RSAT 2018: regulatory sequence analysis tools 20th anniversary

Nga thi thuy Nguyen, Bruno Contreras-Moreira, Jaime Castro-Mondragon, Walter Santana-Garcia, Raul Ossio, Carla daniela Robles-Espinoza, Mathieu Bahin, Samuel Collombet, Pierre Vincens, Denis Thieffry, et al.

Published online 2 May 2018 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* ***W209–W214***
doi: 10.1093/nar/gky317

RSAT 2018: regulatory sequence analysis tools 20th anniversary

Nga Thi Thuy Nguyen[1,†], Bruno Contreras-Moreira[2,3,†], Jaime A. Castro-Mondragon[4,5], Walter Santana-Garcia[6], Raul Ossio[6], Carla Daniela Robles-Espinoza[6,7], Mathieu Bahin[1], Samuel Collombet[1], Pierre Vincens[1], Denis Thieffry[1], Jacques van Helden[4,*], Alejandra Medina-Rivera[6,*] and Morgane Thomas-Chollier[1,*]

[1]Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, PSL Université Paris, 75005 Paris, France, [2]Estación Experimental de Aula Dei-CSIC, Zaragoza, Spain, [3]Fundación ARAID, Zaragoza, Spain, [4]Aix-Marseille Univ, INSERM UMR_S 1090, Theory and Approaches of Genome Complexity (TAGC), F-13288 Marseille, France, [5]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway, [6]Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Campus Juriquilla, Blvd Juriquilla 3001, Santiago de Querétaro 76230, México and [7]Experimental Cancer Genetics, The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK

Received January 31, 2018; Revised April 04, 2018; Editorial Decision April 13, 2018; Accepted April 23, 2018

ABSTRACT

RSAT (Regulatory Sequence Analysis Tools) is a suite of modular tools for the detection and the analysis of *cis*-regulatory elements in genome sequences. Its main applications are (i) motif discovery, including from genome-wide datasets like ChIP-seq/ATAC-seq, (ii) motif scanning, (iii) motif analysis (quality assessment, comparisons and clustering), (iv) analysis of regulatory variations, (v) comparative genomics. Six public servers jointly support 10 000 genomes from all kingdoms. Six novel or refactored programs have been added since the 2015 NAR Web Software Issue, including updated programs to analyse regulatory variants (*retrieve-variation-seq*, *variation-scan*, *convert-variations*), along with tools to extract sequences from a list of coordinates (*retrieve-seq-bed*), to select motifs from motif collections (*retrieve-matrix*), and to extract orthologs based on Ensembl Compara (*get-orthologs-compara*). Three use cases illustrate the integration of new and refactored tools to the suite. This Anniversary update gives a 20-year perspective on the software suite. RSAT is well-documented and available through Web sites, SOAP/WSDL (Simple Object Access Protocol/Web Services Description Language) web services, virtual machines and stand-alone programs at http://www.rsat.eu/.

INTRODUCTION

Initiated in 1998 (1,2), the Regulatory Sequences Analysis Tools (RSAT) project aims at deploying software tools to detect *cis*-regulatory elements in genomic sequences, via a Web interface. RSAT functionalities include *de novo* motif discovery, analyses of motif quality, motif comparisons and clustering, motif scanning to predict transcription factor (TF) binding sites (TFBSs), detection and analysis of regulatory variants, and comparative genomics to discover motifs based on cross-species conservation (Figure 1). Over the last 20 years, the RSAT team has maintained uninterrupted service, while extending developments prompted by the advances in the field of regulatory genomics (Supplementary Figure S1). This Anniversary article gives a 20-year perspective on the software suite, describes its main functionalities, focusing on the novelties since the previous NAR Web server issues (3–6), and presents the various access and training modalities.

