RSAT 2022: regulatory sequence analysis tools

Walter Santana-Garcia, Jaime Castro-Mondragon, Mónica Padilla-Gálvez, Nga Nguyen, Ana Elizondo-Salas, Najla Ksouri, François Gerbes, Denis Thieffry, Pierre Vincens, Bruno Contreras-Moreira, et al.

https://doi.org/10.1093/nar/gkac312

RSAT 2022: regulatory sequence analysis tools

Walter Santana-Garcia[1,†], Jaime A. Castro-Mondragon[2,†], Mónica Padilla-Gálvez[3], Nga Thi Thuy Nguyen[1], Ana Elizondo-Salas[3], Najla Ksouri[4], François Gerbes[5], Denis Thieffry[1], Pierre Vincens[1], Bruno Contreras-Moreira[4,*], Jacques van Helden[5,6,*], Morgane Thomas-Chollier[1,*] and Alejandra Medina-Rivera[3,*]

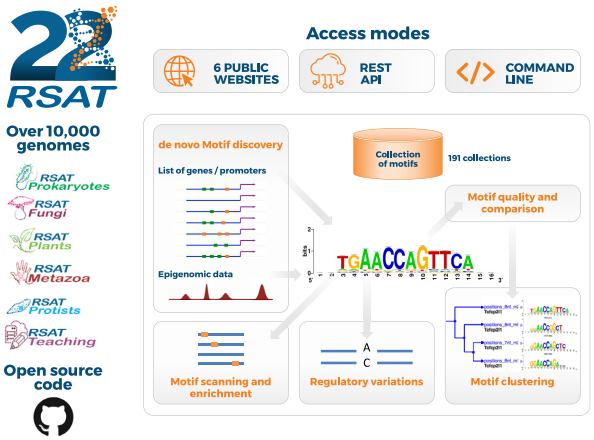
[1]Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, PSL Université Paris, 75005 Paris, France, [2]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway, [3]Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Campus Juriquilla, Blvd Juriquilla 3001, 76230 Santiago de Querétaro, México, [4]Estación Experimental de Aula Dei-CSIC, 50059 Zaragoza, Spain, [5]CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Evry, France and [6]Aix-Marseille Univ, INSERM UMR_S 1090, Lab Theory and Approaches of Genome Complexity (TAGC), F-13288 Marseille, France

Received March 03, 2022; Revised April 12, 2022; Editorial Decision April 13, 2022; Accepted April 20, 2022

ABSTRACT

RSAT (Regulatory Sequence Analysis Tools) enables the detection and the analysis of *cis*-regulatory elements in genomic sequences. This software suite performs (i) *de novo* motif discovery (including from genome-wide datasets like ChIP-seq/ATAC-seq) (ii) genomic sequences scanning with known motifs, (iii) motif analysis (quality assessment, comparisons and clustering), (iv) analysis of regulatory variations and (v) comparative genomics. RSAT comprises 50 tools. Six public Web servers (including a teaching server) are offered to meet the needs of different biological communities. RSAT philosophy and originality are: (i) a multi-modal access depending on the user needs, through web forms, command-line for local installation and programmatic web services, (ii) a support for virtually any genome (animals, bacteria, plants, totalizing over 10 000 genomes directly accessible). Since the 2018 NAR Web Software Issue, we have developed a large REST API, extended the support for additional genomes and external motif collections, enhanced some tools and Web forms, and developed a novel tool that builds or refine gene regulatory networks using motif scanning (network-interactions). The RSAT website provides extensive documentation, tutorials and published protocols. RSAT code is under open-source license and now hosted in GitHub. RSAT is available at http://www.rsat.eu/.

