MIRNA-DISTILLER: a stand-alone application to compile microRNA data from databases

Jessica K. Rieger, Denis A. Bodan and Ulrich M. Zanger*

Dr. Margarete Fischer-Bosch Institute of Clinical Pharmacology, University of Tuebingen, Stuttgart, Germany

MicroRNAs (miRNA) are small non-coding RNA molecules of ~22 nucleotides which regulate large numbers of genes by binding to seed sequences at the 3′-untranslated region of target gene transcripts. The target mRNA is then usually degraded or translation is inhibited, although thus resulting in posttranscriptional down regulation of gene expression at the mRNA and/or protein level. Due to the bioinformatic difficulties in predicting functional miRNA binding sites, several publically available databases have been developed that predict miRNA binding sites based on different algorithms. The parallel use of different databases is currently indispensable, but highly uncomfortable and time consuming, especially when working with numerous genes of interest. We have therefore developed a new stand-alone program, termed MIRNA-DISTILLER, which allows to compile miRNA data for given target genes from public databases. Currently implemented are TargetScan, microCosm, and miRDB, which may be queried independently, pairwise, or together to calculate the respective intersections. Data are stored locally for application of further analysis tools including freely definable biological parameter filters, customized output-lists for both miRNAs and target genes, and various graphical facilities. The software, a data example file and a tutorial are freely available at http://www.ikp-stuttgart.de/content/language1/html/10415.asp

Keywords: database, gene expression, gene regulation, microRNA, non-coding RNA, stand-alone software

INTRODUCTION

MicroRNAs (miRNAs) are single-stranded, non-coding RNA molecules consisting of about 22 nucleotides which generally bind to the 3′-untranslated regions (UTRs) of mRNAs. This usually leads to a down regulation of target gene expression (Bartel, 2009). Following binding to seed sequences of 7–8 nucleotides at the 3′-UTR of target gene transcripts, the target mRNA is either degraded or translation is inhibited, typically resulting in lower expression of protein, although activation of translation has also been reported (Vasudevan et al., 2007; Guo et al., 2010). The large number of miRNAs, estimated at over 1000 different molecules per mammalian species (Berezikov et al., 2005), and their poorly defined binding specificities allow for a vast number of potential miRNA-target gene interactions, which has been estimated to affect expression of one third or even more of all genes in the genome (Lim et al., 2005).

However, systematic investigation is still hampered by the lack of efficient experimental and reliable bioinformatics methods for target identification. Several dedicated public databases are now available, each of them using a specific bioinformatics or computational strategy to predict miRNA-target sites. The different algorithms result in predictions with little overlap regarding number and identity of miRNAs for a given target, or of targets for a given miRNA (Sethupathy et al., 2006). Because of these basic limitations of *in silico* prediction, the parallel use of several databases is necessary to enhance the sensitivity and/or specificity of prediction, and is therefore indispensable in this field. Currently
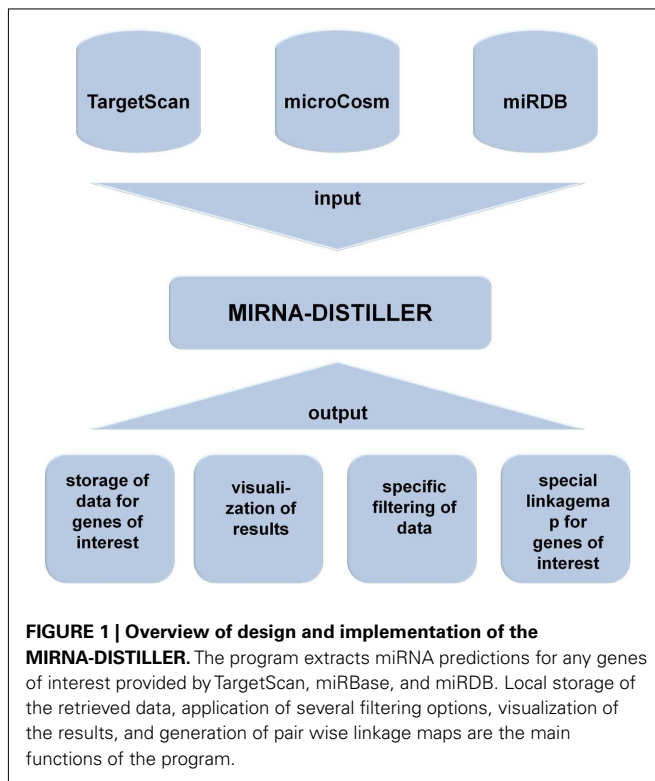
there are at least two databases, miRecords and miRò (Laganà et al., 2009; Xiao et al., 2009) which compile miRNA predictions from different databases. Although these tools are useful for some applications, we noticed a lack of practical solutions for the management of large prediction data sets adapted to the needs of the wet-lab researcher.

