# SigmoID: a user-friendly tool for improving bacterial genome annotation through analysis of transcription control signals

Yevgeny Nikolaichik and Aliaksandr U. Damienikan

Department of Molecular Biology, Belarusian State University, Minsk, Belarus

## ABSTRACT

The majority of bacterial genome annotations are currently automated and based on a 'gene by gene' approach. Regulatory signals and operon structures are rarely taken into account which often results in incomplete and even incorrect gene function assignments. Here we present SigmoID, a cross-platform (OS X, Linux and Windows) open-source application aiming at simplifying the identification of transcription regulatory sites (promoters, transcription factor binding sites and terminators) in bacterial genomes and providing assistance in correcting annotations in accordance with regulatory information. SigmoID combines a user-friendly graphical interface to well known command line tools with a genome browser for visualising regulatory elements in genomic context. Integrated access to online databases with regulatory information (RegPrecise and RegulonDB) and web-based search engines speeds up genome analysis and simplifies correction of genome annotation. We demonstrate some features of SigmoID by constructing a series of regulatory protein binding site profiles for two groups of bacteria: Soft Rot *Enterobacteriaceae* (*Pectobacterium* and *Dickeya* spp.) and *Pseudomonas* spp. Furthermore, we inferred over 900 transcription factor binding sites and alternative sigma factor promoters in the annotated genome of *Pectobacterium atrosepticum*. These regulatory signals control putative transcription units covering about 40% of the *P. atrosepticum* chromosome. Reviewing the annotation in cases where it didn't fit with regulatory information allowed us to correct product and gene names for over 300 loci.

**Subjects** Bioinformatics, Genomics, Microbiology

**Keywords** Transcription factor binding site, Promoter, Terminator, Genome browser, Genome annotation, Sequence logo, *Pectobacterium atrosepticum*

Submitted 26 January 2016
Accepted 29 April 2016
Published 24 May 2016

Corresponding author
Yevgeny Nikolaichik,
nikolaichik@bio.bsu.by

Academic editor
Kay Nieselt

Additional Information and Declarations can be found on page 15

DOI 10.7717/peerj.2056

## INTRODUCTION

With bacterial genome sequencing becoming easily accessible, the demand for quality genome annotation has increased drastically in recent years. Manual genome annotation requires a lot of effort and is time-consuming. Therefore, it is usually beyond the scope of most sequencing projects. As a result, automated approaches are commonly used by researchers while submitting genome sequences to GenBank. The genome annotation pipelines such as Prokka, BASYS, RAST and NCBI Prokaryotic Genome Annotation Pipeline (Van Domselaar et al., 2005; Aziz et al., 2008; Tatusova et al., 2013;

Seemann, 2014) find positions and infer possible functions for most protein and rRNA/tRNA coding sequences, but they are less efficient at finding other structural elements of genome. In particular, none of the popular annotation pipelines cited above identifies regulatory elements such as promoters, transcriptional terminators and transcription factor binding sites (TFBS). However, regulatory elements provide important information about physiological conditions that influence gene activity. Therefore, information about the non-coding sequences of a gene might improve assignment of the gene's function.

Promoters/TFBS and terminators also mark operon borders. Genes within an operon are usually functionally connected and operon structures are normally conserved in related species. Hence, prediction of transcription termination sites, combined with prediction of operon beginnings (marked by promoters and/or TFBSs) gives a good idea about possible operon structure and can help to further improve the quality of functional annotation.

A lot of experimental data on gene regulation by various transcription factors and RNA polymerase sigma factors has been accumulated in recent years, especially on model organisms. A large portion of regulatory information is conveniently accessible from several databases like RegPrecise, RegulonDB and RegTransBase (*Novichkov et al., 2013*; *Gama-Castro et al., 2010*; *Cipriano et al., 2013*; *Kılıç et al., 2013*).

Many existing software packages handle the task of identifying regulatory protein binding sites in DNA sequences. These applications usually rely on binding site representation in the form of positional weight matrix and operate in pair with a tool generating these matrices in the required format. Since their first appearance decades ago, PatSer/Consensus (*Stormo et al., 1982*; *Hertz, Hartzell & Stormo, 1990*), ScanAce/AlignAce (*Hughes et al., 2000*) and MAST/MEME (*Bailey & Elkan, 1994*; *Bailey & Gribskov, 1998*) have been actively used for this purpose.

The vast majority of "traditional" TFBS identification packages do not take into account genome annotation when performing the searches and produce lists of the resulting hit coordinates in the plain text format. The end user then has to manually relate these hit coordinates to the gene/CDS positions to find the corresponding genes. This process becomes inconvenient, cumbersome and slow when many regulators have to be analysed on a whole genome scale. Some GUI tools, including GenomeExplorer (*Mironov, Vinokurova & Gel'fand, 2000*), RegPredict (*Novichkov et al., 2010b*) and Virtual Footprint (*Grote et al., 2009*) simplify this task by displaying the identified TFBSs in genomic context. However, to our knowledge neither of the available TFBS/promoter identification software tools can be used for editing genome annotation. On the other hand, genome browsers like Artemis (*Carver et al., 2012*) are specially designed for editing genome annotation, but none of these browsers can identify regulatory sequences. At the moment, improving genome annotation in accordance with regulatory information requires the use of separate tools for regulatory sequence identification and editing genome annotation, which makes the process laborious and slow. Hence, a tool automating this process would speed it up and improve its consistency and reproducibility.

We developed an application that combines a user-friendly graphical interface to command line tools for regulatory elements identification with a genome browser that has database search and genome editing abilities. Such a combination of features allows to identify annotation inconsistencies and provides an opportunity to correct errors more easily. Another purpose of developing SigmoID was to create a collection of TFBS/promoter search profiles for plant pathogens from the genus *Pectobacteria* and its close relatives from the genus *Dickeya* to facilitate the studies on plant-pathogen interactions involving this important group of pathogenic bacteria. As a step in this direction and to illustrate the ways SigmoID could be utilised, we have updated the annotation of the recently sequenced *Pectobacterium atrosepticum* genome (*Nikolaichik et al., 2014*) by including the information on regulatory sequences while correcting many annotation errors along the way.

# METHODS AND DATA

## Implementation

SigmoID front end is a GUI application written in Xojo (http://xojo.com) which gives it the expected look and usability on all three supported platforms (OS X, Linux and Windows). The GUI displays TFBS in the form of sequence logos (*Schneider & Stephens, 1990*) and provides an interface for three regulatory sequence search engines: nhmmer (*Wheeler & Eddy, 2013*) from the HMMER package (*Eddy, 2011*), MAST (*Bailey & Gribskov, 1998*) from the MEME Suite (*Bailey et al., 2015*) and TransTerm HP (*Kingsford, Ayanbule & Salzberg, 2007*). SigmoID has configuration windows (Fig. S1) for each of the included non-GUI tools. These windows allow to set search parameters and provide default values where possible. Processing of nhmmer, MAST and TransTerm HP outputs, sequence format conversions and addition of regulatory sites to genome annotations are implemented via Python scripts. The scripts are using Biopython modules (*Cock et al., 2009*), hence Python (2.7.x) and BioPython (version 1.64 and above) are required. The scripts are called from the GUI, but could be used separately and integrated into an annotation pipeline if desired. Detailed installation instructions are provided within distributions for each platform. The source code for the whole SigmoID application is available with GPL 3.0 licence (https://github.com/nikolaichik/SigmoID) and is also archived at Zenodo (http://dx.doi.org/10.5281/zenodo.48831). The compiled applications for all three platforms are also available from the same GitHub repository.

## Basic usage

Several scenarios of SigmoID usage are briefly described below while additional documentation is provided with the distribution.

### *Genome search for TFBS/promoters*

Genome search in SigmoID requires only a text file in the FASTA format with aligned sequences for a transcription regulator. The required data can be extracted from the literature or one of the several databases. SigmoID provides an integrated online access to the RegPrecise database (*Novichkov et al., 2013*) which has a large collection of data on transcription factors from a growing number of organisms. Online access is also provided to RegulonDB, the database containing information on most regulators characterised for *Escherichia coli* (*Salgado et al., 2013*). Genome search can also be performed with pre-calibrated TFBS/promoter profiles specific for a certain group of bacteria.

Pre-calibrated profiles can significantly speed up analysis of newly sequenced genomes of bacteria from the same or closely related taxa. The current version of SigmoID includes two sets of such TFBS/promoter profiles optimised for (i) Soft Rot *Enterobacteriaceae* (SRE) including *Pectobacterium* and *Dickeya* genera and (ii) *Pseudomonas* spp. Currently active profile set can be changed in SigmoID preferences by switching to another profile folder distributed with SigmoID or provided by user.

