Published online 27 November 2020 *Nucleic Acids Research, 2021, Vol. 49, Database issue* ***D151–D159***
doi: 10.1093/nar/gkaa1060

DIANA-miRGen v4: indexing promoters and regulators for more than 1500 microRNAs

Nikos Perdikopanis[1,2,3,*,†], Georgios K. Georgakilas[4,†], Dimitris Grigoriadis[1,5], Vasilis Pierros[1,2], Ioannis Kavakiotis[5], Panagiotis Alexiou[4] and Artemis Hatzigeorgiou[1,2,5,*]

[1]Hellenic Pasteur Institute, Athens 11521, Greece, [2]Department of Electrical and Computer Engineering, University of Thessaly, Volos 38221, Greece, [3]Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens 15784, Greece, [4]Central European Institute of Technology, Masaryk University, Kamenice 735/5, 62500 Brno, Czech Republic and [5]Department of Computer Science and Biomedical Informatics, University of Thessaly, Greece

Received September 15, 2020; Revised October 16, 2020; Editorial Decision October 19, 2020; Accepted November 26, 2020

ABSTRACT

Deregulation of microRNA (miRNA) expression plays a critical role in the transition from a physiological to a pathological state. The accurate miRNA promoter identification in multiple cell types is a fundamental endeavor towards understanding and characterizing the underlying mechanisms of both physiological as well as pathological conditions. DIANA-miRGen v4 (www.microrna.gr/mirgenv4) provides cell type specific miRNA transcription start sites (TSSs) for over 1500 miRNAs retrieved from the analysis of >1000 cap analysis of gene expression (CAGE) samples corresponding to 133 tissues, cell lines and primary cells available in FANTOM repository. MiRNA TSS locations were associated with transcription factor binding site (TFBSs) annotation, for >280 TFs, derived from analyzing the majority of ENCODE ChIP-Seq datasets. For the first time, clusters of cell types having common miRNA TSSs are characterized and provided through a user friendly interface with multiple layers of customization. DIANA-miRGen v4 significantly improves our understanding of miRNA biogenesis regulation at the transcriptional level by providing a unique integration of high-quality annotations for hundreds of cell specific miRNA promoters with experimentally derived TFBSs.

INTRODUCTION

Since the discovery of microRNA (miRNA) abundant transcription in mammalian lineages in 2001 (1), this small RNA family has become a subject of intense research, reshaping our understanding of gene regulatory networks and their implications in physiological and pathological conditions. Thousands of studies have unveiled the critical role of miRNAs in fine-tuning gene expression, the mechanisms of miRNA-mediated recognition and targeting of RNAs (2,3), components of the miRNA biogenesis pathway and members of the RNA-induced silencing complex (RISC) which carries out the process of miRNA-guided RNA degradation and/or mRNA translation suppression (4,5). However, the pace for characterizing miRNA gene promoters and their transcription regulators is severely lagging behind, inserting noise in the process of fully integrating miRNAs in gene regulatory networks.

This state of uncertainty primarily arises due to properties of the miRNA biogenesis pathway. Most mammalian miRNAs are transcribed by Polymerase II (Pol2) machinery (6) resulting in the formation of primary transcripts (pri-miRNAs) of variable length that are 5′ capped and polyadenylated. Hairpin-like structures (pre-miRNAs), embedded in pri-miRNAs, are recognized and cleaved in the nucleus by the microprocessing complex which consists of Drosha and Dgcr8 (7). Pre-miRNAs are transported to the cytoplasm where they undergo the maturation process (8) prior to their incorporation in RISC.

The rapid cleavage of pri-miRNAs by Drosha in the nucleus is a key element that hinders the identification of miRNA genes with conventional sequencing techniques. With the advent of several breakthroughs in Next Generation Sequencing (NGS) technology, the sequencing cost limitation has been removed allowing the scientific community to generate deeply sequenced samples that are able to capture the lowly abundant pri-miRNA transcripts (9,10). Cap Analysis of Gene Expression (CAGE) specifically, is an experimental method that was introduced in 2003 (11)

*To whom correspondence should be addressed. Tel: +30 24210 74758; Fax: +30 24210 74997; Email: arhatzig@uth.gr
Correspondence may also be addressed to Nikos Perdikopanis. Tel: +30 24210 74758; Fax: +30 24210 74997; Email: nikosp@di.uoa.gr
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

for capturing and quantifying the 5′ ends of capped RNAs. During the last decade, CAGE has been continuously refined and improved, emerging as a well-established protocol for the identification of transcription start sites (TSS) and promoter regions of transcribed loci. The FANTOM Consortium (12) has extensively applied CAGE on hundreds of tissues and cell lines to produce a high-quality annotation of the human and mouse promoterome and characterize regulatory mechanisms of gene expression.

The recent availability of such datasets as well as the continuous updating of landmark repositories such as the ENCODE (13), allowed the development of computational methods that attempt to identify miRNA gene transcription start sites (TSSs) by combining this wealth of experimental data with Machine Learning (14–16). Such *in silico* approaches are invaluable for the process of creating repositories that index miRNA gene promoters and transcription factors (TFs) that regulate their transcription. DIANA-miRGen v4 aims to explicitly tackle this problem with a holistic approach that includes the combination of multi-omics data with state-of-the-art Machine Learning algorithms and web application technologies, to provide high quality miRNA promoters and TFs that regulate their transcription.

Existing resources for miRNA genes transcription regulation

Publicly available databases addressing the miRNA gene transcription regulation problem such as ChIPBase (17), CircuitsDB (18), TFmiR (19), TMREC (20), TransmiR (21,22) and TSmiR (23) have been extensively reviewed in DIANA-miRGen v3 study (24). They employ various methods for collecting miRNA gene TSSs, with the most frequent strategy involving the grouping of pre-miRNAs into transcriptional units based on their in-between distance (ChIPBase, CircuitsDB), and considering the 5′ end of these units as the group's TSS. However, pri-miRNAs typically exhibit a highly variable size that can span several hundred up to thousands of nucleotides which are difficult to locate with such strategies. Additionally, existing databases index TF:miRNA interactions based on TF motif search in promoters(ChIPBase, CircuitsDB, TFmiR, TMREC, TransmiR, TSmiR), literature text-mining (TMREC, TransmiR) and less frequently on chromatin immunoprecipitation followed by sequencing (ChIP-Seq), an experimental method that provides better accuracy for transcription factor binding site (TFBS) annotation (ChIPBase, CircuitsDB, TSmiR).