GRAPHICAL ABSTRACT



INTRODUCTION

The Regulatory Sequence Analysis Tools (RSAT) provides a wide range of bioinformatics programs enabling the analysis of genomic regulatory sequences in physiological and disease contexts. RSAT enables users to obtain genomic sequences and perform typical analyses, such as *de novo* motif discovery, or motif scanning to predict transcription factor (TF) binding sites (TFBSs). RSAT functionalities also include original analyses, such as motif quality evaluation,

*To whom correspondence should be addressed. Tel: +33 44 32 23 81; 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. Lab. Email: Jacques.van-Helden@univ-amu.fr
Correspondence may also be addressed to Bruno Contreras-Moreira. Tel: +34 976716089; Email: bcontreras@eead.csic.es
†The authors wish it to be known that, in their opinion, these authors should be regarded as Joint First Authors.

motif comparisons and clustering, detection and analysis of regulatory variants, building of control datasets and comparative genomics to discover motifs based on cross-species conservation. Altogether, the RSAT Web site gives access to 50 tools that can be used individually, or sequentially to perform more complex analyses. RSAT has been well-established since its initial development in 1998 (1,2). It has been regularly updated and extended with novel developments stimulated by advances in the field of regulatory genomics. We summarize here the main functionalities, and describe novelties since the previous NAR Web server issues (3–7).

RSAT FUNCTIONALITIES

RSAT tools have been individually described in the previous 2018 NAR update (3), with a historical perspective, as well as by applications (4). We summarize below the main functionalities ordered by data types to analyze, as a useful starting point for novice users (Figure 1). Pointers to the three use cases that exemplify how to combine the tools into routine analysis (3) are indicated.

Epigenomics datasets such as ChIP-seq or ATAC-seq peaks

Genome-wide datasets obtained from epigenomics experiments (e.g. ChIP-seq, ATAC-seq, ChIP-exo, DNaseI, Cut&Run, Cut&Tag) consists of genomic regions—known as peaks—that are likely bound by a given transcription factor (TF), or associated with open chromatin. The prevalent question is 'Which TF binding motifs can be detected in the peaks?'

The peaks can be analyzed with the user-friendly pipeline *peak-motifs* (5,8,9), which relies on *de novo* motif discovery to detect exceptional motifs in a set of sequences. *peak-motifs* runs multiple complementary algorithms [*oligo-analysis* (1), *dyad-analysis* (10), *position-analysis* (11) and *local-word-analysis* (8) that can all be used as independent tools], then compares the predicted motifs with annotated motif databases (*compare-matrices*), and finally predicts the positions of the putative transcription factor binding sites (TFBSs) within the peaks (*matrix-scan* (12) (Figure 2). Two datasets can be provided as input to enable differential analysis.

Alternatively, the peaks can be directly scanned with motifs (e.g. the discovered motifs, or from motif databases such as JASPAR (cf. 'Motifs represented as Position-Scoring Specific Matrices (PSSM) or consensus sequences')) to locate putative TFBSs (*dna-pattern* or *matrix-scan* (12)) or to predict potential cis-regulatory modules (*crer-scan* (3)). The tool *matrix-quality* can measure the enrichment of a specific motif within one or more peak datasets (13).

As input peaks must be provided as FASTA-formatted sequences, RSAT provides two tools to extract sequences from genome-wide peak datasets specified in BED-formated genomic coordinates (cf. 'Genomic coordinates as a BED file').

Control datasets can be built by selecting sequences at random positions from a given genome (*random-genome-fragments*), or by generating simulated sequences matching the size and composition of the peaks (*random-sequences*).

Lists of gene names or identifiers

Genome-wide datasets from transcriptomics experiments (e.g. microarrays, RNA-seq), as well as more targeted *in situ* hybridization experiments, typically results in a list of co-expressed genes. A frequent question is 'Which TFs may co-regulate the expression of these genes?' The typical analysis workflow consists in (i) retrieving sequences relative to these genes (e.g. promoter) and (ii) performing *de novo* motif discovery or motif scanning (cf. 'Epigenomics datasets such as ChIP-seq or ATAC-seq peaks'). Given a list of gene names or identifiers, *retrieve-sequences* extracts promoter sequences of locally-installed genomes, while *retrieve-ensembl-seq* (14) retrieves sequences of promoters or other specified features on-the-fly from Ensembl.

To support comparative genomics analyses, *retrieve-ensembl-seq* can also retrieve sequences from homologous genes. On the Plant server, the tool *get-orthologs-compara* additionally returns detailed information on homologous genes in a set of reference organisms, using precomputed Ensembl Compara data (15,16). On the Fungi and Prokaryotes servers, lists of orthologous genes can be obtained with *get-orthologs*. For the subsequent motif analysis step on these servers, *footprint-discovery* (17,18) and *footprint-scan* directly use cross-species conservation to detect putative regulatory signals in non-coding sequences (phylogenetic footprinting) (Figure 2).

