D190–D195 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 19 November 2015*
doi: 10.1093/nar/gkv1254

DIANA-miRGen v3.0: accurate characterization of microRNA promoters and their regulators

Georgios Georgakilas[1,2,*,†], Ioannis S. Vlachos[1,2,3,†], Konstantinos Zagganas[4,5], Thanasis Vergoulis[4], Maria D. Paraskevopoulou[1,2], Ilias Kanellos[4,6], Panayiotis Tsanakas[6,7], Dimitris Dellis[7], Athanasios Fevgas[8], Theodore Dalamagas[4] and Artemis G. Hatzigeorgiou[1,2,4,*]

[1]DIANA-Lab, Department of Electrical & Computer Engineering, University of Thessaly, 382 21 Volos, Greece, [2]Hellenic Pasteur Institute, 115 21 Athens, Greece, [3]Laboratory for Experimental Surgery and Surgical Research 'N.S. Christeas', Medical School of Athens, University of Athens, 11527 Athens, Greece, [4]'Athena' Research and Innovation Center, 11524 Athens, Greece, [5]University of Peloponnese, Department of Informatics and Telecommunications, 22100 Tripoli, Greece, [6]School of Electrical and Computer Engineering, NTUA, 15773 Zografou, Greece, [7]Greek Research and Technology Network (GRNET), Athens 11527, Greece and [8]Department of Electrical & Computer Engineering, University of Thessaly, 382 21 Volos, Greece

Received September 30, 2015; Revised November 01, 2015; Accepted November 02, 2015

ABSTRACT

microRNAs (miRNAs) are small non-coding RNAs that actively fine-tune gene expression. The accurate characterization of the mechanisms underlying miRNA transcription regulation will further expand our knowledge regarding their implication in homeostatic and pathobiological networks. Aim of DIANA-miRGen v3.0 (http://www.microrna.gr/mirgen) is to provide for the first time accurate cell-line-specific miRNA gene transcription start sites (TSSs), coupled with genome-wide maps of transcription factor (TF) binding sites in order to unveil the mechanisms of miRNA transcription regulation. To this end, more than 7.3 billion RNA-, ChIP- and DNase-Seq next generation sequencing reads were analyzed/assembled and combined with state-of-the-art miRNA TSS prediction and TF binding site identification algorithms. The new database schema and web interface facilitates user interaction, provides advanced queries and innate connection with other DIANA resources for miRNA target identification and pathway analysis. The database currently supports 276 miRNA TSSs that correspond to 428 precursors and >19M binding sites of 202 TFs on a genome-wide scale in nine cell-lines and six tissues of *Homo sapiens* and *Mus musculus*.

INTRODUCTION

microRNAs (miRNAs) are short (18–25 nts) single stranded non-coding RNA molecules that post-transcriptionally regulate gene expression. Since the discovery of their abundant transcription in 2003 (1), miRNAs have been intensely researched for their implication in physiological and pathological conditions. The majority of mammalian miRNAs is transcribed by Polymerase II (Pol2) (2) resulting in the formation of capped, polyadenylated primary transcripts (pri-miRNAs). Pri-miRNAs contain hairpin-like structures (pre-miRNAs) which are processed by the RNase III type enzyme Drosha in the nucleus (3). Pre-miRNAs are subsequently released to the cytoplasm where they undergo their maturation process by Dicer (4). The mature product is loaded onto RNA induced silencing complex (RISC) and acts as a guide in order to mediate mRNA degradation (5) and/or translation suppression (6).

Even though there have been significant advancements in the understanding of the mechanisms underlying their biogenesis, function and role in disease, knowledge regarding the transcription regulation of miRNA genes still remains limited. This has been largely due to the lack of experimental and/or computational methodologies capable of detecting accurately and with high resolution miRNA gene transcription start sites (TSSs).