DIANA-miRGen v3 (24) pushed the envelope further by incorporating miRNA gene TSSs derived from microTSS algorithm (15) for 428 pre-miRNAs and thousands of TF:miRNA connections based on TF footprints in DNase enriched regions and motif scanning. MicroTSS is a computational method providing accurate intergenic miRNA TSS identification by integrating deeply sequenced RNA-Seq, H3K4me3 and Polymerase II ChIP-Seq as well as open chromatin data as identified by DNase-Seq. However, due to microTSS's input data requirements, DIANA-miRGen v3 was limited in a handful of tissues and cell lines, leaving plenty of space for future improvements.

What is new in DIANA-miRGen v4?

Aim of DIANA-miRGen v4 (Figure 1) is to build on top of the available high-quality repository of microTSS (15) predictions and significantly expand the collection of miRNA gene TSSs by utilizing the availability of hundreds of cap analysis of gene expression (CAGE) data produced by the FANTOM consortium (10). To this end, ADAPT-CAGE algorithm (16) has been applied on 1027 CAGE samples corresponding to 133 tissues, cell lines and primary cells, resulting in the identification of cell type specific miRNA gene TSSs for 1534 pre-miRNAs (Figure 1A-C). ADAPT-CAGE is a state-of-the-art Machine Learning framework that is trained to distinguish between CAGE signal derived from TSSs and transcriptional noise. It provides annotation-agnostic, highly accurate and single-nucleotide resolution experimentally derived TSSs on a genome-wide scale by only utilizing aligned CAGE tags and the underlying genomic sequence. To further expand the TF:miRNA repertoire (Figure 1D–F), 368 ChIP-Seq samples for 94 TFs and 61 DNase-Seq datasets were downloaded from the ENCODE (9) repository. DNase-Seq samples were specifically subjected to the TF footprinting identification process prior to motif scanning of footprint loci using position weight matrices (PWMs) from JASPAR (25). Entries from DIANA-miRGen v3 related to TF:miRNA interactions have been imported to the new database and links to the previous release facilitate the interconnection between the two versions.

The database schema was expanded to enable the significant increase of hosted data and support the necessary scalability as well as speed-up for complex queries (Figure 1G). The database access was built on top of the intuitive web-interface of DIANA-miRGen v3, aiming to keep changes to the users' experience as minimal as possible (Figures 2 and 3). At the same time, more metadata were added such as extensive descriptions for all utilized FANTOM samples, TF expression levels based on CAGE, three levels of supported TF binding sites (ChIP-Seq, DNase-Seq footprints and pure motif search) and miRNA or tissue/cell centric queries. Information regarding the function of miRNAs and TFs in physiological and pathological conditions as well as the connection between DIANA-miRGen, UCSC genome browser (26) and other DIANA resources (3,27–30) is still accessible.

METHODS AND RESULTS

Annotation of miRNA TSSs

1027 CAGE samples aligned on hg38, corresponding to 306 cell line, 178 tissue and 543 primary cell samples, were downloaded from FANTOM (10) public repository. In the case of time-course samples, only the 0 h files were downloaded. Each file was enriched with several types of metadata such as the FANTOM sample id, category (tissue, primary cells or cell line), condition (physiological or pathological) and origin (i.e. blood, bladder etc). The latter was also utilized as the piece of information connecting CAGE and ChIP-Seq samples (see transcription factor expression analysis section).

