Published online 21 February 2017

Nucleic Acids Research, 2017, Vol. 45, No. 10 ***e91***
doi: 10.1093/nar/gkx120

RCAS: an RNA centric annotation system for transcriptome-wide regions of interest

Bora Uyar[1,†], Dilmurat Yusuf[1,†], Ricardo Wurmus[1,†], Nikolaus Rajewsky[2], Uwe Ohler[3] and Altuna Akalin[1,*]

[1]Bioinformatics Platform, [2]Systems Biology of Gene Regulatory Elements and [3]Computational Regulatory Genomics, Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, 13092 Berlin, Germany

Received October 20, 2016; Revised January 17, 2017; Editorial Decision February 08, 2017; Accepted February 11, 2017

ABSTRACT

In the field of RNA, the technologies for studying the transcriptome have created a tremendous potential for deciphering the puzzles of the RNA biology. Along with the excitement, the unprecedented volume of RNA related omics data is creating great challenges in bioinformatics analyses. Here, we present the RNA Centric Annotation System (RCAS), an R package, which is designed to ease the process of creating gene-centric annotations and analysis for the genomic regions of interest obtained from various RNA-based omics technologies. The design of RCAS is modular, which enables flexible usage and convenient integration with other bioinformatics workflows. RCAS is an R/Bioconductor package but we also created graphical user interfaces including a Galaxy wrapper and a stand-alone web service. The application of RCAS on published datasets shows that RCAS is not only able to reproduce published findings but also helps generate novel knowledge and hypotheses. The metagene profiles, gene-centric annotation, motif analysis and gene-set analysis provided by RCAS provide contextual knowledge which is necessary for understanding the functional aspects of different biological events that involve RNAs. In addition, the array of different interfaces and deployment options adds the convenience of use for different levels of users. RCAS is available at http://bioconductor.org/packages/release/bioc/html/RCAS.html and http://rcas.mdc-berlin.de.

INTRODUCTION

In one way or another, RNA plays a role in nearly all cellular processes—from the translation of genetic information to the regulation of gene activities (1–3). Anomaly in RNA activity can lead to pathological conditions in organisms (4,5). In recent years, the advance of deep sequencing technologies has provided powerful means for studying the transcriptome, the full span of RNA molecules expressed by an organism. These technologies provide an unprecedented look into functions, regulation, and diversity of RNA molecules. We can now assess the abundance of transcripts and identify previously unknown transcripts with RNA-seq (6). Apart from that, deep-sequencing based techniques provide transcriptome-wide information on many different layers of mRNA regulation and processing. For example, the precise details about transcription initiation, termination, splicing and translation dynamics can be measured using a variety of sequencing techniques. DeepCAGE can give information on the precise usage of transcription start sites (TSSs) (7), whereas NET-Seq can be used for transcription end sites (TESs) (8). Ribo-Seq can be used for ribosome profiling to monitor the translation process (9) and TRAP-Seq for the detection of translating RNAs (10). Additionally, it is critical to understand how RNA is processed, trafficked and localized via RNA-binding proteins. With PAR-CLIP and other CLIP based techniques, one can detect the transcriptome-wide binding sites of the RNA binding proteins (RBPs) (11). One can also survey RNA modifications using deep sequencing, which is also thought to be important for RNA processing and localization. m^6A-seq provides transcriptome-wide location of N^6-methyladenosine (m^6A) sites (12) and m^1A-seq can provide N^1-methyladenosine (m^1A) locations (13). Finally, RNA–RNA interactions and RNA-secondary structure can be mapped via deep sequencing based techniques. ChIRP-seq can locate the genomic binding sites of non-coding RNAs (ncRNAs) (14), CLASH-Seq can map RNA–RNA interactions (15). PARE-Seq can be used for mapping miRNA cleavage sites and degrading RNA (16), and SHAPE-Seq provides RNA structural information (17).

This list is incomplete and will continue to grow with the addition of new techniques and the variations of existing ones. All of these techniques require specialized

[*]To whom correspondence should be addressed. Email: altuna.akalin@mdc-berlin.de
[†]Equal contributions.

analysis and processing workflows. Although not identical and differing in many key aspects, the workflows dedicated to processing these techniques are similar in their output since they usually provide transcriptome-wide regions of interest. For example, PAR-CLIP analysis will provide transcriptome-wide enriched regions for RBP-binding, m^6A-seq and m^1A-seq analysis will provide locations for RNA-modifications, and deepCAGE analysis will provide regions of transcription initiation. The regions of interest obtained from the initial analysis will almost always be further processed by a universal downstream analysis approach. Here, we present 'RNA Centric Annotation System' (RCAS), a tool that performs overlap operations between the regions of interest and the genomic features, producing in-depth annotation summaries with respect to exons, introns, coding sequences (CDSs), 5′/3′ untranslated regions (UTRs), exon–intron boundaries, promoter regions, and whole transcripts. Moreover, RCAS carries out functional annotations for enriched gene sets and GO terms. In addition, RCAS is capable of detecting specific sequence motifs enriched in the targeted regions of the transcriptome. The output of RCAS is a dynamic HTML file embedding interactive figures and tables which are ready for publication purposes. RCAS is now part of the Bioconductor R package library (18). The Guix (19) and Conda (http://conda.pydata.org/docs/index.html) packages are available as alternative means of deployment. For the non-programmer users who are accustomed to graphical user interfaces (GUI), a Galaxy (20) wrapper and a web service have been developed. In essence, RCAS is designed to ease the process of deriving biological insights from large collections of transcriptome-wide regions of interest, offering an automated solution for annotation, summary and functional analysis of the RNAs. The usability of RCAS is ensured by its modular design, user-friendly interactive figures and tables, extensive documentation and testing on the Bioconductor repository, the availability of different packages (Guix and Conda), and by the user interfaces of both command-line and GUI (RCAS web service and Galaxy integration).

To demonstrate the performance and utility of RCAS, we employed four use cases that include the datasets from three high-impact publications (11,13,21) and one in-house dataset which is related to high occupancy target (HOT) regions of RNA-binding proteins (RBPs). The published genomic regions are from the studies using PAR-CLIP (11), RNA-Seq/deepCAGE (6,7) and m^1A-seq (13), respectively. The resulting profiles show that RCAS is not only able to reproduce the published findings but also to generate novel knowledge. The information generated by RCAS can provide a context for understanding different biological events involving RNAs.

MATERIALS AND METHODS

The RCAS R package employs different R functions to perform annotation summarization, GO term and gene set enrichment analysis, and *de novo* sequence motif discovery. For the most up-to-date documentation and demonstration of RCAS functionality, please refer to the Bioconductor repository at http://bioconductor.org/packages/release/bioc/html/RCAS.html for the release version, and at http://bioconductor.org/packages/devel/bioc/html/RCAS.html for the development version.

In this section, we describe the methods used to generate an HTML report using the main RCAS function `RCAS::runReport` (as of RCAS version 1.1.1 available on the development branch of Bioconductor 3.5). The `RCAS::runReport` function utilizes all other RCAS functions in a template 'RMarkdown' script (22) to generate a dynamic HTML report. The scheme of `RCAS::runReport` pipeline to produce a complete HTML report is illustrated in Figure 1. With this function, the users are provided with options to switch on/off individual components. The major steps of the default workflow consists of (i) importing and processing of inputs, (ii) plotting genomic annotation summaries and coverage profiles, (iii) finding enriched motifs among the input query regions, (iv) GO term (23) analysis of targeted transcripts, (v) gene set enrichment analysis of the targeted transcripts.The figures in the final report are generated using the plotly R library (24), thus the users can interactively modify the display of the figures according to their choices and export a snapshot of the modified figures. The tables are generated using the DT R library (25) and are also interactive. For instance, the user can search, filter, and sort the tables; export the tables to Excel, CSV or PDF, and copy or print the contents of the interactively modified table. Thus, besides providing publication quality figures and tables, RCAS can offer the users tools to explore the data further, which may help form new hypotheses. The raw data used to create all the figures and tables can be printed to the current working directory if the printProcessedTables argument in the `RCAS::runReport` function is set to TRUE. Sample HTML reports generated using real experimental data from published datasets can be found in the supplementary files (Supplementary Files 1–6). Below is a more detailed explanation of how the report HTML file is generated and how each figure and table in the report is calculated and plotted.

