Published online 2 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D181–D185***
doi: 10.1093/nar/gkx1035

MSDD: a manually curated database of experimentally supported associations among miRNAs, SNPs and human diseases

Ming Yue†, Dianshuang Zhou†, Hui Zhi†, Peng Wang†, Yan Zhang, Yue Gao, Maoni Guo, Xin Li, Yanxia Wang, Yunpeng Zhang*, Shangwei Ning* and Xia Li*

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

Received August 12, 2017; Revised September 30, 2017; Editorial Decision October 13, 2017; Accepted October 18, 2017

ABSTRACT

The MiRNA SNP Disease Database (MSDD, http://www.bio-bigdata.com/msdd/) is a manually curated database that provides comprehensive experimentally supported associations among microRNAs (miRNAs), single nucleotide polymorphisms (SNPs) and human diseases. SNPs in miRNA-related functional regions such as mature miRNAs, promoter regions, pri-miRNAs, pre-miRNAs and target gene 3′-UTRs, collectively called 'miRSNPs', represent a novel category of functional molecules. miRSNPs can lead to miRNA and its target gene dysregulation, and resulting in susceptibility to or onset of human diseases. A curated collection and summary of miRSNP-associated diseases is essential for a thorough understanding of the mechanisms and functions of miRSNPs. Here, we describe MSDD, which currently documents 525 associations among 182 human miRNAs, 197 SNPs, 153 genes and 164 human diseases through a review of more than 2000 published papers. Each association incorporates information on the miRNAs, SNPs, miRNA target genes and disease names, SNP locations and alleles, the miRNA dysfunctional pattern, experimental techniques, a brief functional description, the original reference and additional annotation. MSDD provides a user-friendly interface to conveniently browse, retrieve, download and submit novel data. MSDD will significantly improve our understanding of miRNA dysfunction in disease, and thus, MSDD has the potential to serve as a timely and valuable resource.

INTRODUCTION

MicroRNAs (miRNAs) are a class of small, endogenous, non-coding RNAs of ∼22 nt in length that post-transcriptionally regulate the cleavage of target mRNAs or participate in translational repression (1,2). To date, increasing evidence has shown that miRNAs are involved in various physiological processes and play important roles in various human diseases (3,4). Among these studies, the role of miRNA-related single nucleotide polymorphisms (SNPs) is gaining increasing attention.

Human miRNA biogenesis and function is a multi-step process. First, an miRNA gene is transcribed to produce a primary miRNA (pri-miRNA) that is then processed into a precursor miRNA (pre-miRNA) and subsequently into a mature miRNA, which ultimately binds to the 3′-UTR of the target messenger RNA (mRNA) (5). Emerging studies have shown that SNPs in pri-miRNAs, pre-miRNAs, mature miRNAs and 3′-UTRs of target mRNA may function as a novel class of regulatory SNPs (commonly called 'miRSNPs'), which can modify miRNA biogenesis and/or target binding and lead to diverse human diseases (6–8). However, most of these studies have each identified one or several disease-associated miRSNPs, so a wealth of information on experimentally supported miRSNPs is buried among the published literature and is not easily accessible. Thus, the need to develop a database to collect and store the latest experimentally supported miRNA–SNP disease associations is urgent.

Due to the important roles and functions of miRSNPs in regulating many aspects of cellular processes relating to disease development, several databases and web tools such as miRNASNP (9), MirSNP (10), PolymiRTS (11), Patrocles (12), SubmiRine (13), MicroSNiPer (14), miRNA–SNiPer (15), Mirsnpscore (16) and mrSNP (17) have been developed. These databases are useful for identifying functional miRSNP candidates. However, most of these databases only

*To whom correspondence should be addressed. Tel: +86 451 86615922; Fax: +86 451 86615922; Email: lixia@hrbmu.edu.cn
Correspondence may also be addressed to Shangwei Ning. Email: ningsw@ems.hrbmu.edu.cn
Correspondence may also be addressed to Yunpeng Zhang. Email: zyp19871208@126.com
†These authors contributed equally to the paper as first authors.

focus on predicting SNP effects on putative miRNA targets or RNA secondary structure or collect human SNPs in predicted miRNA–mRNA binding sites. Other databases, miRNASNP v2.0 (9) and miRdSNP (18), map phenotype-associated SNPs from genome-wide association studies to predicted or experimentally validated miRNA targets and do not provide miRNA dysfunctional patterns or experimental methods. To date, no database has been designed to capture the experimentally supported relationships among miRNAs, SNPs and genes. A manually curated database of experimentally supported miRSNPs that are associated with various human diseases will be useful for researchers and can serve as 'gold standard' data set for accuracy tests, especially for further experimental designs and verification.