Control datasets can be built by randomly selecting genes within a given genome with *random-gene-selection*. Use case 1 (3) combines *get-orthologs-compara, retrieve-sequences* and *matrix-scan* to predict TFBSs of VRN1 within the promoters of the FT1 gene in several plant genomes.

Motifs represented as Position-Scoring Specific Matrices (PSSM) or consensus sequences

Motifs represented as PSSMs or as consensus sequences may be obtained by *de novo* motif analysis, from databases such as JASPAR (19), or directly from the literature. Some typical questions are (i) 'Is the motif of good quality ?', (ii) ' Which sequences contain TFBS matching this motif ?', (iii) 'Does this motif resemble other motifs ?'.

First, *matrix-quality* (13) aims at assessing the quality of a PSSM on sequence datasets provided by the user, by comparing theoretical and empirical score distributions. Second, *matrix-scan* takes as input motifs to locate putative TFBSs in user-provided sequences (cf. 'Epigenomics datasets such as ChIP-seq or ATAC-seq peaks'). Third, *compare-matrices* compares two collections of matrices and returns various similarity statistics along with a PSSMs multi-pairwise alignment. *matrix-clustering* (20) regroups similar PSSMs into clusters, builds consensus PSSMs for each cluster and offers a dynamic visualization of aligned PSSMs. We applied *matrix-clustering* to regroup redundant matrices within and across motifs databases, in order to build the RSAT non-redundant motif collections for insects, plants and vertebrates (20). These collections are accessible with *retrieve-matrix* (3), which conveniently offers additional access to 187 external motifs collections, totalizing 454 524 motifs, all homogenized in TRANSFAC format (Supplementary Table S1). These collections include large databases such as JASPAR (19) and FootprintDB (21), as

