Published online 5 May 2016 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* ***W77–W82***
doi: 10.1093/nar/gkw384

GREAT: a web portal for Genome Regulatory Architecture Tools

Costas Bouyioukos†, François Bucchini†, Mohamed Elati* and François Képès*

iSSB, CNRS, Genopole, UEVE, Université Paris-Saclay, 5 rue Henri Desbruères, Évry 91030 Cedex, France

Received February 23, 2016; Revised April 17, 2016; Accepted April 26, 2016

ABSTRACT

GREAT (Genome REgulatory Architecture Tools) is a novel web portal for tools designed to generate user-friendly and biologically useful analysis of genome architecture and regulation. The online tools of GREAT are freely accessible and compatible with essentially any operating system which runs a modern browser. GREAT is based on the analysis of genome layout -defined as the respective positioning of co-functional genes- and its relation with chromosome architecture and gene expression. GREAT tools allow users to systematically detect regular patterns along co-functional genomic features in an automatic way consisting of three individual steps and respective interactive visualizations. In addition to the complete analysis of regularities, GREAT tools enable the use of periodicity and position information for improving the prediction of transcription factor binding sites using a multi-view machine learning approach. The outcome of this integrative approach features a multivariate analysis of the interplay between the location of a gene and its regulatory sequence. GREAT results are plotted in web interactive graphs and are available for download either as individual plots, self-contained interactive pages or as machine readable tables for downstream analysis. The GREAT portal can be reached at the following URL https://absynth.issb.genopole.fr/GREAT and each individual GREAT tool is available for downloading.

INTRODUCTION

The arrangement of genomic features along chromosomes does not appear to be random. The relative linear order of features -such as genes- which constitutes the genome layout, has been shaped by evolutionary adaptations to accommodate multiple constraints. Transcription regulation, at the genome scale, is among the most crucial of these constraints for cell success. Two main insights indicate non-random genome layout. First, the analysis of contiguous genomic segments between related genomes has highlighted synteny, that is the conservation of short-range gene order (1). Second, the detection of long-range regularities in the positioning of genes which are co-regulated, co-evolved or co-expressed along the genomes of all prokaryotic phyla (2–5), one Archae (6) and one Eukaryote (7).

This general scheme of genome architecture, which highlights both proximity and periodicity in the layout of co-functional genomic features, has been proposed as an organisation principle for global genome regulation (8,9). Indeed, short and long range interactions perform significant roles in shaping the transcriptional landscape of prokaryotic as well as compact eukaryotic genomes (10,11). Genome structure is coupled with genome regulation through the influence of supercoiling (and its associated micro-domains), packing and localised transcriptional activity (12,13). Nevertheless, there is a lack of a unifying framework to explore, study and understand the intricate relationships between genome organisation and regulation. Regularities among co-functional genomic features might reveal potential co-clustering and/or co-regulation of features of interest. An easy to use on-line set of tools can readily provide insights to genome architecture and its relation to regulation, as current experimental and computational modelling techniques are still expensive. Therefore, we set up a web application consisting of tools able to investigate genomic positional regularities, in the context of genome expression regulation.

Here, we present GREAT, an online portal for integrative studies of genome layout and its association with genome regulation. The current release of GREAT includes tools for analysis and visualisation of regular patterns of genomic positions of co-functional genes (or other genomic features). GREAT also includes a machine learning tool which takes advantage of gene positioning to improve transcription factor binding sites (TFBS) prediction. We named the current release of GREAT tools GREAT:SCAN, and the two individual tools Patterns and PreCisIon, respectively.

*To whom correspondence should be addressed. Tel: +33 1 69 47 44 43; Email: mohamed.elati@issb.genopole.fr
Correspondence may also be addressed to François Képès. Tel: +33 1 69 47 44 30; Email: francois.kepes@issb.genopole.fr
†These authors contributed equally to this work.

