presence or absence of PGA in spite of a drastically different global expression level.

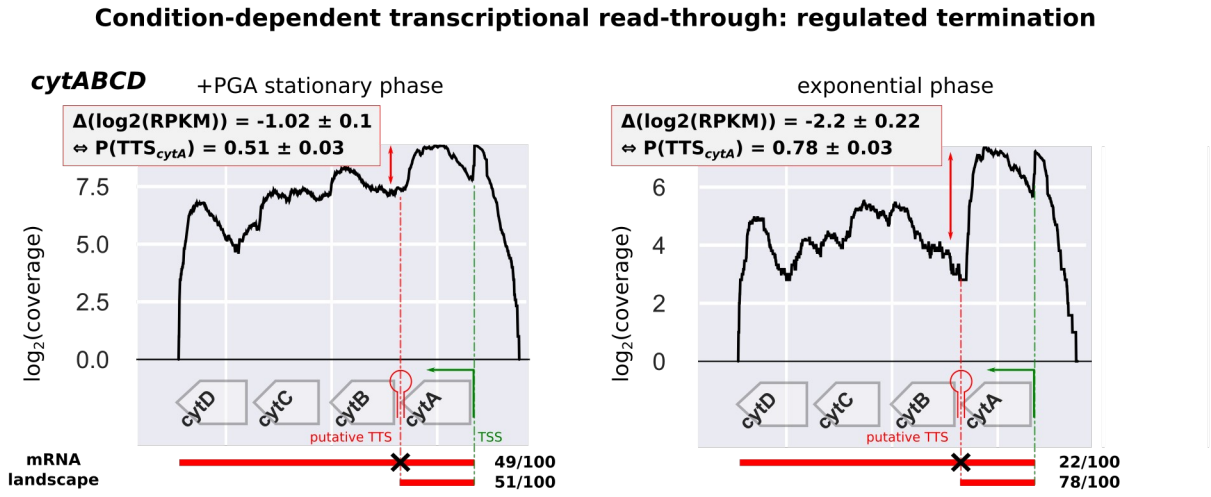

Figure S3. Quantification of transcriptional read-through. Condition-dependent read-through: example of the *cytABCD* TU (74). A putative Rho-independent TTS is identified downstream *cytA* although not validated. Its probability of termination (inferred from the expression variation Δ($\log_2$(RPKM)) of *cytA* compared to the other genes) is regulated and depends both on the growth phase and the presence of PGA (P(TTS$_{cytA}$) = 0.78 ± 0.03 vs 0.51 ± 0.03) besides a global up-regulation of the whole TU.