Databases of miRNA regulation

A common characteristic between available miRNA transcription regulation repositories is the lack of supporting

*To whom correspondence should be addressed. Tel: +30 24210 74758; Fax: +30 24210 74997; Email: arhatzig@inf.uth.gr
Correspondence may also be addressed to Artemis G. Hatzigeorgiou. Tel: +30 24210 74758; Fax: +30 24210 74997; Email: georgakilas@inf.uth.gr
†These authors contributed equally to the paper as first authors.

accurate miRNA TSSs. Most implementations cluster pre-miRNAs into transcriptional units by utilizing heuristic methods based on inter-miRNA distances. However, pri-miRNAs can span dozens of kilobases and such approaches can often be misleading. Other databases utilize the predictions of early miRNA TSS identification algorithms, most of which support low resolution predictions derived from wide transcription signals, such as histone marks. In this section, we provide a brief overview of the current state-of-the-art.

miRGen (7) was initially released in 2007 and focused on exploring the association between the genomic context of miRNAs and their function, providing spatial annotation of miRNAs as well as validated and *in silico*-derived connections between miRNAs and their target-genes. miRGen v2 (8) attempted to expand the scope of the database by incorporating information related to pri-miRNA expression regulation. Promoter regions of miRNAs were derived from computational and experimental sources (9–12), while TF–miRNA associations were calculated by motif scanning of the promoter regions based on position frequency matrices (PFMs). In addition, the database supports miRNA expression profiles in several tissues, as well as single nucleotide variants (SNVs) in TF binding sites and precursor encoding loci.

ChIPBase (13) groups pre-miRNAs downloaded from miRBase v17 (14) into transcriptional units and considers the cluster's TSS as the 5′ end of the first cluster member. The database provides TF-lncRNA as well as TF–miRNA interaction maps in multiple tissues and cell lines. ChIPBase integrates TF ChIP-Seq libraries derived from ENCODE as well as miRNA posttranscriptional regulation information from starBase (15) and Cytoscape-mediated network visualization (16).

CircuitsDB (17) groups miRBase v9.2 (18)-derived pre-miRNAs into clusters and the 5′-most location of each cluster is considered as its TSS. TF–miRNA and TF–gene interactions are detected by overlapping transcription factor binding sites, which are computationally derived from motif scanning analysis in a limited region (1 kb) surrounding miRNA clusters or protein coding gene TSSs, respectively. miRNA–gene interactions are derived using TargetScan (19) and TargetMiner (20) target prediction algorithms.

TFmiR (21) is a web server that integrates TF–gene, TF–miRNA, miRNA–gene and miRNA–miRNA interactions. TF–gene interactions are derived from TRANSFAC[v.2003] (22), OregAnno (23) and TRED (24) databases. TF–miRNA interactions are integrated from TransmiR (25), ChIPBase (13) and relevant literature. miRNA–gene interactions have been incorporated from miRTarBase (26), TarBase (27), miRecords (28) and starBase (15), while miRNA–miRNA interactions originate from PmmR (29).

TMREC (30) and TransmiR (25) are databases that utilize text mining on ∼100 and 5000 publications respectively, in order to extract TF–miRNA interactions in 21 diseases from 16 organisms. The quality of the hosted information is highly dependent on the curation process of the text mining results.

TSmiR (31) provides interaction maps between TFs, miRNAs and their target genes in multiple human tissues. Their TSSs have been obtained from already published *in silico* and experimental methodologies (9–11,32). ChIP-Seq-derived TF binding sites were downloaded from the ENCODE project repository for each of the studied tissues. TSmiR also hosts experimentally verified miRNA–gene interactions from miRTarBase (26) and miRecords (28) and *in silico* predicted interactions from TargetScan (19).

Despite the progress in unveiling the mechanisms of miRNA transcription regulation, the majority of existing studies for the identification of miRNA TSSs rely on low accuracy experimental techniques, *in silico* algorithms that provide low resolution/high false positive rate predictions and heuristics. In addition, the assembly of TF–miRNA interactions frequently is based on literature text-mining, TF motif-assisted scanning of the promoter regions and ChIP-Seq, which limits the search in one TF per experiment. Importantly, as shown by our group (33), previously available genome-wide NGS-based miRNA TSS identification techniques have low accuracy and a low signal to noise ratio.

Aim of DIANA-miRGen v3.0 is to remove the obscurity that surrounds miRNA transcription regulation by providing an accurate genome-wide map of TF–miRNA interactions for multiple tissues and cell-lines (Supplementary Table S1) in *Homo sapiens* and *Mus musculus*. To this end, microTSS (33), a state-of-the-art computational framework was applied on deeply sequenced RNA-, ChIP- and DNase-Seq data resulting in the identification of 276 tissue/cell-line specific TSSs for 428 miRNA precursors in single nucleotide resolution. microTSS is the first algorithm that surpassed the barrier of 54% sensitivity and 64.5% precision in miRNA TSS identification by achieving 93.6% sensitivity and 100% precision, when validated against experimentally identified miRNA TSSs derived from *Drosha null/conditional-null* mouse embryonic stem cells (33). More than 200 TF PFMs were combined with in-house assembled RNA-Seq expression profiles, in order to create sets of motifs specific to each of the studied tissues and cell-lines (34).