RSAT OVER THE LAST 20 YEARS

From yeast-tools to RSAT

The development of RSAT (initially named yeast-tools) (1,2) was prompted by the sequencing of the yeast genome

*To whom correspondence should be addressed. Tel: +33 1 44 32 23 53; Fax: +33 1 44 32 39 41; Email: mthomas@biologie.ens.fr
Correspondence may also be addressed to Alejandra Medina-Rivera. Tel: +52 55 5623 4331; Email: amedina@liigh.unam.mx
Correspondence may also be addressed to Jacques van Helden. Tel: +33 4 91 82 87 49; Fax: +33 4 91 82 87 01; Email: Jacques.van-Helden@univ-amu.fr
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

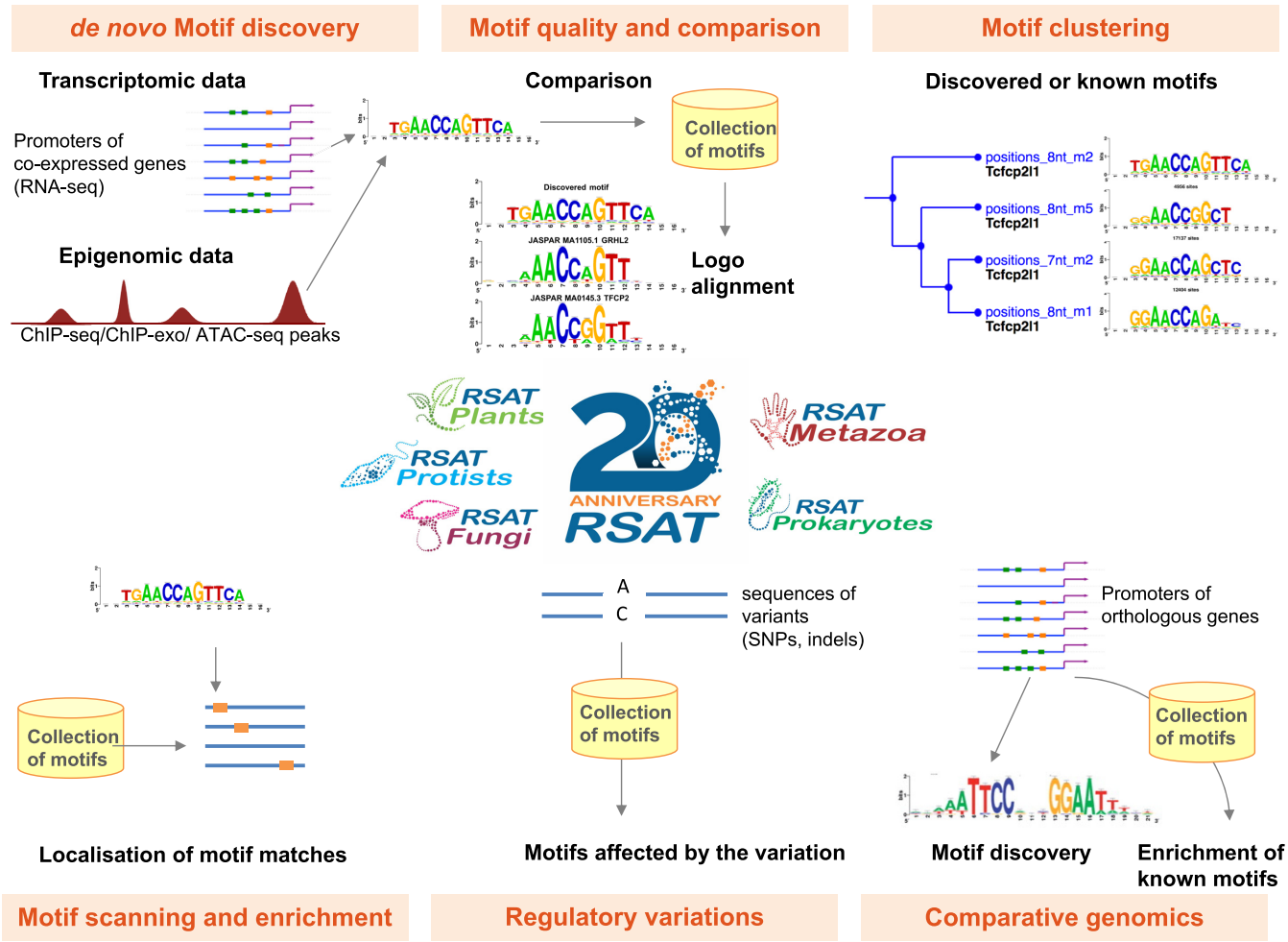

Figure 1. Overview of the main applications of RSAT.