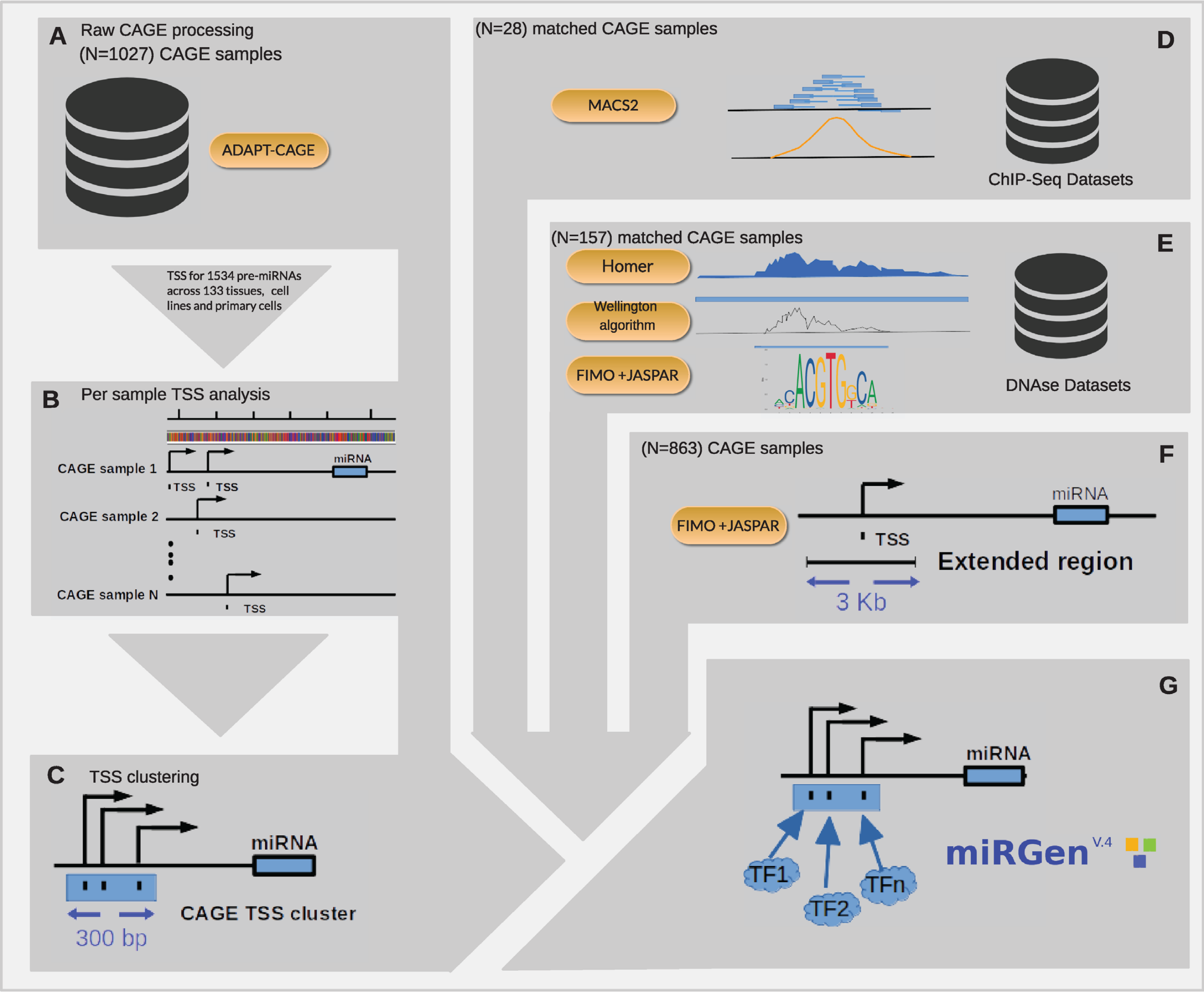
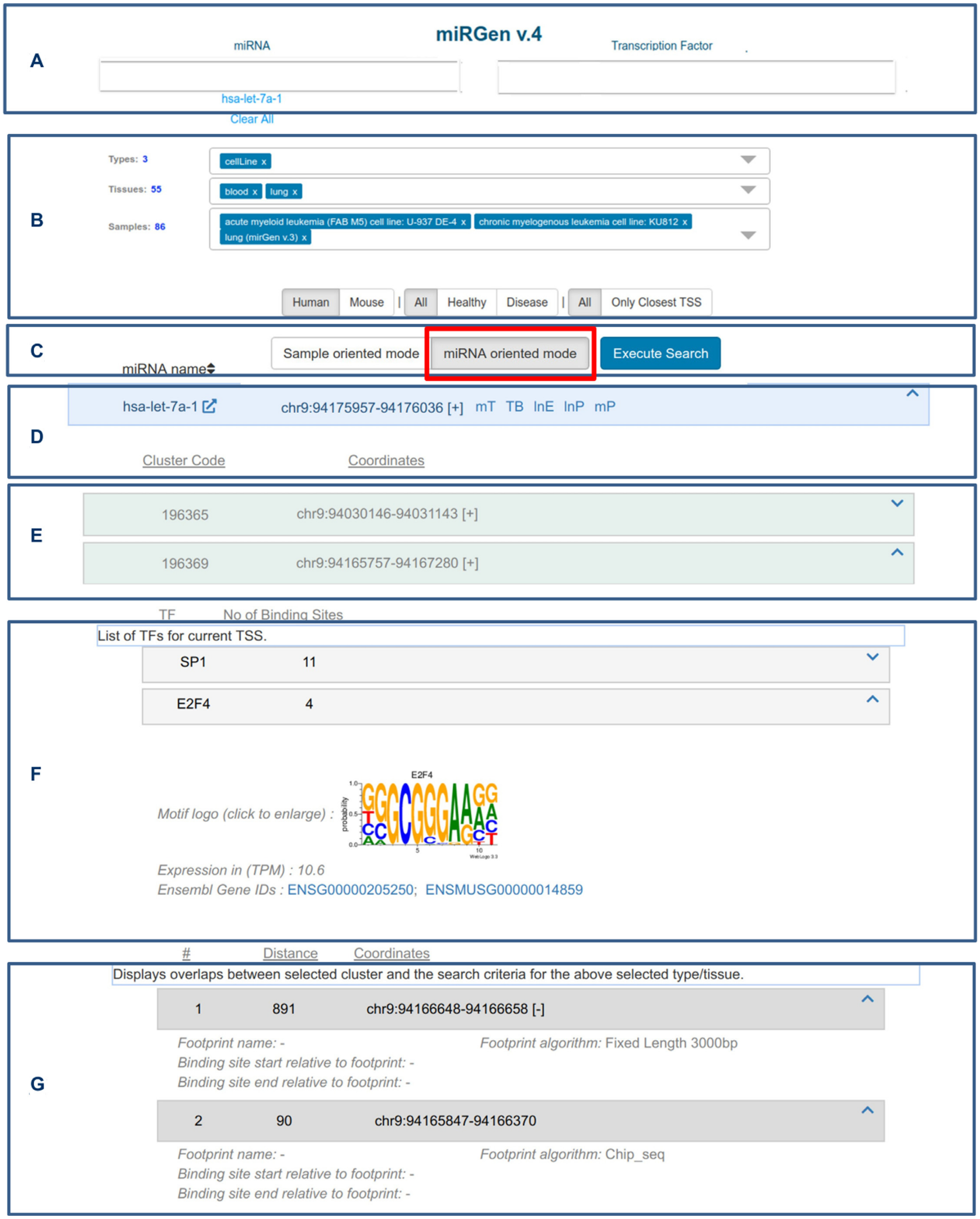


Figure 1. Overview of DIANA-miRGen v4 pipeline based on a bipartite strategy that involves raw CAGE signal filtering and the annotation of TFBSs. In the first part of the pipeline, (**A**) all CAGE samples ($N = 1027$) available in the FANTOM repository were processed with ADAPT-CAGE for filtering out signals corresponding to transcriptional noise from 133 human tissues, cell lines and primary cells. (**B**) Putative miRNA gene TSSs were subsequently selected from the remaining CAGE peaks that localized within a region that spans up to 150 kb upstream of 1534 pre-miRNAs annotated in miRBase v22.1 (**C**) CAGE-derived TSS events with in-between distances <300 bp were merged into TSS clusters, regardless of the originating cell type. In the second part of the pipeline, (**D**) TF ChIP-Seq datasets from ENCODE that match to 28 CAGE samples were processed with macs2 to identify experimentally validated TFBSs in miRNA gene promoters. Additionally, (**E**) DNase-Seq datasets matching 157 CAGE samples were subjected to footprinting extraction with Homer and Wellington to characterize TFBS with experimental support from open chromatin data. Lastly, (**F**) for the remaining 863 CAGE samples, the identification of miRNA gene promoter associated TFBSs was achieved with TF motif scanning within a 3kb window surrounding CAGE-derived TSSs. (**G**) The results from these diverse types of analyses are seamlessly integrated in the DIANA-miRGen v4 database and can be accessed through the redesigned web interface, facilitating ease-of-use, advanced queries and connections to external resources.