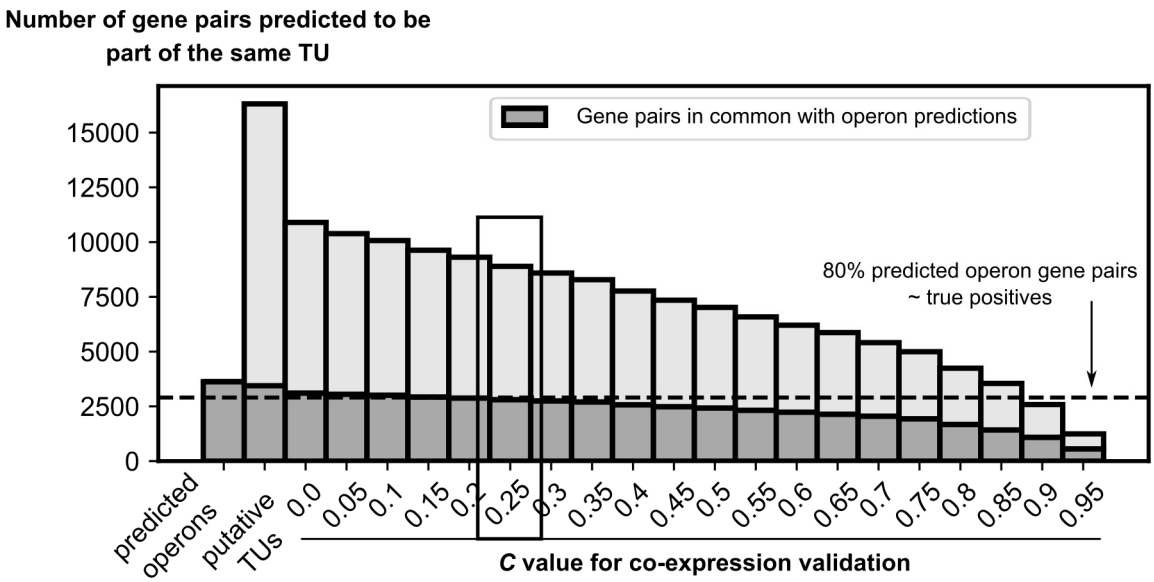

Figure S4. Co-expression validation of transcription units for different correlation thresholds *C*. TUs obtained with high *C* values are more highly correlated but shorter. Putative TUs are obtained from step 1 of the analysis (intergenic signal), without any requirement on the correlation of expression. With the chosen value (*C = 0.25*), TUs group around three times more gene pairs than predicted operons by Rockhopper. The value of *C* was chosen such that 20% of operon predictions were discarded, since it is the number of false predictions of Rockhopper in *E. coli*, a *D. dadantii enterobacterium* relative.

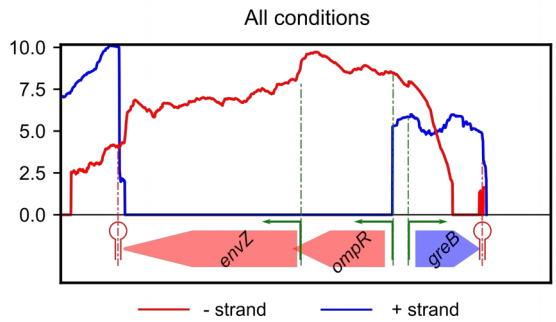

Figure S5. The *greB* and *ompRenvZ* transcription units form a potential divergent excludon. Long 5'UTRs overlapping transcripts are generated by *ompR* and *greB* divergent genes and might form a dsRNA that could prevent each other transcription. *In E. coli, ompR and envZ* are part of the same operon (red), and *greB* is transcribed alone (blue). Such genomic region forming a dsRNA was also identified in *E. coli* (81).

Supplemental tables

Table S1A. *Dickeya dadantii* transcription units defined by our approach.

Table S1B. TSSs across TUs.

Table S1C. TTSs across TUs.

Table S2. Putative TSSs identified by differential RNA-seq (TEX treatment) under a wide range of environmental conditions.

Table S3A. Genomic position and secondary structure of putative TTSs: intrinsic terminators predicted by ARNold (Erpin and RNAmotif algorithms).

Table S3B. Genomic position of putative TTSs: Rho-dependent terminators predicted by RhoTermPredict.

Table S4. Catalogue of transcription unit architecture. Putative TUs are obtained from the first step of the approach (analysis of intergenic signal). Varying the precise value of the correlation threshold *C* for co-expression validation (step 2) does not change the results qualitatively. A larger *C* value results in shorter but more highly correlated TUs. Final TUs obtained with *C = 0.25* are longer than predicted operons and exhibit a similar length distribution as those reported in *E. coli* (13, 18).

Table S5A. Catalogue of putative noncontiguous transcription units.

Table S5B. Catalogue of putative excludons.

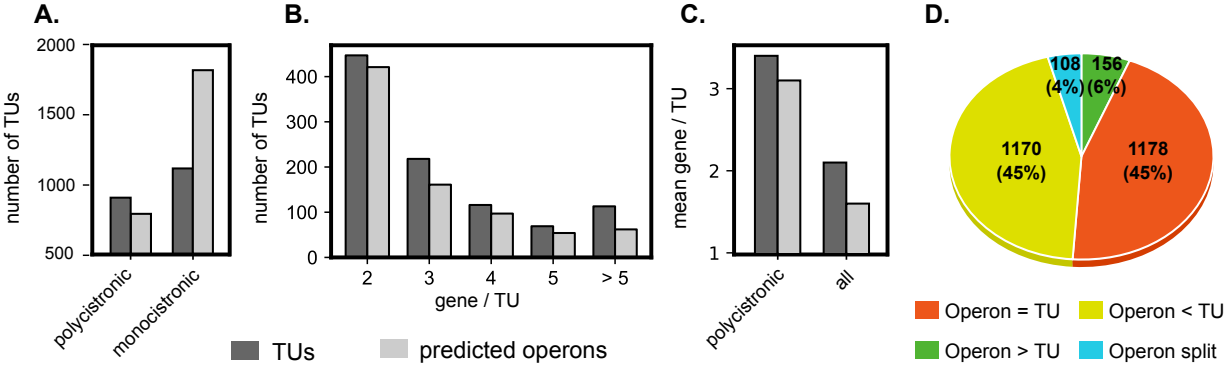

Figure 1. Transcription units comprise predicted operons but are usually longer. (A) Repartition of monocistronic and polycistronic TUs identified by our analysis and comparison to predicted operons. (B) Size distributions. (C) Average number of genes. (D) Fate of predicted operons that are mostly found as or within TUs in our algorithm.

