➢ Description:

High-throughput RNAi screening has been widely used in a spectrum of biomedical research and made it possible to study functional genomics. However, a challenge for authentic biological interpretation of large-scale siRNA or shRNA-mediated loss-of-function studies is the biological pleiotropy resulting from multiple modes of action of siRNA and shRNA reagents. A major confounding feature of these reagents is the microRNA-like translational quelling that can result from short regions (~6 nucleotides) of oligonucleotide complementarity to many different mRNAs. To help identify and correct miRNA-mimic off-target effects, we have developed DecoRNAi (deconvolution analysis of RNAi screening data) for automated quantitation and annotation of microRNA-like off-target effects in primary RNAi screening data sets. DecoRNAi can effectively identify and correct off-target effects from primary screening data and provide data visualization for study and publication. DecoRNAi contains pre-computed seed sequence families for 3 commonly employed commercial siRNA libraries. For custom collections, the tool will compute seed sequence membership from a user-supplied reagent sequence table. All parameters are tunable and output files include global data visualization, the identified seed family associations, the siRNA pools containing off-target seed families, corrected z-scores and the potential miRNAs with phenotypes of interest.

➢ Input file format:

51742	-25.00	CUGAAGACCUCCAAGUUUA	GGACAUGAUCCGAGAGGUG	UGAAGAAAGCCACUGCCAA	GAGCUGAGCAAUGUCCUGG
10660	-23.50	CCAAAGAGCUCAUAAAAGA	CAAGAUGUAUGGUGAGUAU	CCAAUGAGCUCCUGCAAAC	UCUCAAGGACUAUCUGUUA
151648	-23.32	GAACUUGGAUCUCUUGUCU	GAUCACAGAUAUCAUUGAA	CUAAGGAGUUCAUAGAAUU	GAAAUUAAGCCUCCGAAUC
11036	-21.72	GUACUUGGAUAUCGAAAUU	GCUCAAGACAUGGAUAAUA	UAACAGAUGCGAGUUGGUA	CCAUAGAAAUUACUUCGAU
7791	-20.63	GAAGUGAUAUAUCGCCUUG	GAAACUCGCUGGCGGGUGU	CACAGGAGUUUGUAGGAUU	GUUCUCACCUUAUAGAGCA
5272	-18.36	GAAAGUCGCGCACGGCCUU	UCGAAGACCUCGCCGCUGU	UGGCCGAACUCGAGCAGAA	CGUCUGUGCGGAGAAGUUA
7314	-18.24	CAUCUUAGCCUGAAGGAUA	UGAAAGAAGCCCAAGAUAU	CAGCCAGCGUGAACUAUAA	AAACGCAGGUCUUUUAUAG
341405	-17.61	UCAACAGCAUCAUUAGUUA	GAUUUAAGCAUUCGGGUUA	UGGCCAGCCUUCAGUAAUA	GAAUUGAAACCGUGCUACA
9183	-17.38	GACAAGAACUUCCACAUGA	GAAUGUGGCUGUCAACGAA	GACCAAGAAUGAUCCUUUC	GGUGAGCAGUAUUGAUUUG
687	-16.20	GGACCAAGCCAGACUGUAU	GGAAUGAACCGUUUGACGA	CCGGGUAGCUCAAUUGAUG	CAAGAUAACCCUUCGCACA

Note 1: Input file has to be a Comma-separated values (CSV) file whose extension name is ".csv".

Note 2: Input file has **NO HEADER** in the first row.

Note 3: The first two column are Gene Entrez ID and phenotypic readouts (it is robust Z score in all of our example datasets), respectively. In following columns, users have to provide siRNA sense strand sequences if they are using custom siRNA library and make sure one sequence per cell, in which case the first column doesn't have to be Gene Entrez ID and can be any format. If users choose our built-in siRNA library for analysis, sequences are not needed.

Note 4: For user-provided siRNA sense strand sequences, please make sure the sequences don't contain 3' overhanging nucleotides because we are using sense strand to generate antisense strand sequence. For example, usually, siRNAs contain 21 nt and the last 2 are overhanging nt, in which case please only include the first 19 nt in the input file.

➢ Parameters:

Input File: Name of input file for analysis. This file should contain Gene ID (for example, Gene Entrez ID name), normalized screening data and sense strand siRNA sequences. Default format is a csv (comma separated value) file, in which the first column contains Gene ID name, the second column contains normalized screening data and the following columns are the sense strand siRNA sequences (one sequence per column, i.e., for example, there would be 4 separate sequence columns if 4 oligos are present in a pool). See the user's manual for details.