We have therefore developed MIRNA-DISTILLER, a stand-alone program which allows to automatically extract selected data for miRNAs predicted to interact with a given set of target genes from several publically available databases (in the current version including TargetScan, microCosm, and miRDB). The program calculates intersections and stores data locally for application of further analysis tools (**Figure 1**). For example, adjustment of binding scores provided by each database and freely definable filter definitions for additional miRNA properties (e.g., tissue-specific expression profiles) allows effective selection of candidate miRNAs for a given gene of interest. The MIRNA-DISTILLER thus facilitates miRNA bioinformatics and helps to focus experimental validation on the most promising candidates.

EXPERIMENTAL METHODS

ISOLATION AND QUANTIFICATION OF miRNAs

For profiling expression of miRNAs four pools of liver total RNA, each consisting of 10 different liver donor tissues were prepared. The origin of the human liver samples was described before (Klein et al., 2010). In brief, liver tissues were previously collected from patients undergoing liver surgery at the Department of General,

FIGURE 1 | Overview of design and implementation of the MIRNA-DISTILLER. The program extracts miRNA predictions for any genes of interest provided by TargetScan, miRBase, and miRDB. Local storage of the retrieved data, application of several filtering options, visualization of the results, and generation of pair wise linkage maps are the main functions of the program.

Visceral, and Transplantation Surgery (A. K. Nuessler, P. Neuhaus, Campus Virchow, University Medical Center Charité, Humboldt University Berlin, Germany). The study was approved by the ethics committees of the medical faculties of the Charité, Humboldt University, and of the University of Tuebingen and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from each patient. All tissue samples had been examined by a pathologist and only histologically normal liver tissue was collected and stored at $-80°$C. Total RNA was extracted using the mirVANA kit from Ambion (Austin, TX, USA). Reverse transcription of total RNA ($1\,\mu$g) was performed by using stem-loop RT primer pool A v 2.0 from Applied Biosystems (Foster City, CA, USA) following the protocol provided by the manufacturer. qPCR for miRNA expression was performed with micro fluidic cards TaqMan® Low Density Array A v 2.0 for humans (Applied Biosystems) and detected with the ABI 7900HT Fast real time RT-PCR system. Relative quantification (RQ) was calculated by normalization to the endogenous control MammU6 $RQ = 2\exp(-\Delta Ct)$.

PROGRAM DESCRIPTION

GENERAL IMPLEMENTATION

MIRNA-DISTILLER was programmed in Visual Basic.Net (Microsoft). The software, a data example file and a tutorial are freely available at http://www.ikp-stuttgart.de/content/language1/html/10415.asp

SYSTEM REQUIREMENTS

The MIRNA-DISTILLER is an executable stand-alone software which runs without installation in the Microsoft Windows operating systems windows vista and windows 7 (XP and previous versions need to have windows framework.net 2.0 or higher installed).

Database selection: the current program version (MIRNA-DISTILLER Version 1.0.0.5) allows to connect to TargetScan[1] (Lewis et al., 2005; Grimson et al., 2007; Friedman et al., 2009), microCosm[2] (Griffiths-Jones et al., 2006, 2008), and miRDB[3] (Wang, 2008). These databases have been selected because they are among the most comprehensive ones and because they share a similar searching format which allowed development of a search mask that accepts the official gene name, thus avoiding the necessity to copy/paste sequences or gene IDs.

QUERY INTERFACE

On the main interface genes of interest are entered by their official gene name with the "New Gene" button into the "Genes" tab (**Figure 2**). Entering only one or several hundred genes of interest is possible. The program then automatically downloads miRNA prediction data and corresponding binding scores for each selected database for local storage. Selecting one of the previously loaded genes in the left window provides a summary table for the predicted miRNAs, where miRNAs predicted from all of the three databases are highlighted in blue. The total number of miRNAs predicted by each database are visualized in a bar diagram.