To bridge this gap, we have developed **M**iRNA **S**NP **D**isease **D**atabase (**MSDD**), a manually curated database, to collect and integrate experimentally supported disease-associated miRSNPs into a high quality, comprehensive resource (Figure 1). The current version of MSDD documents 525 manually curated relationships between 182 human miRNAs, 197 SNPs, 153 genes and 164 human diseases. We expect that this elaborate database specifically designed for miRNAs, SNPs and human diseases will serve as an important catalyst for future research.

DATA COLLECTION AND DATABASE CONTENT

To ensure the quality of the database, all MSDD entries were manually collected through several steps that were used to assemble the databases miRTarBase (19), HMDD v2.0 (20) and Lnc2Cancer (21) in the collection process. First, we searched the PubMed database (22) with a list of keywords, such as 'SNP miRNA,' 'polymorphism miRNA,' 'SNP microRNA' and 'polymorphism microRNA.' All published literature expounding miRNA–SNP interactions that described associations with human diseases or traits was downloaded to extract the key information. It this step, 2387 published literature were downloaded from the PubMed database (22) (before May 2017). Second, we extracted experimentally supported miRNA–SNP disease associations by manually curating information from published papers. All selected studies were reviewed by at least two researchers. In this step, we retrieved the miRNA, SNP, miRNA target gene, disease name and all identifiers were manually annotated using the controlled vocabularies (e.g. controlled vocabulary for ncRNA classes, MeSH), the information of population, number of samples and minor allele frequency of disease-associated SNPs, the SNP location (relative position of the SNP, such as within the 3′-UTR or pre-miRNA) and allele, the miRNA dysfunctional pattern (the effect of the miRNA on the expression of the target gene, e.g. increase, decrease, loss or gain), the experimental methods used (e.g. western blot, quantitative reverse transcriptase-polymerase chain reaction (qRT-PCR)), the experimental samples (cell line and/or tissue), hyperlinks to the PubMed database and a brief functional description of the miRNA–SNP disease regulation mechanism from the original study. We referred to previous studies and selected miRNA–SNP disease associations for manual curation using strict criteria (23). We only collected high-quality associations with multiple lines of strong experimental evidence, including confirmation by genotyping, western blot, qRT-PCR or luciferase reporter assays.

After completing this process, a total of 525 associations between 182 human miRNAs, 197 SNPs, 153 genes and 164 human diseases were manually curated from 397 published papers. Each curated association was given a unique accession number (accession ID: MSDD00###). Moreover, We extracted the annotation information of each miRNA from miRBase (24), including the genome context, stem–loop structure and sequence of mature miRNA. We also downloaded the annotation information of each SNP from dbSNP (25), including ancestral allele (wild-type) and contextual information.

Finally, all data in MSDD is stored and managed using MySQL (version 5.7.18). The web interfaces were built in JSP. The data processing programs are written in Java (version 1.8.0), and the web services are built using Apache Tomcat. The MSDD database is freely available at http://www.bio-bigdata.com/msdd/ and http://www.bio-bigdata.net/msdd/.

USER INTERFACE

MSDD provides a user-friendly web interface that enables users to browse, search and retrieve all miRNA–SNP disease associations in the database (Figure 2). In the 'Browse' page, users can click on a specific gene, miRNA, SNP or disease name, and a list of matched entries is returned. In the 'Search' page, MSDD allows users to search by gene name, miRNA name, SNP ID, disease name or or combinations of these categories. MSDD offers fuzzy keyword searching capabilities, facilitating searches by returning the closest possible matching records. MSDD provides an option in the 'Search' page that allows users to filter associations by experimental method and SNP position. MSDD also offers a submission page that enables researchers to submit novel experimentally supported miRNA–SNP disease associations. Once approved by the submission review committee, the submitted record will be included in an updated release. In addition, all data in the database can be downloaded in the 'Download' page. MSDD also provides two visualization maps on the 'Download' page that enables users to download data by clicking on the appropriate area. One is a human body map that classifies data according to the organ and another map is a tag cloud that displays hotspot data. Finally, a detailed tutorial showing users how to use MSDD is available on the 'Help' page.