The aforementioned wealth of information is hosted in a re-designed database schema and is freely accessible through an intuitive and easy-to-use interface (Figure 1) that incorporates rich meta-data regarding the function of miRNAs and TFs as well as their implication in physiological conditions and diseases. The interconnection between miRGen v3.0 and other DIANA resources enables users to perform miRNA pathway analyses with miRPath (35), identify miRNA predicted targets on protein coding genes with microT (36) and validated targets with TarBase (37) or *in silico* as well as experimentally verified miRNA targets on lncRNAs with LncBase (38).

METHODS AND RESULTS

Analysis of raw RNA- and ChIP-Seq data sets

Raw RNA-Seq as well as H3K4me3 and Pol2 ChIP-Seq data, corresponding to nine cell-lines and six tissues in human and mouse (Supplementary Table S1), have been derived from the key integrative ENCODE publications (39,40) and downloaded from public online repositories (41,42). Quality control has been performed using FastQC (43). Contaminants were detected and removed utilizing a

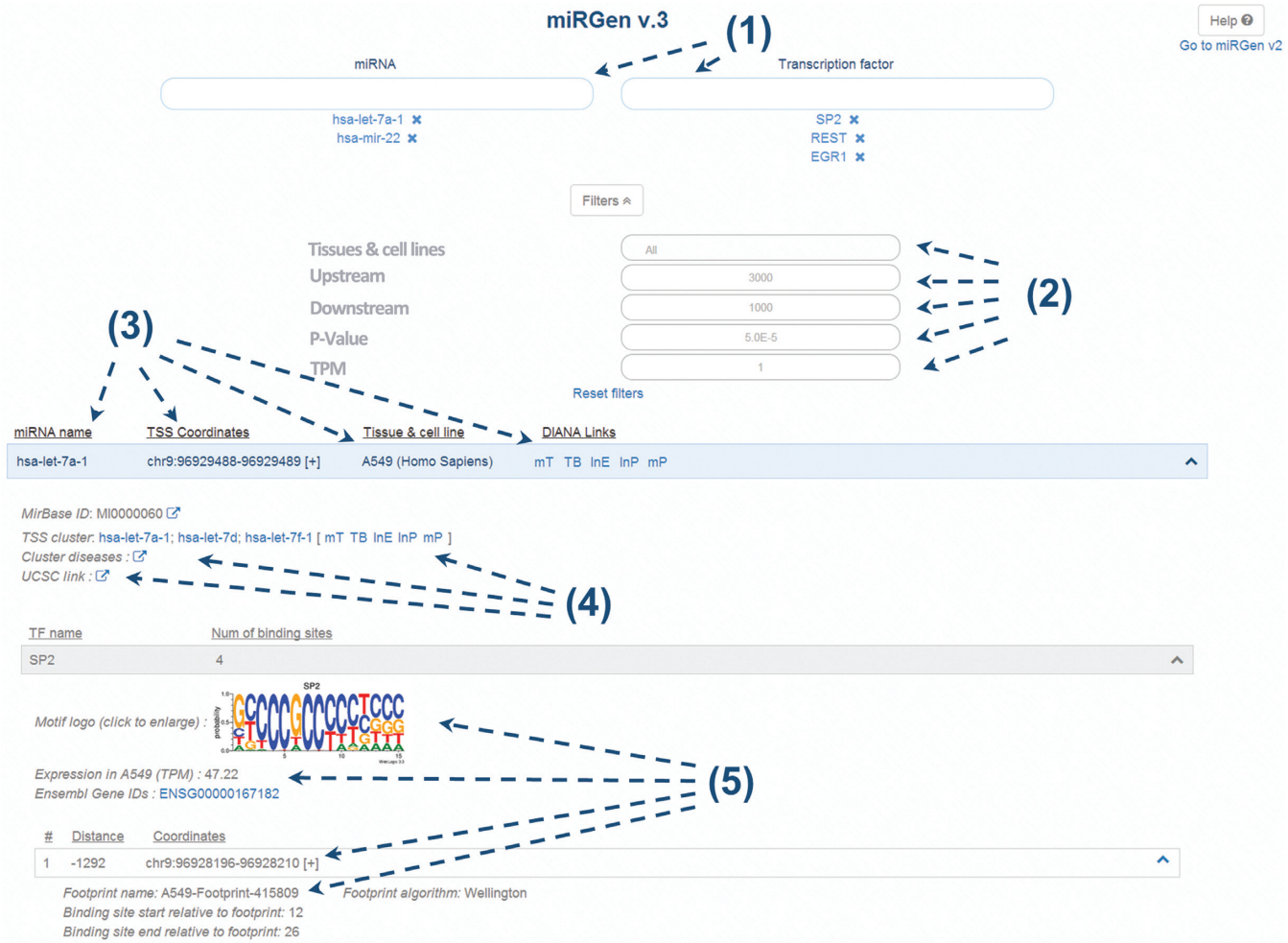

Figure 1. DIANA-miRGen v3.0 interface. Users are able to query (**1**) the database by entering pre-miRNA and/or TF names in the relevant search fields. An expandable advanced filtering menu (**2**) enables the selection of specific tissues/cell-lines, the TF binding site search space surrounding each miRNA TSSs as well as different thresholds on TF expression (transcripts per million normalization) and FIMO-derived *P*-value of TF's binding strength on the underlying footprint. The interface supports extensive miRNA-related information (**3**–**4**) such as the genomic location of the TSS, the associated cluster members, external links to miRBase, microT-CDS, TarBase v7.0, LncBase, miRPath v3.0 and a graphical representation of the queried regulatory region in the UCSC genome browser. Additionally, users are able to explore information (**5**) related to the expression of the TFs found to regulate each miRNA, genomic location and graphical representation of each binding site.

combination of an in-house developed algorithm and already available tools such as Minion (44), Reaper (44), Trimgalore (45) and Trimmomatic (46). Following preprocessing, GSNAP spliced aligner (47) was utilized to map the RNA-Seq reads against the reference genomes (GRCm38/mm10 and GRCh37/hg19 genome assemblies). GSNAP has been appropriately parameterized in order to detect novel and known