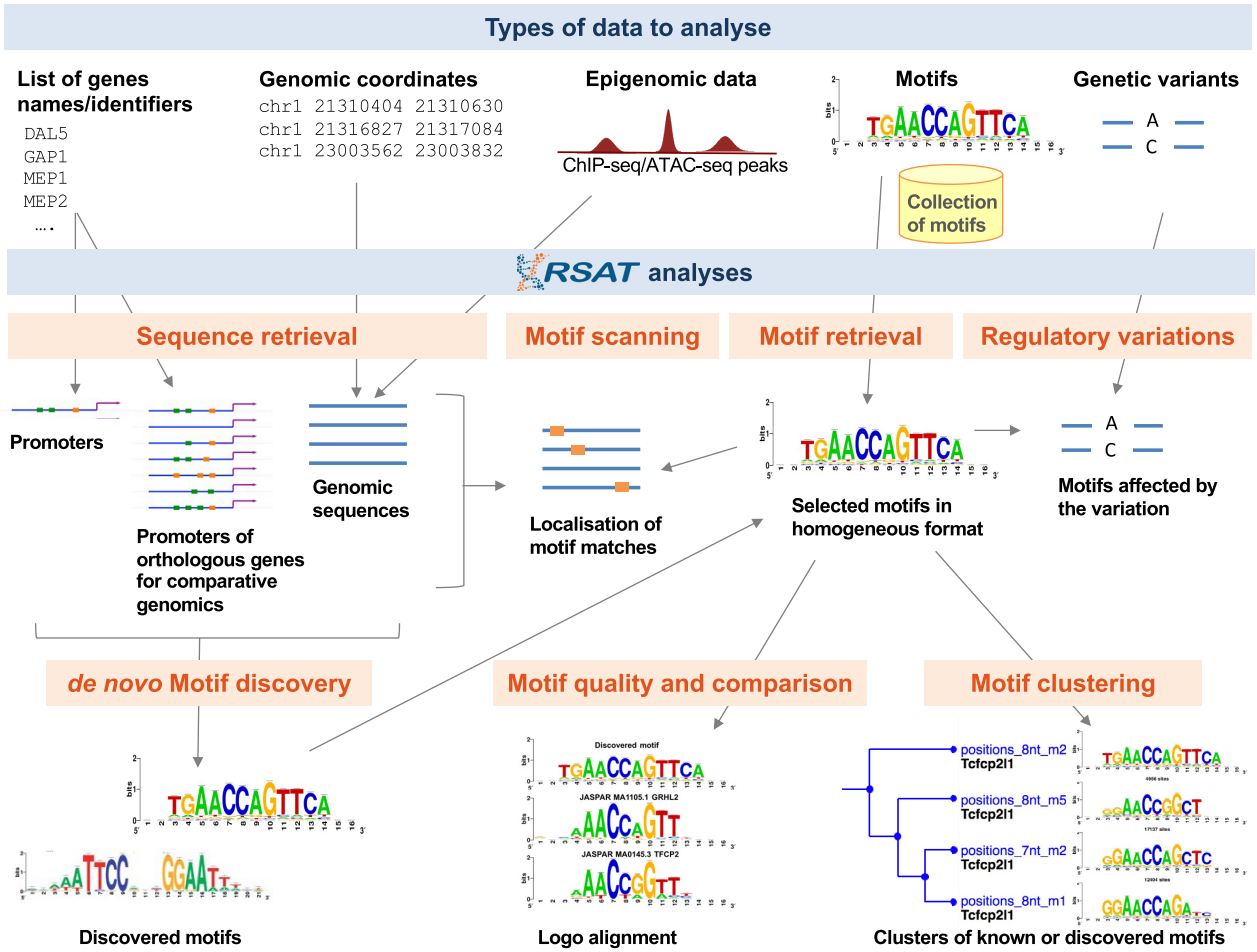


Figure 1. Overview of the main applications of RSAT, with associated input data types.

well as more specific ones such as ANISEED (22), RegulonDB (23) or RNA binding motifs, covering all kingdoms (Metazoa, Prokaryotes, Fungi, Plants). JASPAR (19) provides matrix-clustering results for each release, to provide information on the redundancy of motifs (https://jaspar.genereg.net/matrix-clusters/).

As there is no standard format for the PSSMs files, the tool *convert-matrix* performs interconversion between multiple motifs formats, and generates graphical representations of motifs in the form of logos. This allows users to focus on their scientific questions rather than formatting issues.

Control datasets can be built by generating permuted versions of PSSMs with *permute-matrix* or simulated matrix with *random-motif*.

Genomic coordinates as a BED file

Lists of features (e.g. peaks, predicted TFBSs) with their genomic coordinates are conventionally encoded in BED-formatted files (or GFF/GTF). The usual question is 'How to identify TFBSs within these regions?' The first step is to extract the corresponding genomic sequences; we provide user friendly tools with web interfaces to facilitate this task. Sequences can be automatically extracted from the UCSC genome browser with *fetch-sequences-from-UCSC* (3) or from locally -installed genomes with *sequences-from-BED/GFF/VCF*, which internally uses BEDTools and supports repeat-masking (24). Use case 2 (3) combines *retrieve-matrix*, *matrix-clustering*, *sequences-from-BED/GFF/VCF* and *matrix-scan* to generate a non-redundant AP1 motif from multiple annotated motifs, and predict TFBSs of AP1 within ChIP-seq peaks.

Lists of genetic variants as VCF files

Lists of genetic variants (SNPs, indels) can be retrieved from Genome-wide Association Studies (GWAS) and from databases such as Ensembl. A standard question is 'Which non-coding variants are affecting TF binding on cis-regulatory elements?' RSAT provides *variation-tools* (25), a series of programs to obtain information on individual variants, extract their flanking sequences, scan these flanking sequences with motif collections and predict which variants may affect TF binding.