Once the TFBS/promoter data are loaded and displayed in the main window, one of the two algorithms (nhmmer or MAST) can be used for searching genome sequences in GenBank format. The search results are displayed in the main window in the plain text format. SigmoID can add newly discovered regulatory elements to the annotation. At this stage the search results are filtered to remove sites with the wrong orientation relatively to the closest ORFs, sites located downstream of converging ORFs, and (as an option) sites within ORFs. In addition, because nhmmer reports hits on both strands, one duplicate for the palindromic TFBSs is removed. This processing is done by python scripts (separate for each search algorithm) that can be called via the "Annotate Current Sites…" command. The output is saved in a separate GenBank file which is opened in the integrated genome browser to present filtered results visually in the genomic context. Search thresholds and filtering options can be set by the user via the configuration windows (Figs. S1A and S1B); calibrated profiles have all required settings pre-configured.

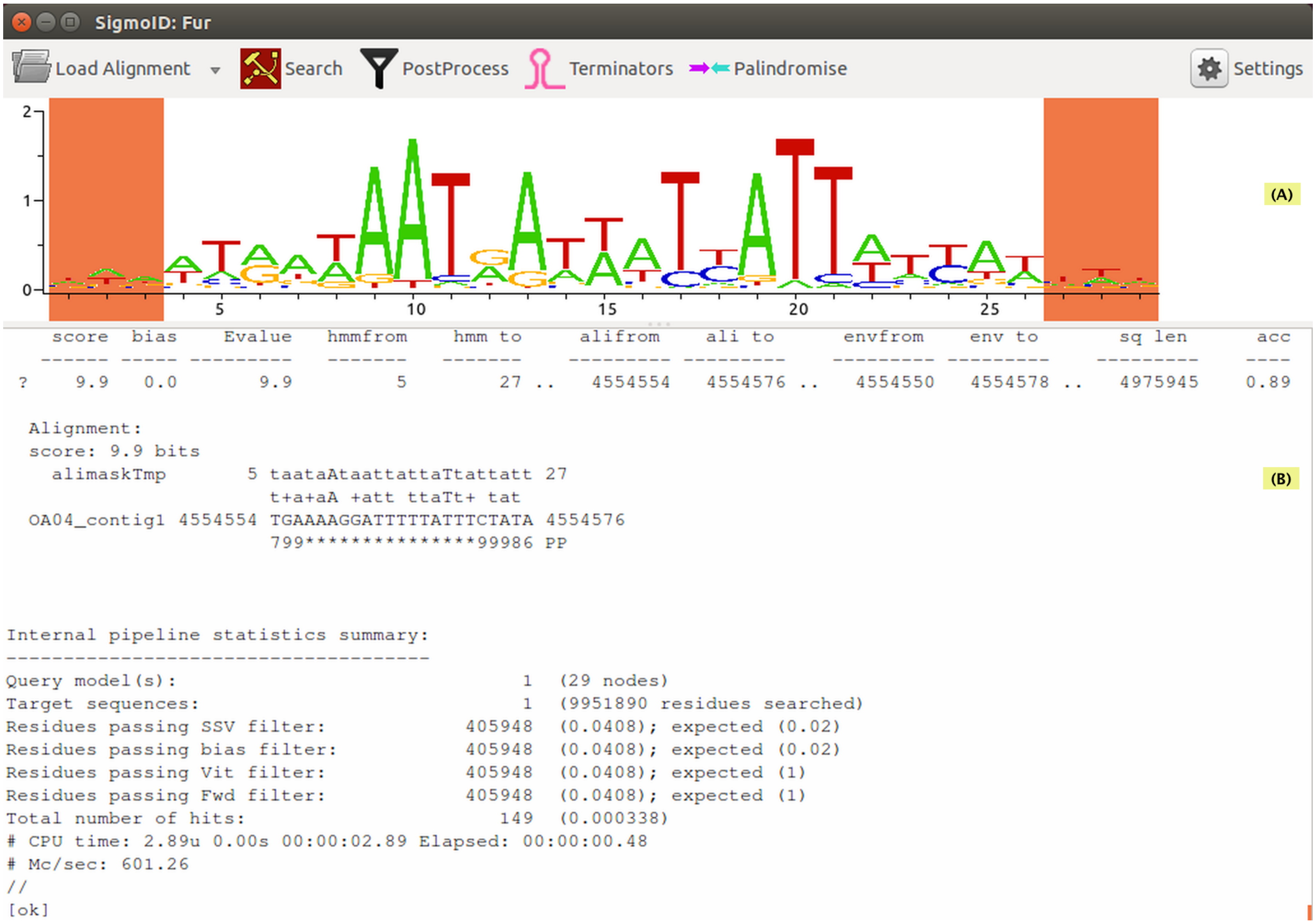
### Construction of optimised TFBS/promoter profiles

Utility functions within SigmoID are provided for extension/shortening/partial masking of an existing alignment of TFBS/promoter sequences. Displaying an alignment sequence logo helps to choose correct boundaries of the site (Fig. 1). Genome browser provides an opportunity to display nhmmer/MAST search hits in the genomic context; the newly discovered regulatory sites could then be copied or saved in a new file. Finally, a calibrated profile could be saved with the help of the Profile Wizard window (Fig. S1C). Calibration thresholds have to be set by the user, but choosing the right calibration values is simplified with a function that identifies the minimal score for individual sequences in the training set. The cut-off values are stored in the final calibrated hidden Markov model (HMM) profile produced by hmmbuild and are applied by default for further searches with the stored profile. For MAST searches, the profiles are converted into position specific scoring matrices and stored in the MEME format. For these matrices, a single $p$-value threshold should be set by the user. Alongside the calibrated profiles, SigmoID also stores the settings for running nhmmer/MAST and the post-processing scripts as well as the description of the profile.

### Transcription terminator search

This function provides a simple GUI for configuring TransTerm HP, performs necessary sequence format conversions and adds transcriptional terminator annotations to genome files in the GenBank format. The configuration window (Fig. S1D) allows setting TransTerm confidence score and terminator stem/loop parameters. The default values provided should be appropriate to use in most cases.

This function can also be called during the genome scan.



```
    score  bias    Evalue   hmmfrom    hmm to     alifrom    ali to      envfrom    env to       sq len      acc
   ------ ----- ---------   -------   -------   --------- ---------    --------- ---------    ---------     ----
 ?    9.9   0.0       9.9         5        27 ..   4554554   4554576 ..   4554550   4554578 ..   4975945     0.89

  Alignment:
  score: 9.9 bits
    alimaskTmp       5 taataAtaattattaTtattatt 27
                       t+a+aA +att ttaTt+ tat
  OA04_contig1 4554554 TGAAAAGGATTTTTATTTCTATA 4554576
                       799**************99986 PP



Internal pipeline statistics summary:
-------------------------------------
Query model(s):                            1  (29 nodes)
Target sequences:                          1  (9951890 residues searched)
Residues passing SSV filter:          405948  (0.0408); expected (0.02)
Residues passing bias filter:         405948  (0.0408); expected (0.02)
Residues passing Vit filter:          405948  (0.0408); expected (1)
Residues passing Fwd filter:          405948  (0.0408); expected (1)
Total number of hits:                    149  (0.000338)
# CPU time: 2.89u 0.00s 00:00:02.89 Elapsed: 00:00:00.48
# Mc/sec: 601.26
//
[ok]
```