splice junctions. RSEM (48) has been utilized to quantify gene expression. RSEM has been used with a reference transcript annotation derived from Ensembl (49) 75 (GRCh37) and 81 (GRCm38) for human and mouse, respectively. The alignment of the pre-processed ChIP-Seq reads against the reference genomes was performed with Bowtie v1 (50). The analysis resulted in ∼2.7 billion RNA-Seq uniquely mapped paired-end (PE) reads and ∼529M ChIP-Seq uniquely mapped single-end (SE) reads. Reads aligned to more than one genomic location have been discarded from subsequent analyses.

Analysis of DNase-Seq data sets

In order to assemble a genome-wide map of transcription factor binding sites in the studied tissues and cell-lines, approximately 4.1 billion DNase-Seq reads (39,40) were analyzed (Supplementary Table S1). HOMER (51) was utilized for the detection of genomic regions enriched in DNase-Seq signal (hotspots) with a false discovery rate threshold of 1%. Hotspots were subsequently processed with Wellington algorithm (52) in order to identify transcription factor footprints using a *P*-value threshold of 10^{-8}. The liftOver algorithm from UCSC repository (53) was utilized to transfer *Mus musculus* NCBI37/mm9 footprints to GRCm38/mm10 genome assembly.

Approximately 200 non-redundant TF binding motifs were downloaded from JASPAR core (34). Tissue/cell-line specific sets of TFs were created by filtering JASPAR-derived PFMs with the analyzed RNA-Seq expression data using a threshold of one transcript per million (TPM). For

each tissue/cell-line, footprints identified by Wellington and PFMs of expressed TFs were combined with FIMO (54) in order to create a genome-wide map of TF binding sites using a robust *P*-value threshold of 10^{-5}. TF motif logos have been generated with WebLogo v3.3 (55).

Identification of miRNA TSSs

The accurate identification of miRNA TSSs in the studied tissues and cell-lines was enabled with microTSS algorithm (33) and miRBase (56) database. microTSS can accurately identify miRNA TSSs in single nucleotide resolution by integrating deeply sequenced RNA samples with ChIP-Seq derived H3K4me3 and Pol2 occupancy data as well as open chromatin domains identified by DNase-Seq. microTSS was appropriately parameterized to maximize sensitivity and applied on the previously described RNA-, ChIP- and DNase-Seq datasets resulting in the identification of 276 tissue/cell-line specific miRNA TSSs that correspond to 428 pre-miRNAs. miRNAs with common TSS were grouped into clusters.