(7). The motivation was to ease the extraction of non-coding sequences upstream of genes, and the prediction of TF binding sites (TFBSs). The programs for *ab initio* motif discovery, still today at the core of the suite, are based on a variety of criteria: over-represented oligonucleotides (*oligo-analysis* (1)), of spaced pairs (*dyad-analysis* (8)), or positionally biased oligonucleotides (*position-analysis* (9)). In 2000, a series of bacterial genomes were added to the suite, leading to the first RSAT release in 2003, supporting 100 genomes (6) (Supplementary Figure S1).

At that time RSAT started to support Position Specific Scoring Matrices (PSSMs) motif representation, in particular within the core tool *matrix-scan*, which scans sequences to locate putative TFBSs (10). As more and more prokaryotic genomes were sequenced, it became possible to use cross-species conservation to detect putative regulatory signals in non-coding sequences (phylogenetic footprinting) with *footprint-discovery* (11,12) (available for Prokaryotes and Fungi). In the 2008 server update (5), the number of tools almost doubled, programmatic access was offered as SOAP Web Service, and almost 700 genomes were supported, with six mirror servers across Europe and Mexico being available.

RSAT in the high-throughput sequencing era

The microarray technology that enabled measuring full transcriptomes and its ChIP-on-chip application opened the field of regulatory genomics to genome-wide analyses. In 2007, the ChIP-seq approach (13,14), taking advantage of high-throughput sequencing technology, revolutionised the field with an unprecedented level of precision and quantity of experimentally-detected functional TF binding regions. In contrast with alternative tools, RSAT programs coped well with datasets obtained from this technology, prompting the development of the user-friendly integrated pipeline *peak-motifs* (4,15,16), which enabled the online analysis of full datasets without size restriction. To better cope with the increasing size of datasets, multiple RSAT tools have been optimized over the years to reduce execution time. The 2011 RSAT server update (4) comprised new tools to retrieve sequences on-the-fly from Ensembl (*retrieve-ensembl-seq* (17)), to generate control sets, to discover motifs (*info-gibbs* (18)), to evaluate the quality of PSSMs (*matrix-quality* (19)) and to compare them (*compare-matrices*).

In 2015, after a drastic increase in available genomes in RSAT (~3300 at that time, with support for Ensembl

Genomes), it became necessary to reorganise the public mirrors into taxon-specific servers, concomitantly better accommodating the specific needs of user communities. The NAR server update (3) further presented novel tools, such as *matrix-clustering* (3,20), a clustering tool to regroup similar PSSMs and offers a dynamic visualisation of aligned PSSMs. We also introduced a first series of tools (*variation-seq*, *retrieve-variation-seq*) to predict the impact of non-coding variants on *cis*-regulatory elements. To facilitate a local installation of the suite, Virtual Machines were made available for download.

Currently, the RSAT Website includes six novel or refactored programs (tagged with asterisks in Table 1) for a total of 52 programs. As of January 2018, RSAT public servers support 10 032 locally-installed genomes (including 9 451 Prokaryotes, 238 Fungi, 91 Metazoa, 66 Plants and 186 Protists). For Prokaryotes, we now only support NCBI genome assemblies classified as 'Complete Genome' and 'Chromosome', leaving out any genome project classified as 'Contig' or 'Scaffold'. Nevertheless, additional genomes can be installed on request.

RSAT 2018 NOVELTIES

Many RSAT functionalities are described in the previous 2015 NAR update (3). We focus on the main novelties below (Table 1), situating the new tools in the global context of the suite.

Obtaining sequences and homologous genes

RSAT maintains locally-installed genomes and integrates on-the-fly access to external databases (Ensembl and UCSC). It offers tools to retrieve sequences relative to annotated genomic features (*retrieve-seq* for promoter sequences of local genomes, *retrieve-ensembl-seq* (17) for Ensembl vertebrate species). For genome-wide epigenomic datasets where genomic coordinates are usually specified in BED format, corresponding sequences can be extracted from UCSC (*fetch-sequences from UCSC*) and from local genomes with a new program supporting repeat-masking (*retrieve-seq-bed*). In addition, *retrieve-ensembl-seq* supports the retrieval of sequences from homologous genes. A new tool also relies on Ensembl Compara (21) to return detailed information on homologous genes in a set of reference organisms (*get-orthologs-compara*, currently only for Plants). In Fungi and Prokaryotes, lists of orthologous genes can be obtained with *get-orthologs*.