Control datasets can be built by generating permuted versions of PSSMs with *permute-matrix*. Use case 3 (3) combines *convert-variations, retrieve-variation-seq* and *variation-*

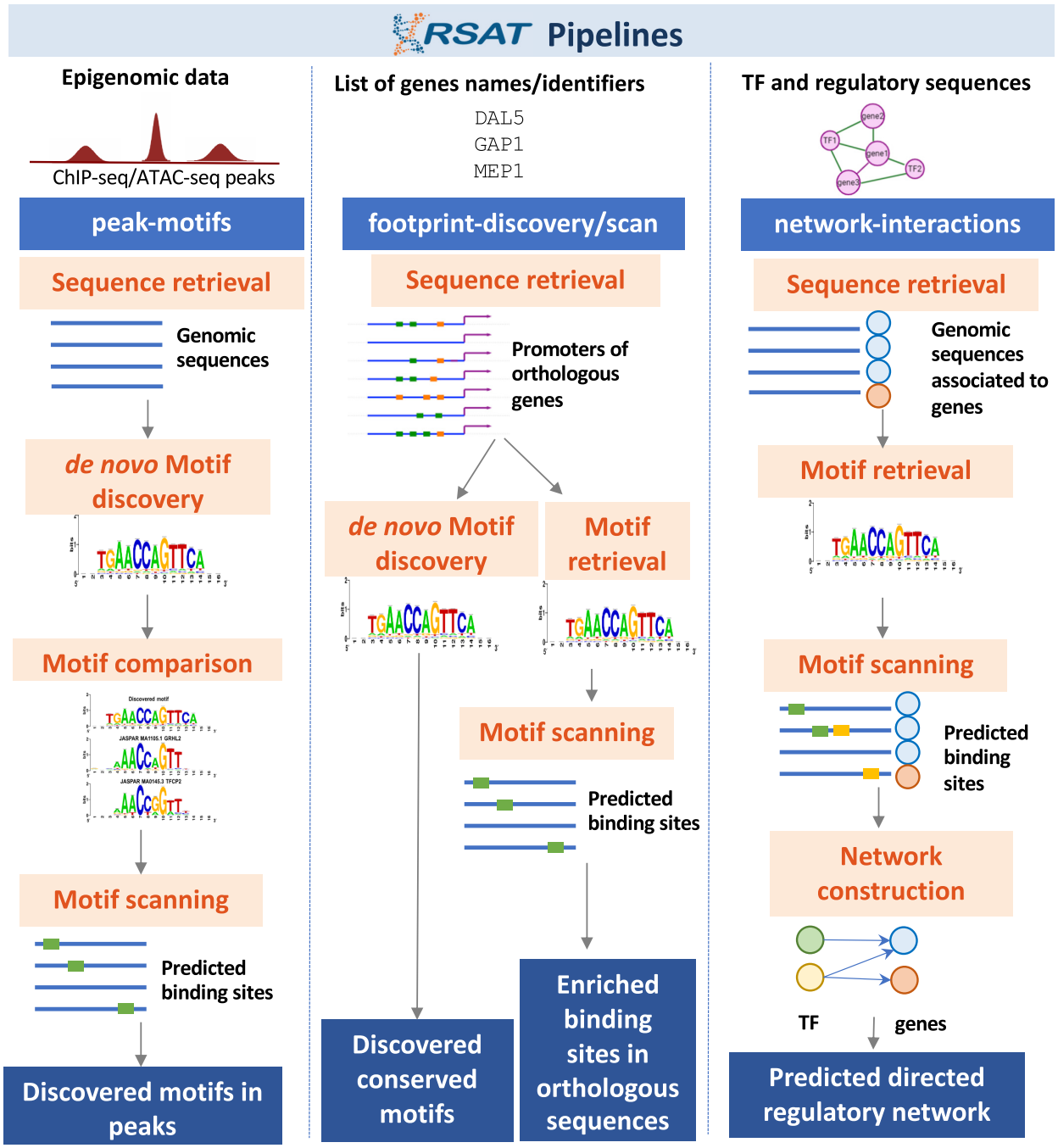


Figure 2. Three pipelines offering pre-defined combinations of RSAT tools (*peak-motifs*, *footprint-scan* and *footprint-discovery*, *network-interactions*).

scan on a VCF-formatted file specifying allelic variants detected in melanoma. It illustrates how scanning the surrounding sequences of the variants with the AP1 motif enables the identification of potential regulatory variants affecting AP1 binding.