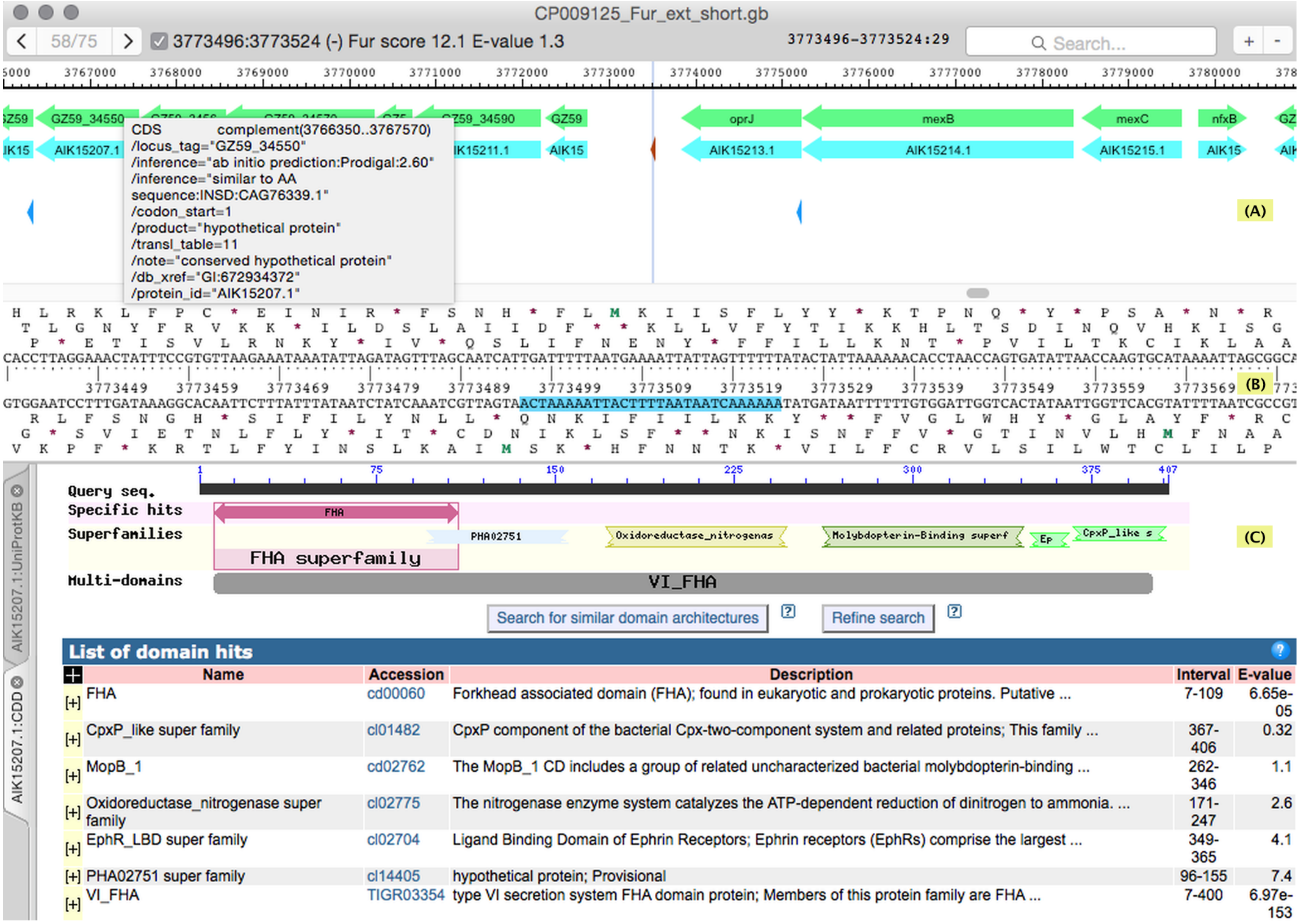
**Figure 1 The main SigmoID window.** The window is divided into two panes: (A) and (B). (A) shows the various pieces of information about the currently opened alignment of TFBS sequences: the sequence logo (on this figure), raw sequences, alignment description, profile data and search settings. Dragging across the logo selects part of the alignment which can be exported in FASTA format; shift-dragging can select several parts of the alignment. Starting a search with part(s) of the logo highlighted (as shown here) masks the highlighted positions with alimask and then performs a nhmmer search with the masked profile. (B) displays standard output from the search programmes and various information messages. The very end of nhmmer output is shown.

### Genome scan

The Genome Scan function searches a chosen genome TFBS/promoters with a number (or all) of the provided profiles (or calibrated profiles created by the user) and transcriptional terminators and adds corresponding annotations to the GenBank file. This function can only be used with the pre-calibrated HMM profiles. nhmmer is used as the search engine with this function and the corresponding post-processing script is run with the settings stored in the calibrated profiles.

### Editing genome annotation

Genome browser (Fig. 2) could be opened after a search to skim quickly through the newly found sites and inspect their genomic context. Running a terminator search beforehand



3773496:3773524 (-) Fur score 12.1 E-value 1.3

```
CDS            complement(3766350..3767570)
/locus_tag="GZ59_34550"
/inference="ab initio prediction:Prodigal:2.60"
/inference="similar to AA
sequence:INSD:CAG76339.1"
/codon_start=1
/product="hypothetical protein"
/transl_table=11
/note="conserved hypothetical protein"
/db_xref="GI:672934372"
/protein_id="AIK15207.1"
```

```
 H  L  R  K  L  F  P  C  *  E  I  N  I  R  *  F  S  N  H  *  F  L  M  K  I  I  S  F  L  Y  Y  *  K  T  P  N  Q  *  Y  *  P  S  A  *  N  *  R
  T  L  G  N  Y  F  R  V  K  K  *  I  L  D  S  L  A  I  I  D  F  *  *  K  L  L  V  F  Y  T  I  K  K  H  L  T  S  D  I  N  Q  V  H  K  I  S  G
   P  *  E  T  I  S  V  L  R  N  K  Y  *  I  V  *  Q  S  L  I  F  N  E  N  Y  *  F  F  I  L  L  K  N  T  *  P  V  I  L  T  K  C  I  K  L  A  A
CACCTTAGGAAACTATTTCCGTGTTAAGAAATAAATATTAGATAGTTTAGCAATCATTGATTTTTAATGAAAATTATTAGTTTTTTATACTATTAAAAAACACCTAACCAGTGATATTAACCAAGTGCATAAAATTAGCGGCA
        3773449   3773459   3773469   3773479   3773489   3773499   3773509   3773519   3773529   3773539   3773549   3773559   3773569   773
GTGGAATCCTTTGATAAAGGCACAATTCTTTATTTATAATCTATCAAATCGTTAGTAACTAAAAATTACTTTTAATAATCAAAAAATATGATAATTTTTTGTGGATTGGTCACTATAATTGGTTCACGTATTTTAATCGCCGT
   R  L  F  S  N  G  H  *  S  I  F  I  L  Y  N  L  L  *  Q  N  K  I  F  I  I  L  K  K  Y  *  *  F  V  G  L  W  H  Y  *  G  L  A  Y  F  *  R  C
  G  *  S  V  I  E  T  N  L  F  L  Y  *  I  T  *  C  D  N  I  K  L  S  F  *  *  N  K  I  S  N  F  F  V  *  G  T  I  N  V  L  H  M  F  N  A  A
 V  K  P  F  *  K  R  T  L  F  Y  I  N  S  L  K  A  I  M  S  K  *  H  F  N  N  T  K  *  V  I  L  F  C  R  V  L  S  I  L  W  T  C  L  I  L  P
```

List of domain hits

| | Name | Accession | Description | Interval | E-value |
|---|---|---|---|---|---|
| [+] | FHA | cd00060 | Forkhead associated domain (FHA); found in eukaryotic and prokaryotic proteins. Putative ... | 7-109 | 6.65e-05 |
| [+] | CpxP_like super family | cl01482 | CpxP component of the bacterial Cpx-two-component system and related proteins; This family ... | 367-406 | 0.32 |
| [+] | MopB_1 | cd02762 | The MopB_1 CD includes a group of related uncharacterized bacterial molybdopterin-binding ... | 262-346 | 1.1 |
| [+] | Oxidoreductase_nitrogenase super family | cl02775 | The nitrogenase enzyme system catalyzes the ATP-dependent reduction of dinitrogen to ammonia. ... | 171-247 | 2.6 |
| [+] | EphR_LBD super family | cl02704 | Ligand Binding Domain of Ephrin Receptors; Ephrin receptors (EphRs) comprise the largest ... | 349-365 | 4.1 |
| [+] | PHA02751 super family | cl14405 | hypothetical protein; Provisional | 96-155 | 7.4 |
| [+] | VI_FHA | TIGR03354 | type VI secretion system FHA domain protein; Members of this protein family are FHA ... | 7-400 | 6.97e-153 |