The gene-linked "miRNA Ranking" tab by default provides a list with all miRNAs ranked according to the number of their target genes (**Figure 3**). Additional sorting options, including alphabetical, are also supported. Once the miRNA data have been retrieved for a given gene set, the gene targets predicted for each miRNA can also be displayed with corresponding binding scores and listed according to the database. This feature thus allows the user to identify possible target genes for a given miRNA.

FILTERING OPTIONS

Two different filters are available to be used separately or in combination. The "Actual Minimum Score" filter limits all data according to the selected minimum score. "Filter Definitions" allows the definition of filters in a user-specific way based on additional miRNA properties, e.g., such obtained from the researchers own investigations. A typical application of this filter is to delete miRNAs not expressed in the tissue of interest, based on experimental data.

User-specific filtering of the data with the "Filter-Definitions" option provides a broad spectrum of adoptions by using classical operators (e.g., bigger, smaller, is part of). The combination of user-defined miRNA properties and filter-definitions supports creative data management.

EXAMPLE OF GENE-miRNA INTERACTIONS

We analyzed 62 selected ADME genes (i.e., genes involved in the absorption, distribution, metabolism, or excretion of drugs and foreign chemicals) for miRNA binding sites. In order to filter the prediction data only for liver expressed miRNAs, we measured miRNA expression in pooled human liver samples. The ADME

---

[1]http://www.targetscan.org/vert_50
[2]http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5
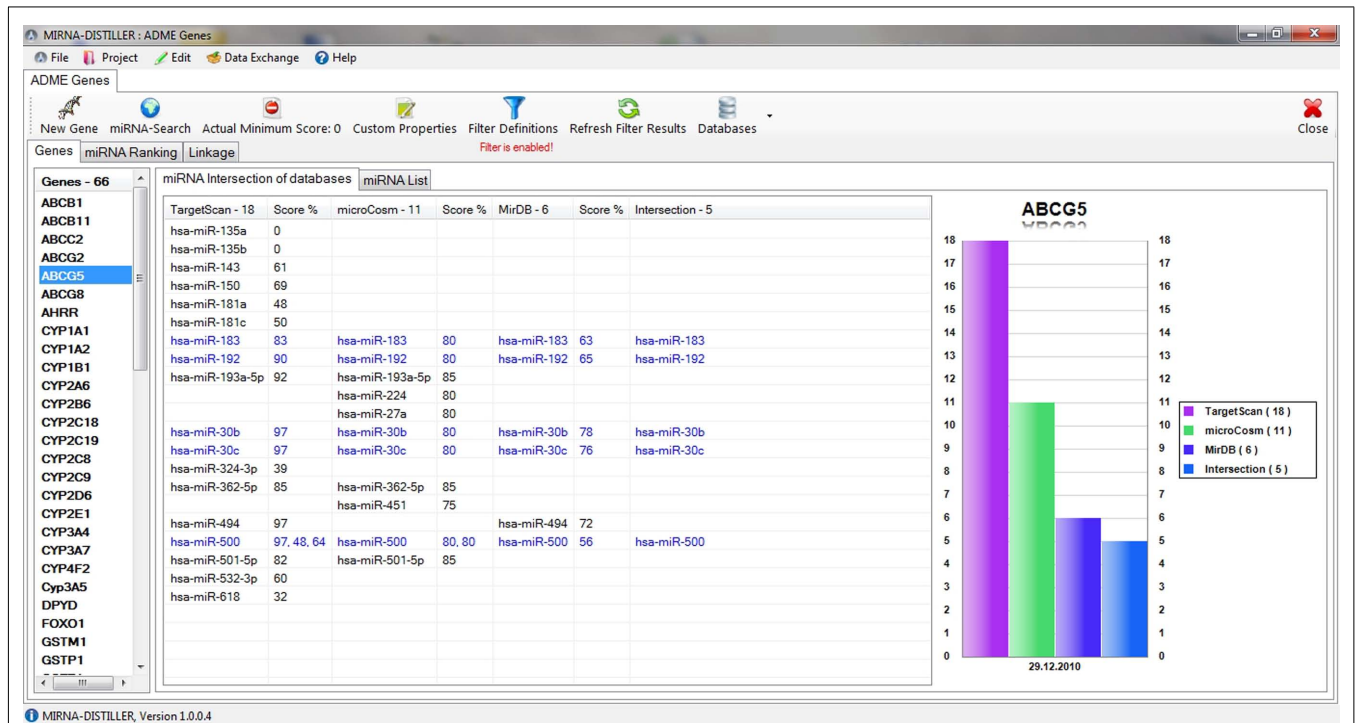[3]http://mirdb.org/miRDB