Importing and processing of inputs

The input BED file (containing query genomic regions) and input GTF file (containing reference genomic regions) are imported using the rtracklayer library (26). Then, the GTF file is further processed using the GenomicFeatures library (27) in order to extract the genomic regions of transcript features such as (i) transcripts (using the `GenomicFeatures::transcripts` function), (ii) promoters (using the `GenomicFeatures::promoters` function) where a promoter region is defined as the region encompassing from 2000 bp upstream and 200 bp downstream of the TSS, (iii) exons (using the `GenomicFeatures::exonsBy` function where exons are grouped by 'transcripts'), (iv) introns (using the `GenomicFeatures::intronsByTranscript` function), (v) CDSs (using the `GenomicFeatures::cdsBy` function where coding exons are grouped by transcripts), (vi) 5′/3′ UTRs (using the `GenomicFeatures::fiveUTRsByTranscript` and `GenomicFeatures::threeUTRsByTranscript` functions, respec-

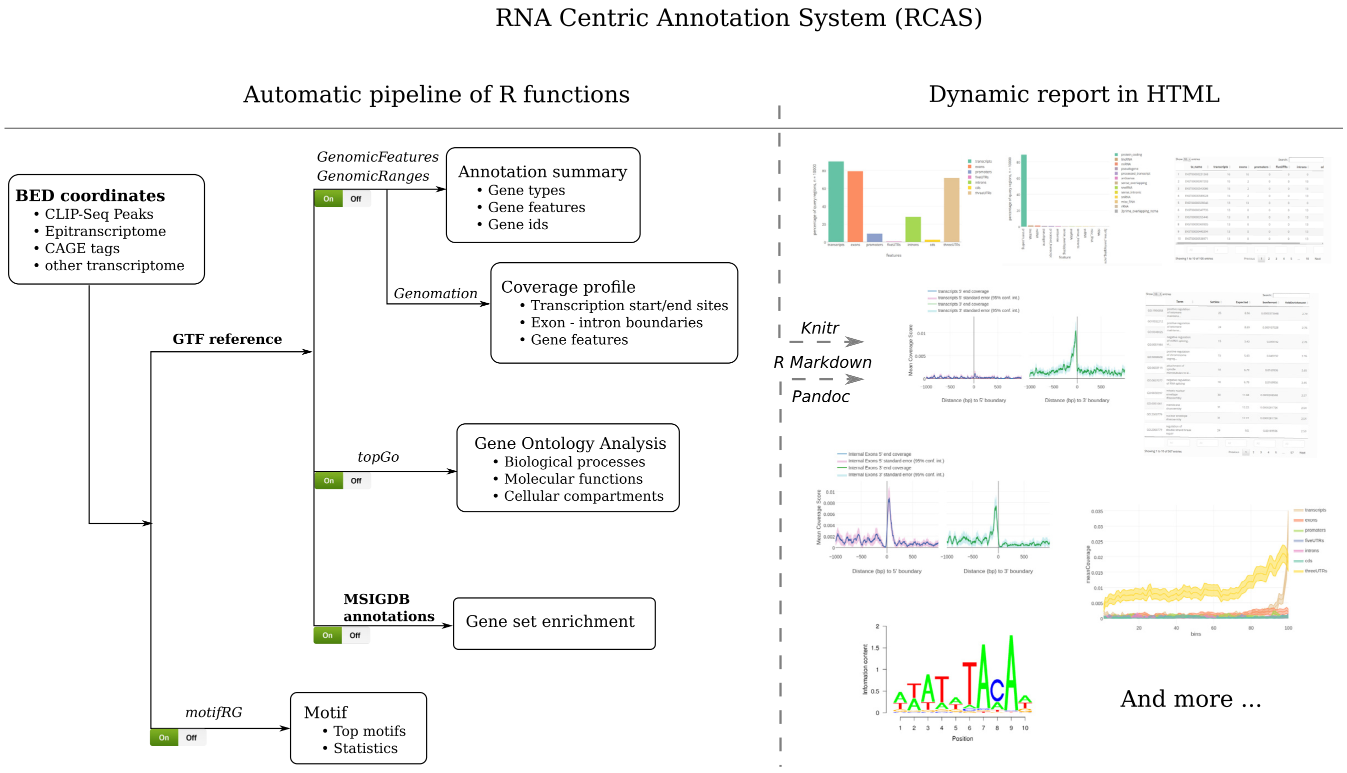


Figure 1. The scheme of RCAS. RCAS is composed of several components: annotation summary, coverage profile, GO term analysis, gene set enrichment and motif discovery. The individual components employ various R functions. The modular design provides options for switching on/off the individual components. The output of RCAS is a dynamic HTML consisting of interactive figures and tables, which are downloadable and ready for the purpose of publication.

tively). Functions to retrieve the UTR regions return GRanges objects which contain one or more exons (thus zero or more introns). Such UTRs with multiple genomic intervals (one interval for each exon) are reduced to a single interval by using the `GenomicRanges::range` function to find the 5′ most start and the 3′ most end of the UTR. Currently, the supported genome versions are human hg19, mouse mm9, worm ce10 and fly dm3.

Summary of genomic annotations

RCAS employs the GenomicRanges library (27) to perform overlap operations between the query regions in the BED file and the reference genomic features in the GTF file. The number of query regions that overlap different kinds of gene features are counted. The first plot in the HTML report output of the `RCAS::runReport` function displays the distribution of query regions across gene features. In the figure, the 'x' axis denotes the types of gene features included in the analysis and the 'y' axis denotes the percentage of query regions (out of the total number of query regions denoted with '*n*') that overlap at least one genomic interval that host the corresponding feature. Notice that the sum of the percentage values for different features do not add up to 100%, because some query regions may overlap multiple kinds of features. For each transcript in the GTF file, the number of query regions overlapping the different types of transcript features are counted and sorted by total number of overlaps in the whole transcripts. An interactive table of top 100 transcripts is provided. In addition, the number of query regions that overlap different kinds of gene types are counted. In the resulting figure, the 'x' axis denotes the types of genes included in the analysis and the 'y' axis denotes the percentage of query regions (out of total number of query regions denoted with 'n') that overlap at least one genomic interval that host the corresponding gene type.

Producing coverage profiles of query regions

RCAS employs the genomation library's (28) ScoreMatrix and ScoreMatrixBin functions to generate coverage profiles. A coverage profile in the context of RCAS means the depth of coverage/signal (i.e. the number of query regions) observed at a given genomic segment or a collection of overlaid genomic segments. Coverage profiles are useful for observing the relative location of the query regions with respect to target regions. For instance, an RNA binding protein's (RBP) preferred binding location and the signal strength at the UTR regions relative to the TSSs/TESs might be important to know to understand how the RBP might be regulating its target mRNAs; or an increased signal at exon - intron boundaries may suggest roles for an RBP in alternative splicing regulation. There are two main types of such coverage profiles:

1) A coverage profile of query regions at/around feature boundaries (flanking regions at start and end positions of the genomic segments of the given fea-

tures): in the report HTML output, coverage profiles at the TSSs/TESs and the coding exon - intron boundaries are provided. However, in principle, this type of coverage profile can be obtained for any other kind of feature or any collection of genomic segments using the `RCAS::getFeatureBoundaryCoverage` and `RCAS::getFeatureBoundaryCoverageBin`. The getFeatureBoundaryCoverage function extracts the flanking regions of 5′ and 3′ boundaries of a given set of genomic features and computes the per-base coverage of query regions across these boundaries. On the other hand, getFeatureBoundaryCoverageBin extracts the flanking regions of 5′ and 3′ boundaries of a given set of genomic features, splits them into 100 equally sized bins and computes the per-bin coverage of query regions across these boundaries.

2) A coverage profile of query regions across the length of different gene features: To generate this plot, the query regions are overlaid with the genomic regions of features. Each entry corresponding to a feature is divided into 100 bins of equal length and for each bin the number of query regions that cover the corresponding bin is counted. Features shorter than 100 bp are excluded, because such features cannot be divided into 100 equal integer-sized bins. Thus, a coverage profile is obtained based on the distribution of the query regions. The strandedness of the features are taken into account. The coverage profile is plotted in the 5′ to 3′ direction.

When generating a report using the runReport function, the target genomic features are by default randomly downsampled to 10 000 genomic intervals in order to speed up the process of obtaining coverage profiles. Although this number is fixed when generating reports, it can be tuned in the RCAS functions calculateCoverageProfile, calculateCoverageProfileList, or getFeatureBoundaryCoverage by changing the 'sampleN' argument passed to these functions.

Motif analysis

The genomic sequences of the input query regions are extracted using the BSgenome-associated R libraries (29) and a randomly down-sampled set of these sequences (randomly selected 10 000 intervals) is used as input to the R package motifRG (30). As motifRG is a discriminatory motif discovery tool, besides the query region sequences, a background set of sequences with the same length distribution and similar sequence content is used as input (by choosing the sequences from the neighborhood of the query regions). To enable motif discovery for query regions that are too short (<15 bp long), those query regions are resized to 15 bp using the `rtracklayer::resize(fix = 'center')` function. The output consists of a figure of the motif logos and a table displaying the statistics of motif discovery.