**Figure 2 The genome browser window.** The window is divided into three panes: (A), (B), (C). (A) shows the map of the annotated features, (B)—nucleotide/amino acid sequence of the currently highlighted part of the feature map, and the bottom one—the results of database searches. (C) is a tabbed web browser with navigation commands available via contextual menus. This allows to investigate the details of a particular search hit quickly. Hovering the mouse pointer over a feature in the top pane pops up a small window displaying its current annotation (shown). Clicking on a feature selects it, right-clicking displays a contextual menu allowing to launch a number of web-based searches with a selected feature as a query; the menu also has commands for feature copying and editing. The toolbar on top of the window contains a navigator control (on the left) allowing to scroll the sequence display from one hit to another. The information on the currently selected hit is shown to the right of the navigator. The search field distinguishes three query types automatically: nucleotide sequence, coordinate and annotation text. The rightmost control allows to zoom the feature map in or out. This example shows the browser window in the middle of analysis of a search with Fur binding site profile. The currently selected (and highlighted in the top and middle panes) Fur binding site is located in front of a large operon on the bottom strand, most of the frames in which were originally annotated as "hypothetical protein." Searching the Uniprot and especially the Conserved Domain Database (the actual result of CDD search with one protein coded by this operon is shown) revealed that all these frames code for components of the type VI secretion system, which is known to be controlled by Fur.

might be helpful to delimit putative operons more clearly. Alternatively, the browser could be used to view an existing GenBank file independently of any search function. SigmoID allows editing of the existing features and their qualifiers and provides integrated web-based hmmsearch, hmmscan, blastn and blastp search functions for verifying the existing annotation. The search is possible against protein, nucleotide and protein domain databases. The access to these functions is provided via the corresponding APIs (*NCBI, 2016*; *Finn, Clements & Eddy, 2011*). The genome browser can also be used to search for a sequence, a coordinate or a feature text within a genome file.

### *Extracting regulon information from the annotated genome*

The function "List regulons" allows listing putative transcription units controlled by either all of the regulators or by just one of them. An annotated TFBS or a promoter is considered by this function as the start of a putative transcription unit. The configuration window (Fig. S1E) allows the user to set the allowed interval for the location of promoter/TFBS relative to the first gene of a putative operon. Two cases are considered being an operon end: a non-coding gap of the specified (and user-configurable) size and a transcription terminator. The divergently transcribed neighbouring operons which might be controlled by the same TFBS could be listed as a single divergon unit. The function also outputs the counts of putative transcription units controlled by each regulator.

### *Data plots*

SigmoID can display simple plots of genome-related data. The "Add Plot…" command from the "Genome" menu loads the data from text files that have tab-separated pairs of genome coordinate and data value (one pair per line). It can also load the data produced by the samtools depth command (*Li, 2011*). Up to four overlapping plots can be added. This, among other possible options, allows to simultaneously display read coverage for each strand of two samples from a strand-specific RNA-Seq experiment.

RNA-Seq coverage plots could be helpful for verifying positions of regulatory sequences relatively to transcription start sites. At the moment, SigmoID does not include all functions required and can only load and display read count values. These can be produced in different ways, for example using Bowtie 2 (*Langmead & Salzberg, 2012*) and samtools as described in detail in the SigmoID help file.

## TFBS, promoter and genome data

The RegPrecise database (*Novichkov et al., 2013*) version 3 was used as the main source of data for TFBSs for different transcription factors. An integrated access to this database is provided within SigmoID via the web services API (*Novichkov et al., 2012*). The data for the regulators not present in RegPrecise (which includes all promoters) were extracted either from original publications or from RegulonDB (*Salgado et al., 2013*).

For the construction of optimised search profiles, the complete genome sequences of *Pectobacterium atrosepticum* SCRI1043 (*Bell et al., 2004*), *Pectobacterium carotovorum* PCC21 (*Park et al., 2012*), *Pectobacterium wasabiae* SCC3193 (*Koskinen et al., 2012*) and *Dickeya dadantii* 3937 (*Glasner et al., 2011*) were downloaded from the GenBank database (*Clark et al., 2016*). They were used as the main reference sequences for the soft rot *Enterobacteriaceae* group. For some profiles with few binding sites, additional genome sequences were employed (accession numbers NC_012917, CP001790, CP001836, CP001654, CP007744). Verification and calibration of *Pseudomonas*-optimised profiles were done with the genome sequences of strains referenced by RegPrecise (accession numbers AE016853, CP000076, AE004091, CT573326, CP003677, AE015451, CP000680).

# RESULTS AND DISCUSSION

## Search engine integration

Two different approaches have been routinely utilised for representation and identification of regulatory protein binding sites. Position specific scoring matrices (PSSM) are traditionally used for representing both TFBSs and promoters (*Stormo, 2000*). Profile hidden Markov models (HMM) have been often used for promoter identification (*Yada et al., 1997*; *Pedersen et al., 1996*; *Fouts et al., 2002*), but rarely applied to TFBS searches. Since we aimed at both promoter and TFBS identification, we looked for the available tools implementing both approaches. Among the open-source cross-platform tools we have chosen the HMMER and MEME Suite packages as freely available and widely used (*Eddy, 2011*; *Bailey et al., 2015*). We have found HMM-based and PSSM-based search (as implemented in nhmmer and MAST) to perform similarly in terms of specificity, sensitivity and speed (data not shown). Nevertheless, there were minor differences between the two search algorithms in detecting "weak" binding sites (with no clear advantage shown by either algorithm), therefore both nhmmer and MAST are provided as search engines within SigmoID. Subsequent filtering of search results and annotation additions can be done regardless of the search algorithm.

A similar approach has been taken for terminator identification. The actual search is performed by TransTerm HP; SigmoID implements the required sequence format conversions and adds terminator annotations to genome file in GenBank format.

The results produced by all three search engines are displayed in genome browser (Fig. 2).

## Using SigmoID to construct optimised HMM profiles

Experimental data on TF/sigma factor binding sites are available for a small fraction of the sequenced genomes. As this sort of experiments is often technically challenging and time-consuming, bioinformatics could be used to infer possible regulatory networks in related species using already available experimental data. Therefore, we aimed at collecting the regulatory information, construction of the optimised TFBS/promoter search profiles and making them easily accessible with the help of a user-friendly tool.

Before proceeding with profile construction, we verified the presence of the corresponding regulatory gene in each genome by FASTA search (*Pearson, 2004*). General profile optimisation algorithm was similar to the one used by RegPrecise (*Novichkov et al., 2010a*), but utilised nhmmer as the main search engine. For the purpose of this work, the TFBS/promoter sequences extracted from the original data source were used to build a primary HMM profile and search for similar sites in the genomes of related bacterial species. Orthologous sites found by nhmmer were always retained. Contrary to the conservative approach utilised by RegPrecise for automated regulon propagation (*Novichkov et al., 2010a*), additional conserved high scoring sites were considered and retained if located in front of the genes known or likely to be involved in the same physiological process as the characterised members of the regulog. This procedure resulted in secondary alignment composed exclusively of target genomes sequences. At this point, we evaluated the possibility to improve signal/noise ratio by verifying the boundaries of the alignments. To do this, we applied the "Extend binding sites" function of SigmoID to add bases on either side of each sequence in the current profile, and, if additional conserved bases were discovered beyond the original borders, saved the new extended alignment (via the "Save alignment selection" function). Once the profile boundaries were determined, the target genomes were re-scanned and additional sites, if found, were added to the alignment. Finally, the profiles were calibrated to simplify their further use. In most cases the lowest nhmmer score of a specific hit matching the original data source was used for the trusted cut-off value, the highest score of a nonspecific hit—for the noise cut-off and the mean of these two values—for the gathering cut-off. The gathering cut-off is used by default in further searches with the calibrated profile.

## Calibrated profiles included with SigmoID

SigmoID includes sets of calibrated HMM profiles for two groups of bacteria. The SRE group includes plant pathogens from the *Pectobacterium* and *Dickeya* genera. Genome sequences of the two strains from this group, *Pectobacterium atrosepticum* SCRI1043 and *Dickeya dadantii* 3937, have a high-quality curated annotation (*Bell et al., 2004*; *Glasner et al., 2011*) and were used throughout this work for search and extraction of the TFBS data. To better cover sequence diversity within this group of plant pathogens, genome sequences from representatives of the other two species, *P. carotovorum* and *P. wasabiae*, were also included.

A number of transcription factors and their binding sites have been experimentally characterised in *Dickeya dadantii* while *Pectobacterium spp*. is less studied in this respect. On the other hand, the RegPrecise database (*Novichkov et al., 2013*) contains the information on TFBS for many important regulators in *P. atrosepticum* SCRI1043, which allowed us to build the corresponding search profiles. These profiles, however, performed sub-optimally when applied to other *Pectobacterium* species and especially *Dickeya*. Therefore, we have adjusted some of the profiles to perform more or less uniformly for the whole SRE group, as described above.