Obtaining motifs (PSSMs)

The ChIP-seq 'revolution' gave rise to a dramatic increase in the number of PSSMs stored in established motif databases, such as JASPAR (22), and a multiplication of independent motif collections. To facilitate the access to motifs, RSAT now locally hosts 50 external motif databases (JASPAR (22), Cis-Bp (23), FootprintDB (24), etc.) (Supplementary Table S1), covering DNA and RNA binding motifs in a wide range of organisms (Metazoa, Prokaryotes, Fungi, Plants). These collections have been homogenised in TRANSFAC format to alleviate format conversion. A new tool enables the extraction of particular motifs from these collections, based on identifiers or names (*retrieve-matrix*). The selection menu for motif collections is organised by color-coded taxons and is searchable to simplify access. This menu is also integrated in the tools using PSSMs as input, so that motifs can now be directly selected, rather than copy/pasted from external databases. To cope with the problem of motif redundancy within and across collections, we have established non-redundant motif collections by automatically clustering all the motifs from these collections (20).

Detecting regulatory variations

Population genomics and Genome-Wide Association Studies (GWAS) projects produce information on genetic variants (SNPs, indels), many of which are located in non-coding regions of the genome, and may thus affect *cis*-regulatory elements. To predict the impact of sequence variations on TF binding, variants and their flanking sequences can be extracted (*retrieve-variation-seq*) and scanned with a collection of motifs (*variation-scan*). This tool has been refactored to support multiallelic variants and indels, and optimized for time efficiency. A new tool further eases file format conversion between VCF, GVF and varBed (*convert-variations*). Altogether, RSAT sequence variation analysis tools (*variation-info, retrieve-variation-seq, convert-variations* and *variation-scan*) enable users to input any motif collection, and either retrieve Ensembl annotated variants, using either IDs or genomic coordinates, or input their own variants of interest (manuscript in prep).

Enhanced Web interface and visualisation

The home page has been extensively redesigned to simplify navigation and facilitate access to the tutorials, training material, and to the question-based menu guiding new users to the appropriate tools depending on their aims. A box to search the tools has been added, while tools that are not available on certain servers now appear deactivated in the menu. The number of organisms supported on each server is now clearly displayed. To accommodate the increasing number of supported organisms (especially in the Prokaryotes server), the selection menu has been replaced by a search engine implemented in Ajax. The visualisation tool *feature-map* has been re-implemented using modern libraries (d3) (*feature-map2*). A twitter account @RSATools is now alive with the feed displayed on the main page.

USE CASES

We present three use cases that exemplify applications integrating the novel tools into routine analysis (Supplementary Use Cases).

Use Case 1: Identify the binding motif for VRN1 in the promoters of the Flowering Locus T-like 1 orthologous genes. This use case on the Plant server integrates the usage of the novel tools *get-orthologs-compara* and *feature-map2,* with *matrix-scan and retrieve-sequence.*

Use Case 2: Select the motifs of transcription factors that conform AP-1 heterodimers, identify and reduce redundancy within this set of motifs, and detect AP1 binding

Table 1. Main tools available on RSAT Web servers (2018 update)