GO term analysis

Based on the overlap operations between the query regions and the reference transcript features, the list of gene ids corresponding to the target transcripts is obtained. This list of genes is used as input to the R package topGO (31) to discover enriched GO terms for biological processes, molecular functions, and cellular components. The results are reported in interactive tables containing P-value calculations of the enriched terms based on classical Fisher's exact test and filtered according to adjusted P-values (cutoff of adjusted P-value = 0.1) based on multiple testing correction using the Benjamini–Hochberg method, and the fold increase relative to the background. The GO term enrichment analysis is not yet supported for the worm ce10 genome version.

Gene set enrichment analysis

If the user has provided a gene set annotation collection downloaded from the Molecular Signatures Database (32), where each gene in the collection is represented by Entrez gene ids, a gene-set enrichment analysis is carried out (GMT format file contains two columns: the first column contains the name of the gene set and the second column contains a comma-separated list of gene ids). Based on the overlap operations between the query regions and the target transcript features, the list of gene ids corresponding to the target transcripts is obtained. For each gene set, a 2 × 2 contingency table is constructed for the genes overlapping the query regions and the background set of genes (the whole list of genes in the given gene set). A Fisher's exact test is applied to find out if the targeted genes are enriched in that gene set relative to the background. The results are reported in interactive tables containing P-value calculations of the enriched terms based on classical Fisher's exact test, adjusted P-values based on multiple testing correction using the Bonferroni method, Benjamini-Hochberg (BH) method, and sorted in decreasing order according to the fold change relative to the background. The results are filtered according to adjusted P-values (BH > 0.1).

If the `RCAS::runReport` function is initiated for a species other than human ('hg19'), the provided gene sets from the MSIGDB are mapped to the corresponding species via orthology relationships retrieved from the Ensembl database (33) via the biomaRt library (34), for which an internet connection must be available. The gene-set enrichment analysis is not yet supported for the worm ce10 genome version.

Use cases

For annotation summaries, the GTF file for *Homo sapiens* (genome version GRCh37 (hg19)) was downloaded from the Ensembl database (35). For gene set enrichment analysis, we used the curated pathways gene set collection (containing 1330 gene sets from various pathway databases such as KEGG (36), BioCarta (37), PID (38) and Reactome (39)) downloaded from the Molecular Signatures Database (40,41). We provide the download links for the respective reference files and BED files at the github repository of RCAS (https://github.com/BIMSBbioinfo/RCAS). For the four use cases, we enabled all components of RCAS to generate the complete profiles.

Use case 1: CLIP analysis. The PAR-CLIP experiment's (11) aim was to study several RBPs, some of which are Pumilio 2 (PUM2), Quaking (QKI) and insulin-like growth factor 2 mRNA-binding proteins 1, 2 and 3 (IGF2BP1-3).

The corresponding BED files were obtained from the doRiNA database (42).

Use case 2: CAGE analysis. The deepCAGE experiment (21) showed the association of tiny RNAs (tiRNAs) with TSSs in animals. The corresponding BED file, which contains genomic coordinates according to the human genome hg18, was downloaded from the FANTOM consortium (43). Since RCAS currently does not yet support hg18, using Utility (Batch Coordinate Conversion) (44), we converted the genomic coordinates from hg18 to hg19.

Use case 3: epitranscriptome analysis. The m^{1}A-seq experiment's (13) aim was to study N^1-methyladenosine methylome in eukaryotic messenger RNA. We obtained the corresponding human methylation sites from Gene Expression Omnibus (45) with the accession number GSE70485. We extended the m^{1}A peak middle position to cover the flanking regions (25 bp on each side).

Use case 4: HOT regions of RBPs. Using in-house R scripts, the HOT regions (Supplementary File 7) were determined among the peak regions of human RBPs, which are available in the doRiNA database. Briefly, RBP-binding sites returned by all available CLIP experiments are used for calculating the density of the binding sites over the genome using 500 bp sliding windows. We calculated the local maxima of the density vector for each chromosome. We made sure local maxima of the density vector are the only maxima in the 2000 bp region surrounding the maxima. This is necessary to remove sub-optimal maxima around the real maxima. We then ranked these maxima based on the density scores, which are effectively a number of overlapping peaks. We used the 99th percentile to define HOT regions, meaning anything above the 99th percentile is declared a HOT region for RBPs.

Software packages

RCAS is developed as an R/Bioconductor package. The R library is also packaged using the package managers Conda and Guix. The source code of the latest development of RCAS is available at https://github.com/BIMSBbioinfo/RCAS.

The R commands to install the release version of RCAS from Bioconductor:

```
> source('http://bioconductor.org/biocLite.R')
> biocLite('RCAS')
```

The R commands to install the development version of RCAS from GitHub:

```
> library('devtools')
> devtools::install_github('BIMSBbioinfo/RCAS')
```

Conda is a cross-platform package manager and environment management system that deploys binaries of software packages. Various dedicated Conda channels have been established to host specialized packages, among which the 'bioconda' channel (https://bioconda.github.io/) provides bioinformatics packages. The 'r' channel is dedicated to R packages and the 'conda-forge' channel for general purpose packages. We have integrated the RCAS Conda package to the 'bioconda' channel. Additionally, the Conda packages are also available in the 'bimsbbioinfo' channel which is a custom channel for our group. Users can install RCAS by specifying these channels. For example, to install the latest package for RCAS 1.1.1, users can issue the following shell command:

```
> conda install bioconductor-rcas -c bimsbbioinfo -c r -c conda-forge
```

Guix is a package and environment manager for the GNU system. As a functional package manager, every Guix package expression closes over the complete dependency graph of a given piece of software, thereby enabling reproducible and portable software environments.

The shell command to install RCAS along with R using Guix:

```
> guix package -i r r-rcas
```

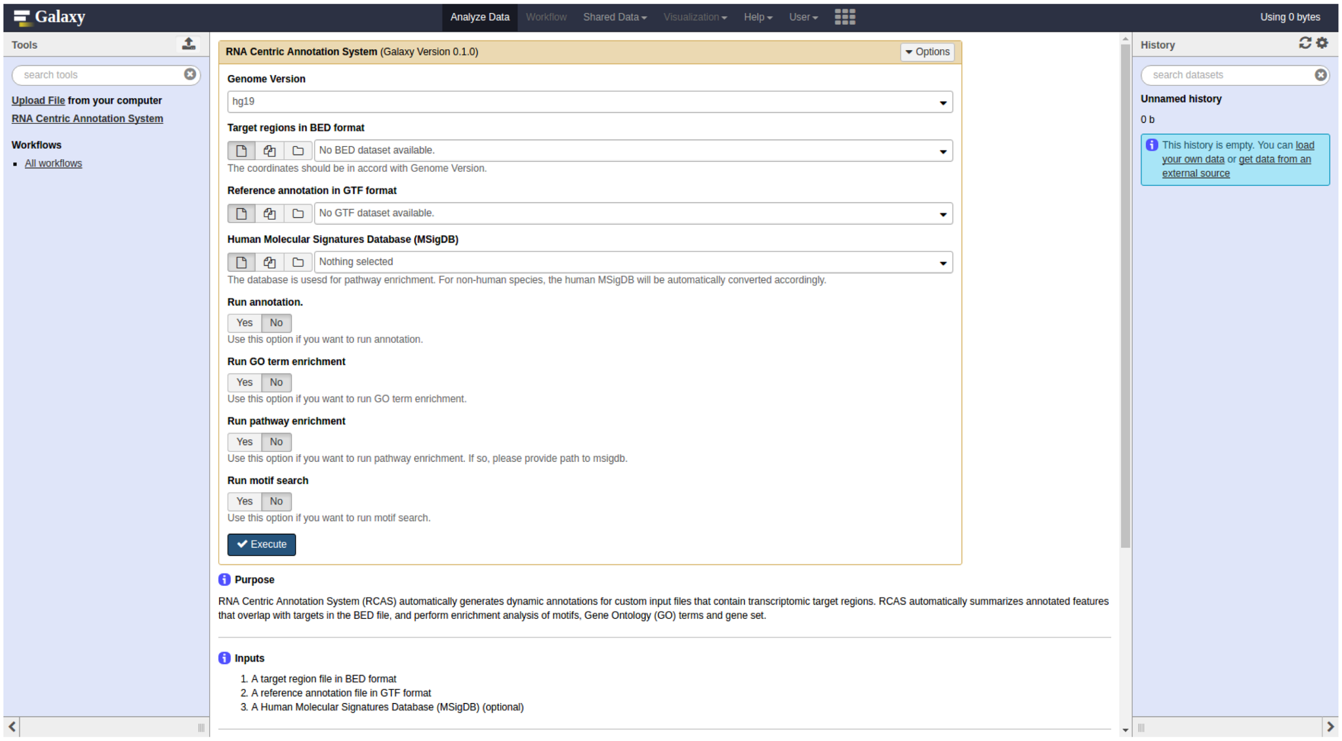
User interfaces

In addition to the command-line interface of R, we also provide graphical user interfaces via a Galaxy wrapper as well as a stand-alone web service. The Galaxy wrapper of RCAS integrates the annotation system with the Galaxy framework. The GUI provides options to specify BED file, GTF file, GMT file and the genome version. In addition, we provide widgets for enabling and disabling individual components. The RCAS dependencies of the Galaxy wrapper are installed via Conda which has been recruited by Galaxy (since version 16.01) to handle tool dependencies. The Galaxy wrapper is available at https://testtoolshed.g2.bx.psu.edu. Figure 2 is a screenshot of the RCAS Galaxy interface.