There is increasing evidence in the literature that besides transcription initiation events, CAGE also identifies capping sites along various locations of transcribed loci such as different splicing products, isoforms and capped molecules that can be summed up as transcriptional noise (31–33). Only a portion of regions enriched in CAGE signal typically overlap with the surrounding region of annotated TSSs. Therefore, for the purposes of this study, an intermediate step for pre-processing CAGE datasets is required prior to miRNA gene TSS identification. To this end, ADAPT-CAGE (16) algorithm was applied on every CAGE dataset that was downloaded from FANTOM's repository. ADAPT-CAGE is able to accurately distinguish between CAGE tag-clusters that represent transcription initiation events and tag-clusters that belong to the transcriptional noise category. The algorithm utilizes Machine Learning to take advantage of the observed differences in the profile of structural DNA features (34–37) and Polymerase II related TF motifs between promoter and non-promoter genomic regions. After applying ADAPT-CAGE, a score cutoff of 0.5 was used to produce the final results that can be summed up to 29 806 351 TSSs for all 1027 FANTOM samples. Tag-clusters with a distance less than 300 bp from one-another were merged into a single transcription initiation event (TSS



hsa-let-7a-1

Types: 3 — cellLine x

Tissues: 55 — blood x lung x

Samples: 86 — acute myeloid leukemia (FAB M5) cell line: U-937 DE-4 x chronic myelogenous leukemia cell line: KU812 x lung (mirGen v.3) x

miRNA name	
hsa-let-7a-1	chr9:94175957-94176036 [+]

Cluster Code	Coordinates
196365	chr9:94030146-94031143 [+]
196369	chr9:94165757-94167280 [+]

List of TFs for current TSS.

TF	No of Binding Sites
SP1	11
E2F4	4

Motif logo (click to enlarge) :

Expression in (TPM) : 10.6

Ensembl Gene IDs : ENSG00000205250; ENSMUSG00000014859

Displays overlaps between selected cluster and the search criteria for the above selected type/tissue.

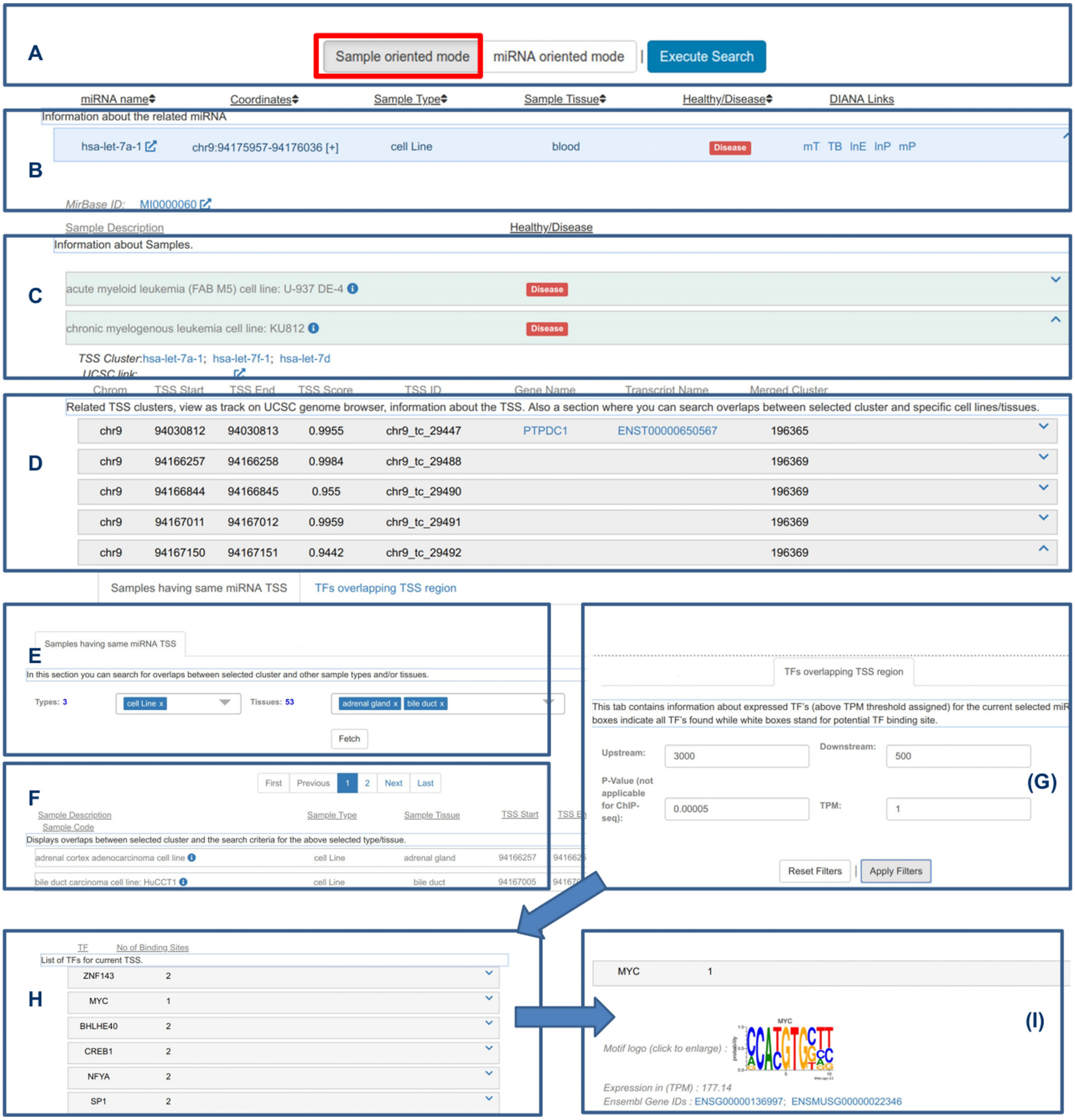
#	Distance	Coordinates
1	891	chr9:94166648-94166658 [-]

Footprint name: - *Footprint algorithm:* Fixed Length 3000bp
Binding site start relative to footprint: -
Binding site end relative to footprint: -

#	Distance	Coordinates
2	90	chr9:94165847-94166370

Footprint name: - *Footprint algorithm:* Chip_seq
Binding site start relative to footprint: -
Binding site end relative to footprint: -

Figure 2. Web interface visualization of the miRNA oriented query mode results. (**A**) Search fields to query for single or multiple miRNAs and/or TFs. The underlying miRNA nomenclature is based on the hairpin miRBase naming standards. TF names were derived JASPAR 2018. (**B**) Options for filtering results based on supported species, sample type (tissue, cell line and primary cell), healthy/disease status and retrieval ofall TSSs or only the ones closest to the corresponding pre-miRNAs. (**C**) Options for selecting between the two query modes. Results can be retrieved after pressing the 'Execute Search' button. (**D**) Results' main panel (one per hairpin) with pre-miRNA related information and interconnection with other DIANA resources. (**E**) Sub panels with TSS cluster related information. (**F**) Panel with metadata related to TFs that were found to bind in the region surrounding the corresponding TSS cluster. The number next to the TF's name indicates the number of identified binding sites. Motif logo of TF binding preference, expression level and links to Ensembl can be viewed directly below the sub panel. (**G**) Detailed information related to each TFBS such as absolute and relative coordinates to the cluster and level of support (ChIP-Seq, DNase-Seq footprinting and motif scanning).