MATERIALS AND METHODS

The GREAT portal currently comprises two independent yet interrelated tools (both grouped under the label GREAT:SCAN). These tools enable systematic analyses and visualisations of genome layout and potential 3D organisation based on periodic or proximal positioning of co-functional genes, as well as exploring the interplay between these genomic features and genome expression regulation.

Input for both tools consists of genome length and the coordinates of genomic features of the organism of interest. These features include, but are not limited to, co-regulated, co-functional or co-evolved genes, genes encoding pathways of interest, ChIP peaks, nucleosome positions or replication origins from any organism that these are available.

PATTERNS performs a systematic analysis of regular patterns in the features of interest and PreCisIon uses information from position regularities together with promoter sequence information, to improve the prediction of TFBSs. PreCisIon requires additionally promoter sequences (or instance like the ones that can be obtained from the RSAT database (14)) as input.

Figure 1 represents a graphical overview of the procedures and outcomes of the GREAT:SCAN tools. It illustrates the full complement of the computational steps which start from input genomic positions and lead to the generation of rich and informative visualisations and machine readable output files. The relationships between the two different tools is also depicted to highlight the integrative approach in the analysis of genome layout, architecture and regulation which is introduced by the GREAT portal.

Each of the two tools in GREAT:SCAN is described below.

GREAT:SCAN:Patterns

PATTERNS performs an exhaustive and systematic analysis of periodic patterns in the positions of genomic features, and reports informative visualisations. The software runs in three steps: detection of genomic periods on the full genome; genomic features clustering and intuitive visualisation; mapping of periodic regions on the genome.

Input. The main input requirement is a file containing the genomic coordinate and an arbitrary ID for each feature of interest (for a single chromosome). The tool performs the analysis in three steps, and each step consists of a calculation and a visualisation part.

Workflow. STEP I: Exhaustively search periodic patterns in the genomic coordinates. Patterns detects all possible periods by using the Solenoid Coordinates Model (SCM), merges extremely similar ones and evaluates periods by their p-values. The exact algorithms are described in (6,15). *P* values are corrected for multiple testing inversely proportional to the period length. An important parameter for the calculation of periodicity statistics is the "Average gene-to-gene distance" which controls the replacement of proximal genomic features by their mean position. Proximal features can generate artifactually low p values and are routinely substituted by their mean for any analysis of periodicities.

Step I, reports the observed periods and their respective corrected p values in a "periodogram", a plot inspired from spectral analysis methods.

STEP II: A clustering of the genomic features is performed for each significant period based on the "phase coordinates". That is the remainder of the modulo division over the period length. Then for each feature the positional score is calculated. That is an information based measure of how much each individual feature has contributed to the particular periodical pattern (15). Finally, an automatically generated "cluster gram" reports the "in-phase" arrangement and provides indications for potential 3D co-localization. Features replaced due to proximity will appear to have identical positional score in the clustergram. The sensitivity of the clustering algorithm is controlled by the "clustering exponent" user parameter which can be found in the "Advanced settings" of the parameters panel of the tool.

STEP III: A sliding window, of size varying up to the whole genome, is employed in this step. The window rolls along the whole genome and the process described in "step I" detects periods within its borders. This approach, described in (6), allows the mapping of regions with potential high regularity on the whole chromosome. The resulting graph is a genomic map of regions with the observed significant periodicities and their significance.

Results. A detailed description of the elements of each visual output can be found in the working example section. A full description of the user parameters to control each of the above three steps can be found in the online help documentation page for GREAT.

GREAT:SCAN:MULTIPATTERNS: is an extension of the Patterns tool for multi-chromosome organisms. GREAT:SCAN:MultiPatterns performs the same systematic analysis for periodic patterns in the positions of genomic features but applies to organisms with more than one chromosomes and is also able to repeat the search for multiple regulators (TFs) or conditions.

Input:. Two files, one containing a set of regulators/conditions and their associated genomic features, the other containing the chromosome number and coordinate of these features.