FIGURE 2 | Main interface. The left window lists the genes of interest that have been entered before, the middle window contains the predicted miRNAs with corresponding scores for each selected database and the intersection list, and the right window displays the selected intersection graphically.
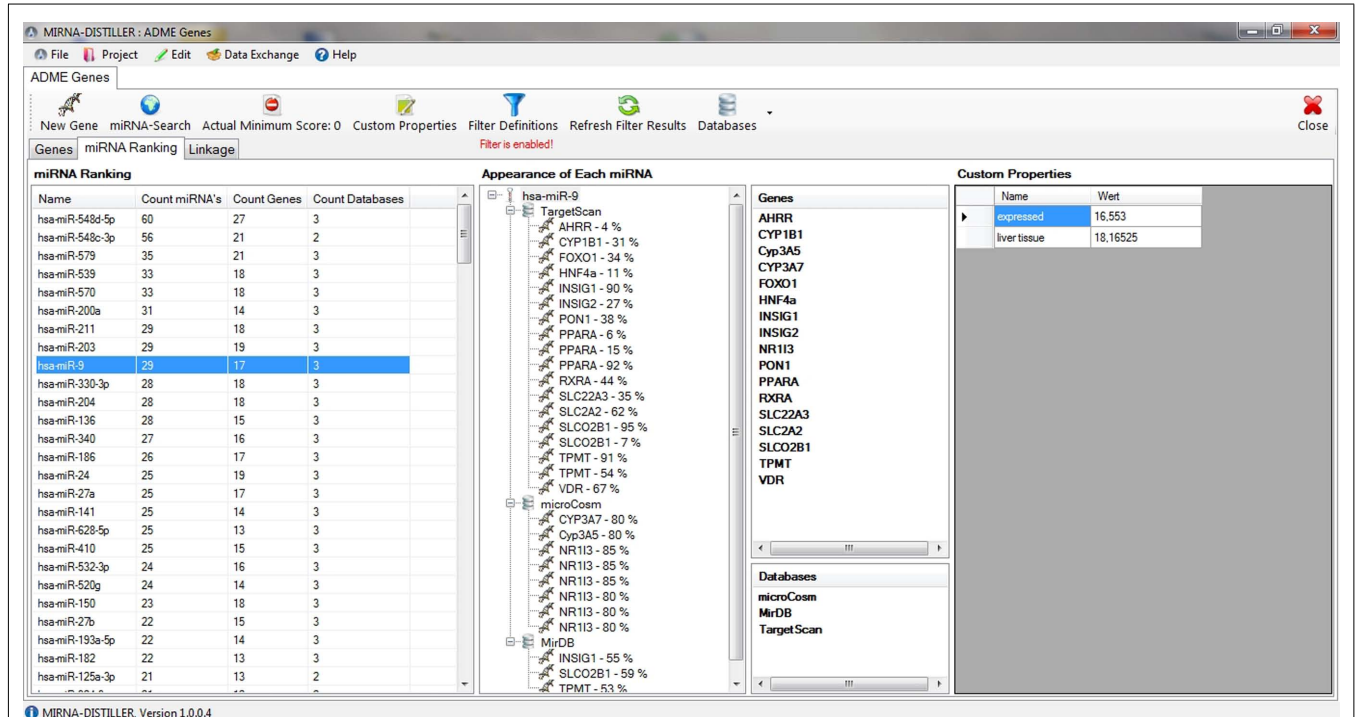

FIGURE 3 | miRNA Ranking. All miRNAs are ranked by their occurrence in the preloaded gene set. It is also possible to sort the miRNAs alphabetically. Furthermore the program allows to predict target genes for all previously selected miRNAs.