FUTURE EXTENSIONS

More recently, high-throughput technologies, such as next-generation sequencing, have produced extensive data on human disease biology, and the number of validated disease-associated miRNA–SNP interactions will continue to increase in the future. These advances in research will provide an opportunity to further extend MSDD. We will continue to manually curate newly validated miRNA–SNP disease associations and update the database every 2 months. We will incorporate new tools and functional annotations as well as more data sources to improve the utility and content coverage of this database.

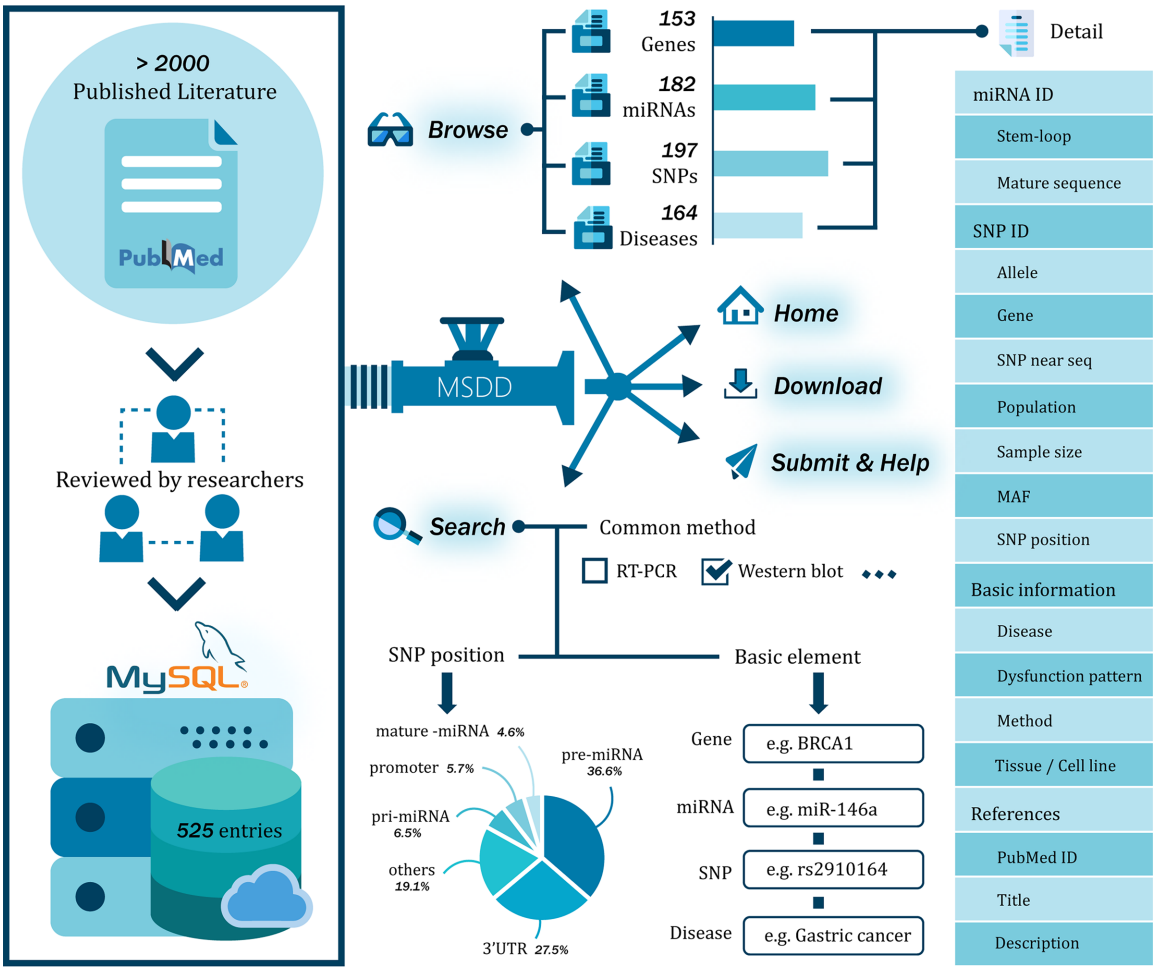

Figure 1. Data sources and the structure of MSDD.