Sample oriented mode | miRNA oriented mode | Execute Search

miRNA name	Coordinates	Sample Type	Sample Tissue	Healthy/Disease	DIANA Links
hsa-let-7a-1	chr9:94175957-94176036 [+]	cell Line	blood	Disease	mT TB InE InP mP

Information about the related miRNA

MirBase ID: MI0000060

Sample Description	Healthy/Disease
acute myeloid leukemia (FAB M5) cell line: U-937 DE-4	Disease
chronic myelogenous leukemia cell line: KU812	Disease

Information about Samples.

TSS Cluster: hsa-let-7a-1; hsa-let-7f-1; hsa-let-7d

UCSC link:

Chrom	TSS Start	TSS End	TSS Score	TSS ID	Gene Name	Transcript Name	Merged Cluster
chr9	94030812	94030813	0.9955	chr9_tc_29447	PTPDC1	ENST00000650567	196365
chr9	94166257	94166258	0.9984	chr9_tc_29488			196369
chr9	94166844	94166845	0.955	chr9_tc_29490			196369
chr9	94167011	94167012	0.9959	chr9_tc_29491			196369
chr9	94167150	94167151	0.9442	chr9_tc_29492			196369

Related TSS clusters, view as track on UCSC genome browser, information about the TSS. Also a section where you can search overlaps between selected cluster and specific cell lines/tissues.

Samples having same miRNA TSS | TFs overlapping TSS region

Samples having same miRNA TSS

In this section you can search for overlaps between selected cluster and other sample types and/or tissues.

Types: 3 — cell Line x

Tissues: 53 — adrenal gland x, bile duct x

Fetch

First | Previous | 1 | 2 | Next | Last

Sample Description Sample Code	Sample Type	Sample Tissue	TSS Start	TSS E
adrenal cortex adenocarcinoma cell line	cell Line	adrenal gland	94166257	941662
bile duct carcinoma cell line: HuCCT1	cell Line	bile duct	94167005	94167

Displays overlaps between selected cluster and the search criteria for the above selected type/tissue.

TFs overlapping TSS region

This tab contains information about expressed TF's (above TPM threshold assigned) for the current selected miR
boxes indicate all TF's found while white boxes stand for potential TF binding site.

Upstream: 3000
Downstream: 500
P-Value (not applicable for ChIP-seq): 0.00005
TPM: 1

Reset Filters | Apply Filters

TF	No of Binding Sites
ZNF143	2
MYC	1
BHLHE40	2
CREB1	2
NFYA	2
SP1	2

List of TFs for current TSS.

MYC 1

Motif logo (click to enlarge) :

Expression in (TPM) : 177.14
Ensembl Gene IDs : ENSG00000136997; ENSMUSG00000022346

Figure 3. Web interface visualization of the sample oriented query mode results. (**A**) Results can be viewed after selecting the ‘sample oriented mode’ filter and pressing the ‘Execute Search’ button. See Figure 2 for information related to selecting miRNAs, TFs, cell types and healthy/disease status. (**B**) Results’ main panel (one per hairpin) with information related to the pre-miRNA, type and source tissue as well as healthy/disease status of samples with at least one TSS identified in the 150kb search space upstream of the hairpin, and interconnection with other DIANA resources. (**C**) Expandable sub panels with information about each sample, pre-miRNAs sharing the same TSS and link to view the results on UCSC genome browser. (**D**) Upon expanding the sample’s sub panel, the inner tier of TSS panels can be accessed, hosting the genomic coordinates, expression level and hosting protein coding gene in case of intragenic pre-miRNAs. The expansion of each TSS panel leads to two tabs. (**E**, **F**) One assists users in retrieving other samples that also support a transcription initiation event closeby to the expanded TSS. (**G–I**) The other, similar to Figure 2F, G, provides access to TFBSs in a user-defined region around the expanded TSS based on TF expression level and FIMO *P*-value filtering options.

cluster). The information regarding each member of TSS clusters is retained and it is accessible through the web interface.

For the characterization of miRNA gene TSSs, all human pre-miRNAs from miRBase v22.1 (38) were downloaded. It has been previously shown that pri-miRNA transcripts can sometimes exhibit sizes of more than 100 kb (15). In DIANA-miRGen v4, all putative CAGE-derived miRNA gene TSSs are selected within a region starting at the location of each pre-miRNA and spanning up to 150 kb upstream. All CAGE tag-clusters passing ADAPT-CAGE's score cutoff and located in this search space are flagged as putative miRNA gene TSSs in the database. The CAGE tag-cluster that is closest to each pre-miRNA is highlighted as the corresponding miRNA gene TSS in the web interface, by default. However, our strategy of reporting all CAGE tag-clusters within the 150 kb search space facilitates studies related to alternative transcription initiation and/or tissue-specific promoter usage for miRNA genes. Pre-miRNAs located in protein coding gene introns may share a promoter with the host gene or have their promoter located in the same intron. The aforementioned search space strategy of 150kb is capturing both cases.

Analysis of ChIP- and DNase-Seq datasets

The generation of the TF:miRNA atlas hosted in DIANA-miRGen v4 was based on a three-step strategy. Initially, ChIP-Seq datasets for 94 TFs (Supplementary Table S1) were downloaded from the ENCODE public repository (9). For cases where peak files have not been generated by the ENCODE Consortium, we utilized Macs2 (39) with default parameters to identify loci enriched in ChIP-Seq signal. The transformation of genomic coordinates from earlier versions of the human genome to GRCh38 was achieved with UCSC liftover software (26).

Even though the number of TF ChIP-Seq samples in ENCODE is high, the tissue diversity of samples for each TF is low. Typically, a handful of TFs are studied exhaustively, while the majority of the remaining TFs in ENCODE are studied in a single tissue, cell line or primary cell type. To avoid this limitation and further expand our TF:miRNA collection, we also analysed chromatin accessibility data. We downloaded 61 DNase-seq samples from ENCODE (9) covering 25 tissues, 14 cell lines and 22 primary cell types (Supplementary Table S2). Homer (40) (size = 1000 minDist = 2500 o = auto tbp = 0) was utilized to characterize genomic loci enriched in DNase-Seq signal with a false discovery rate (FDR) threshold of 1%. Subsequently, Wellington algorithm (41) was applied on the identified accessible regions and the raw DNase-Seq signal to identify TF footprints using a P-value threshold of 10^{-8}. The connection between footprints and the occupying TF was achieved by scanning the sequence corresponding to footprints with the non-redundant TF motifs in JASPAR (25) using FIMO algorithm from MEME suite (42) with a P-value threshold of 10^{-5}.