We provide a simple stand-alone web interface to RCAS. It consists of an application server through which users can request the generation of RCAS reports and a worker to process queued requests in the background. Both are written in Guile Scheme. The source code is available in a separate repository at https://github.com/BIMSBbioinfo/rcas-web. An instance of the web interface has been deployed and is publicly available at http://rcas.mdc-berlin.de for demonstration purposes. It can be installed together with RCAS via Guix using the following shell command:

```
> guix package -i rcas-web
```

The web interface allows users to upload a single BED file, which is used as the main input to RCAS and to select which of the four analysis modules to run. Users can select from one of four reference genome assemblies and select one annotation database to be used for the gene set enrichment analysis module. The intervals in the BED file can optionally be downsampled. Upon submission, a job is enqueued to run RCAS in the background and generate the specified HTML report. Once RCAS has generated the report, the requester can access it online or download it in a bundle along with any produced output files. Figure 3 is a screenshot of the RCAS web service.



Upload File from your computer

RNA Centric Annotation System

Workflows

- All workflows

Unnamed history

0 b

This history is empty. You can load your own data or get data from an external source

RNA Centric Annotation System (Galaxy Version 0.1.0)

Options

Genome Version

hg19

Target regions in BED format

No BED dataset available.

The coordinates should be in accord with Genome Version.

Reference annotation in GTF format

No GTF dataset available.

Human Molecular Signatures Database (MSigDB)

Nothing selected

The database is usesd for pathway enrichment. For non-human species, the human MSigDB will be automatically converted accordingly.

Run annotation.

Yes [✓ No]

Use this option if you want to run annotation.

Run GO term enrichment

Yes [✓ No]

Use this option if you want to run GO term enrichment.

Run pathway enrichment

Yes [✓ No]

Use this option if you want to run pathway enrichment. If so, please provide path to msigdb.

Run motif search

Yes [✓ No]

Use this option if you want to run motif search.

✓ Execute

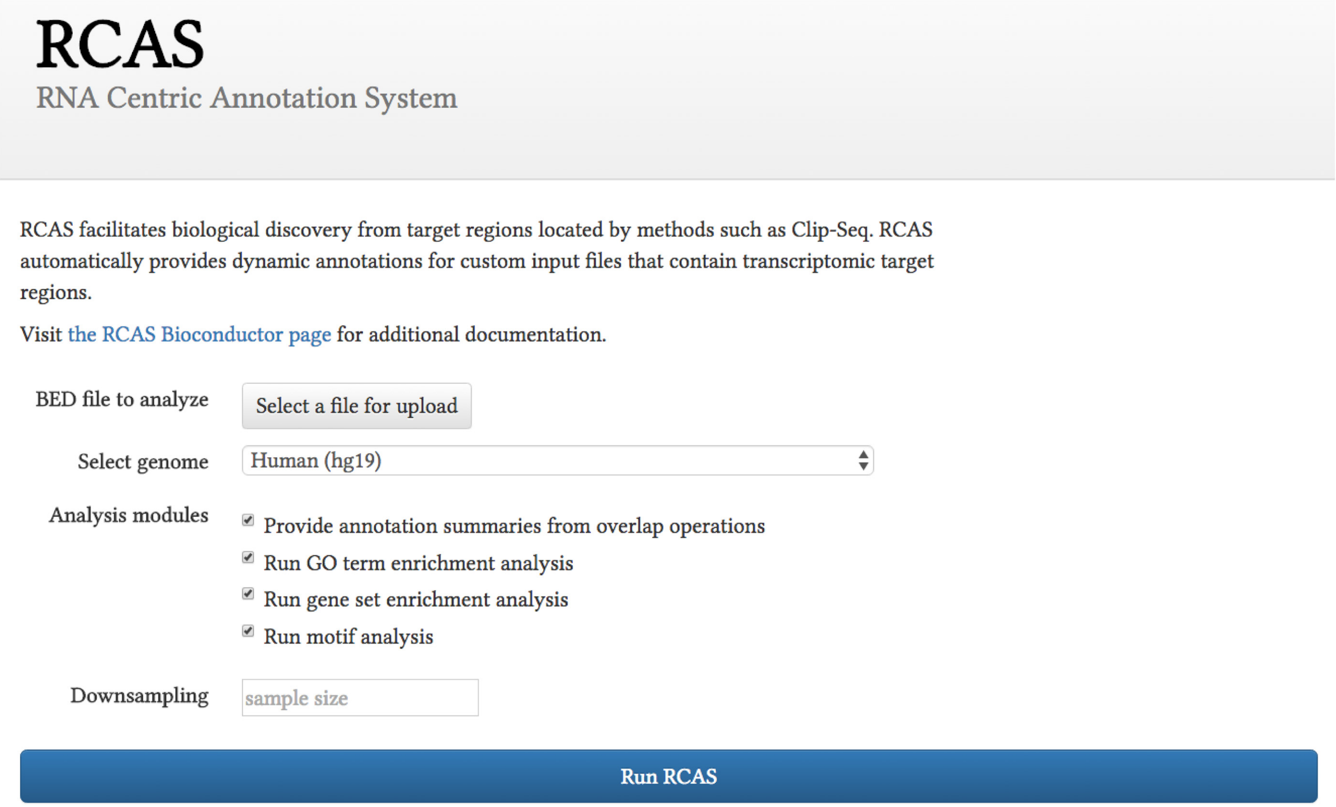
Purpose

RNA Centric Annotation System (RCAS) automatically generates dynamic annotations for custom input files that contain transcriptomic target regions. RCAS automatically summarizes annotated features that overlap with targets in the BED file, and perform enrichment analysis of motifs, Gene Ontology (GO) terms and gene set.

Inputs

1. A target region file in BED format
2. A reference annotation file in GTF format
3. A Human Molecular Signatures Database (MSigDB) (optional)

Figure 2. A screenshot of the Galaxy interface for RCAS. Options are provided for inputs and switches for running individual components.



RCAS

RNA Centric Annotation System

RCAS facilitates biological discovery from target regions located by methods such as Clip-Seq. RCAS automatically provides dynamic annotations for custom input files that contain transcriptomic target regions.

Visit the RCAS Bioconductor page for additional documentation.

BED file to analyze: Select a file for upload

Select genome: Human (hg19)

Analysis modules:

☑ Provide annotation summaries from overlap operations

☑ Run GO term enrichment analysis

☑ Run gene set enrichment analysis

☑ Run motif analysis

Downsampling: sample size

Run RCAS

Figure 3. A screenshot of the interface of the RCAS web service.

RESULTS

Use cases

In order to assess the performance of RCAS, we employed four use cases with the datasets from three high-impact publications and an in-house dataset. The results from the three use cases are used to benchmark the RCAS annotations against the well-studied peak regions which are derived from the respective experiments using PAR-CLIP (11), m^{1}A-seq (13), and deepCAGE (21). For the last use case, we used RCAS to annotate the features associated with the potential HOT regions of RBPs.