Table 1 | Illustration of predicted miRNA binding sites with and without specific filter options.

ADME Gen	Total miRNAs	Liver expressed miRNAs	Percentage in liver expressed	ADME Gen	Total miRNAs	Liver expressed miRNAs	Percentage in liver expressed
ABCB1	44	11	25	NAT2	78	6	8
ABCB11	54	10	19	NME1	74	24	32
ABCG2	71	20	28	NME2	3	2	67
ABCC2	91	32	35	NR1H3	30	4	13
ABCG5	44	11	25	NR1I2	59	24	41
ABCG8	36	12	33	NR1I3	82	21	26
AHRR	171	38	22	NR3C1	87	28	32
CYP1A1	40	8	20	PON3	25	3	12
CYP1A2	39	4	10	POR	70	11	16
CYP1B1	88	25	28	PPARa	117	41	35
CYP2A6	8	3	38	RXRa	27	10	37
CYP2B6	43	10	23	SLC15A2	62	13	21
CYP2C18	40	8	20	SLC22A1	32	4	13
CYP2C19	59	13	22	SLC22A2	48	10	21
CYP2C8	53	8	15	SLC22A3	71	14	20
CYP2C9	63	18	29	SLC22A6	47	10	21
CYP2D6	41	9	22	SLC2A2	98	26	27
CYP2E1	32	6	19	SLC47A1	38	9	24
CYP3A4	77	15	19	SLCO1B1	36	11	31
CYP3A5	20	6	30	SLCO1B3	46	10	22
CYP3A7	77	18	23	SLCO2B1	43	12	28
CYP4F2	45	12	27	SREBF2	35	13	37
DPYD	92	25	27	SULT1A1	38	4	11
GSTM1	62	12	19	TPMT	68	16	24
GSTP1	18	6	33	UGT1A1	76	12	16
GSTT1	53	9	17	UGT1A3	38	8	21
HMGCR	76	25	33	UGT1A5	38	8	21
HNF4a	103	19	18	UGT2B15	60	12	20
INSIG1	80	30	38	UGT2B17	43	7	16
INSIG2	82	17	21	UGT2B7	34	5	15
NAT1	49	10	20	VDR	62	10	16

For 62 selected ADME genes the total number of miRNA binding sites predicted from all three databases was compiled with the MIRNA-DISTILLER. After filtering for liver expressed miRNAs approximately 25% of miRNAs remained for each gene.

genes are listed in **Table 1** with their total number of predicted miRNA binding sites compiled with the MIRNA-DISTILLER from all three databases. After defining and applying a filter for liver specific miRNAs based on the experimental expression data, between 10 and 67% (average 25%) of the predictions remained, thus effectively reducing the number of relevant predictions for each gene.

EXAMPLE OF PUBLISHED GENE-miRNA INTERACTIONS
To further demonstrate the usefulness of parallel database searches as a proof of concept, 42 published targets of miR-122, the major liver specific miRNA (Gramantieri et al., 2007; Filipowicz and Grosshans, 2011), were loaded in the MIRNA-DISTILLER from miRTarBase[4]. This database has stored more than 3000 miRNA-target interactions collected and curated manually from literature

[4]http://mirtarbase.mbc.nctu.edu.tw/index.html

to represent functionally relevant miRNAs. Thus, the miRNA-target interactions in miRTarBase have been experimentally validated by different experimental approaches (Hsu et al., 2011). For these validated miR-122 target genes we would expect that the ranking of miRNAs provided by miRNA-DISTILLER should result in miR-122 as the most common miRNA among this set of genes. As shown in **Figure 4**, miR-122 was indeed found to be on the top of the ranking list and was predicted to bind to 34 of the 42 genes. Some but not all of the binding sites were predicted by more than one database.

LINKAGE TAB
The "Linkage" tab (**Figure 5**) is an additional feature that automatically displays a graphical matrix (heatmap) that compares any two genes with respect to their predicted miRNAs, facilitating identification of potentially co-regulated genes. The calculation of this graph takes filter settings automatically into account.