Workflow. Multiplexing steps I and III described in Patterns for many chromosomes and regulators or conditions.

Results:. Visualisations equivalent to the output of steps I and III of Patterns as well as machine readable database tables of all the observed periods.

GREAT:SCAN:PreCisIon

PRECISION is a multi-view boosting machine learning tool for TFBS prediction. It is based on a previously published learning algorithm and tool (6,16). PreCisIon reports an ensemble classifier comprising sequence and position classifiers (16) and performs a multivariate statistical analysis of the interplay between position and sequence (6).

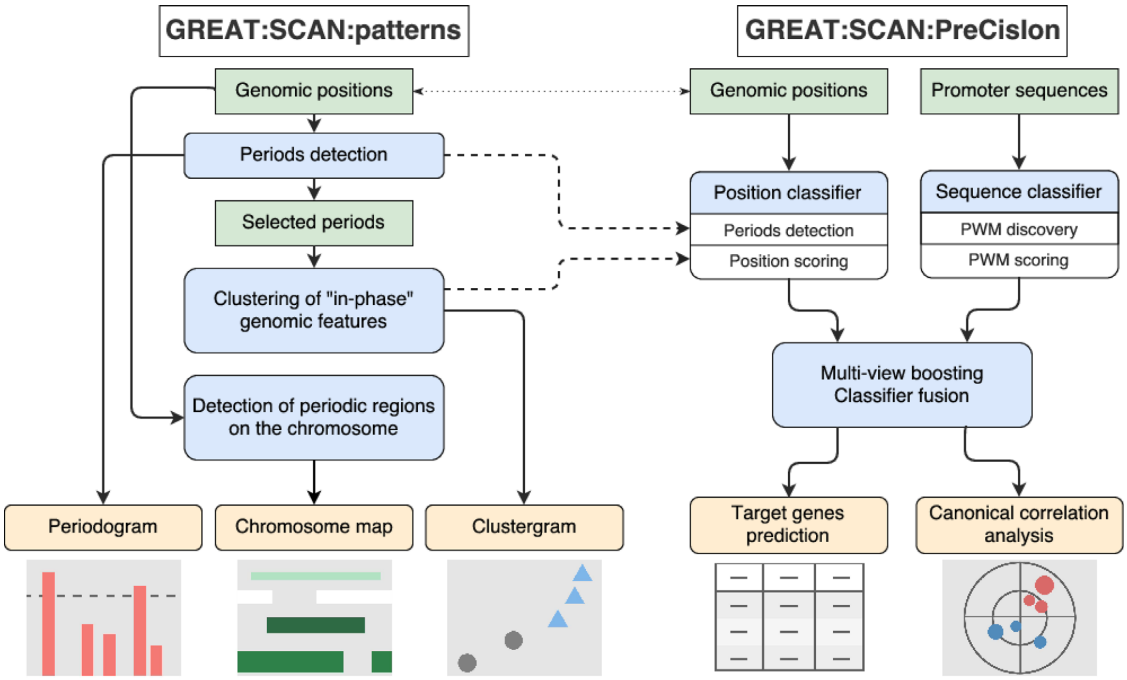

Figure 1. Overview of `GREAT:SCAN` tools. Input and intermediate data are represented by green boxes. Analysis steps and computations performed by the tools are represented by blue boxes. Visual outputs produced by the tools are represented by orange boxes and corresponding thumbnails depict caricatures of the original visualisations. Solid lines depict successive computation steps, dashed lines depict links between the output of one tool and input to other, the thin doubled-arrowed dotted line represents interchangeability between input files. The first and second step inside each `GREAT:SCAN:PreCisIon` classifier correspond to the learning and the training stage.

Input. A file with promoter sequences of a set of co-regulated genes and a file with their respective genomic coordinates, the latter can be identical with the input file of `Patterns` described in the respective section. Currently the GREAT portal runs `PreCisIon` using ready-made preloaded input files for 24 *Escherichia coli* TFs with the highest number of targets.