Case 1: Peak regions of RBPs derived from PAR-CLIP. In the landmark study (11), Hafner *et al.* applied the PAR-CLIP technology to study the respective binding characteristics of different RBPs. Among them, there are PUM2, QKI and IGF2BP1-3. It is known that the RNA binding of PUM2 is featured by high sequence-specificity (46) and the motifs are well defined. This feature is particularly relevant for the assessment of the motif discovery module of RCAS, in which the motifRG library (30) is utilized. QKI was implicated in the processes of pre-mRNA splicing (47) and thus it likely prefers to bind intronic regions (11). IGF2BP regulates mRNA stability, transport, and translation. The RCAS outputs are in agreement with the published binding characteristics of each RBP in terms of the binding preferences and the motifs. RCAS reports 73.8% of the PUM2 binding sites derived from 3′ UTR (Supplementary File 1), 77.6% of the QKI binding sites from introns (Supplementary File 2), and 66.0% of the IGF2BP1–3 binding sites from 3′ UTR (Supplementary File 3). In regards to motif discovery by RCAS, for PUM2, three of the top motifs (UGUAUA, UGUAAA, UGUACA) (Supplementary File 1) reasonably match the known recognition element - UGUANAUA (N = A, C, G, U). For QKI, two of the top motifs (ACUAAC, ACUAAU) (Supplementary File 2) fit the previously reported consensus, AYUAAY (Y = C, U). For IGF2BP1-3, one top motif CACAUC (Supplementary File 3) reflects the published consensus, CAUH (H = A, U, C).

In addition to the genomic features and motifs which were reported by the previous study (11), RCAS provides additional annotations such as the coverage profiles related to TSS, TES, and exon-intron boundaries. According to the RCAS report, there is one characteristic which is common among PUM2, QKI, and IGF2BP1-3—the strong binding preference to the TESs when compared to the TSSs (Figure 4A–C). At the exon-intron boundaries, IGF2BP proteins show a strong increase of signal at the exons relative to the neighboring introns on both 5′ and 3′ ends of internal exons. On the other hand, neither PUM2 nor QKI show a significant signal on either end of internal exons (Supplementary Files 1–3).

Case 2: peak regions of m^{1}A methylation derived from m^{1}A-seq. In the study (13), the authors found the m^{1}A peaks mostly enriched in 5′ UTRs and cluster around the AUG start codon. Furthermore, it was found that the detected motifs are GC-rich. Moreover, their GO analysis revealed the enrichment of biological processes which are related to translation and RNA metabolism. RCAS successfully reproduces the findings in the previous study. In the RCAS report, there are 56.8% of m^{1}A methylation sites at 5′ UTR in comparison with 12.1% at 3′ UTR. In addition, RCAS reports 75.9% of the methylation sites associated with promoters (within 2000 bp upstream and 200 bp downstream of the TSS), 43.6% with introns and 55.8% with CDSs (Supplementary File 4, note that the promoter, intron, and UTRs can overlap and there is no precedence between them when calculating overlap statistics). Moreover, RCAS reports the GC-rich top motifs (CCAUGG, GGCGGC, CGCUGC, CCGCCG) (Supplementary File 4). The GO analysis of RCAS (Supplementary File 5) suggests that the transcripts with this modification mark are associated with biological processes (when sorted by adjusted *P*-values) that are related to the regulation of mRNA splicing and processing, and translational initiation.

Besides recovering the published results, RCAS provides novel insights such as the coverage profiles at various genomic features. At the promoters, the peak of coverage is at the 3′ end of the regions. At the 5′ UTRs, the peak appears at the 5′ end of the regions (Figure 5A). Moreover, the coverage is also enriched at the TSSs with a slightly downstream shift (Figure 5B). In comparison, no enrichment is observed at TES. The peak coverage at TSS conforms with the previous result that m^{1}A peaks cluster around the AUG start codon (also notice that the top reported motif contains ‘AUG’). At both 5′ and 3′ boundaries, internal exons show a clear increase for m^{1}A modification sites compared to neighboring introns (Figure 5C). Together with the result of GO analysis, the enrichment at the boundaries may indicate a potential role of m^{1}A in alternative splicing.

Case 3: loci of tiRNAs derived from deepCAGE. In the study of (21), combining small RNA-Seq and DeepCAGE, the authors have identified the genomic regions of tiRNAs—10–30 bp downstream of TSSs in human, chicken and fly. These tiRNAs are preferentially associated with GC-rich promoters in highly expressed genes. We used RCAS to annotate the genomic regions of human tiRNAs. The RCAS summaries succeed in replicating the previous findings. In the RCAS report, the originated regions are associated with promoters by 93.9%, with 5′ UTR by 78.5%, with introns by 32.4% and with CDSs by 17.8% (Supplementary File 5). Our coverage profiles show the enrichment of tiRNAs at the 3′ ends of promoters and the 5′ ends of the 5′ UTRs (Figure 6A) with a distinguished peak at the TSSs (Figure 6B). The detected motifs are featured with high GC content (CGGCUG, UGGCGG, GGCUGC, GCGGCC) (Supplementary File 5).

Case 4: HOT regions of RBPs. HOT regions are segments of the genome occupied by a large number of different transcription factors (TFs) (48). These regions appear to be a common feature in human and invertebrate model organisms.

To our knowledge, it is not yet known whether the comparable HOT regions of RBPs exist. We applied RCAS to annotate the candidate HOT regions of RBPs which were identified by our in-house methods. The report shows the majority of the regions are exonic by 98.3%, associated with

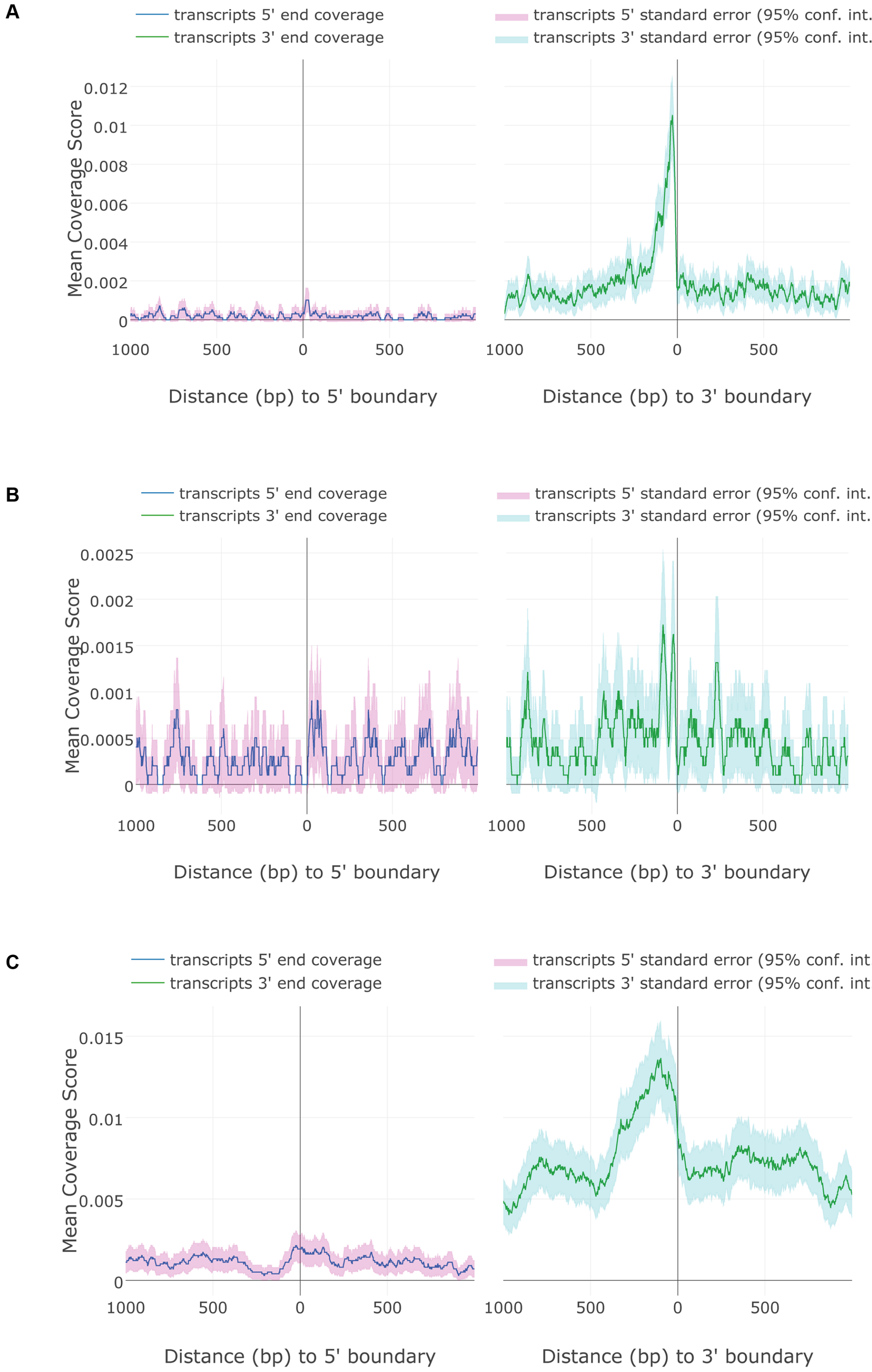


Figure 4. The coverage profiles of PUM2 (**A**), QKI (**B**), IGF2BP1–3 (**C**) at TSS/TES. The three types of RBPs all appear to have the stronger binding preference at TES. Each coverage profile is represented with a ribbon where a solid line passing through the middle area. The solid lines represent the mean coverage score distribution and the thickness of the ribbons represents the 95% confidence interval (equal to 1.96 times the standard error of the mean).