RSAT 2022 NOVELTIES

RSAT locally installed organisms and motif collections

Since the last NAR Web server issue, we have further extended the number of supported organisms on the public servers, notably for Plants (+25 genomes) and Prokaryotes (+195 genomes). Some organisms were installed upon user request. As of February 2022, RSAT public servers support 10 076 locally installed genomes, including 9 646 Prokaryotes, 245 Fungi, 186 Protists, 91 Metazoa and 93 Plants. Besides, we have extended the number of external motif databases directly accessible in the common TRANSFAC format, from 50 to 187 external databases (cf. 'Motifs represented as Position-Scoring Specific Matrices (PSSM) or consensus sequences') (Supplementary Table S1). Some motif collections were added upon user request. Adding

new collections can now be made directly by a pull request on GitHub. All collections are freely downloadable to be used independently of RSAT (https://github.com/rsa-tools/motif_databases).

Users genome installation requests for servers are welcomed. In order to get a genome installed users have to contact the RSAT team through email 'rsat-contact@list01.biologie.ens.fr' with the information of the requested genome: organism name, genome version, source (*i.e.* NCBI, ENSEMBL) and url link to the genome data. In the case of motif collections, users can also request additions by providing: name, data, URL link and version information.

Furthermore, interested users can install genomes locally in their own RSAT instances. The documentation at https://rsa-tools.github.io/managing-RSAT contains detailed manuals to install genomes from different sources, such as RSAT servers, Ensembl, NCBI and from original FASTA and GTF data files.

Programmatic REST API access

Our programmatic SOAP/WSDL access is being replaced by the increasingly popular Web service REST API. It provides access to a large set of 49 tools of the RSAT suite. The REST API has been developed with the flask library; its documentation is generated with Swagger UI. Example clients in Python have been written to further help users using this API.

Updated web interface and tools

Some tools are highly parameterisable, thereby complexifying the corresponding Web forms. We have started to redesign these forms to simplify usage: we are now better separating the mandatory inputs/parameters from the optional ones (see *retrieve-sequence*, *matrix-clustering* and *network-interactions*). Several tools have been updated with additional functionalities or increased efficiency. This is the case of *variation-tools* (cf. 'Lists of genetic variants as VCF files'), for which haplotype scanning has been improved to assess the regulatory effect in TFBSs of haplotypes with large number of variants (SNPs and indels) in Metazoa and Plants.

Prediction of TF-gene interactions to build and refine gene regulatory networks

Many efforts have been made to infer gene regulatory networks (GRN) from transcriptomic data, with approaches based on coexpression, orthology or sequence motifs (26), but there is no consensus on a single best method. To further improve the inferred GRNs, it is common to apply motif scanning (pattern-matching) as a second step upon inferred interactions. We introduce *network-interactions*, a new user-friendly GRN reconstruction pipeline based on pattern-matching, which can help refine GRNs generated by other tools (Figure 2). It takes as input two lists: (i) the TFs of interests specified as a list of TF names and (ii) a list of genomic regions associated with gene names (typically promoter/enhancer regions of genes) provided as BED coordinates. A seed network, previously generated from other tools (i.e. based on co-expression), can optionally be provided. *network-interactions* runs *matrix-scan* using one of the motif collections available in RSAT (default is JASPAR's 2022 vertebrates motif collection (19)) to predict TF-gene interactions. *network-interactions* thereby generates several networks: (i) a complete network for all TF-gene interactions, (ii) another network focusing on TF-TF interactions, (iii) one with 3-step TFs indirect interactions (TF-TF-gene) and (iv) when provided with an input GRN, the overlap and the complements between the input network and the network generated by *network-interactions*, where the overlap includes the putative TF binding information. This novel tool extends RSAT's suite and offers a straightforward and flexible method to expand and refine GRNs.

RSAT source code on GitHub and Docker container

The RSAT source code, under AGPL-3.0 open-source license, has been transferred to GitHub, to stimulate community-wise participation in its development: https://github.com/rsa-tools. Additional RSAT documentation is available there as well. A Docker container has been built to analyze the promoters of coexpressed genes in plants (27): https://github.com/eead-csic-compbio/coexpression_motif_discovery.