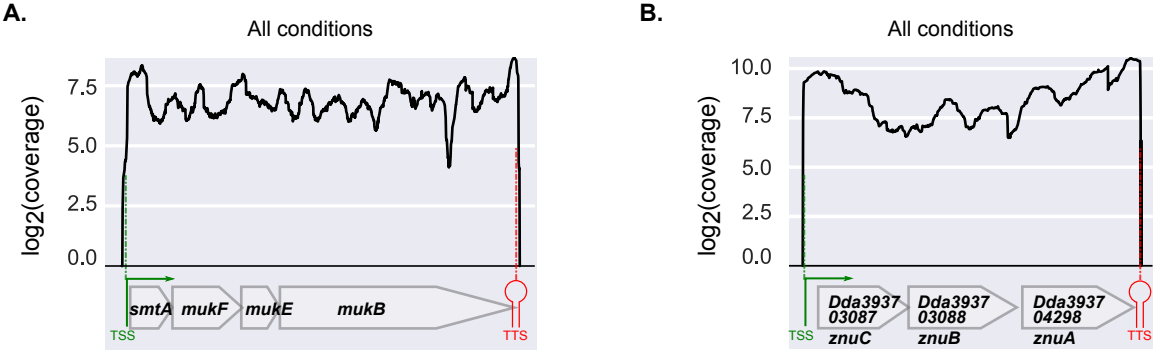

Figure 2. Many transcription units identified by our approach coincide with operons. (A) Example of a known operon (*smtAmukFEB*). (B) Identification of a new TU exhibiting uniform read coverage and strong internal cross-correlations (Fig. S1), clearly indicative of an operon. Its function was unknown but a homology analysis revealed that it corresponds to the cluster of genes *znuCBA*, a Zn^{2+} uptake system (43).