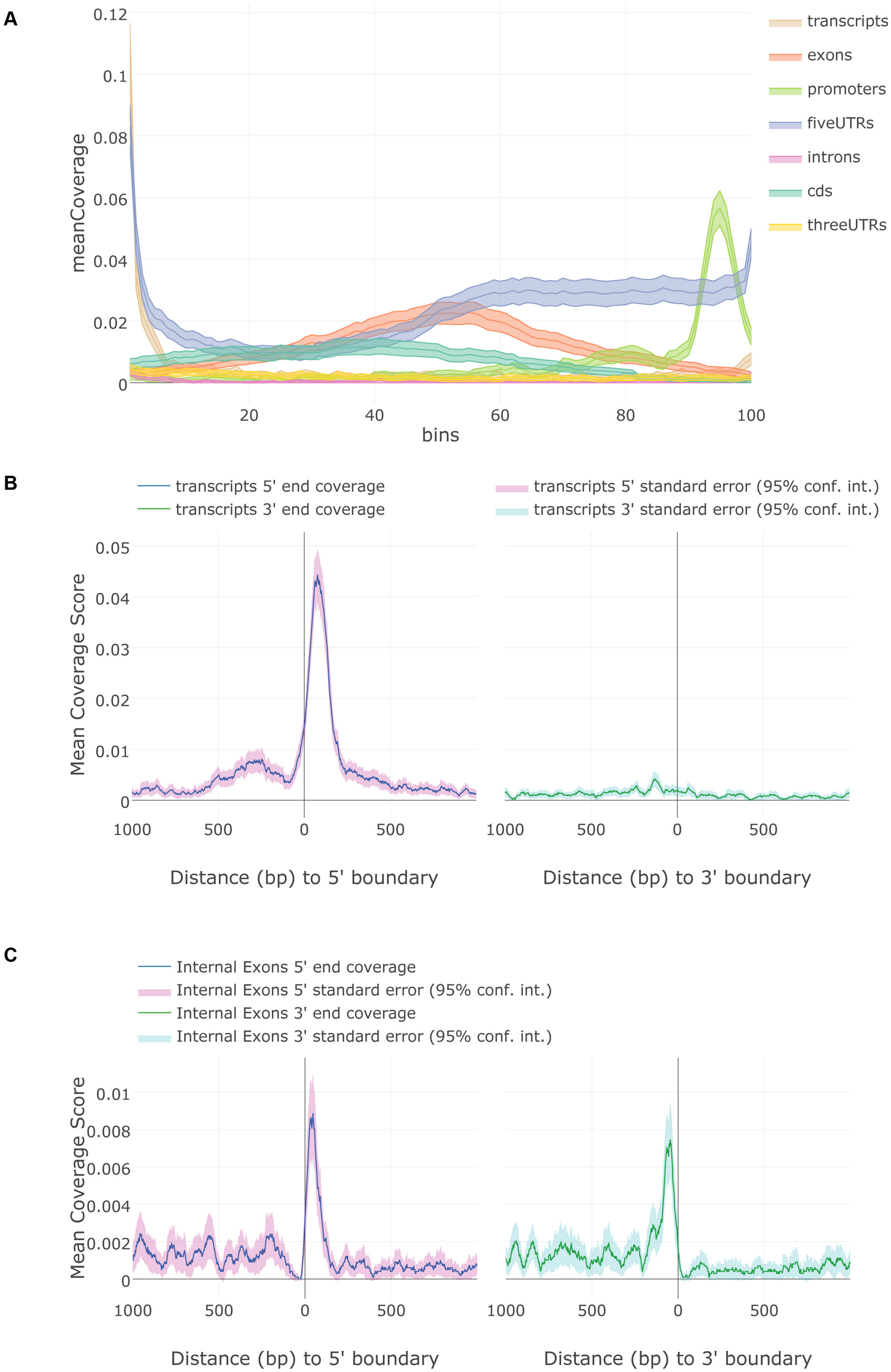


Figure 5. The coverage profiles of m^1A across different features (**A**), at TSS/TES (**B**), at exon-intron boundaries (**C**). (A) At the promoters, the peak of coverage is at the 3′ end of the regions. At 5′ UTRs, the peak is at the 5′ end of the regions. (B) m^1A is clearly enriched at TSS. (C) m^1A show a stark contrast between adjacent exonic and intronic regions. Each coverage profile is represented with a ribbon with a solid line passing through the middle of the ribbon. The solid lines in the middle of each coverage profile represent the mean coverage score distribution and the thickness of the ribbons encapsulating the solid line represents the 95% confidence interval (equal to 1.96 times the standard error of the mean).

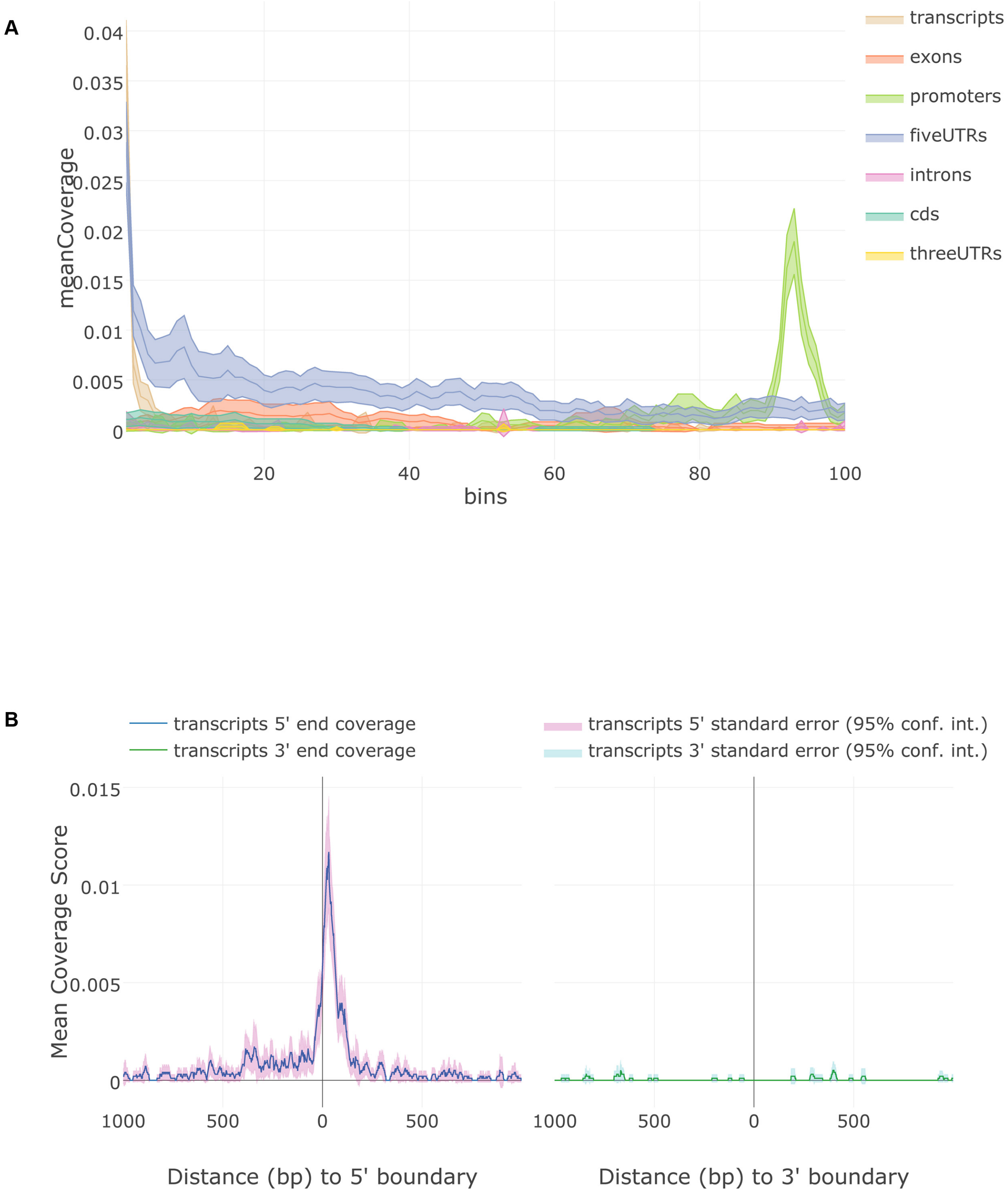


Figure 6. The coverage profiles of tiRNAs across different genomic features (**A**), at TSS/TES (**B**). (A) tiRNs are enriched at the 3′ end of promoters while peak at 5′ end of the regions of 5′ UTRs. (B) tiRNAs are clearly enriched at TSS. Each coverage profile is represented with a ribbon where a solid line passing through the middle area. The solid lines represent the mean coverage score distribution and the thickness of the ribbons represents the 95% confidence interval (equal to 1.96 times the standard error of the mean).