Strand: For identification and quantification of off-target effects, DecoRNAi can employ sense strand only, antisense strand only or both strands. The default setting is using both strands.

Lambda: Penalty parameter in the model for identification of off-target effects. Default is 0.001.

Seed: Specify which hexamer is used to define the seed sequence for analysis. For the most part, a siRNA oligo contains 21 nt. We can therefore assign any of 14 different hexamers as the seed sequence. For example, 1 means nucleotides 5′ 1~6 hexamer and 2 means nucleotides 5′ 2~7 hexamer and etc. Default is 2.

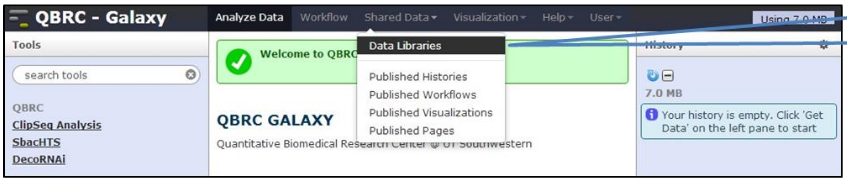
Library: Users can specify the siRNA library for analysis. Seed families are pre-computed for the Dharmacon siGenome (version history 0), Dharmacon siGenome (version history 2), and Ambion. Gene Entrez ID is necessary to map between input data and stored sequences. Users can also upload custom library-wide sequence information for each oligonucleotide or processed siRNA, in which case Gene ID is free of format and type. Default is Custom*.*

Strength of seed-linked effect: Users can specify the cutoff for strength of seed-linked effect. Must be positive value and default is 1. A smaller value will select more off-target seed families.

Significance (P value): Users can specify the cutoff for significance (P value). Default is 0.01. In summary report, False Discovery Rate (FDR) will be provided to control multiple testing issues.

➢ Graphical demonstration of work flow

1) Download publicly available demo datasets



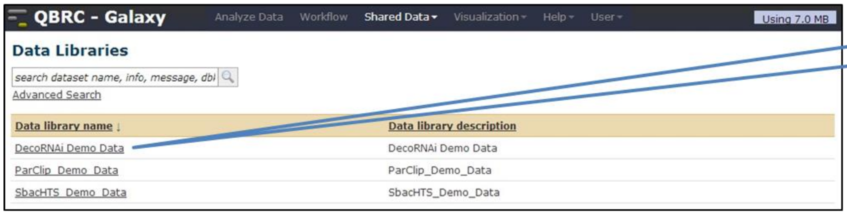
Welcome to QBRC

QBRC GALAXY

Quantitative Biomedical Research Center @ UT Southwestern

Your history is empty. Click 'Get Data' on the left pane to start

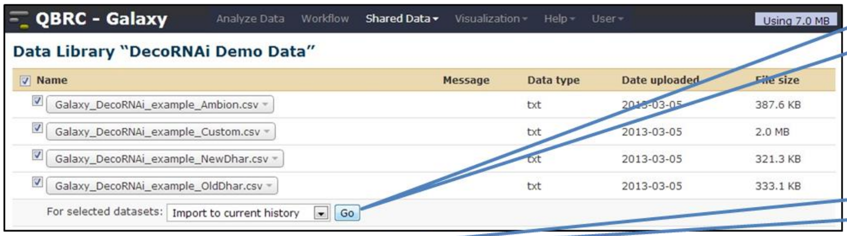
Step 1:
Click "Shared Data" on the top of the middle and select "Data Libraries" tab.



Data Libraries

Data library name ↓	Data library description
DecoRNAi Demo Data	DecoRNAi Demo Data
ParClip_Demo_Data	ParClip_Demo_Data
SbacHTS_Demo_Data	SbacHTS_Demo_Data

Step 2:
Select "DecoRNAi_Demo_Data" and click.

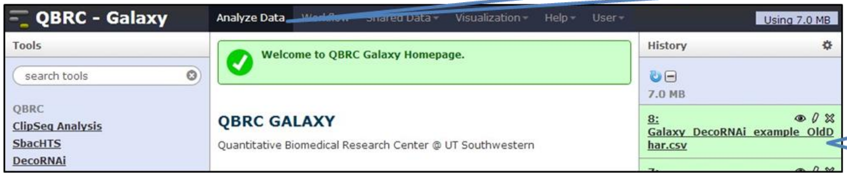


Data Library "DecoRNAi Demo Data"

☑ Name	Message	Data type	Date uploaded	File size
☑ Galaxy_DecoRNAi_example_Ambion.csv		txt	2013-03-05	387.6 KB
☑ Galaxy_DecoRNAi_example_Custom.csv		txt	2013-03-05	2.0 MB
☑ Galaxy_DecoRNAi_example_NewDhar.csv		txt	2013-03-05	321.3 KB
☑ Galaxy_DecoRNAi_example_OldDhar.csv		txt	2013-03-05	333.1 KB

For selected datasets: Import to current history Go

Step 3:
Check the boxs associated with "Galaxy_DecoRNAi_example" CSV files and click "Go".