Learning to use RSAT

In addition to the above-mentioned use cases, RSAT provides extensive documentation, tutorials and published protocols (4). To target non-expert users, including biologists and biomedical practitioners, the main tools are accessible through web forms with DEMO buttons and tutorials. The latest protocols (28,29) and application (27) focuses on motif discovery in plant genomes; the described approaches can generally be applied to other organisms. Most of our previously published protocols (9,12,30) are still relevant to learn about the underlying algorithms, choosing the relevant parameters and interpreting the results, despite updates in the Web interfaces. Users may also contact us via email or via our Twitter account @RSATools.

CONCLUSIONS

Compared to alternative programs, RSAT is unique for its wide range of functionalities, extensive motifs collections and >10 000 supported organisms from all kingdoms. The main alternatives are the MEME suite (31), which mainly focuses on motif analyses, and HOMER (32), which primarily focuses on motif discovery. Deep-learning methods are more focused in discovering context-specific TFBS, whereas RSAT aims at providing a complete environment for motif analysis. We aim for RSAT to be usable in combination with other programs (including MEME and HOMER); RSAT thus offers several file format conversion utility tools (*convert-matrix*, *convert-background-models*, *convert-features*, ...). After 20 years of existence, RSAT remains one of the most used tools in regulatory genomics. Looking forward, we aim at (i) continuing to enhance the suite in particular to cope with the challenges

posed by single cell technologies in terms of data analysis efficiency, and (ii) continuing to ensure long-term maintenance, with packaging in conda, a non-plant docker container and continuous integration on GitHub.

DATA AVAILABILITY

RSAT public 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. For bioinformatician users, RSAT is accessible (i) as a command-line suite for installation on a local server or on a computer cloud, from its source code https://github.com/rsa-tools, or (ii) via the REST API web programmatic access. RSAT is part of the Service Delivery Plan of the Elixir-France node (European distributed infrastructure for life-science information): https://elixir-europe.org/services/list?field_scientific_domain_tid=All&field_elixir_badge_tid=All&field_type_of_service_tid=All&field_elixir_node_target_id=981&combine=.

RSAT code and documentation is available through GitHub https://github.com/rsa-tools. The Docker container for plants is located at: https://github.com/eead-csic-compbio/coexpression_motif_discovery. Motif collections can be found at https://github.com/rsa-tools/motif_databases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

RSAT is managed by an international team (France, Mexico, Spain, Norway). We are particularly thankful to the colleagues who help us installing and maintaining RSAT servers: Victor del Moral Chavez, Alfredo José Hernández Alvarez (Centro de Ciencias Genomicas, Cuernavaca, Mexico), Luis Alberto Aguilar Bautista and Jair Garcia Sotelo (Laboratorio Nacional de Visualización Científica Avanzada, Mexico), Aurora Martín (Estación Experimental de Aula dei, Zaragoza, Spain), along with the ABIMS platform in Roscoff, France and the mutualized task force of the Institut Français de Bioinformatique. We also thank Olivier Sand and Matthieu Defrance for regularly answering RSAT-related questions. We thank Ieva Rauluseviciute for finding and reporting bugs in RSAT programs. We thank Mauricio Guzman for designing all logos for RSAT and styling the figures. The testing squad of LIIGH trainees provided tremendous help: Ana Laura Hernández-Ledesma, Juan Manuel Martinez-Villalobos, Paula R. Reyes-Pérez. We especially acknowledge Julio Collado-Vides, who impulsed the project and supported it during the last 25 years.