The profile optimisation procedure resulted in changing TFBS/promoter boundaries for the majority of profiles (27 out of 35) and more often than not increased their information content (Table 1, corresponding sequence logos are shown in Fig. S2). The cases where the information content decreased were usually the result of a small number of sequences in the original data source. Despite of the decreased information content, these profiles represent the corresponding regulog better due to the inclusion of TFBS/promoters controlling divergent regulon members.

The second set of HMM profiles was constructed for *Pseudomonas* spp. This genus is represented in RegPrecise by seven species with significant differences in genome organisation and includes plant and animal associated species, both pathogenic and non-pathogenic. Most of the characterised TFs and their binding sites appeared to be well conserved across the genus, which in many cases allowed direct use of the data present in RegPrecise for TFBS identification in *Pseudomonas* genomes. This approach appeared to be possible for the majority of TFs (25 out of 30). For only two of the TFs with data in RegPrecise (RutR and Zur) a universal profile wasn't feasible due to significant

**Table 1** Characteristics of the calibrated SRE-specific profiles included with SigmoID.[a]

| Regulatory protein | Binding site length/ information content | | Number of binding sites used for profile construction | Source of the original binding site data |
|---|---|---|---|---|
| | Original | Final | | |
| ArcA | 15/16.9 | 19/21.0 | 33 | *Ravcheev (2009)* |
| ArgR | 18/17.8 | 24/23.2 | 24 | RegPrecise |
| AscG | 20/30.4 | 20/23.3 | 7 | RegPrecise |
| BaeR | 20/26.2 | 27/30.4 | 13 | *Nishino, Honda & Yamaguchi (2005)* |
| CpxR | 15/15.0 | 15/20.3 | 16 | RegulonDB |
| CRP | 16/10.2 | 24/15.8 | 100 | RegPrecise |
| FliA | 27/23.5 | 26/23.5 | 96 | *Ide, Ikebe & Kutsukake (1999)* |
| FNR | 14/12.4 | 22/20.7 | 34 | RegulonDB |
| FruR | 16/17.8 | 18/19.4 | 25 | RegPrecise |
| Fur | 19/20.2 | 23/18.8 | 98 | RegPrecise |
| GalR | 20/24.0 | 18/19.5 | 39 | RegPrecise |
| HrpL | 28/19.0 | 31/25.9 | 63 | *McNally et al. (2011)* |
| KdgR | 21/22.1 | 25/22.4 | 85 | *Rodionov (2004)* |
| LexA | 20/20.3 | 20/21.2 | 34 | RegPrecise |
| MetJ | 16/20.2 | 24/23.9 | 21 | RegPrecise |
| MetR | 15/17.3 | 21/19.2 | 18 | RegPrecise |
| NarP | 16/17.9 | 18/17.6 | 54 | RegPrecise |
| NifA | 16/14.6 | 18/19.5 | 18 | RegPrecise |
| NsrR | 19/27.9 | 19/23.2 | 20 | RegPrecise |
| NtrC | 17/22.0 | 17/17.8 | 56 | RegPrecise |
| OxyR | 17/13.3 | 15/18.1 | 19 | *Seo et al. (2015)* |
| PhoP1 | 19/24.4 | 16/20.4 | 24 | *Harari et al. (2010)* |
| PhoP2 | 19/16.4 | 16/20.3 | 31 | *Harari et al. (2010)* |
| PmrA | 20/23.1 | 19/22.5 | 19 | *Ogasawara et al. (2012)* |
| PurR | 16/18.4 | 16/17.6 | 23 | RegPrecise |
| RcsB | 14/19.6 | 18/17.4 | 39 | *Andresen et al. (2010)* |
| RhaS | 51/42.7 | 51/40.2 | 21 | RegPrecise |
| Rob | 17/15.6 | 20/22.7 | 45 | RegulonDB |
| RpoE | 27/18.4 | 30/20.3 | 57 | *Rhodius et al. (2006)* |
| RpoH1 | 28/27.5 | 33/29.6 | 20 | *Nonaka et al. (2006)* |
| RpoH2 | 29/24.7 | 30/30.8 | 19 | *Nonaka et al. (2006)* |
| RpoN | 16/14.7 | 16/21.8 | 105 | *Barrios, Valderrama & Morett (1999)* |
| RpoS | 26/22.8 | 35/22.9 | 25 | RegulonDB |
| SlyA | 12/10.7 | 20/19.2 | 55 | *Stapleton et al. (2002)*, *Haque et al. (2009)* and *McVicker et al. (2011)* |
| UxuR | 18/27.6 | 22/23.7 | 17 | RegPrecise |

**Notes.**

[a] The table does not include the data on additional 9 profiles for TFs controlling just one or very few operons. These profiles are still distributed with SigmoID.

interspecies variation in regulon content and binding site composition. In this case profiles for plant-associated species were constructed.

Six *Pseudomonas*-specific profiles are based on the data not present in RegPrecise. The profiles for the virulence regulator RhpR as well as for the HrpL- and PvdS-dependent promoters were derived from experimentally determined binding sites (*Deng et al., 2010*; *Deng et al., 2014*; *Ferreira et al., 2006*; *Swingle et al., 2008*). FliA-, RpoN- and RpoH-dependent promoters appeared to be sufficiently conserved between *Enterobacteriaceae* and *Pseudomonadaceae*, which allowed us to create the corresponding promoter profiles via the standard optimisation procedure described above with SRE profiles as the starting point.

The resulting profile collection distributed with SigmoID includes 37 *Pseudomonas*-specific profiles (Table S1).

## TFBS/promoter position verification with RNA-Seq data

SigmoID can simultaneously display regulatory elements and RNA-Seq coverage graphs which helps to verify TFBS/promoter positions relative to transcription start sites. At the time of writing, one large-scale RNA-Seq experimental dataset was available from public databases for the bacterial genera discussed in this work (*Kwenda et al., 2016*). The experiment had samples prepared from *P. atrosepticum* cultures grown to stationary phase in two media. Since the difference between the samples is multifactorial, this dataset could only have limited use for TFBS/promoter verification. However, one of the alternative sigma factors, RpoS, is strongly expressed in both samples, which allowed us to define RpoS-dependent promoter profile for this bacterium.

To our knowledge, the RpoS regulon has not been characterised in SRE species. Because of this, we had difficulty constructing the HMM profile for RpoS-dependent promoters in SRE starting with *E. coli* data. nhmmer search scores were low and high uncertainty with the potential RpoS regulon members complicated selecting true hits for profile construction. The RNA-Seq data allowed us to select nhmmer hits correctly located in front of transcription start sites of the genes expressed in the stationary phase (Fig. 3A) and to construct the HMM profile for RpoS-dependent promoters in SRE. This profile has identified 24 transcriptional units as potentially controlled by RpoS in *P. atrosepticum* (Table S2). Sequence logo of their promoters (Fig. 3B) shows features typical of the RpoS-dependent promoters characterised recently for *E. coli* by ChIP-Seq analysis (*Peano et al., 2015*).

## Applying SigmoID to improving the annotation of *Pectobacterium atrosepticum* 21A genome

We have recently sequenced the genome of *P. atrosepticum* strain 21A (*Nikolaichik et al., 2014*). The version of this genome submitted to GenBank had mostly automated annotation produced by using the Prokka pipeline (*Seemann, 2014*). While testing SigmoID on this genome sequence, we noticed some inconsistencies in the annotation (known genes missing from or unlikely genes present in regulons). Therefore, we tested if SigmoID could aid to systematically improve functional annotation of this genome using all the available regulatory information. Putative terminators in this genome