FANTOM (10) repository hosts 1027 CAGE samples. Consequently, a big portion ($N = 863$) of these samples will not match to any ChIP- or DNase-Seq dataset in ENCODE (9). To generate TF:miRNA connections for such cases as well as for samples with ChIP- and/or DNase-Seq support (for completeness of the computational pipeline), we utilized FIMO and the previously aforementioned JASPAR TF motifs to scan a predefined area of 3 kb centered on TSS clusters as identified by ADAPT-CAGE algorithm (see analysis of CAGE samples and miRNA gene TSS identification section). The results of the three-step strategy described in this section are fully accessible by the user through the web interface as well as appropriate metadata that highlight the level of support (ChIP-Seq, footprinting and motif scanning) for each reported TFBS.

Transcription factor expression analysis

GRCh38 reference annotation was downloaded from Ensembl (43). For every sample in FANTOM, the normalized (tags per million - TPM) number of CAGE reads overlapping the regions centered around (±500 bp (44)) the annotated TF gene TSSs was utilized to quantify TF expression levels (12). This information is a crucial component of DIANA-miRGen v4 functionality since the appropriate tissue, cell line or primary cell type specific TF:miRNA connections depend on it. Additionally, users are able to filter queried results based on the desired TF expression level.

Database and web interface development

The existing relational schema implemented for DIANA-miRGen v3 was extended to facilitate the new data. New indices were created to ensure fast query execution. PostgreSQL RDBMS v10 is used to host the database. A new backend was developed using Java Spring framework and .NET Core 2.2. The interface was implemented using Angular v8 with bootstrap 3 CSS. A special effort was made to keep the main look and feel as close as possible to DIANA-miRGen v3 to allow the unhindered transition between the two releases for existing users. Explanatory tooltips and legends were added in various places of the tool so that users can have a quick reference on the part of data they are working on. The hairpin and TF search fields were re-designed to extend the existing functionality. The proposed ids do not include those already selected, thus speeding up the search process by providing less and less options. Furthermore, users can add multiple ids to those search fields by pasting a list of ids separated with commas. We added paging and sorting capabilities to all levels of information. Finally, we moved secondary sample and TF fields to the inner levels to provide a finer search experience, by allowing different search options for each TSS/Cluster.

Tissue-specific analysis of TFs regulating miRNA gene expression

Aim of miRGen is to aggregate as much information as possible regarding miRNA gene promoters and their regulators, and organize this plethora of data intuitively, thus enabling the research community to study gene regulatory networks with increased resolution. As an example of a miRGen use-case, we proceeded with a meta-analysis of intergenic and intragenic miRNA gene regulators in 11 cell types where we found a match between CAGE and TF ChIP-Seq

Table 1. Groups with matching CAGE and ChIP-Seq data from 11 cell types

Group	FANTOM id	#TFs	Cell type	Description
Group 1	CNhs12331 CNhs12332 CNhs12333	50	GM12878	B lymphoblastoid cell line
Group 2	CNhs11943	29	MCF7	Breast carcinoma cell line
Group 3	CNhs11250 CNhs12334 CNhs12335 CNhs12336	60	K562	Chronic myelogenous leukemia cell line
Group 4	CNhs11047 CNhs11046	1	NA	Embryonic kidney cell line
Group 5	CNhs10872 CNhs11967 CNhs12010	5	NA	Endothelial Cells - Umbilical
Group 6	CNhs12325 CNhs12327 CNhs12326	24	HelaS3	Epitheloid carcinoma cell line
Group 7	CNhs14067 CNhs14068 CNhs13964	23	H1	H1 embryonic stem cells differentiation to CD34+ HSC
Group 8	CNhs12328 CNhs12329 CNhs12330	32	HepG2	Hepatocellular carcinoma cell line
Group 9	CNhs12458 CNhs12684 CNhs12786	60	K562	K562 erythroblastic leukemia response to hemin
Group 10	CNhs10624 CNhs11798	10	NA	Liver adult
Group 11	CNhs11275	15	A549	Lung adenocarcinoma cell line

datasets (Supplementary Figure S1, Supplementary Table S3).

Out of the 1027 CAGE samples, 28 exhibit matching TF ChIP-Seq data from ENCODE and were arranged in eleven (11) cell-line/tissue groups (Table 1). The miRNA gene promoters were defined as a −3 kb/+1 kb window around the most proximal CAGE peak that was identified as valid TSS by ADAPT-CAGE. Subsequently, the promoters were divided into intergenic and intragenic categories based on the miRNA genomic localization. TFBSs were enumerated by calculating the overlap between TF ChIP-Seq peaks with the CAGE derived promoters. Since group 4 only has 1 TF ChIP-Seq dataset, we did not include it in the figures.

Based on this analysis,observe similarities in the mixture of top-10 TFs between groups that belong in the same lineage such as the hematopoietic cluster (groups 1, 3, 5 and 9) (Supplementary Figure S1). Especially group 3 (granulocytes from chronic myelogenous leukemia) and group 9 (erythroblasts from chronic myelogenous leukemia), they exhibit an almost identical top-10 TF binding profile (Supplementary Figures S1B and F). Additionally, CREB, MAX and MYC were found to be the top regulators of both intergenic and intragenic miRNA gene promoters in MCF7 cells, as expected by surveying the literature (45,46). Interestingly, CTCF was found in the top-10 TFs of all groups except 2 and 6. CTCF is a TF known for its multiple roles and especially for its function as an anchor for the creation of topologically associating domains and chromatin conformation (47).

INTERFACE

Formulating queries to the database

The significant increase in data, compared to the previous version, is hosted in the database schema of DIANA-miRGen v3 (24) that has been slightly modified to facilitate new data and metadata as well as more complex queries. The plethora of information that is hosted in the database includes miRNA metadata regarding implications in physiological and pathological conditions from the literature, miRNA localization, CAGE-derived miRNA gene TSSs filtered by ADAPT-CAGE algorithm, computationally and experimentally derived TFBS in miRNA promoter regions as well as TF expression levels.