ADVANTAGE OVER ALREADY EXISTING TOOLS

MIRNA-DISTILLER is focused on high-throughput screening of miRNA binding sites for a defined gene set in order to find new possible miRNA-gene interactions. Several web-based tools for compiling miRNA data from different databases already exist, e.g., the above mentioned miRecords (Xiao et al., 2009) and miRò (Laganà et al., 2009). Although these tools also allow to retrieve and to compile data from various miRNA databases, they lack some features we consider highly useful in our program. Thus, the features to store data locally and to combine them with own

experimental data to allow specific retrievals and continuous work with the data are missing in the other tools. Also immediate graphical display of results has not been implemented. In this sense our program is unique as none of the already existing tools has implemented similar options. MIRNA-DISTILLER is therefore a useful tool for miRNA researcher dealing with huge data sets to manage and interactively work with them.

CONCLUSION

MIRNA-DISTILLER facilitates systematic analyses of miRNA predictions for given gene set by combining several useful tools in one application. It allows simultaneous data download from up to three common publicly available databases, their comparison and local data storage for further analysis, including several flexible filter options. The user friendly interface permits automatic data processing for large numbers of genes and in combination with the provided tools helps to prioritize potential miRNA binding sites. In this way, miRNA-DISTILLER assists to collect and analyze complex miRNA data and helps to generate hypotheses for experimental validation.
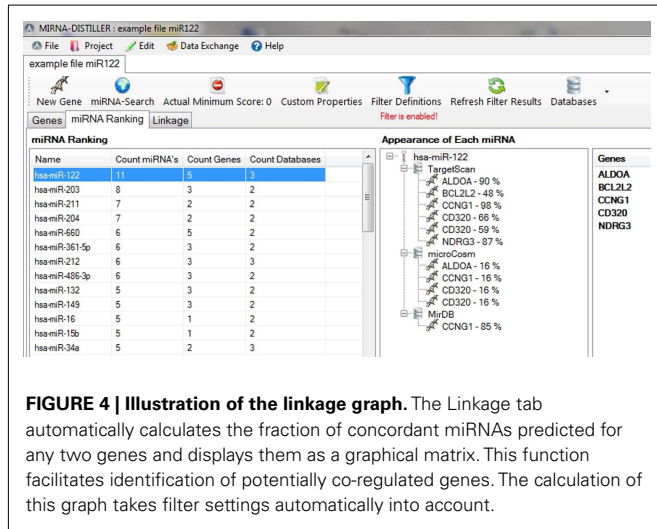