Workflow. In each boosting step the classifier with the lowest prediction error is selected, either the position or the sequence. At the end of boosting an ensemble classifier is returned. The performance of the `PreCisIon` prediction algorithm has been extensively assessed by using all the TFs from two well studied benchmark organisms *E. coli* and *Bacillus subtilis* (16). In several cases the position classifier alone performs better than the sequence one and in many other cases the combined boosted classifier outperforms any single one (Figure 2 of (16)).

Results. A table with the predicted TF targets is returned. Then, the prediction scores from each selected classifier are analysed using canonical correlation analysis (CCA) (17) to explore associations between position information and sequence information. A CCA correlation circle plot intuitively illustrates the two main variates which might indicate potential interplay between sequence and position.

IMPLEMENTATION

Tools

GREAT:SCAN command-line tools are developed using a variety of languages (such as R (18), Python and Java) and are released using docker (19), a virtualisation software allowing the deployment of applications inside software containers. This implementation brings the following advantages. It makes deployment easy, as nothing has to be installed on the host apart from docker. Furthermore, as every process runs in its own environment, which contains all necessary dependencies, the typical problems of missing libraries or version conflicts do not occur. This is a significant advantage as bioinformatics tools require a broad variety of software dependencies.

Server

The `GREAT` portal is a web application implemented using Flask (20), a python framework, as back-end. Interactive visualisations available on the portal are developed using two mainstream open-source JavaScript libraries, Highcharts (21) and D3.js (22). Static visualisations are rendered as Scalable Vector Graphics (SVG) images. The web portal itself is deployed as a docker container. Each job also runs in a distinct tool-specific docker container created upon submission request. After completion, job's results are processed to generate the visualisations and populate the results pages.