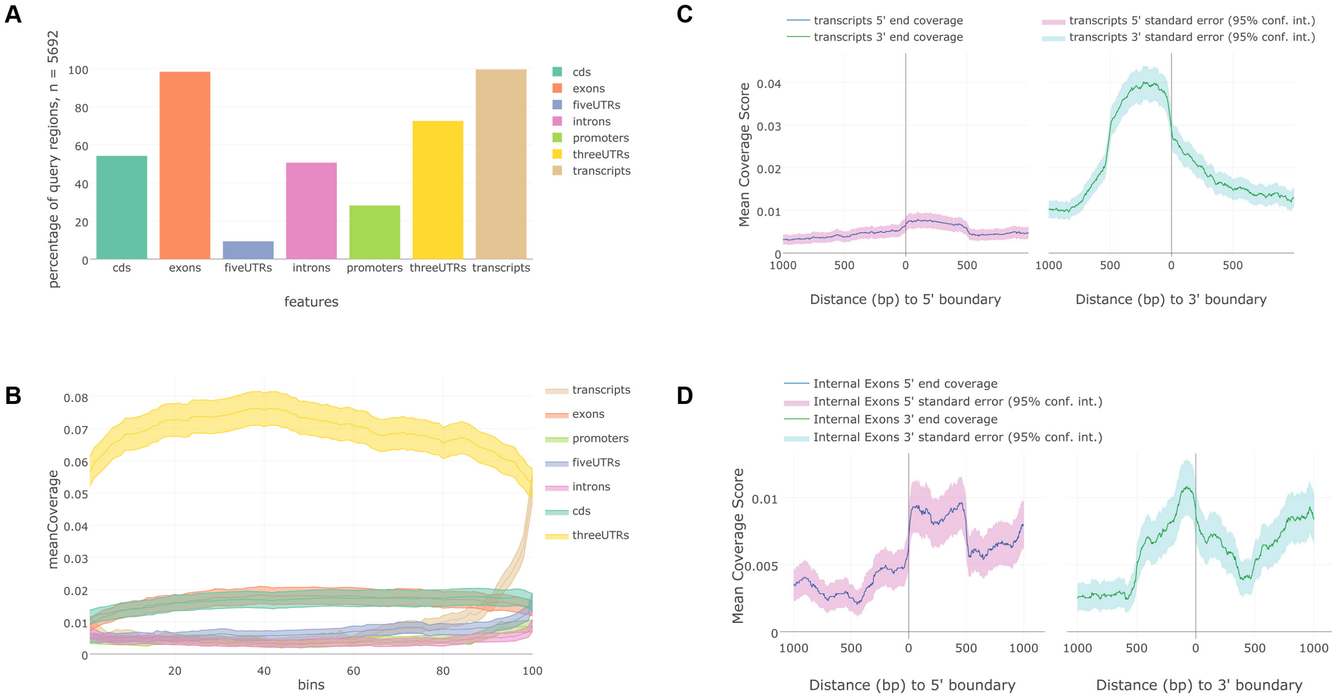


Figure 7. (**A**) The distribution of HOT regions across gene features. (**B**) Binned coverage profiles of HOT regions overlaid on gene features. (**C**) Nucleotide resolution coverage profiles of HOT regions overlaid on TSS/TES boundaries. (**D**) Nucleotide resolution coverage profiles of HOT regions overlaid on 5′ and 3′ boundaries of internal exons. Each coverage profile is represented with a ribbon where a solid line passing through the middle area. The solid lines represent the mean coverage score distribution and the thickness of the ribbons represents the 95% confidence interval (equal to 1.96 times the standard error of the mean).

3′ UTRs by 72.5%, and intronic by 50.6% (Figure 7A). However, when looking at the depth of coverage profiles, 3′ UTRs show the highest signal (Figure 7B). Moreover, we observe the distinguished peaks at TESs (Figure 7C). This characteristics is in agreement with the common binding preference observed in use case 1 that includes PUM2, QKI and IGF2BP1-3. The signal strength at the intron-exon junctions does not display a strong difference, particularly at the 3′ end boundaries of internal exons (Figure 7D).

The observed trends in the coverage profiles could be either biologically meaningful or can be explained by the inherent biases of the studied RBPs, for which CLIP-Seq peak regions are available in the doRiNA database. Sixteen out of 38 RBPs in doRiNA have GO terms associated with 'splicing', which could explain the observed signal at exon–intron junctions; 20 out of 38 RBPs in doRiNA have GO terms associated with 'miRNA', 'gene silencing', or 'translation', which could explain the increased signal at the 3′ UTRs (Supplementary File 8).

GO term analysis of the transcripts that harbor HOT regions also suggests that the targets of RBPs have similar enriched GO terms as the RBPs themselves. For instance, 'gene silencing by miRNA' and 'regulation of RNA splicing' show up as significant biological process terms for these transcripts. Again the gene set enrichment analysis of these transcripts also reveals significant Reactome pathways such as '3′ UTR mediated translational regulation' and 'mRNA splicing'. On the other hand, no sequence motifs have been discovered for the HOT region sequences (Supplementary File 6).

DISCUSSION

The advance of RNA-based omics technologies has created unprecedented opportunities for biological discovery centered around RNA molecules. However, there have been challenges to interpreting the gigantic amount of information yielded by the emerging omics methods. Although data generated from each different RNA-based omics method require particularly tailored workflows, most of these workflows will contain common steps such as experimental design, sample preparation, sequencing, quality control and pre-processing of sequencing data, obtaining transcriptome-wide regions of interest (these genomic regions may represent binding sites of RBPs, methylation sites, loci of RNA species, etc.), and as the final step, functional annotation and analysis of these regions followed by reporting the analysis results. This last step may provide the final results of an experiment, suggest intriguing points for further experimental validation, or could reveal novel hypotheses for follow-up studies. With the aim to ease the process of finding biological meaning within the large datasets of transcriptome-wide regions of interest, we developed RCAS, an RNA-centric functional annotation and reporting tool, which is designed to help users quickly summarise their experiment results, find functional associations, interactively explore the results, and eventually have a standalone HTML report with exportable figures and tables for further analysis or for publication purposes.

In order to make RCAS as accessible as possible to a wide spectrum of users with computational or experimental backgrounds, we have developed multiple means of deploying the tool. At the core is a cross-platform R package in the Bioconductor repository, where the package is tested to work on Windows, Linux, and Mac OS X operating systems. To enable different means of installation, the R library and its dependencies have been packaged with the Conda and Guix package management systems. To help users with limited experience with the R command line, we have developed a web service and a wrapper script to integrate RCAS with Galaxy.

To assess the accuracy of RCAS, we employed four use cases, mainly using the example published datasets. The resulting benchmarks show that RCAS successfully reproduces the published results. RCAS correctly identifies the underlying preferences of different biological events (RBP binding, methylation, and RNA regulation). The generated motifs reasonably match the published ones. In addition, RCAS is capable of generating novel insights which are not present in the publications, including the RBP binding preference on TESs and the m^1A methylation preference on exon–intron boundaries. The RCAS output is a dynamic HTML file which is composed of interactive figures and tables that are of high-quality, ready for the purpose of publication.

The outputs of RCAS can be combined for meta-analysis to compare the characteristics of genomic regions of different biological contexts. This would assist the discovery of common patterns shared by different events that carry similar functional implications. For instance, in the case of m^1A methylation and tiRNA regulation, there are common preferences on promoters and 5′ UTRs. The coverage enrichment sites are both at TSS. The detected motifs are both GC-rich. Even though m^1A and tiRNA regulation are separate events, the common patterns shared by the two are not unexpected given the context that both events are related to the regulation of gene transcription.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

German Federal Ministry of Education and Research (BMBF) as part of the RNA Bioinformatics Center of the German Network for Bioinformatics Infrastructure (de.NBI) [031 A538C RBC (de.NBI)]. Funding for open access charge: Helmholtz Association.