Database interface development

A new relational schema was designed to host all miRGen v3.0 data. Indices were created to guarantee the efficient execution of the system and foreign keys were added to avoid integrity violations in the data. PostgreSQL was utilized to implement the hosting database. MiRGen's Web interface (Figure 1) was designed around the new database schema and effort was made to be adaptable to a wide variety of screen formats and devices (PCs, tablets, smartphones, etc.). The interface has been developed using the Yii 2.0 PHP framework. The precursor and transcription factor search fields were designed as auto-complete search boxes to assist users in selecting the proper search keywords. Finally, useful filters were implemented to facilitate focusing on particular data that match the interests of the user.

INTERFACE

Querying the database

DIANA-miRGen v3.0 supports miRBase and Ensembl nomenclatures allowing for any combination of pre-miRNA and TF names as valid search terms for querying the database (Figure 1). The hierarchy of the results is organized in three main sections: 'miRNA', associated 'TFs' and their corresponding 'binding sites'. The first section hosts miRNA-related metadata such as the genomic coordinates of the miRNA TSS, the relevant tissue or cell-line, other members of the cluster in case of polycistronic miRNAs, direct links to miRBase, microT-CDS, TarBase, LncBase and miRPath as well as information on the implication of the miRNA in pathological conditions in the form of tag clouds derived from MeSH disease terms. In the TF section, users can explore the expression levels of each TF in the queried tissues or cell lines and other TF-related information such as the TF motif logo, the source of the utilized PFM and links to external databases for pursuing additional evidence. The last section supports information related to each TF binding site and the associated TF footprint. Users can identify the genomic coordinates of the binding site, its relative distance from the TSS, the strength of its binding in the form of a p-value as calculated from FIMO, the associated TF footprint and the algorithm utilized for its identification.

DIANA-miRGen v3.0 offers multiple ways to filter the displayed results in the form of an easily accessed and expandable menu. The provided search and filtering options include: combinations of pre-miRNA/TF names, the length of the region surrounding each TSS that will be investigated for TF binding sites, the supported tissues and cell lines, a threshold on the *P*-value provided by FIMO for the strength of the binding sites and a threshold of the expression value for each TF.

CONCLUSION

DIANA-miRGen v3.0 miRNA regulatory repository has significantly widened the scope of previous releases with a completely redesigned database schema and web interface as well as the wealth of supporting information. During the last decade, numerous NGS techniques have emerged enabling the implementation of accurate and robust miRNA TSS identification algorithms. The latest version of miRGen database utilizes miRNA TSS predictions derived from microTSS, which up to this date is considered as the most accurate algorithm for identifying the transcription start site of miRNA genes. In addition, experimental protocols have been specifically developed for assembling binding sites of single (ChIP-Seq) or multiple (DNase-Seq) TFs on a genome-wide scale. DIANA-miRGen v3.0 supports state-of-the-art tissue/cell-line specific miRNA TSS predictions and DNase-Seq mediated TF binding sites from 9 cell-lines and 6 tissues in *Homo sapiens* and *Mus musculus*. The database schema and web interface were designed from the ground up to support ease-of-use, advanced queries and filtering of the results as well as to facilitate the integration of additional experimental evidence and meta-data in the future. The volume and quality of information will enable researchers to add more pieces to the puzzle of biological networks by incorporating the regulatory mechanisms of miRNA transcription.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The majority of the analyses presented in this study were performed in the National HPC Infrastructure of Greek Research and Technology Network.

FUNDING

'TOM' [2862], 'ARISTEIA' Action of the 'OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING', General Secretariat for Research and Technology, Ministry of Education, Greece, European Social Fund (ESF); National Resources and a Fondation Santé grant to Artemis Hatzigeorgiou. Funding for open access charge: a Fondation Santé grant to Artemis Hatzigeorgiou.

Conflict of interest statement. None declared.