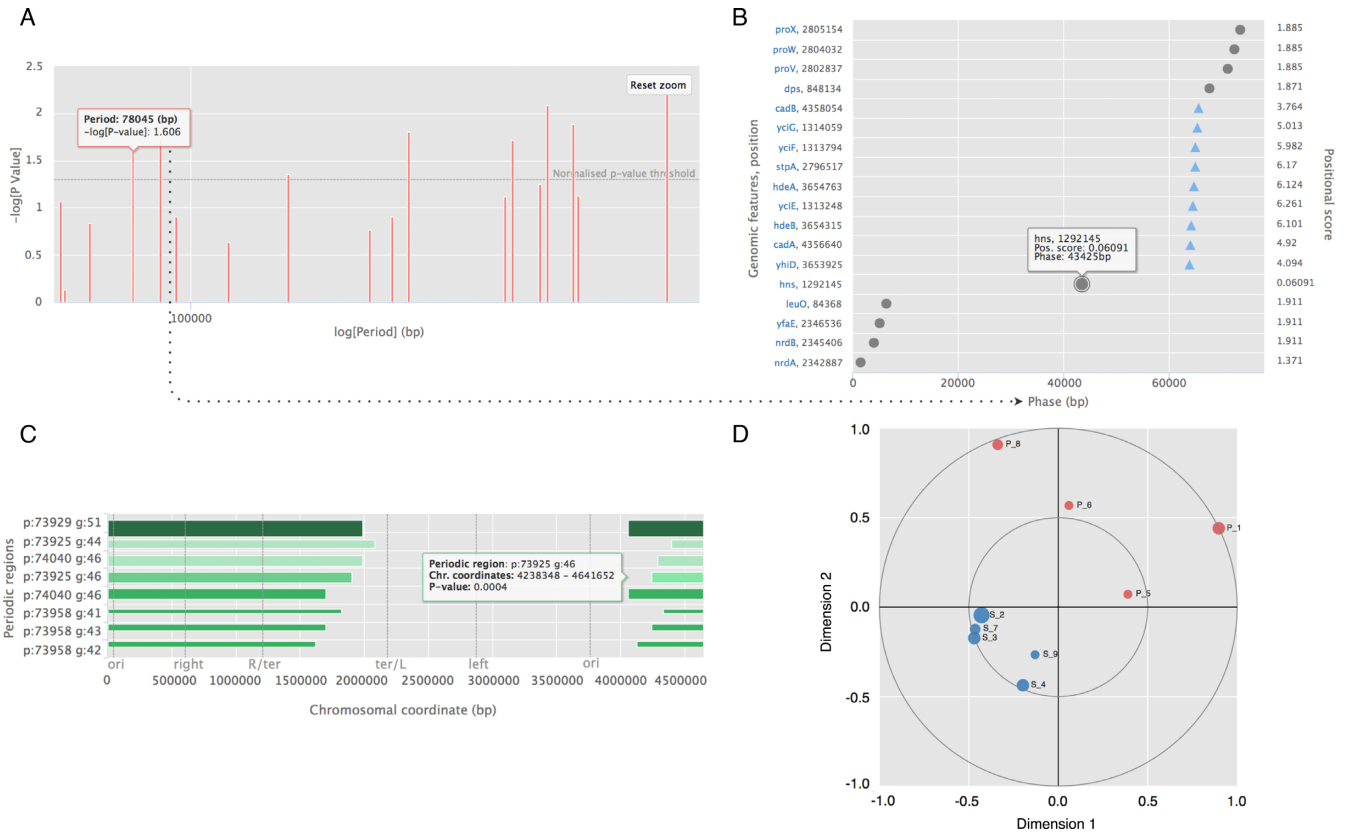

Figure 2. Graphical outputs of `Patterns`. (**A**) Periodogram: The period length (in log scale) is plotted in the horizontal axis, the vertical axis represents the antilogarithm of the *P*-values. (**B**) Clustergram: Period length (or 'phase') is plotted on the horizontal axis, names and chromosomal coordinates on the left vertical axis and and the positional score (details in the text) on the right. The dashed line indicates that the highlighted period on (A) was used to calculate the 'clustergram' (B) (details on the step I and II description of `Patterns`). (**C**) Chromosome map: The horizontal axis spans the genome length and each coloured bar, stacked from bottom to top, represents a region with detected periodic pattern. All three visualizations provide interactive mouse over information as shown by the tooltips. Graphical output of `GREAT:SCAN:PreCisIon` (**D**): interactive, correlation circle plot of the prediction scores. The axes represent the CCA variates (i.e. the two components which capture the highest correlation between variables). Iterations where the position classifier performed better are represented as red dots and where the sequence classifier performed better as blue dots. The bigger the dot, the smaller the classification error. Clicking on a point will load visualizations of the respective classifier at the particular iteration.

USING GREAT

Online portal

GREAT:SCAN tools can be accessed in different ways: via the web portal or using the command-line version of the tools, available as docker images. The simplest way to use `GREAT:SCAN` is through the web portal, which provides a user-friendly interface (while maintaining the full functionality and parameter choices of the command line) with interactive visualisations and does not require any additional software installation.

Command line (Docker images)

For advanced users wanting to run jobs locally from the command-line or include `GREAT` tools in pipelines, we packed each tool in docker and the images can be downloaded from the "Help" section of the web portal.

Results availability

GREAT is free and open to all users, there is no login requirement. After submission of a job, a unique job-specific URL is provided which can be bookmarked and accessed at any time within seven days. Links to the submitted jobs are also displayed on the loading page. Note that the results can be downloaded as zip archive and are available for seven days.

Documentation

GREAT documentation is available online and can be accessed using the "Help" link in the top navigation bar. It contains a brief description of each tool, along with examples of the input data and parameters, links to sample input data and results, explanations on how to interpret the results given by each tool and references. In addition to this, users have at their disposal mouse-over tooltips that give concise explanations of many critical elements of the portal.