Conflict of interest statement. None declared.

REFERENCES

1. Lodish,H., Berk,A., Lawrence Zipursky,S., Matsudaira,P., Baltimore,D. and Darnell,J. (2000) *The Three Roles of RNA in Protein Synthesis*, W. H. Freeman.
2. Morris,K.V. and Mattick,J.S. (2014) The rise of regulatory RNA. *Nat. Rev. Genet.*, **15**, 423–437.
3. Cech,T.R. and Steitz,J.A. (2014) The noncoding RNA revolution—trashing old rules to forge new ones. *Cell*, **157**, 77–94.
4. Castello,A., Fischer,B., Hentze,M.W. and Preiss,T. (2013) RNA-binding proteins in Mendelian disease. *Trends Genet.*, **29**, 318–327.
5. Strobel,E.J., Watters,K.E., Loughrey,D. and Lucks,J.B. (2016) RNA systems biology: uniting functional discoveries and structural tools to understand global roles of RNAs. *Curr. Opin. Biotechnol.*, **39**, 182–191.
6. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
7. de Hoon,M. and Hayashizaki,Y. (2008) Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *Biotechniques*, **44**, 627–632.
8. Churchman,L.S. and Weissman,J.S. (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, **469**, 368–373.
9. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R.S. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
10. Jiao,Y. and Meyerowitz,E.M. (2010) Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control. *Mol. Syst. Biol.*, **6**, 419.
11. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.-C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
12. Dominissini,D., Moshitch-Moshkovitz,S., Schwartz,S., Salmon-Divon,M., Ungar,L., Osenberg,S., Cesarkas,K., Jacob-Hirsch,J., Amariglio,N., Kupiec,M. *et al.* (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, **485**, 201–206.
13. Dominissini,D., Nachtergaele,S., Moshitch-Moshkovitz,S., Peer,E., Kol,N., Ben-Haim,M.S., Dai,Q., Di Segni,A., Salmon-Divon,M., Clark,W.C. *et al.* (2016) The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA. *Nature*, **530**, 441–446.
14. Chu,C., Qu,K., Zhong,F.L., Artandi,S.E. and Chang,H.Y. (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell*, **44**, 667–678.
15. Kudla,G., Granneman,S., Hahn,D., Beggs,J.D. and Tollervey,D. (2011) Cross-linking, ligation, and sequencing of hybrids reveals RNA–RNA interactions in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 10010–10015.
16. German,M.A., Pillay,M., Jeong,D.-H., Hetawal,A., Luo,S., Janardhanan,P., Kannan,V., Rymarquis,L.A., Nobuta,K., German,R. *et al.* (2008) Global identification of microRNA–target RNA pairs by parallel analysis of RNA ends. *Nat. Biotechnol.*, **26**, 941–946.
17. Lucks,J.B., Mortimer,S.A., Trapnell,C., Luo,S., Aviran,S., Schroth,G.P., Pachter,L., Doudna,J.A. and Arkin,A.P. (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 11063–11068.
18. Gentleman,R., Robert,G. and Vincent,C. (2005) Bioconductor: software and development strategies for statistical genomics. *Encyclopedia of Genet. Genomics Proteomics Bioinformatics*.
19. Courtès,L. and Wurmus,R. (2015) Reproducible and user-controlled software environments in HPC with Guix. In: Hunold,S, Costan,A, Giménez,D, Iosup,A, Ricci,L, Requena,MEG, Scarano,V, Varbanescu,AL, Scott,SL and Lankes,S (eds). *Euro-Par 2015: Parallel Processing Workshops*. Lecture Notes in Computer Science. Springer International Publishing, pp. 579–591.
20. Afgan,E., Baker,D., van den Beek,M., Blankenberg,D., Bouvier,D., Čech,M., Chilton,J., Clements,D., Coraor,N., Eberhard,C. *et al.* (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, doi:10.1093/nar/gkw343.
21. Taft,R.J., Glazov,E.A., Cloonan,N., Simons,C., Stephen,S., Faulkner,G.J., Lassmann,T., Forrest,A.R.R., Grimmond,S.M., Schroder,K. *et al.* (2009) Tiny RNAs associated with transcription start sites in animals. *Nat. Genet.*, **41**, 572–578.
22. Allaire,J.J., Joe,C., Yihui,X., Jonathan,M., Winston,C., Jeff,A., Hadley,W., Aron,A. and Rob,H. (2016) rmarkdown: dynamic documents for R. *R package version 1.0*.
23. The Gene Ontology Consortium (2014) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.

24. Carson,S., Chris,P., Toby,H., Scott,C., Karthik,R., Marianne,C. and Pedro,D. (2016) plotly: create interactive web graphics via 'plotly.js'. R package version 4.5.2.
25. Xie,Y. (2016) DT: a Wrapper of the JavaScript Library 'DataTables'. R package version 0.2.
26. Lawrence,M., Gentleman,R. and Carey,V. (2009) rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, **25**, 1841–1842.
27. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.
28. Akalin,A., Franke,V., Vlahoviček,K., Mason,C.E. and Schübeler,D. (2015) Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics*, **31**, 1127–1129.
29. Pages,H. (2012) BSgenome: Infrastructure for Biostrings-based genome data packages. R package version.
30. Yao,Z., Macquarrie,K.L., Fong,A.P., Tapscott,S.J., Ruzzo,W.L. and Gentleman,R.C. (2014) Discriminative motif analysis of high-throughput dataset. *Bioinformatics*, **30**, 775–783.
31. Alexa,A. and Rahnenfuhrer,J. (2010) topGO: enrichment analysis for gene ontology. R package version.
32. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
33. Yates,A., Andrew,Y., Wasiu,A., Ridwan Amode,M., Daniel,B., Konstantinos,B., Denise,C.-S., Carla,C., Peter,C., Stephen,F. *et al.* (2015) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.
34. Durinck,S., Steffen,D., Spellman,P.T., Ewan,B. and Wolfgang,H. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.
35. Aken,B.L., Ayling,S., Barrell,D., Clarke,L., Curwen,V., Fairley,S., Fernandez Banet,J., Billis,K., García Girón,C., Hourlier,T. *et al.* (2016) The Ensembl gene annotation system. *Database*, doi:10.1093/database/baw093.
36. Qiu,Y.-Q. and Yu-Qing,Q. (2013) KEGG Pathway Database. *Encyclop. Syst. Biol.*, 1068–1069.
37. Nishimura,D. and Darryl,N. (2001) BioCarta. *Biotech Softw. Internet Rep.*, **2**, 117–120.
38. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
39. Fabregat,A., Sidiropoulos,K., Garapati,P., Gillespie,M., Hausmann,K., Haw,R., Jassal,B., Jupe,S., Korninger,F., McKay,S. *et al.* (2016) The reactome pathway Knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.
40. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdottir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
41. Liberzon,A., Arthur,L., Chet,B., Helga,T., Mahmoud,G., Mesirov,J.P. and Pablo,T. (2015) The molecular signatures database hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
42. Blin,K., Dieterich,C., Wurmus,R., Rajewsky,N., Landthaler,M. and Akalin,A. (2014) DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.*, **43**, D160–D167.
43. Kawaji,H., Severin,J., Lizio,M., Forrest,A.R.R., van Nimwegen,E., Rehli,M., Schroder,K., Irvine,K., Suzuki,H., Carninci,P. *et al.* (2011) Update of the FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Res.*, **39**, D856–D860.
44. Speir,M.L., Zweig,A.S., Rosenbloom,K.R., Raney,B.J., Paten,B., Nejad,P., Lee,B.T., Learned,K., Karolchik,D., Hinrichs,A.S. *et al.* (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, **44**, D717–D725.
45. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
46. Wang,X., McLachlan,J., Zamore,P.D. and Hall,T.M.T. (2002) Modular recognition of RNA by a human pumilio-homology domain. *Cell*, **110**, 501–512.
47. Chénard,C.A. and Richard,S. (2008) New implications for the QUAKING RNA binding protein in human disease. *J. Neurosci. Res.*, **86**, 233–242.
48. Foley,J.W. and Sidow,A. (2013) Transcription-factor occupancy at HOT regions quantitatively predicts RNA polymerase recruitment in five human cell lines. *BMC Genomics*, **14**, 720.