DISCUSSION AND CONCLUSION

Over time, studies have increasingly indicated that variants in miRNA sequences may trigger disease by altering the expression or maturation of miRNAs or by interfering with miRNA interactions with mRNA (26,27). In the past few years, many databases have been published to aid researchers in exploring the impact of SNPs on miRNA genes and their targets. For SNPs in miRNA genes, it is convenient to map SNPs to miRNA regions from resources like miRBase (24). For SNPs in miRNA targets, identifying potential binding sites for a given miRNA in genomic sequences is important. However, these studies and databases emphasize the importance of prediction tools in the identification of potential miRNA–SNP relationships (Supplementary Table S1). To the best of our knowledge, none of these resources were developed to specifically collect experimentally supported miRNA–SNP association data in various human diseases. Thus, we developed MSDD, a disease-association database that provides a comprehensive resource on miRNA dysregulation with the modulation of SNPs.

In addition to collecting a great number of experimentally supported miRNA–SNP disease associations, MSDD may provide mechanistic insight and experimental evidence into future research. For example, by searching MSDD using 'miR-34,' a well-known human miRNA gene family, we found that the present study clearly demonstrates the SNP rs4938723, located in the promoter region of miR-34a/b/c, can significantly affect miRNA expression in various human cancers. This miRNA–SNP association has the potential to be an effective biomarker for different human cancers. In another example, by searching MSDD using 'gastric cancer', one of the most common cancer types worldwide, we found that several SNPs in miRNA genes or target sites have been proved to be associated with this cancer by affecting the miRNA-mediated regulatory function. More importantly, some functional miRNA–SNP associations, such as the SNP rs2910164 in the pre-miRNA of miR-146a, are supported both in blood and in tissue, which may be especially useful for cancer specialists who focusing on circulating miRNA cancer biomarker.

Data from MSDD can facilitate the understanding of important principles and future research trends. For example, despite that miRSNPs in many functional regions such as the promoter regions of miRNAs are associated with diseases, we found that most of the disease-associated miRSNPs are located four main functional regions, including pri-miRNAs, pre-miRNAs, mature-miRNAs and