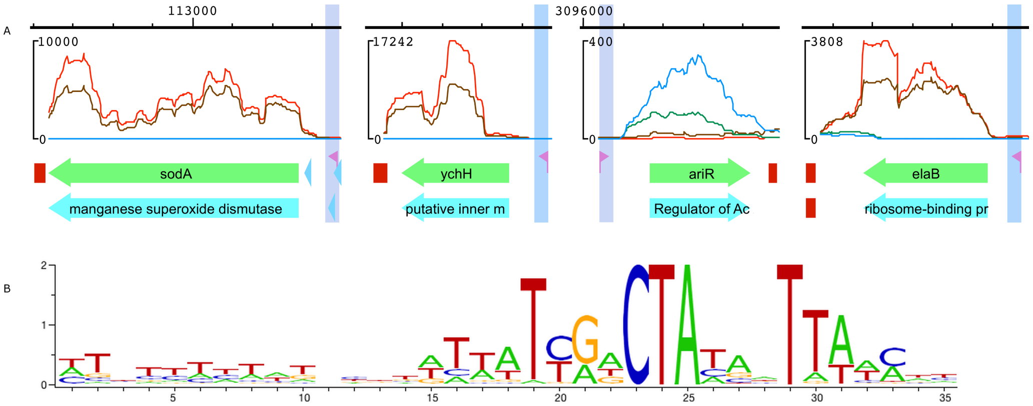


**Figure 3 RpoS promoter verification with RNA-Seq data.** (A) Putative RpoS-dependent promoter positions relative to selected genes expressed in stationary phase. RNA-Seq reads from GEO accession GSE68547 were mapped onto *P. atrosepticum* 21A chromosome with bowtie 2 resulting in over 95% of uniquely mapped reads. Read counts were calculated with samtools, displayed by SigmoID and resulting pictures were exported as vector images. RpoS-dependent promoter is highlighted and marked by a flag sign. The counts for reads originating from the culture grown in minimal medium are represented by a blue line for the sense strand and by a brown line for the antisense strand. The counts for reads originating from the rich medium grown culture are represented by a green line for the sense strand and by a red line for the antisense strand. (B) Sequence logo of the RpoS-dependent promoter in *P. atrosepticum*.

were identified first, followed by sequential scanning of the genome using (i) all the calibrated profiles described above and (ii) uncalibrated profiles for other *P. atrosepticum* regulators present in RegPrecise. This enabled us to delimit putative transcription units controlled by known regulators. The annotation of CDS within these transcription units was verified (and edited if necessary) according to the regulatory information. At this stage the database search abilities interfaced by SigmoID within the genome browser (Fig. 2) were used extensively.

It should be noted that two independent automated annotations are available for the *P. atrosepticum* 21A chromosome, produced by the Prokka and NCBI pipelines (accession numbers CP009125.1 and NZ_CP009125.1). Both automated annotations contain errors, some of which could be easily corrected if regulatory information was taken into account. For example, three neighbouring open reading frames (locus tags GZ59_45330, GZ59_45340 and GZ59_45350) are annotated in NZ_CP009125.1 as "molecular chaperone TorD," "cytochrome C biogenesis protein CcmF" and "cytochrome C biogenesis protein CcmE." However, the inference of a CRP site in front of the first gene suggests that the operon might be related to the catabolism of secondary carbon sources. The presence of the AscG binding site near the CRP site gives a hint as to what such a source could be, as AscG is a specialised regulator of beta-glucoside utilisation genes (*Ishida, Kori & Ishihama, 2009*). Searching the Conserved Domain Database shows a strong similarity of the protein products coded for by the GZ59_45330, GZ59_45340 and GZ59_45350 loci to PTS components. The results of NCBI non-redundant protein database searches, however, contain a mixture of PTS components and unrelated TorD/CcmF/CcmE hits. A direct check of homology between GZ59_45330/GZ59_45340/GZ59_45350 and experimentally characterised TorD/CcmF/CcmE from *E. coli* shows no significant similarity. Further analysis of NZ_CP009125.1 revealed that TorD and CcmF were present in the annotation four times and CcmE—three times, and in all cases except two (one copy each of CcmE and CcmF) these annotations were incorrect: these frames should be annotated as cellobiose-specific PTS system subunits IIA, IIC and IIB.

Although most often the errors were only present in one of the annotations, in some cases both annotations were either incorrect or significantly incomplete. Table S3 gives some of the examples where these errors were noticed and corrected with the regulatory information taken into account.

Once the *P. atosepticum* 21A genome annotation has been updated, we have used the "List regulons" function of SigmoID for the final check and resolved few issues with missing or inconsistent annotation within putative transcription units.

The updated annotation of *P. atrosepticum* 21A genome (accession number CP009125.1) contains about three hundred changes in gene/CDS features introduced in accordance with regulatory information (Table 2). Most of the changes replace a "hypothetical protein" with a more meaningful product name. Other changes, however, improve incomplete annotation or replace previous versions entirely.

Overall, the SigmoID-assisted review of *P. atosepticum* 21A genome annotation adds 930 transcription/sigma factor binding sites for 74 regulatory proteins and over 1,000 transcriptional terminators. A total of 592 putative transcription units with identified regulator binding sites suggestive of their regulation and physiological role (Data S1) include approximately 40 percent of the genes in the *P. atosepticum* 21A chromosome.

## CONCLUSIONS

The growing accessibility of genomic sequencing is opening up genome sequence analysis to more and more bench scientists, and not all of them are comfortable with the command line tools. SigmoID was designed to speed up extraction of regulatory information from genomic sequences and databases, present it visually and to provide an easy and convenient access to important command line search tools for a larger number of biologists.

Binding of a regulatory protein to DNA is rarely studied in more than one or very few species. DNA binding sites of orthologous proteins may diverge significantly, sometimes even between the genera of the same family. Hence, a TFBS profile built with the data for one taxonomic group will perform sub-optimally in a more or less distant species that has orthologous regulator. SigmoID provides helpful utilities for optimising TFBS/promoter profiles for a new taxonomic group. This allows to get a better understanding of transcriptional regulation in a wider range of bacteria.

Graphical display of regulatory information in genomic context suggests possible operon structure/regulation and highlights annotation problems while database search and annotation editing possibilities of SigmoID in many cases allow correcting these problems. Therefore, we hope that SigmoID might be a useful complement to automated annotation pipelines at the final stages of genome annotation.

There are other potential uses for SigmoID that are beyond the scope of this paper. The current SigmoID distribution contains two large collections of optimised TFBS/promoter

**Table 2** Summary of the updates made to the annotation of *P. atrosepticum* 21A genome by using SigmoID.

| Regulatory protein | Binding sites found | Edited downstream loci | | |
|---|---|---|---|---|
| | | CDS feature changed | Gene name changed | Total edited loci |
| ArcA | 31 | 5 | 4 | 8 |
| AscG | 4 | 3 | 6 | 7 |
| BaeR | 3 | 1 | 1 | 1 |
| CpxR | 35 | 5 | 9 | 9 |
| CRP | 110 | 28 | 31 | 44 |
| DcuR | 10 | 1 | 3 | 3 |
| FliA | 33 | 8 | 7 | 8 |
| FNR | 92 | 23 | 24 | 29 |
| FruR | 14 | 5 | 5 | 5 |
| Fur | 85 | 45 | 36 | 50 |
| GalR | 10 | 3 | 0 | 3 |
| HrpL | 6 | 5 | 3 | 5 |
| KdgR | 43 | 10 | 8 | 13 |
| LexA | 20 | 1 | 4 | 4 |
| MetJ | 14 | 4 | 6 | 6 |
| NarP | 22 | 2 | 5 | 5 |
| NsrR | 12 | 1 | 1 | 1 |
| OxyR | 13 | 3 | 1 | 3 |
| PhoB | 10 | 0 | 2 | 2 |
| PhoP | 24 | 3 | 6 | 6 |
| PurR | 12 | 1 | 2 | 2 |
| RcsB | 24 | 4 | 6 | 6 |
| Rob | 25 | 13 | 8 | 13 |
| RpoE | 19 | 6 | 2 | 8 |
| RpoH | 9 | 0 | 3 | 3 |
| RpoN | 38 | 32 | 22 | 34 |
| RpoS | 24 | 9 | 9 | 10 |
| SlyA | 22 | 13 | 11 | 14 |
| UxuR | 5 | 0 | 4 | 4 |
| Zur | 4 | 3 | 1 | 3 |

profiles that may be useful in studies of transcriptional regulation in SRE and *Pseudomonas*. Importantly, this work resulted in constructing optimised profiles for 41 transcriptional regulator for *Pectobacterium* genus. Some of these regulators have little or no experimental data in pectobacteria (ArcA, BaeR, CpxR, OxyR, PhoB, PhoP, PmrA, RpoS, Rob, SlyA). We also propose profiles for three regulators (HrpS, VasH and RtcR) with unknown binding sites. A preliminary assessment of the inferred regulons controlled by Rob, SlyA, PhoP and other regulators implies their importance for pathogenicity and should facilitate experimental characterisation of their role in host-pathogen interaction.

## ACKNOWLEDGEMENTS

We thank Alexander Nikolaichik for advice on compiling the HMMER package for Windows and Sveta Makovets for critical reading of the manuscript and helpful suggestions.

## ADDITIONAL INFORMATION AND DECLARATIONS

### Funding

This work was supported in part by the State Research Programme "Biotechnology" within projects 2.52 and 2.24. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Grant Disclosures

The following grant information was disclosed by the authors:
The State Research Programme "Biotechnology".