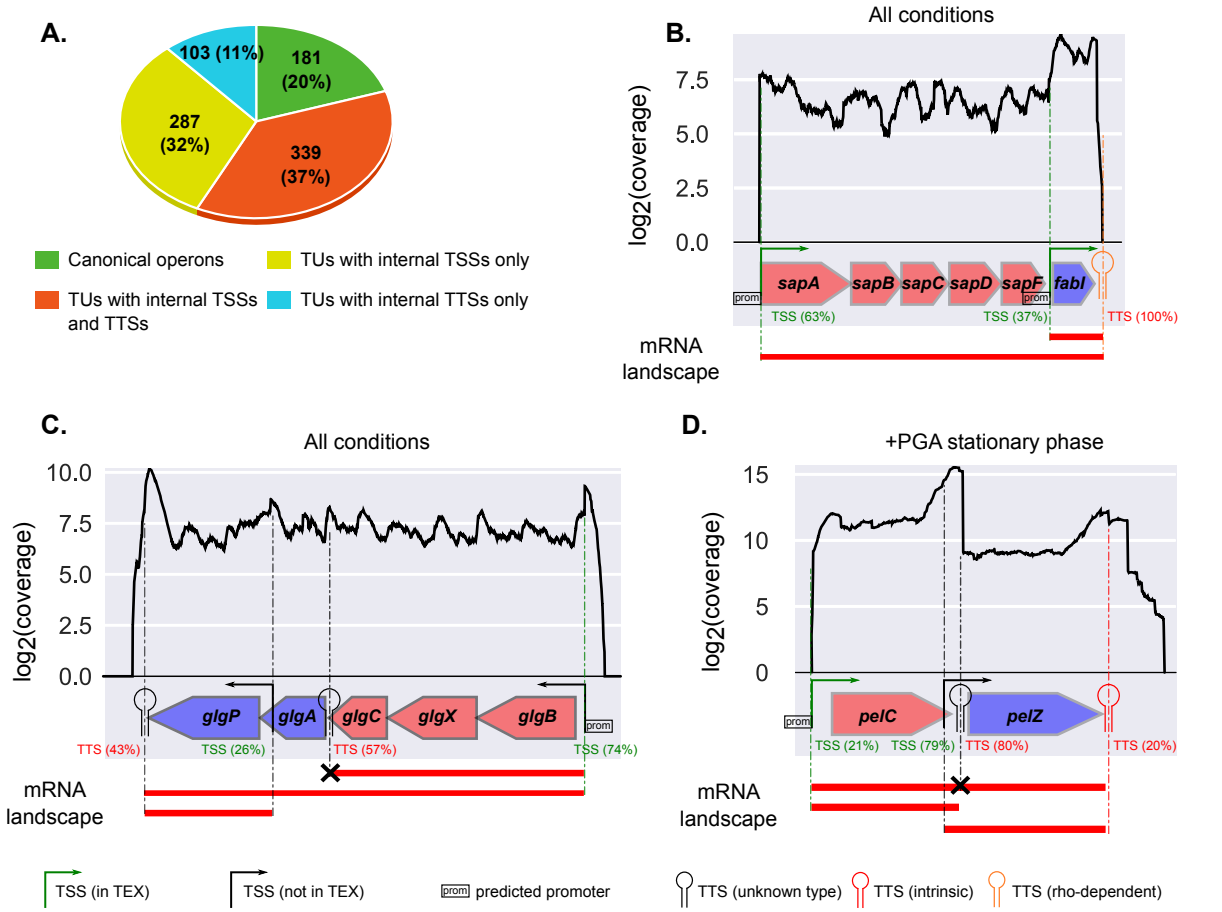

Figure 3. Most transcription units are complex with internal TSSs and/or TTSs that generate alternative transcripts. (A) Characteristics of TUs. (B) The *sapABCDF* and *fabI* genes, predicted by Rockhopper (36) as two separate red and blue operons respectively, were identified as a single TU, with a strong internal TSS (37% total magnitude) expressing *fabI* alone. (C) The *glg* genes were identified as a single TU (involving several isoforms) containing two separate predicted operons (blue and red genes), as suggested by the uniform read coverage and in line with results in *E. coli* (18, 20, 41). (D) Identification of the *pelCZ* TU with different isoforms depending on the condition, as previously determined (50). The two genes are split into different operons by Rockhopper. A strong internal TSS, followed by a strong TTS, contributes to the complexity of its expression (see text).