Application	Program name	Input	Output	Description
Obtaining sequences (Sequence Tools)	retrieve-seq	Gene names	Sequences	Given a set of gene names, returns upstream, downstream (relative to ORF start) or unspliced ORF sequences. Segments overlapping an upstream ORF can be excluded or included.
	fetch-sequences (from UCSC)	Genomic coordinates	Sequences	From a set of genomic coordinates (bed file), collects the sequences from the UCSC genome browser.
	*retrieve-seq-bed	Genomic coordinates	Sequences	From a set of genomic coordinates (bed/gff/vcf file), collects the sequences from installed organisms. Supports repeat masking option.
	retrieve-ensembl-seq	Gene names	Sequences	Returns upstream, downstream, intronic, exonic, UTR, mRNA or CDS for a list of genes from EnsEMBL vertebrates.
Motif discovery	oligo-analysis	Sequences	Over/under-represented oligonucleotides + PSSM	Analyses oligonucleotide occurrences in a set of sequences, and detects over/under-represented oligonucleotides, using various background models and scoring statistics.
	dyad-analysis	Sequences	Over/under-represented dyads + PSSM	Detects over-represented dyads (spaced pairs of oligonucleotides) within a set of sequences.
NGS ChIP-seq	peak-motifs	Sequences	Discovered motifs + predicted sites	Discovers motifs in ChIP-seq peak sequence sets, and returns detailed information on sequence composition and discovered motifs, with correspondence in databases and predicted binding sites.
Pattern matching	crer-scan	Transcription factor binding sites	Cis-regulatory enriched regions (CRER)	Given a set of cis-regulatory elements (predicted sites, annotated sites, ChIP-seq peaks), detects regions presenting a significant enrichment in CRERs.
	matrix-scan (-quick)	Sequences + PSSMs	Matching positions in input sequences	Scans sequences with one or several PSSMs to identify instances of the corresponding motifs (putative sites). Supports a variety of background models (Bernoulli, Markov chains of any order).
Motif quality and comparisons (Matrix Tools)	*retrieve–matrix	Motif collection + motif name/ID	Motif (PSSM)	From a chosen motif collection (supported external database), extract the PSSMs specified by the provided name or identifier.
	matrix-quality	Motif (PSSM) + sequence set(s)	Score distribution statistics + ROC curves	Evaluates the quality of a PSSM by comparing score distributions obtained with this matrix in control sequence sets.
	compare-matrices	Two sets of PSSM	Similarity scores + matrix alignments	Compares two collections of PSSMs, and returns various similarity statistics + matrix alignments.
	matrix-clustering	One or several sets of PSSM	Clusters of matrices + similarity trees	Clusters similar PSSMs and builds consensus matrices for each cluster.
Comparative genomics	get-orthologs	Gene names + taxon	List of homologous genes with percentage of identity, alignment length, and e-value	Given a list of genes from a query organism, and a reference taxon, returns the orthologs of the query gene(s) in all the organisms belonging to the reference taxon.
	*get-orthologs-compara	Ensembl gene ids	Ensembl gene ids + homology relation information	Given a list of Ensembl stable gene IDs from one or more query organisms, returns orthologs (optionally paralogs and homologs). Relies on primary data from Ensembl Compara.
	footprint-discovery	Sequences	Conserved dyads + PSSM	Detects phylogenetic footprints by applying *dyad-analysis* in promoters of a set of orthologous genes.
	footprint-scan	Sequences + PSSM	Conserved motifs + binding sites	Scans promoters of orthologous genes with one or several PSSMs to detect enriched motifs and predict phylogenetically conserved target genes.
Regulatory variants (GeneticVariation Tools)	retrieve-variation-seq	Identifier of variations	Sequences of the variants	Given a set of IDs for genetic variations, returns the corresponding variants and their flanking sequences. The output file can be scanned with the tool *variation-scan*.

Table 1. Continued

Application	Program name	Input	Output	Description
	*variation-scan	Variant sequences	Regulatory variants	Scans variant sequences with PSSM and report variations that affect the binding score, in order to predict regulatory variants. Faster version with novel support for indels.
	*convert-variations	File with genetic variants	File with genetic variants in the specified format	Converts between different file formats that store genetic variation information. The most commonly used formats are: VCF and GVF, varBed format presents several advantages for scanning variations with matrices using *variation-scan*.
Visualisation	*feature-map2	Coordinates (relative or absolute)	Image depicting features over lines representing sequences	Generates a graphical map of features localized on one or several sequences. Several maps can be drawn in parallel, allowing to detect conserved positions. Exports in svg, png/jpeg.