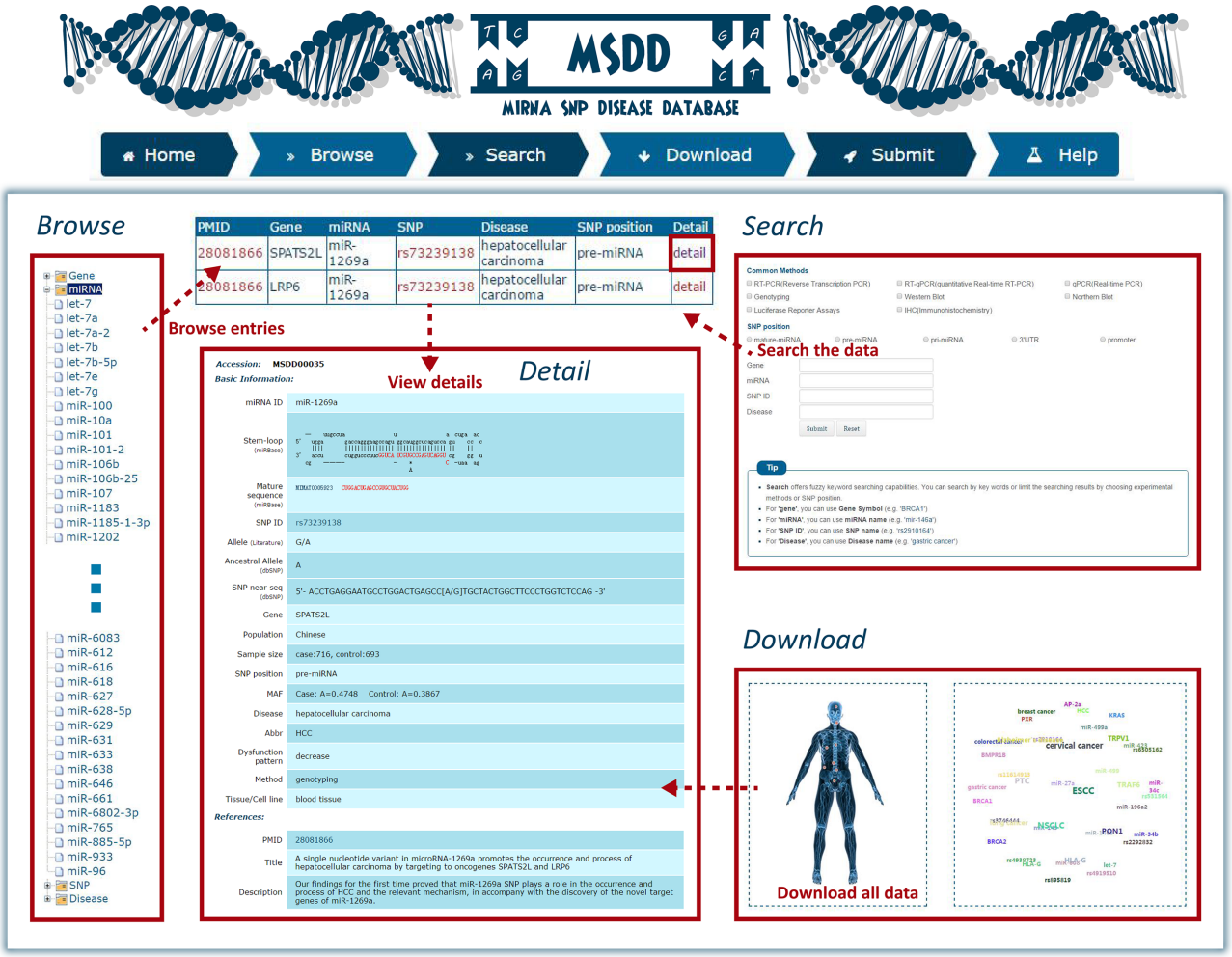

Figure 2. A schematic workflow of MSDD.

miRNA target genes (Supplementary Figure S1A). Additionally, we list the top 10 miRNAs, SNPs and disease in the database (Supplementary Figure S2A–C). Remarkably, we found that miR-146a and the SNP rs2910164 make up the most common miRNA–SNP pair in the MSDD, which is involved in 52 diseases. The disease with the highest connectivity is hepatocellular carcinoma, which is associated with 26 miRNAs and 25 SNPs, thus potentially providing further insight into this disease. Finally, we quantified the number of published papers each year that reported miRNA–SNP disease associations and found that the number of publications has generally been increasing dramatically (Supplementary Figure S1B). In particular, from 2012 to 2014, the number of publications has increased in an exponential manner, suggesting that research on miRNA–SNP disease associations has become a hot topic in recent years, thus highlighting the timeliness of developing a special-purpose repository to document these valuable data.

In summary, MSDD not only provides a comprehensive miRNA–SNP disease database with experimental support but also presents a more global view on miRNA functions in human diseases. MSDD will serve as a valuable resource for researchers interested in determining the role of miRNA-related SNPs in human diseases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National High Technology Research and Development Program of China [863 Program, 2014AA021102]; National Program on Key Basic Research Project [973 Program, 2014CB910504]; National Natural Science Foundation of China [91439117, 61473106, 31401090, 31501038, 31601080]; Harbin Special Funds for Innovative Talents of Science and Technology Research Project [RC2016QN003028]; Harbin Medical University (Yu Weihan Outstanding Youth Training Fund). Funding for open access charge: National Natural Science Foundation of China [91439117, 61473106, 31401090, 31501038, 31601080].