WORKING EXAMPLES

Exemplary cases of the usage of the `GREAT` portal are provided to demonstrate the capabilities of genome regulatory architecture analysis and explain in details the outputs that the `GREAT:SCAN` tools generate.

GREAT:SCAN:Patterns

The graphical output of all three analysis steps of Patterns are depicted in Figure 2. The periodogram, the clustergram and the chromosome mapping plots are included. The periodogram on Figure 2(A) illustrates all the periods that have passed a first user defined *P*-value threshold and identifies the (also user-defined) significance p-value cutoff. Periodograms represent a fast yet informative way for immediate exploration of periodic patterns. Each clustergram Figure 2(B) pertains to a particular significant period and illustrates the phase arrangement of the genomic features. This intuitive visualization allows detection of genes sharing the same position on the phase diagram which potentially corresponds to actual 3D clusters of co-localized genes. Finally, the output of the third step of Patterns is a map of periodic regions along the whole chromosome Figure 2(C). The extremities of each region are depicted and a rich plot is generated offering information for the period, the number of genes and the significance of the finding.

The GREAT:SCAN:MultiPatterns extension reports in tables all the detected periodic regions for many regulators and chromosomes at the same time, and generates graphical output entirely equivalent with the first and last step of Patterns.

GREAT:SCAN:PreCisIon

The main output of PreCisIon is an ensemble classifier for the prediction of novel target genes of a given TF. This classifier is used to predict targets of the TF and predictions are reported in an interactive table on the output page of the tool. In addition to the predictions, GREAT:SCAN:PreCisIon features an additional step, the multivariate analysis of the interplay between position and sequence (introduced in (23)), which is depicted by a Canonical Correlation Analysis (CCA) circle plot, presented in Figure 2(D) CCA circle plots (explained in (24)) provide exploratory information of potential correlation or anti-correlation between position and sequence in a TFBS.

More generally, for every step in each tool and visualization, the equivalent machine-readable files are also available. The GREAT portal renders them as interactive HTML tables which can be browsed according to user needs and downloaded for further analysis.

CONCLUSIONS

GREAT is, to our knowledge, the first suite of tools which provides an integrated framework to study patterns of regularities among genomic features of interest and connect these observations with genome regulation. The input of GREAT can be any set of co-functional genes but it is not restricted to that. Any set of coordinate:ID pair can be used as an input and we envisage to see GREAT:SCAN tools be used for the analysis of patterns of a series of genomic features including but not limited to ChIP peaks data, nucleosome positioning and replication start sites (to name a few). We contemplate that the usage of GREAT will provide bioscience researchers with novel and informative insights regarding genome organization, regulation and function.

PERSPECTIVE

In the near future, GREAT will acquire a capacity to deal, in an integrated manner, not only with multiple chromosomes but also with multiple datasets (e.g. all TFs of a particular organism). This integration will allow to compare and interpret the observed patterns of multiple regulons (or different genomic features of interest) with higher order descriptors such as physiological states.

ACKNOWLEDGEMENTS

We thank Joan Hérisson for his help with the deployment of the tools on the abSYNTH platform at iSSB. We would also like to thank our first users, the wet lab biologists of the iSSB MEGA team for extensive testing and feedback about the tools and portal.

FUNDING

CNRS (to F.K.), UEVE (to M.E.), Genopole project 'Robust' (to F.K.); ANR project 'Synaptic' (to F.K., M.E. and C.B.); ITMO cancer project 'LIONS' N° BIO2015-04 (to M.E. and F.B.). Funding for open access charge: ANR project SYNPATHIC-ROBUST.

Conflict of interest statement. None declared.