This table presents a selection of key tools equipped with a Web interface. Connect to the RSAT Web site to obtain the complete list of available tools. Novel tools and major updates since the 2015 Web software issue are emphasized by an asterisk (*).

sites in JUNB keratinocyte ChIP-seq peaks. This use case introduces the usage of *retrieve-matrix*, *matrix-clustering*, *sequences-from-bed* and *feature-map2,* along with *matrix-scan.*

Use Case 3: Identify genetic variants associated with melanoma, which could affect AP1 binding sites. Exemplifies the usage of the complete refactored variation tools on the Metazoa server: *retrieve-variation-seq* and *variation-scan,* with *convert-variations*.

ACCESSING AND LEARNING TO USE RSAT

In addition to the public Web sites, RSAT can be remotely accessed via SOAP Web services. RSAT can also be used via Unix command-line, after installation of the suite on a local server or on a computer cloud, either from source code or with a Virtual Machine (on any operating system supporting VirtualBox, including Windows) (3).

To learn how to use the RSAT suite, extensive documentation material is available (3). The latest protocols (25,26) describe motif discovery in plant genomes, but the approach can be applied to any organisms supported on the other RSAT servers. Although the Web interfaces are being continually updated, most of our previously published protocols (10,15,27) are still usable to gain experience in understanding the underlying algorithms, choosing the relevant parameters and interpreting the results.

CONCLUSIONS

The RSAT suite is unique for its broad range of functionalities and supported organisms from all kingdoms. The main alternative is the MEME suite (28), which mainly focuses on motif analyses. Since the beginning of the project, RSAT strives to facilitate inter-connections with complementary programs (including MEME) and motif databases, thanks to a series of utility tools to convert alternative file formats (*convert-background-models*, *convert-features*, *convert-matrix*, etc.). Celebrating its 20th Anniversary, RSAT is gearing up for more interoperability with REST programmatic standards, better packaging with conda associated with a Docker image, and centralised documentation on GitHub.

DATA AVAILABILITY

All public RSAT servers are accessible from the RSAT portal at http://www.rsat.eu/. RSAT Web servers can be freely accessed by all users without login requirement.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are particularly thankful to the colleagues who help us installing and maintaining RSAT servers: Victor del Moral Chavez, Romualdo Zayas-Lagunas, Alfredo José Hernández Alvarez (Centro de Ciencias Genomicas, Cuernavaca, Mexico), Laboratorio Nacional de Visualización Científica Avanzada (Mexico) specially Luis Alberto Aguilar Bautista and Jair Garcia Sotelo, along with the ABims platform in Roscoff, France. We thank Najla Ksouri and Chesco Montardit for providing feedback on the installation of Prunus genomes; Olivier Sand, Matthieu Defrance and Céline Hernandez for regularly answering to RSAT-related questions; Gabriel Moreno-Hagelsieb for helping with the Prokaryote genomes. We thank Mauricio Guzman for designing all logos for RSAT and styling the figures. The testing squad of LIIGH trainees provided tremendous help: Karen J. Nuñez-Reza, Lucia Ramirez-Navarro, Molina-Aguilar Christian, Ana V. Altamirano, Castañeda-García C, Aldo Hernandez-Corchado, Omar Isaac García-Salinas. We especially acknowledge Julio Collado-Vides, who impulsed the project and supported it during the last 20 years.

FUNDING

French Government implemented by RENABI-IFB program [ANR-11-INSB-0013] to N.T.T.N.; ANR [ANR-14-

CE11-0006-02] to M.T.C. and D.T.; A.M.-R.'s laboratory is supported by a CONACYT grant [269449]; Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica – Universidad Nacional Autónoma de México (PAPIIT-UNAM) grant [IA206517]; M.T.-C., A.M.-R and D.T. further acknowledge SEP-CONACYT – ECOS-ANUIES support. J.A.C.-M. benefited from a PhD grant from the Ecole Doctorale des Sciences de la Vie et de la Santé, Aix-Marseille Université, and is supported by Norwegian Research Council [187615]; Helse Sør-Øst, and University of Oslo through the Centre for Molecular Medicine Norway (NCMM); B.C.M. was funded by Spanish MINECO [AGL2016-80967-R] and by Aix-Marseille Université as Chercheur Invité in 2015; C.D.R.-E.'s laboratory is supported by a Wellcome Trust Seed Award [204562/Z/16/Z]; PAPIIT-UNAM grant [IA200318]; R.O. is supported by a PhD studentship from CONACYT. Funding for open access charge: Agence Nationale de la Recherche.