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
2. Leung,A.K. and Sharp,P.A. (2007) microRNAs: a safeguard against turmoil? *Cell*, **130**, 581–585.
3. Sayed,D. and Abdellatif,M. (2011) MicroRNAs in development and disease. *Physiol. Rev.*, **91**, 827–887.
4. Mendell,J.T. and Olson,E.N. (2012) MicroRNAs in stress signaling and human disease. *Cell*, **148**, 1172–1187.
5. Shukla,G.C., Singh,J. and Barik,S. (2011) MicroRNAs: processing, maturation, target recognition and regulatory functions. *Mol. Cell. Pharmacol.*, **3**, 83–92.
6. Duan,J., Shi,J., Fiorentino,A., Leites,C., Chen,X., Moy,W., Chen,J., Alexandrov,B.S., Usheva,A., He,D. *et al.* (2014) A rare functional noncoding variant at the GWAS-implicated MIR137/MIR2682 locus might confer risk to schizophrenia and bipolar disorder. *Am. J. Hum. Genet.*, **95**, 744–753.
7. Ryan,B.M., Robles,A.I., McClary,A.C., Haznadar,M., Bowman,E.D., Pine,S.R., Brown,D., Khan,M., Shiraishi,K., Kohno,T. *et al.* (2015) Identification of a functional SNP in the 3′UTR of CXCR2 that is associated with reduced risk of lung cancer. *Cancer Res.*, **75**, 566–575.
8. Gu,S., Rong,H., Zhang,G., Kang,L., Yang,M. and Guan,H. (2016) Functional SNP in 3′-UTR MicroRNA-binding site of ZNF350 confers risk for age-related cataract. *Hum. Mutat.*, **37**, 1223–1230.
9. Gong,J., Liu,C., Liu,W., Wu,Y., Ma,Z., Chen,H. and Guo,A.Y. (2015) An update of miRNASNP database for better SNP selection by GWAS data, miRNA expression and online tools. *Database (Oxford)*, **2015**, bav029.
10. Liu,C., Zhang,F., Li,T., Lu,M., Wang,L., Yue,W. and Zhang,D. (2012) MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. *BMC Genomics*, **13**, 661.
11. Bhattacharya,A., Ziebarth,J.D. and Cui,Y. (2014) PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. *Nucleic Acids Res.*, **42**, D86–D91.
12. Hiard,S., Charlier,C., Coppieters,W., Georges,M. and Baurain,D. (2010) Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. *Nucleic Acids Res.*, **38**, D640–D651.
13. Maxwell,E.K., Campbell,J.D., Spira,A. and Baxevanis,A.D. (2015) SubmiRine: assessing variants in microRNA targets using clinical genomic data sets. *Nucleic Acids Res.*, **43**, 3886–3898.
14. Barenboim,M., Zoltick,B.J., Guo,Y. and Weinberger,D.R. (2010) MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets. *Hum. Mutat.*, **31**, 1223–1232.
15. Zorc,M., Skok,D.J., Godnic,I., Calin,G.A., Horvat,S., Jiang,Z., Dovc,P. and Kunej,T. (2012) Catalog of microRNA seed polymorphisms in vertebrates. *PLoS One*, **7**, e30737.
16. Thomas,L.F., Saito,T. and Saetrom,P. (2011) Inferring causative variants in microRNA target sites. *Nucleic Acids Res.*, **39**, e109.
17. Deveci,M., Catalyurek,U.V. and Toland,A.E. (2014) mrSNP: software to detect SNP effects on microRNA binding. *BMC Bioinformatics*, **15**, 73.
18. Bruno,A.E., Li,L., Kalabus,J.L., Pan,Y., Yu,A. and Hu,Z. (2012) miRdSNP: a database of disease-associated SNPs and microRNA target sites on 3′UTRs of human genes. *BMC Genomics*, **13**, 44.
19. Chou,C.H., Chang,N.W., Shrestha,S., Hsu,S.D., Lin,Y.L., Lee,W.H., Yang,C.D., Hong,H.C., Wei,T.Y., Tu,S.J. *et al.* (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*, **44**, D239–D247.
20. Li,Y., Qiu,C., Tu,J., Geng,B., Yang,J., Jiang,T. and Cui,Q. (2014) HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.*, **42**, D1070–D1074.
21. Ning,S., Zhang,J., Wang,P., Zhi,H., Wang,J., Liu,Y., Gao,Y., Guo,M., Yue,M., Wang,L. *et al.* (2016) Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.*, **44**, D980–D985.
22. Coordinators, N.R. (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **44**, D7–D19.
23. 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.
24. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
25. Smigielski,E.M., Sirotkin,K., Ward,M. and Sherry,S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.*, **28**, 352–355.
26. Yang,P.W., Huang,Y.C., Hsieh,C.Y., Hua,K.T., Huang,Y.T., Chiang,T.H., Chen,J.S., Huang,P.M., Hsu,H.H., Kuo,S.W. *et al.* (2014) Association of miRNA-related genetic polymorphisms and prognosis in patients with esophageal squamous cell carcinoma. *Ann. Surg. Oncol.*, **21**, S601–S609.
27. Ganzinelli,M., Rulli,E., Caiola,E., Garassino,M.C., Broggini,M., Copreni,E., Piva,S., Longo,F., Labianca,R., Bareggi,C. *et al.* (2015) Role of KRAS-LCS6 polymorphism in advanced NSCLC patients treated with erlotinib or docetaxel in second line treatment (TAILOR). *Sci. Rep.*, **5**, 16331.