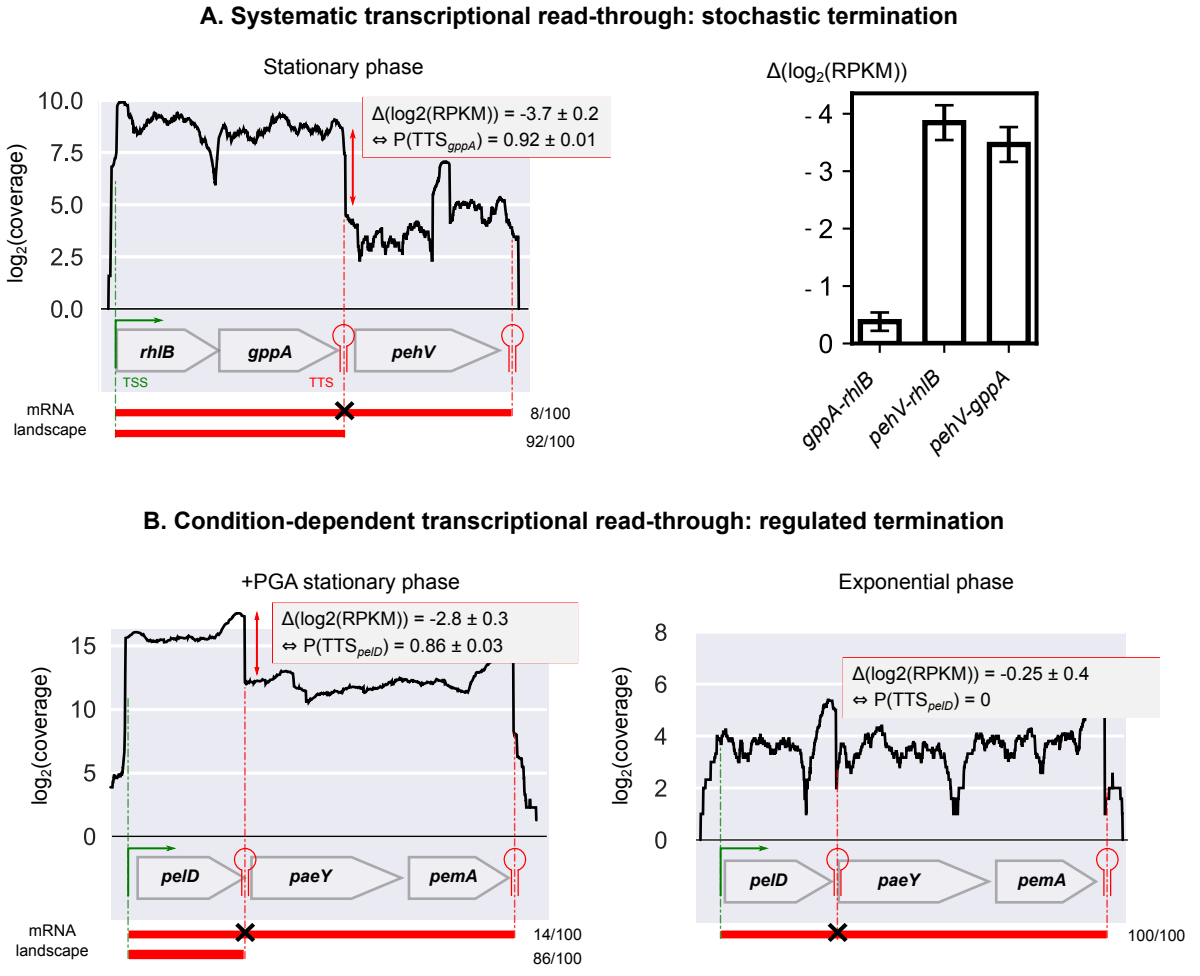

Figure 4. Quantification of transcriptional read-through. (A) Non-conditional read-through: example of the *rhlB-gppA-pehV* TU. The first two genes are homogeneously transcribed among conditions, resulting in an expression variation Δ(log2(RPKM)) close to 0 (95% confidence intervals are shown). The internal TTS downstream of *gppA* (TSS_{gppA}) is occasionally overstepped, resulting in two transcript isoforms. This partial read-through exhibits a constant rate across investigated conditions and results in a systematic and regular decrease of *pehV* expression compared to the other genes while their expression level changes by up to 50%. The probability of stochastic termination ($P(TSS_{gppA})$ = 0.92 ± 0.01) is inferred from this expression variation. (B) Condition-dependent read-through: example of the *pelD-paeY-pemA* TU. A TTS is identified downstream of *pelD* in agreement with previous studies (71). Its termination probability is regulated and depends on growth phase and presence of PGA (0.86 ± 0.03 vs 0), besides a global up- or down-regulation of the whole TU.

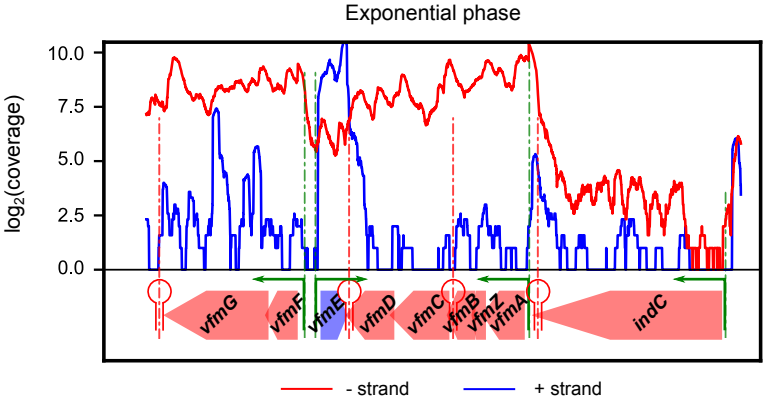

Figure 5. The *vfm* locus forms a potential noncontiguous transcription unit. The *vfmE* gene is transcribed in the opposite direction of the *indCvfmAZBCDFG* TU, generating two overlapping mRNAs that might be involved in a mutual regulation (see text).