FIGURE 4 | Illustration of the linkage graph. The Linkage tab automatically calculates the fraction of concordant miRNAs predicted for any two genes and displays them as a graphical matrix. This function facilitates identification of potentially co-regulated genes. The calculation of this graph takes filter settings automatically into account.
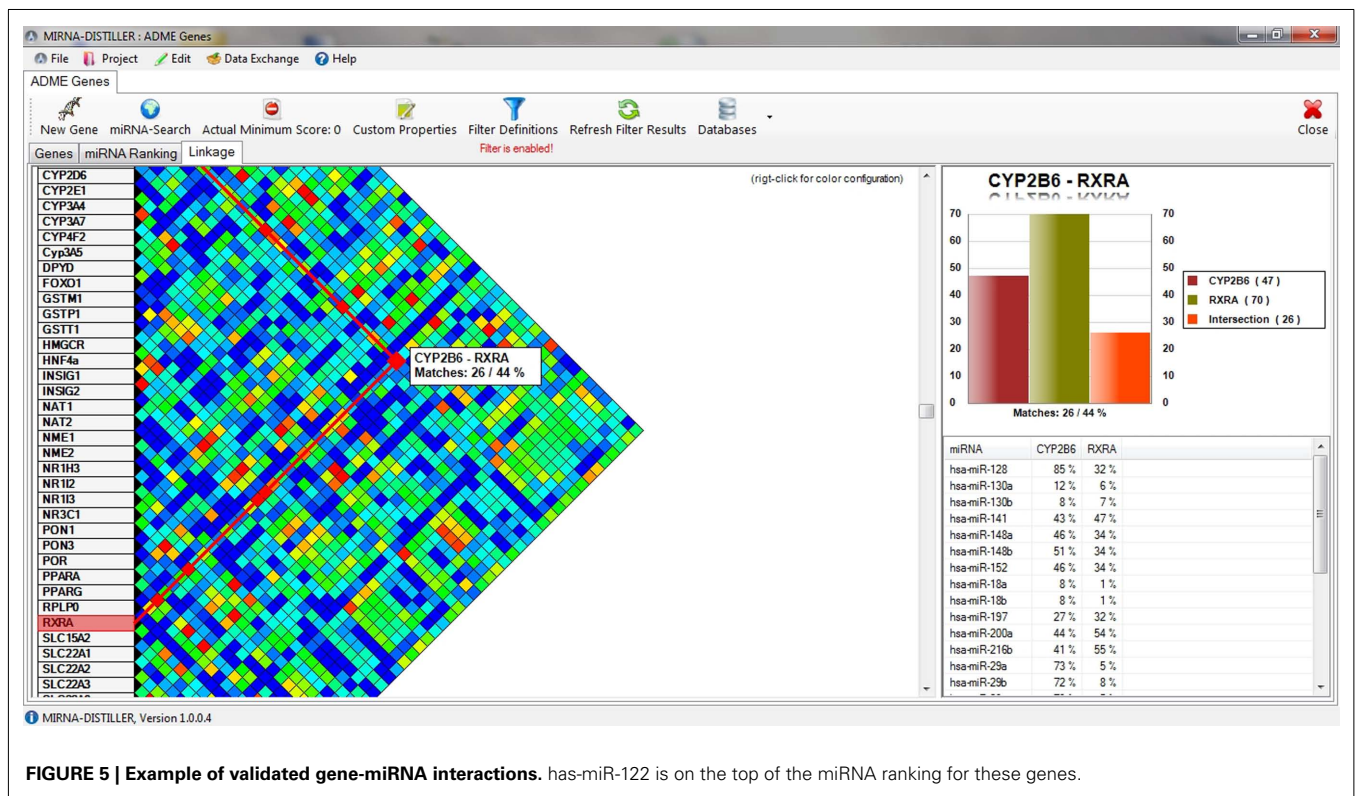
FIGURE 5 | Example of validated gene-miRNA interactions. has-miR-122 is on the top of the miRNA ranking for these genes.

REFERENCES

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136, 215–233.

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R. H. A., and Cuppen, E. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* 120, 21–24.

Filipowicz, W., and Grosshans, H. (2011). The liver-specific microRNA miR-122: biology and therapeutic potential. *Prog. Drug. Res.* 67, 221–238.

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19, 92–105.

Gramantieri, L., Ferracin, M., Fornari, F., Veronese, A., Sabbioni, S., Liu, C.-G., Calin, G. A., Giovannini, C., Ferrazzi, E., Grazi, G. L., Croce, C. M., Bolondi, L., and Negrini, M. (2007). Cyclin G1 is a target of miR-122a, a microRNA frequently down-regulated in human hepatocellular carcinoma. *Cancer Res.* 67, 6092–6099.

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.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res.* 36, D154–D158.

Grimson, A., Farh, K. K.-H., 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.

Guo, H., Ingolia, N. T., Weissman, J. S., and Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466, 835–840.

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., Chien, C.-H., Wu, M.-C., Huang, C.-Y., Tsou, A.-P., and Huang, H.-D. (2011). miR-TarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 39, D163–D169.

Klein, K., Winter, S., Turpeinen, M., Schwab, M., and Zanger, U. M. (2010). Pathway-targeted pharmacogenomics of CYP1A2 in human liver. *Front. Pharmacol.* 1:129. doi: 10.3389/fphar.2010.00129

Laganà, A., Forte, S., Giudice, A., Arena, M. R., Puglisi, P. L., Giugno, R., Pulvirenti, A., Shasha, D., and Ferro, A. (2009). miRò: a miRNA knowledge base. *Database (Oxford)* 2009, doi: 10.1093/database/bap008

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15–20.

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S., and Johnson, J. M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 433, 769–773.

Sethupathy, P., Megraw, M., and Hatzigeorgiou, A. G. (2006). A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat. Methods* 3, 881–886.

Vasudevan, S., Tong, Y., and Steitz, J. A. (2007). Switching from repression to activation: microRNAs can up-regulate translation. *Science* 318, 1931–1934.

Wang, X. (2008). miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* 14, 1012–1017.

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.