### Competing Interests

The authors declare there are no competing interests.

### Author Contributions

- Yevgeny Nikolaichik conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Aliaksandr U. Damienikan performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, reviewed drafts of the paper.

### DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

GenBank accession number CP009125.1.

### Data Availability

The following information was supplied regarding data availability:

Data is available at Github:
https://github.com/nikolaichik/SigmoID.

### Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.2056#supplemental-information.

## REFERENCES

**Andresen L, Sala E, Koiv V, Mae A. 2010.** A role for the Rcs phosphorelay in regulating expression of plant cell wall degrading enzymes in *Pectobacterium carotovorum* subsp. carotovorum. *Microbiology* **156**:1323–1334 DOI 10.1099/mic.0.033936-0.

**Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008.** The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**:75 DOI 10.1186/1471-2164-9-75.

**Bailey TL, Elkan C. 1994.** Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* **2**:28–36.

**Bailey TL, Gribskov M. 1998.** Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**:48–54 DOI 10.1093/bioinformatics/14.1.48.

**Bailey TL, Johnson J, Grant CE, Noble WS. 2015.** The MEME suite. *Nucleic Acids Research* **43**:W39–W49 DOI 10.1093/nar/gkv416.

**Barrios H, Valderrama B, Morett E. 1999.** Compilation and analysis of sigma(54)-dependent promoter sequences. *Nucleic Acids Research* **27**:4305–4313 DOI 10.1093/nar/27.22.4305.

**Bell KS, Sebaihia M, Pritchard L, Holden MTG, Hyman LJ, Holeva MC, Thomson NR, Bentley SD, Churcher LJC, Mungall K, Atkin R, Bason N, Brooks K, Chillingworth T, Clark K, Doggett J, Fraser A, Hance Z, Hauser H, Jagels K, Moule S, Norbertczak H, Ormond D, Price C, Quail MA, Sanders M, Walker D, Whitehead S, Salmond GPC, Birch PRJ, Parkhill J, Toth IK. 2004.** Genome sequence of the enterobacterial phytopathogen Erwinia carotovora subsp. atroseptica and characterization of virulence factors. *Proceedings of the National Academy of Sciences of the United States of America* **101**:11105–11110 DOI 10.1073/pnas.0402424101.

**Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2012.** Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**:464–469 DOI 10.1093/bioinformatics/btr703.

**Cipriano MJ, Novichkov PN, Kazakov AE, Rodionov DA, Arkin AP, Gelfand MS, Dubchak I. 2013.** RegTransBase—a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. *BMC Genomics* **14**:213 DOI 10.1186/1471-2164-14-213.

**Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016.** GenBank. *Nucleic Acids Research* **44**:D67–D72 DOI 10.1093/nar/gkv1276.

**Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, De Hoon MJL. 2009.** Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**:1422–1423 DOI 10.1093/bioinformatics/btp163.

**Deng X, Lan L, Xiao Y, Kennelly M, Zhou J-M, Tang X. 2010.** Pseudomonas syringae two-component response regulator RhpR regulates promoters carrying an inverted repeat element. *Molecular Plant-Microbe Interactions* **23**:927–939 DOI 10.1094/MPMI-23-7-0927.

**Deng X, Liang H, Chen K, He C, Lan L, Tang X. 2014.** Molecular mechanisms of two-component system RhpRS regulating type III secretion system in Pseudomonas syringae. *Nucleic Acids Research* **42**:11472–11486 DOI 10.1093/nar/gku865.

**Eddy SR. 2011.** Accelerated profile HMM searches. *PLoS Computational Biology* **7**:e1002195 DOI 10.1371/journal.pcbi.1002195.

**Ferreira AO, Myers CR, Gordon JS, Martin GB, Vencato M, Collmer A, Wehling MD, Alfano JR, Moreno-Hagelsieb G, Lamboy WF, DeClerck G, Schneider DJ, Cartinhour SW. 2006.** Whole-genome expression profiling defines the HrpL regulon of Pseudomonas syringae pv. tomato DC3000, allows de novo reconstruction of the Hrp cis clement, and identifies novel coregulated genes. *Molecular Plant-Microbe Interactions* **19**:1167–1179 DOI 10.1094/MPMI-19-1167.

**Finn RD, Clements J, Eddy SR. 2011.** HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**:W29–W37 DOI 10.1093/nar/gkr367.

**Fouts DE, Abramovitch RB, Alfano JR, Baldo AM, Buell CR, Cartinhour S, Chatterjee AK, D'Ascenzo M, Gwinn ML, Lazarowitz SG, Lin N-C, Martin GB, Rehm AH, Schneider DJ, Van Dijk K, Tang X, Collmer A. 2002.** Genomewide identification of Pseudomonas syringae pv. tomato DC3000 promoters controlled by the HrpL alternative sigma factor. *Proceedings of the National Academy of Sciences of the United States of America* **99**:2275–2280 DOI 10.1073/pnas.032514099.

**Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muniz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, Garcia-Sotelo JS, Lopez-Fuentes A, Porron-Sotelo L, Alquicira-Hernandez S, Medina-Rivera A, Martinez-Flores I, Alquicira-Hernandez K, Martinez-Adame R, Bonavides-Martinez C, Miranda-Rios J, Huerta AM, Mendoza-Vargas A, Collado-Torres L, Taboada B, Vega-Alvarado L, Olvera M, Olvera L, Grande R, Morett E, Collado-Vides J. 2010.** RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Research* **39**:D98–D105 DOI 10.1093/nar/gkq1110.

**Glasner JD, Yang C-H, Reverchon S, Hugouvieux-Cotte-Pattat N, Condemine G, Bohin J-P, Gijsegem FV, Yang S, Franza T, Expert D, Plunkett G, Francisco MJS, 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. *Journal of Bacteriology* **193**:2076–2077 DOI 10.1128/JB.01513-10.

**Grote A, Klein J, Retter I, Haddad I, Behling S, Bunk B, Biegler I, Yarmolinetz S, Jahn D, Münch R. 2009.** PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. *Nucleic Acids Research* **37**:D61–D65 DOI 10.1093/nar/gkn837.

**Haque MM, Kabir MS, Aini LQ, Hirata H, Tsuyumu S. 2009.** SlyA, a MarR family transcriptional regulator, is essential for virulence in Dickeya dadantii 3937. *Journal of Bacteriology* **191**:5409–5418 DOI 10.1128/JB.00240-09.

**Harari O, Park S-Y, Huang H, Groisman EA, Zwir I. 2010.** Defining the plasticity of transcription factor binding sites by deconstructing DNA consensus sequences: the PhoP-binding sites among gamma/enterobacteria. *PLoS Computational Biology* **6**:e1000862 DOI 10.1371/journal.pcbi.1000862.

**Hertz GZ, Hartzell GW, Stormo GD. 1990.** Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Computer Applications in the Biosciences* **6**:81–92 DOI 10.1093/bioinformatics/6.2.81.

**Hughes JD, Estep PW, Tavazoie S, Church GM. 2000.** Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology* **296**:1205–1214 DOI 10.1006/jmbi.2000.3519.

**Ide N, Ikebe T, Kutsukake K. 1999.** Reevaluation of the promoter structure of the class 3 flagellar operons of *Escherichia coli* and Salmonella. *Genes & Genetic Systems* **74**:113–116 DOI 10.1266/ggs.74.113.

**Ishida Y, Kori A, Ishihama A. 2009.** Participation of regulator AscG of the $\beta$-glucoside utilization operon in regulation of the propionate catabolism operon. *Journal of Bacteriology* **191**:6136–6144 DOI 10.1128/JB.00663-09.

**Kılıç S, White ER, Sagitova DM, Cornish JP, Erill I. 2013.** CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. *Nucleic Acids Research* **42**:D156–D160 DOI 10.1093/nar/gkt1123.

**Kingsford CL, Ayanbule K, Salzberg SL. 2007.** Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biology* **8**:R22 DOI 10.1186/gb-2007-8-2-r22.

**Koskinen JP, Laine P, Niemi O, Nykyri J, Harjunpää H, Auvinen P, Paulin L, Pirhonen M, Palva T, Holm L. 2012.** Genome sequence of *pectobacterium* sp. strain SCC3193. *Journal of Bacteriology* **194**:6004–6004 DOI 10.1128/JB.00681-12.