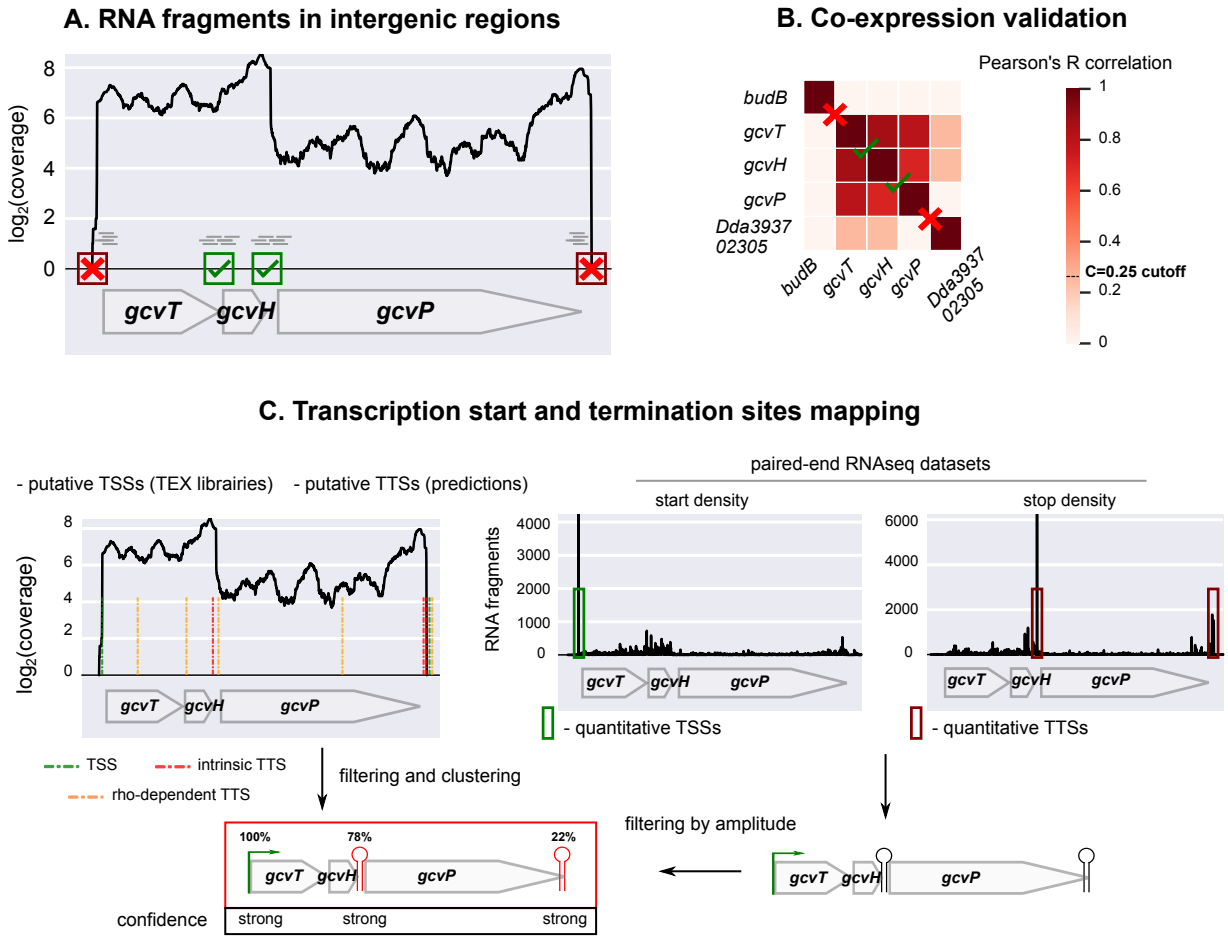

Figure 5. Algorithm for the characterization of *D. dadantii* transcription units. (A) Definition of putative TUs based on RNA-seq coverage (dataset 1) in intergenic regions, with isodirectional genes being split when the coverage drops to zero (here before *gcvT* and after *gcvP*). (B) Validation based on correlation of expression across 32 conditions (dataset 2). The genes of identified putative TUs are correlated, in contrast to surrounding isodirectional genes (*budB* and *Dda3937_02305*). (C) TSS and TTS mapping based on dRNA-seq (TEX libraries, dataset 3), TTS predictions, and paired-end RNA-seq data (dataset 1). First, putative TSSs of high resolution but qualitative strength were defined from an analysis of TEX libraries, and putative TTSs were predicted. Second, a list of TSSs and TTSs of quantitative strength (for the investigated experimental conditions) but poorer resolution was defined from the enrichment of RNA-seq paired-end fragment starts (start density) and stops (stop density) upstream of gene starts and downstream of gene stops, respectively. Third, only TSSs and TTSs with sufficient strength were retained, and compared to the closest (independently obtained) TEX TSS / predicted hairpin loop, in order to define their exact position and level of confidence. As a result of the analysis, this TU included the *gcvTHP* genes, the first two genes being expressed both as part of the entire transcript and as an independent transcript generated from a strong internal TTS (76% of total magnitude), explaining the lower correlation between *gcvP* and the remaining genes.