FUNDING

Institut Universitaire de France (to M.T.C., N.T.T.N.); ANR [ANR-14-CE11-0006-01 to M.T.C., N.T.T.N.]; ITMO Cancer [20CM114-00 to D.T., W.S.G.]; CONACYT [11311 to A.M.R., M.P.G.]; Universidad Nacional Autónoma de México PAPIIT (UNAM) [IA203021]; Spanish State Research Agency, EUROPEAN REGIONAL DEVELOPMENT FUND (FEDER) [AGL2017-83358-R to N.K.; AEI/FEDER, UE]; Gobierno de Aragón [A08_17R, A09_20R, Phd Contract to N.K.]; PRIMA [PCI2019-103526 to B.C.M.; Programación Conjunta Internacional, Programa Estatal de l+D+i Orientada a los Retos de la Sociedad]; Erasmus+ (to N.K.); Norwegian Research Council [288404 to J.A.C.M.]; Institut Français de Bioinformatique (IFB) [ANR-11-INBS-0013]; SEP-CONACYT-ECOS-ANUIES [291235 to A.M.R., M.T.C. and D.T.] . Funding for open access charge: Institut Universitaire de France.
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 Chichester Engl.*, **16**, 177–187
3. Nguyen,N.T.T., Contreras-Moreira,B., Castro-Mondragon,J.A., Santana-Garcia,W., Ossio,R., Robles-Espinoza,C.D., Bahin,M., Collombet,S., Vincens,P., Thieffry,D. *et al.* (2018) RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res.*, **46**, W209–W214.
4. 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.
5. 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.
6. Thomas-Chollier,M., Sand,O., Turatsinze,J.-V., Janky,R., Defrance,M., Vervisch,E., Brohée,S. and van Helden,J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, W119–W127.
7. van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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. *Bioinforma. Oxf. Engl.*, **25**, 2739–2740.
15. 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 J. Biol. Databases Curation*, **2016**, baw053.

16. Yates,A.D., Allen,J., Amode,R.M., Azov,A.G., Barba,M., Becerra,A., Bhai,J., Campbell,L.I., Carbajo Martinez,M., Chakiachvili,M. *et al.* (2022) Ensembl genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Res.*, **50**, D996–D1003.
17. Janky,R. and van Helden,J. (2008) Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. *BMC Bioinf.*, **9**, 37.
18. 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.
19. Castro-Mondragon,J.A., Riudavets-Puig,R., Rauluseviciute,I., Berhanu Lemma,R., Turchi,L., Blanc-Mathieu,R., Lucas,J., Boddie,P., Khan,A., Manosalva Pérez,N. *et al.* (2022) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **50**, D165–D173.
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. Contreras-Moreira,B. and Sebastian,A. (2016) FootprintDB: analysis of plant cis-regulatory elements, transcription factors, and binding interfaces. *Methods Mol. Biol. Clifton NJ*, **1482**, 259–277.
22. Brozovic,M., Dantec,C., Dardaillon,J., Dauga,D., Faure,E., Gineste,M., Louis,A., Naville,M., Nitta,K.R., Piette,J. *et al.* (2018) ANISEED 2017: extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets. *Nucleic Acids Res.*, **46**, D718–D725.
23. Santos-Zavaleta,A., Salgado,H., Gama-Castro,S., Sánchez-Pérez,M., Gómez-Romero,L., Ledezma-Tejeida,D., García-Sotelo,J.S., Alquicira-Hernández,K., Muñiz-Rascado,L.J., Peña-Loredo,P. *et al.* (2019) RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in e. coli K-12. *Nucleic Acids Res.*, **47**, D212–D220.
24. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
25. Santana-Garcia,W., Rocha-Acevedo,M., Ramirez-Navarro,L., Mbouamboua,Y., Thieffry,D., Thomas-Chollier,M., Contreras-Moreira,B., van Helden,J. and Medina-Rivera,A. (2019) RSAT variation-tools: an accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding. *Comput. Struct. Biotechnol. J.*, **17**, 1415–1428.
26. Mercatelli,D., Scalambra,L., Triboli,L., Ray,F. and Giorgi,F.M. (2020) Gene regulatory network inference resources: a practical overview. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.*, **1863**, 194430.
27. Ksouri,N., Castro-Mondragón,J.A., Montardit-Tarda,F., van Helden,J., Contreras-Moreira,B. and Gogorcena,Y. (2021) Tuning promoter boundaries improves regulatory motif discovery in nonmodel plants: the peach example. *Plant Physiol.*, **185**, 1242–1258.
28. 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. Clifton NJ*, **1482**, 279–295.
29. 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. Clifton NJ*, **1482**, 297–322.
30. 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.
31. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME suite. *Nucleic Acids Res.*, **43**, W39–W49.
32. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol. Cell*, **38**, 576–589.