**Kwenda S, Gorshkov V, Ramesh AM, Naidoo S, Rubagotti E, Birch PRJ, Moleleki LN. 2016.** Discovery and profiling of small RNAs responsive to stress conditions in the plant pathogen *Pectobacterium atrosepticum*. *BMC Genomics* **17**:47 DOI 10.1186/s12864-016-2376-0.

**Langmead B, Salzberg SL. 2012.** Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359 DOI 10.1038/nmeth.1923.

**Li H. 2011.** A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**:2987–2993 DOI 10.1093/bioinformatics/btr509.

**McNally R, Toth IK, Cock PJ, Pritchard L, Hedley PE, Morris JA, Zhao Y, Sundin GW. 2011.** Genetic characterization of the HrpL regulon of the fire blight pathogen Erwinia amylovora reveals novel virulence factors. *Molecular Plant Pathology* **13**:160–173 DOI 10.1111/j.1364-3703.2011.00738.x.

**McVicker G, Sun L, Sohanpal BK, Gashi K, Williamson RA, Plumbridge J, Blomfield IC. 2011.** SlyA protein activates fimB gene expression and type 1 fimbriation in *Escherichia coli* K-12. *Journal of Biological Chemistry* **286**:32026–32035 DOI 10.1074/jbc.M111.266619.

**Mironov AA, Vinokurova NP, Gel'fand MS. 2000.** Software for analyzing bacterial genomes. *Molecular Biology* **34**:253–262.

**NCBI. 2016.** QBLAST's URL API. User's Guide. *Available at http://www.ncbi.nlm.nih.gov/BLAST/Doc/urlapi.html* (accessed 13 May 2016).

**Nikolaichik Y, Gorshkov V, Gogolev Y, Valentovich L, Evtushenkov A. 2014.** Genome sequence of *Pectobacterium atrosepticum* strain 21A. *Genome Announcements* **2**:e00935–14 DOI 10.1128/genomeA.00935-14.

**Nishino K, Honda T, Yamaguchi A. 2005.** Genome-wide analyses of *Escherichia coli* gene expression responsive to the BaeSR two-component regulatory system. *Journal of Bacteriology* **187**:1763–1772 DOI 10.1128/JB.187.5.1763-1772.2005.

**Nonaka G, Blankschien M, Herman C, Gross CA, Rhodius VA. 2006.** Regulon and promoter analysis of the *E. coli* heat-shock factor, σ32, reveals a multifaceted cellular response to heat stress. *Genes & Development* **20**:1776–1789 DOI 10.1101/gad.1428206.

**Novichkov PS, Brettin TS, Novichkova ES, Dehal PS, Arkin AP, Dubchak I, Rodionov DA. 2012.** RegPrecise web services interface: programmatic access to the transcriptional regulatory interactions in bacteria reconstructed by comparative genomics. *Nucleic Acids Research* **40**:W604–W608 DOI 10.1093/nar/gks562.

**Novichkov PS, Kazakov AE, Ravcheev DA, Leyn SA, Kovaleva GY, Sutormin RA, Kazanov MD, Riehl W, Arkin AP, Dubchak I, Rodionov DA. 2013.** RegPrecise 3.0—a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics* **14**:745 DOI 10.1186/1471-2164-14-745.

**Novichkov PS, Laikova ON, Novichkova ES, Gelfand MS, Arkin AP, Dubchak I, Rodionov DA. 2010a.** RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Research* **38**:D111–D118 DOI 10.1093/nar/gkp894.

**Novichkov PS, Rodionov DA, Stavrovskaya ED, Novichkova ES, Kazakov AE, Gelfand MS, Arkin AP, Mironov AA, Dubchak I. 2010b.** RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Research* **38**:W299–W307 DOI 10.1093/nar/gkq531.

**Ogasawara H, Shinohara S, Yamamoto K, Ishihama A. 2012.** Novel regulation targets of the metal-response BasS-BasR two-component system of *Escherichia coli*. *Microbiology* **158**:1482–1492 DOI 10.1099/mic.0.057745-0.

**Park T-H, Choi B-S, Choi A-Y, Choi I-Y, Heu S, Park B-S. 2012.** Genome sequence of *Pectobacterium carotovorum* subsp. carotovorum strain PCC21, a pathogen causing soft rot in chinese cabbage. *Journal of Bacteriology* **194**:6345–6346 DOI 10.1128/JB.01583-12.

**Peano C, Wolf J, Demol J, Rossi E, Petiti L, De Bellis G, Geiselmann J, Egli T, Lacour S, Landini P. 2015.** Characterization of the *Escherichia coli* σS core regulon by Chromatin Immunoprecipitation-sequencing (ChIP-seq) analysis. *Scientific Reports* **5**:10469 DOI 10.1038/srep10469.

**Pearson W. 2004.** Finding protein and nucleotide similarities with FASTA. *Current Protocols in Bioinformatics* **8:3.9**:3.9.1–3.9.23 DOI 10.1002/0471250953.bi0309s04.

**Pedersen AG, Baldi P, Brunak S, Chauvin Y. 1996.** Characterization of prokaryotic and eukaryotic promoters using hidden Markov models. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology* **4**:182–191.

**Ravcheev DA. 2009.** The study of prokaryotic regulatory systems evolution with comparative genomic analysis methods. PhD Thesis, Moscow, A.A. Kharkevich Institute for Information Transmission Problems.

**Rhodius VA, Suh WC, Nonaka G, West J, Gross CA. 2006.** Conserved and variable functions of the $\sigma$E stress response in related genomes. *PLoS Biology* **4**:e2 DOI 10.1371/journal.pbio.0040002.

**Rodionov DA. 2004.** Comparative genomics of the KdgR regulon in Erwinia chrysanthemi 3937 and other gamma-proteobacteria. *Microbiology* **150**:3571–3590 DOI 10.1099/mic.0.27041-0.

**Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, Weiss V, Solano-Lira H, Martínez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernández S, Alquicira-Hernández K, López-Fuentes A, Porrón-Sotelo L, Huerta AM, Bonavides-Martínez C, Balderas-Martínez YI, Pannier L, Olvera M, Labastida A, Jiménez-Jacinto V, Vega-Alvarado L, Del Moral-Chávez V, Hernández-Alvarez A, Morett E, Collado-Vides J. 2013.** RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research* **41**:D203–D213 DOI 10.1093/nar/gks1201.

**Schneider TD, Stephens RM. 1990.** Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* **18**:6097–6100 DOI 10.1093/nar/18.20.6097.

**Seemann T. 2014.** Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**:2068–2069 DOI 10.1093/bioinformatics/btu153.

**Seo SW, Kim D, Szubin R, Palsson BO. 2015.** Genome-wide reconstruction of OxyR and SoxRS transcriptional regulatory networks under oxidative stress in *Escherichia coli* K-12 MG1655. *Cell Reports* **12**:1289–1299 DOI 10.1016/j.celrep.2015.07.043.

**Stapleton MR, Norte VA, Read RC, Green J. 2002.** Interaction of the Salmonella typhimurium transcription and virulence factor SlyA with target DNA and identification of members of the SlyA regulon. *Journal of Biological Chemistry* **277**:17630–17637 DOI 10.1074/jbc.M110178200.

**Stormo GD. 2000.** DNA binding sites: representation and discovery. *Bioinformatics* **16**:16–23 DOI 10.1093/bioinformatics/16.1.16.

**Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. 1982.** Use of the "Perceptron" algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Research* **10**:2997–3011 DOI 10.1093/nar/10.9.2997.

**Swingle B, Thete D, Moll M, Myers CR, Schneider DJ, Cartinhour S. 2008.** Characterization of the PvdS-regulated promoter motif in Pseudomonas syringae pv. tomato DC3000 reveals regulon members and insights regarding PvdS function in other pseudomonads. *Molecular Microbiology* **68**:871–889 DOI 10.1111/j.1365-2958.2008.06209.x.

**Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Ciufo S, Li W. 2013.** *Prokaryotic genome annotation pipeline*. Bethesda: National Center for Biotechnology Information. *Available at http://www.ncbi.nlm.nih.gov/genome/annotation_prok/*.

**Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P, Szafron D, Greiner R, Wishart DS. 2005.** BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Research* **33**:W455–W459 DOI 10.1093/nar/gki593.

**Wheeler TJ, Eddy SR. 2013.** nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**:2487–2489 DOI 10.1093/bioinformatics/btt403.

**Yada T, Totoki Y, Ishii T, Nakai K. 1997.** Functional prediction of B. subtilis genes from their regulatory sequences. *Proceedings of International Conference on Intelligent Systems for Molecular Biology* **5**:354–357.