Conflict of interest statement. None declared.

REFERENCES

1. van Helden,J., André,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.
2. van Helden,J., André,B. and Collado-Vides,J. (2000) A web site for the computational analysis of yeast regulatory sequences. *Yeast*, **16**, 177–187.
3. Medina-Rivera,A., Defrance,M., Sand,O., Herrmann,C., Castro-Mondragon,J.A., Delerce,J., Jaeger,S., Blanchet,C., Vincens,P., Caron,C. *et al.* (2015) RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res.*, **43**, W50–W56.
4. 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.
5. Thomas-Chollier,M., Sand,O., Turatsinze,J.-V., Janky,R. 's, Defrance,M., Vervisch,E., Brohée,S. and van Helden,J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, W119–W127.
6. van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.
7. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 563–567.
8. van Helden,J., Rios,A.F. and Collado-Vides,J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, **28**, 1808–1818.
9. van Helden,J., del Olmo,M. and Pérez-Ortín,J.E. (2000) Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.*, **28**, 1000–1010.
10. Turatsinze,J.-V., Thomas-Chollier,M., Defrance,M. and van Helden,J. (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat. Protoc.*, **3**, 1578–1588.
11. Janky,R. and van Helden,J. (2008) Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. *BMC Bioinformatics*, **9**, 37.
12. Brohée,S., Janky,R., Abdel-Sater,F., Vanderstocken,G., André,B. and van Helden,J. (2011) Unraveling networks of co-regulated genes on the sole basis of genome sequences. *Nucleic Acids Res.*, **39**, 6340–6358.
13. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
14. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
15. Thomas-Chollier,M., Herrmann,C., Defrance,M., Sand,O., Thieffry,D. and van Helden,J. (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.*, **40**, e31.
16. Thomas-Chollier,M., Darbo,E., Herrmann,C., Defrance,M., Thieffry,D. and van Helden,J. (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nat. Protoc.*, **7**, 1551–1568.
17. Sand,O., Thomas-Chollier,M. and van Helden,J. (2009) Retrieve-ensembl-seq: user-friendly and large-scale retrieval of single or multi-genome sequences from Ensembl. *Bioinformatics*, **25**, 2739–2740.
18. Defrance,M. and van Helden,J. (2009) info-gibbs: a motif discovery algorithm that directly optimizes information content during sampling. *Bioinformatics*, **25**, 2715–2722.
19. Medina-Rivera,A., Abreu-Goodger,C., Thomas-Chollier,M., Salgado,H., Collado-Vides,J. and van Helden,J. (2011) Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res.*, **39**, 808–824.
20. Castro-Mondragon,J.A., Jaeger,S., Thieffry,D., Thomas-Chollier,M. and van Helden,J. (2017) RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.*, **45**, e119.
21. Herrero,J., Muffato,M., Beal,K., Fitzgerald,S., Gordon,L., Pignatelli,M., Vilella,A.J., Searle,S.M.J., Amode,R., Brent,S. *et al.* (2016) Ensembl comparative genomics resources. *Database*, **2016**, baw053.
22. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D1284.
23. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
24. Contreras-Moreira,B. and Sebastian,A. (2016) FootprintDB: Analysis of plant Cis-Regulatory elements, transcription factors, and binding interfaces. *Methods Mol. Biol.*, **1482**, 259–277.
25. Castro-Mondragon,J.A., Rioualen,C., Contreras-Moreira,B. and van Helden,J. (2016) RSAT::Plants: Motif discovery in ChIP-Seq peaks of plant genomes. *Methods Mol. Biol.*, **1482**, 297–322.
26. Contreras-Moreira,B., Castro-Mondragon,J.A., Rioualen,C., Cantalapiedra,C.P. and van Helden,J. (2016) RSAT::Plants: Motif discovery within clusters of upstream sequences in plant genomes. *Methods Mol. Biol.*, **1482**, 279–295.
27. Defrance,M., Janky,R., Sand,O. and van Helden,J. (2008) Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat. Protoc.*, **3**, 1589–1603.
28. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME suite. *Nucleic Acids Res.*, **43**, W39–W49.