REFERENCES

1. Ambros,V. (2001) microRNAs: tiny regulators with great potential. *Cell*, **107**, 823–826.
2. 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.
3. Lee,Y., Ahn,C., Han,J., Choi,H., Kim,J., Yim,J., Lee,J., Provost,P., Radmark,O., Kim,S. *et al.* (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415–419.
4. Hutvagner,G., McLachlan,J., Pasquinelli,A.E., Balint,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.
5. 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.
6. Kloosterman,W.P., Wienholds,E., Ketting,R.F. and Plasterk,R.H. (2004) Substrate requirements for let-7 function in the developing zebrafish embryo. *Nucleic Acids Res.*, **32**, 6284–6291.
7. Megraw,M., Sethupathy,P., Corda,B. and Hatzigeorgiou,A.G. (2007) miRGen: a database for the study of animal microRNA genomic organization and function. *Nucleic Acids Res.*, **35**, D149–D155.
8. Alexiou,P., Vergoulis,T., Gleditzsch,M., Prekas,G., Dalamagas,T., Megraw,M., Grosse,I., Sellis,T. and Hatzigeorgiou,A.G. (2010) miRGen 2.0: a database of microRNA genomic information and regulation. *Nucleic Acids Res.*, **38**, D137–D141.
9. Corcoran,D.L., Pandit,K.V., Gordon,B., Bhattacharjee,A., Kaminski,N. and Benos,P.V. (2009) Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. *PLoS One*, **4**, e5279.
10. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
11. Ozsolak,F., Poling,L.L., Wang,Z., Liu,H., Liu,X.S., Roeder,R.G., Zhang,X., Song,J.S. and Fisher,D.E. (2008) Chromatin structure analyses identify miRNA promoters. *Genes Dev.*, **22**, 3172–3183.
12. Marson,A., Levine,S.S., Cole,M.F., Frampton,G.M., Brambrink,T., Johnstone,S., Guenther,M.G., Johnston,W.K., Wernig,M., Newman,J. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.
13. 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.
14. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
15. Yang,J.H., Li,J.H., Shao,P., Zhou,H., Chen,Y.Q. and Qu,L.H. (2011) starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res.*, **39**, D202–D209.
16. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
17. Friard,O., Re,A., Taverna,D., De Bortoli,M. and Cora,D. (2010) CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse. *BMC Bioinformatics*, **11**, 435.
18. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
19. Grimson,A., Farh,K.K., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
20. Bandyopadhyay,S. and Mitra,R. (2009) TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics*, **25**, 2625–2631.
21. 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.
22. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
23. Griffith,O.L., Montgomery,S.B., Bernier,B., Chu,B., Kasaian,K., Aerts,S., Mahony,S., Sleumer,M.C., Bilenky,M., Haeussler,M. *et al.* (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.*, **36**, D107–D113.
24. 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.
25. Wang,J., Lu,M., Qiu,C. and Cui,Q. (2010) TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Res.*, **38**, D119–D122.
26. Hsu,S.D., Lin,F.M., Wu,W.Y., Liang,C., Huang,W.C., Chan,W.L., Tsai,W.T., Chen,G.Z., Lee,C.J., Chiu,C.M. *et al.* (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **39**, D163–D169.
27. Sethupathy,P., Corda,B. and Hatzigeorgiou,A.G. (2006) TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, **12**, 192–197.
28. Xiao,F., Zuo,Z., Cai,G., Kang,S., Gao,X. and Li,T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.
29. Sengupta,D. and Bandyopadhyay,S. (2011) Participation of microRNAs in human interactome: extraction of microRNA-microRNA regulations. *Mol. bioSystems*, **7**, 1966–1973.
30. 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.
31. 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. *Scientific Rep.*, **4**, 5150.
32. Chien,C.H., Sun,Y.M., Chang,W.C., Chiang-Hsieh,P.Y., Lee,T.Y., Tsai,W.C., Horng,J.T., Tsou,A.P. and Huang,H.D. (2011) Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. *Nucleic Acids Res.*, **39**, 9345–9356.
33. 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.
34. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.Y., Chou,A., Ienasescu,H. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
35. 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.
36. 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.
37. 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.
38. Paraskevopoulou,M.D., Georgakilas,G., Kostoulas,N., Reczko,M., Maragkakis,M., Dalamagas,T.M. and Hatzigeorgiou,A.G. (2013) DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res.*, **41**, D239–D245.
39. Yue,F., Cheng,Y., Breschi,A., Vierstra,J., Wu,W., Ryba,T., Sandstrom,R., Ma,Z., Davis,C., Pope,B.D. *et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, **515**, 355–364.

40. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
41. Karolchik,D., Barber,G.P., Casper,J., Clawson,H., Cline,M.S., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, **42**, D764–D770.
42. 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.
43. Andrews,S. (2015) FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
44. Davis,M.P., van Dongen,S., Abreu-Goodger,C., Bartonicek,N. and Enright,A.J. (2013) Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods*, **63**, 41–49.
45. Krueger,F. (2015) Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
46. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
47. Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.
48. Li,B. and Dewey,C. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
49. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.
50. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
51. 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.
52. 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.
53. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
54. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
55. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
56. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.