REFERENCES

1. Huynen,M.A. and Snel,B. (2000) Gene and context: integrative approaches to genome analysis. *Adv. Protein Chem.*, **54**, 345–379.
2. Képès,F. (2004) Periodic transcriptional organization of the *E.coli* genome. *J. Mol. Biol.*, **340**, 957–964.
3. Junier,I., Hérisson,J. and Képès,F. (2012) Genomic organization of evolutionarily correlated genes in bacteria: limits and strategies. *J. Mol. Biol.*, **419**, 369–386.
4. Wright,M.A., Kharchenko,P., Church,G.M. and Segré,D. (2007) Chromosomal periodicity of evolutionarily conserved gene pairs. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 10559–10564.
5. Jeong,K.S., Ahn,J. and Khodursky,A.B. (2004) Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biol.*, **5**, R86.
6. Bouyioukos,C., Elati,M. and Képès,F. (2015) Protocols for probing genome architecture of regulatory networks in hydrocarbon and lipid microorganisms. In: McGenity,TJ, Timmis,KN and Fernández B,Nogales (eds). *Hydrocarbon and Lipid Microbiology Protocols*. Springer, Protocols Handbooks, Heidelberg, pp. 1–16.
7. Képès,F. (2003) Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. *J. Mol. Biol.*, **329**, 859–865.
8. Képès,F. and Vaillant,C. (2003) Transcription-Based Solenoidal Model of Chromosomes. *ComPlexUs*, **1**, 171–180.
9. Dorman,C.J. (2013) Genome architecture and global gene regulation in bacteria: making progress towards a unified model? *Nat. Rev. Microbiol.*, **11**, 349–355.
10. Carpentier,A.-S., Torrésani,B., Grossmann,A. and Hénaut,A. (2005) Decoding the nucleoid organisation of *Bacillus subtilis* and *Escherichia coli* through gene expression data. *BMC Genomics*, **6**, 84.
11. Li,S. and Heermann,D.W. (2013) Transcriptional regulatory network shapes the genome structure of *Saccharomyces cerevisiae*. *Nucleus*, **4**, 216–228.
12. Allen,T.E., Price,N.D., Joyce,A.R. and Palsson,B.O. (2006) Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization. *PLoS Comput. Biol.*, **2**, e2.
13. Weng,X. and Xiao,J. (2014) Spatial organization of transcription in bacterial cells. *Trends Genetics*, doi:10.1016/j.tig.2014.04.008.
14. Thomas-Chollier,M., Defrance,M., Medina-Rivera,A., Sand,O., Herrmann,C., Thieffry,D. and van Helden,J. (2011) RSAT 2011:

Regulatory Sequence Analysis Tools. *Nucleic Acids Res.*, **39**, W86–W91.
15. Junier,I., Hérisson,J. and Képès,F. (2010) Periodic pattern detection in sparse boolean sequences. *Algorithms Mol. Biol.*, **5**, 31.
16. Elati,M., Nicolle,R., Junier,I., Fernández,D., Fekih,R., Font,J. and Képès,F. (2013) PreCisIon: PREdiction of CIS-regulatory elements improved by gene's positION. *Nucleic Acids Res.*, **41**, 1406–1415.
17. Lê Cao,K.-A., Martin,P.G.P., Robert-Granié,C. and Besse,P. (2009) Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, **10**, 34.
18. R Core Team. (2015) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, www.r-project.org.
19. Docker, Build, Ship, Run any app anywhere. (2016) www.docker.com/.
20. Flask (A Python Microframework). (2016) flask.pocoo.org/.
21. Highcharts, Interactive javascript charts for your webpage. (2016) www.highcharts.com/.
22. D3.js Data-Driven-Documents. (2016) d3js.org/.
23. Bouyioukos,C., Elati,M. and Képès,F. (2016) Analysis tools for the interplay between genome layout and regulation. *BMC Bioinformatics*, in press.
24. González,I., Lê Cao,K.-A., Davis,M.J. and Déjean,S. (2012) Visualising associations between paired 'omics' data sets. *BioData Mining*, **5**, 19.