Two different approaches were implemented for querying the information hosted in the database (Figures 2 and 3). The first approach is miRNA and/or TF oriented (Figure 2). Users can type their favorite miRNA and/or TF (Figure 2A) and apply several filters prior to search execution such as selecting for human or mouse results, disease or healthy samples and the retrieval of all or just the closest TSS within the 150 kb region upstream of each miRNA (Figure 2B). The results can be further processed in real-time based on three additional filters related to sample type (cell line, primary cells and tissue), tissue specificity and biological or technical replicate number (Figure 2B). The idea behind this query approach is that TSSs are clustered across samples based on their in-between distance to highlight putative alternative miRNA gene promoter usage. The information regarding samples that participated in creating the across-samples TSS clusters is retained in the interface and can be further manipulated to include or exclude samples from this process.

The second query approach is sample oriented (Figure 3). The same input terms and filtering options, as described above, need to be used prior to search execution. However, in this query mode, the information regarding every sample specific TSS is returned and can be accessed independently from the web interface (Figure 3A–C). Within each sample specific TSS panel, users are still able to access the information that is generated by the first query strategy, that is groups of samples with TSSs localized in small distances. Essentially, users are able to access all the hosted information regardless of the chosen query mode. However, each mode provides a unique search path that answers different biological questions, or in other terms, investigating the same data from different perspectives.

Query results and available options

In the previous section, the two operating modes were introduced, highlighting two different query mechanisms. Regardless of the selected mode, the information is grouped and visualized in similar ways in the web interface. Pre-miRNAs sit at the top of the results' hierarchy, each one in its own unique panel (main results panel) with information related to hairpin name, genomic coordinates and interconnection to other DIANA Lab web servers and databases (Figure 2D). In the case of sample oriented query mode, the main panel also hosts information related to sample type, tissue name and healthy or disease status (Figure 3B).

In the miRNA oriented query mode, users can expand the main panel for each miRNA to further explore the results. For every TSS cluster there is a different sub panel hosting the cluster's name and genomic coordinates (Figure 2E). These sub panels can also be expanded and users can access information related to the samples contributing to the formation of a given cluster as well as TFs that are localized in a user-defined region around each identified TSS (Figure 2F, G). TF binding events can be specifically filtered based on their distance from TSSs, level of gene expression and statistical significance of the footprinting process and/or motif scanning.

In the sample oriented query mode, users can expand the main panel described above to explore each cell-specific TSS individually (Figure 3D). Sub panels hosting sample related information can be expanded to access the final layer of panels, one for every identified TSS, showing their genomic coordinates, host gene name and ensembl id (for intragenic pre-miRNAs), other samples with TSSs in the local vicinity and nearby TFBSs as previously described (Figure 3E–I).

CONCLUSION

DIANA-miRGen v4 has evolved into the most comprehensive repository for miRNA gene regulation, hosting thousands of experimental data that connect miRNA genes and TFs regulating their transcription. The latest version of miRGen database is based on a state-of-the-art Machine Learning algorithm (16) to remove transcriptional noise from >1027 CAGE samples available in FANTOM (10), and characterize cell type specific TSSs for 1534 pre-miRNAs. In addition, the majority of TF ChIP- ($N = 368$) and DNase-Seq ($N = 61$) ENCODE (9) samples were reanalyzed and incorporated in the database facilitating a tier-based experimental evidence for TF:miRNA interactions. The database schema from the previous release was modified accordingly to support the increase in hosted data and more complex queries and the web interface was completely redesigned aiming to offer ease-of-use, advanced filtering, scalability and most importantly two mechanisms that enable the exploration of available data from different perspectives related to different biological questions. DIANA-miRGen v4 has surpassed previous releases as well as other miRNA gene regulation databases by offering high quality experimentally derived miRNA gene TSSs and TFBSs from multiple sources in thousands of samples, bringing the community one step closer to the complete integration of miRNAs in gene regulatory networks.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

'ELIXIR-GR: The Greek Research Infrastructure for Data Management and Analysis in Life Sciences' [MIS 5002780], which is implemented under the Action 'Reinforcement of the Research and Innovation Infrastructure', and 'Metalasso: A platform for computational meta-analysis of Genome Wide Association Studies-GWAS' [MIS 5032832], which is implemented under the Action 'Research-Create-Innovate', both funded by the Operational Programme 'Competitiveness, Entrepreneurship and Innovation' [NSRF 2014–2020] and co-financed by Greece and the European Union (European Regional Development Fund); Postdoc@MUNI [CZ.02.2.69/0.0/0.0/16_027/0008360 to G.K.G.]; I.K.s postdoctoral research was carried out under the call 'call for interest for postdoctoral researchers, scholarship for postdoctoral research' of University of Thessaly, that is implemented by University of Thessaly and funded by the 'Stavros Niarchos Foundation'. Funding for open access charge: project 'Metalasso'.
Conflict of interest statement. None declared.

REFERENCES

1. Ambros,V. (2001) microRNAs: tiny regulators with great potential. *Cell*, **107**, 823–826.
2. Agarwal,V., Bell,G.W., Nam,J.-W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.
3. Paraskevopoulou,M.D., Georgakilas,G., Kostoulas,N., Vlachos,I.S., Vergoulis,T., Reczko,M., Filippidis,C., Dalamagas,T. and Hatzigeorgiou,A.G. (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res.*, **41**, W169–W173.
4. Jones-Rhoades,M.W., Bartel,D.P. and Bartel,B. (2006) MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.*, **57**, 19–53.
5. Kloosterman,W.P. (2004) Substrate requirements for let-7 function in the developing zebrafish embryo. *Nucleic Acids Res.*, **32**, 6284–6291.
6. Lee,Y., Kim,M., Han,J., Yeom,K.-H., Lee,S., Baek,S.H. and Kim,V.N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, **23**, 4051–4060.
7. Lee,Y., Ahn,C., Han,J., Choi,H., Kim,J., Yim,J., Lee,J., Provost,P., Rådmark,O., Kim,S. *et al.* (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415–419.
8. Hutvágner,G., McLachlan,J., Pasquinelli,A.E., Bálint,E., Tuschl,T. and Zamore,P.D. (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, **293**, 834–838.
9. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
10. Lizio,M., Harshbarger,J., Shimoji,H., Severin,J., Kasukawa,T., Sahin,S., Abugessaisa,I., Fukuda,S., Hori,F., Ishikawa-Kato,S. *et al.* (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.*, **16**, 22.
11. Shiraki,T., Kondo,S., Katayama,S., Waki,K., Kasukawa,T., Kawaji,H., Kodzius,R., Watahiki,A., Nakamura,M., Arakawa,T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 15776–15781.
12. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest,A.R.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M.J.L., Haberle,V., Lassmann,T., Kulakovskiy,I.V., Lizio,M. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
13. ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
14. Marsico,A., Huska,M.R., Lasserre,J., Hu,H., Vucicevic,D., Musahl,A., Orom,U. and Vingron,M. (2013) PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. *Genome Biol.*, **14**, R84.
15. Georgakilas,G., Vlachos,I.S., Paraskevopoulou,M.D., Yang,P., Zhang,Y., Economides,A.N. and Hatzigeorgiou,A.G. (2014) microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. *Nat. Commun.*, **5**, 5700.
16. Georgakilas,G.K., Perdikopanis,N. and Hatzigeorgiou,A. (2020) Solving the transcription start site identification problem with ADAPT-CAGE: a Machine Learning algorithm for analysis of CAGE data. *Sci. Rep.*, **10**, 877.

17. Yang,J.-H., Li,J.-H., Jiang,S., Zhou,H. and Qu,L.-H. (2013) ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. *Nucleic Acids Res.*, **41**, D177–D187.
18. Friard,O., Re,A., Taverna,D., De Bortoli,M. and Corá,D. (2010) CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse. *BMC Bioinformatics*, **11**, 435.
19. Hamed,M., Spaniol,C., Nazarieh,M. and Helms,V. (2015) TFmiR: a web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks. *Nucleic Acids Res.*, **43**, W283–W288.
20. Wang,S., Li,W., Lian,B., Liu,X., Zhang,Y., Dai,E., Yu,X., Meng,F., Jiang,W. and Li,X. (2015) TMREC: a database of transcription factor and MiRNA regulatory cascades in human diseases. *PLoS One*, **10**, e0125222.
21. Wang,J., Lu,M., Qiu,C. and Cui,Q. (2010) TransmiR: a transcription factor–microRNA regulation database. *Nucleic Acids Res.*, **38**, D119–D122.
22. Tong,Z., Cui,Q., Wang,J. and Zhou,Y. (2019) TransmiR v2.0: an updated transcription factor-microRNA regulation database. *Nucleic Acids Res.*, **47**, D253–D258.
23. Guo,Z., Maki,M., Ding,R., Yang,Y., Zhang,B. and Xiong,L. (2014) Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. *Sci. Rep.*, **4**, 5150.
24. Georgakilas,G., Vlachos,I.S., Zagganas,K., Vergoulis,T., Paraskevopoulou,M.D., Kanellos,I., Tsanakas,P., Dellis,D., Fevgas,A., Dalamagas,T. *et al.* (2016) DIANA-miRGen v3.0: accurate characterization of microRNA promoters and their regulators. *Nucleic Acids Res.*, **44**, D190–D195.
25. 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–D1284.
26. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
27. Vlachos,I.S., Paraskevopoulou,M.D., Karagkouni,D., Georgakilas,G., Vergoulis,T., Kanellos,I., Anastasopoulos,I.-L., Maniou,S., Karathanou,K., Kalfakakou,D. *et al.* (2015) DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res.*, **43**, D153–D159.
28. Karagkouni,D., Paraskevopoulou,M.D., Chatzopoulos,S., Vlachos,I.S., Tastsoglou,S., Kanellos,I., Papadimitriou,D., Kavakiotis,I., Maniou,S., Skoufos,G. *et al.* (2018) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic Acids Res.*, **46**, D239–D245.
29. Vlachos,I.S., Zagganas,K., Paraskevopoulou,M.D., Georgakilas,G., Karagkouni,D., Vergoulis,T., Dalamagas,T. and Hatzigeorgiou,A.G. (2015) DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res.*, **43**, W460–W466.
30. Paraskevopoulou,M.D., Vlachos,I.S., Karagkouni,D., Georgakilas,G., Kanellos,I., Vergoulis,T., Zagganas,K., Tsanakas,P., Floros,E., Dalamagas,T. *et al.* (2016) DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Res.*, **44**, D231–D238.
31. Carninci,P. (2010) RNA dust: where are the genes? *DNA Res.*, **17**, 51–59.
32. Takahashi,H., Lassmann,T., Murata,M. and Carninci,P. (2012) 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.*, **7**, 542–561.
33. Fejes-Toth,K., Sotirova,V., Sachidanandam,R., Assaf,G., Hannon,G.J., Kapranov,P., Foissac,S., Willingham,A.T., Duttagupta,R., Dumais,E. *et al.* (2009) Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs: affymetrix/cold Spring Harbor Laboratory ENCODE Transcriptome Project. *Nature*, **457**, 1028.
34. Gan,Y., Guan,J. and Zhou,S. (2012) A comparison study on feature selection of DNA structural properties for promoter prediction. *BMC Bioinformatics*, **13**, 4.
35. Fukue,Y., Sumida,N., Nishikawa,J.-I. and Ohyama,T. (2004) Core promoter elements of eukaryotic genes have a highly distinctive mechanical property. *Nucleic Acids Res.*, **32**, 5834–5840.
36. Kanhere,A. and Bansal,M. (2005) Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Res.*, **33**, 3165–3175.
37. Abeel,T., Saeys,Y., Bonnet,E., Rouze,P. and Van de Peer,Y. (2008) Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.*, **18**, 310–323.
38. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
39. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
40. 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.
41. Piper,J., Elze,M.C., Cauchy,P., Cockerill,P.N., Bonifer,C. and Ott,S. (2013) Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.*, **41**, e201.
42. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME Suite. *Nucleic Acids Res.*, **43**, W39–W49.
43. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
44. Kawaji,H., Lizio,M., Itoh,M., Kanamori-Katayama,M., Kaiho,A., Nishiyori-Sueki,H., Shin,J.W., Kojima-Ishiyama,M., Kawano,M., Murata,M. *et al.* (2014) Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res.*, **24**, 708–717.
45. Xu,J., Chen,Y. and Olopade,O.I. (2010) MYC and breast cancer. *Genes Cancer*, **1**, 629–640.
46. Steven,A. and Seliger,B. (2016) Control of CREB expression in tumors: from molecular mechanisms and signal transduction pathways to therapeutic target. *Oncotarget*, **7**, 35454–35465.
47. Kentepozidou,E., Aitken,S.J., Feig,C., Stefflova,K., Ibarra-Soria,X., Odom,D.T., Roller,M. and Flicek,P. (2020) Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol.*, **21**, 5.