Welcome to QBRC Galaxy Homepage.

QBRC GALAXY

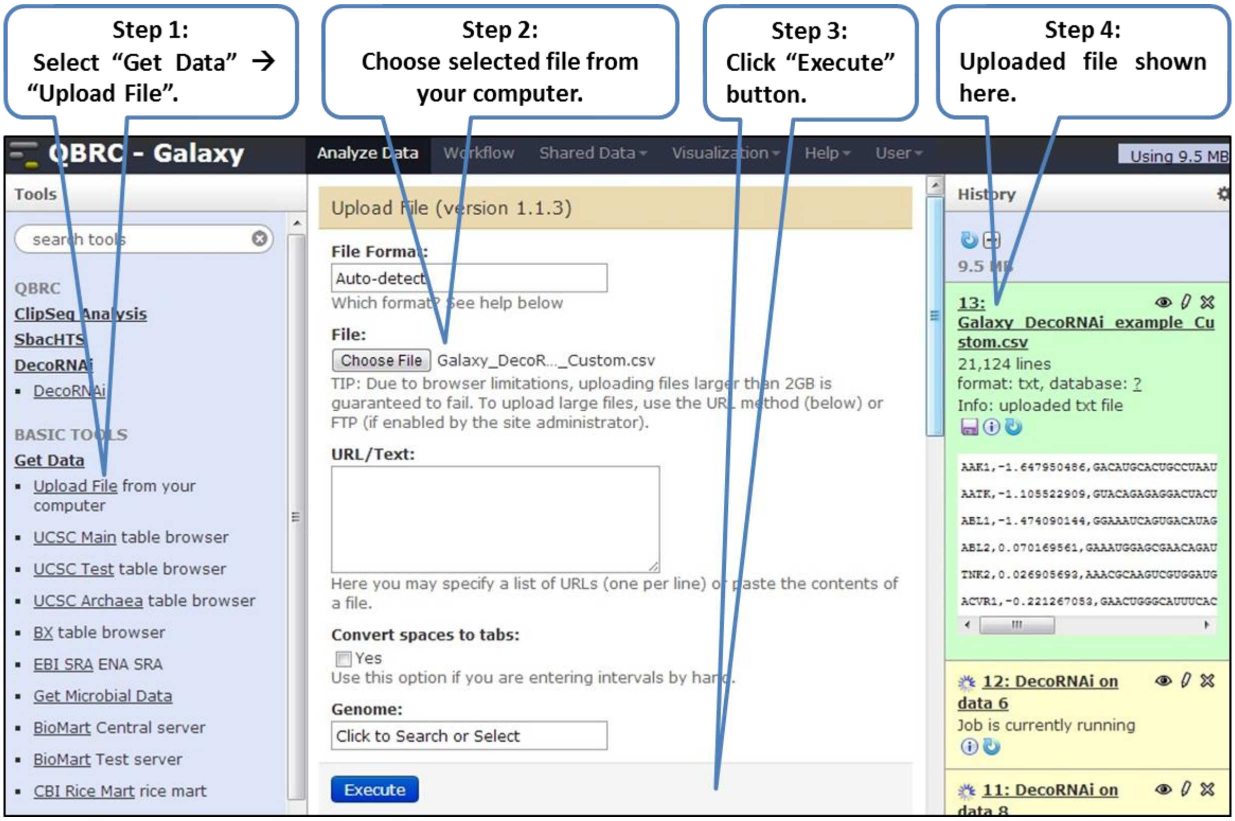
Quantitative Biomedical Research Center @ UT Southwestern

8: Galaxy_DecoRNAi_example_OldDhar.csv

Step 4:
Click "Analyze Data" button on the left of the middle top.

Step 5:
Downloaded data is available for analysis.

2) Upload custom data



Step 1: Select "Get Data" → "Upload File".

Step 2: Choose selected file from your computer.

Step 3: Click "Execute" button.

Step 4: Uploaded file shown here.

Upload File (version 1.1.3)

File Format: Auto-detect

Which format? See help below

File: Galaxy_DecoR..._Custom.csv

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Convert spaces to tabs:

□ Yes

Use this option if you are entering intervals by hand.

Genome: Click to Search or Select

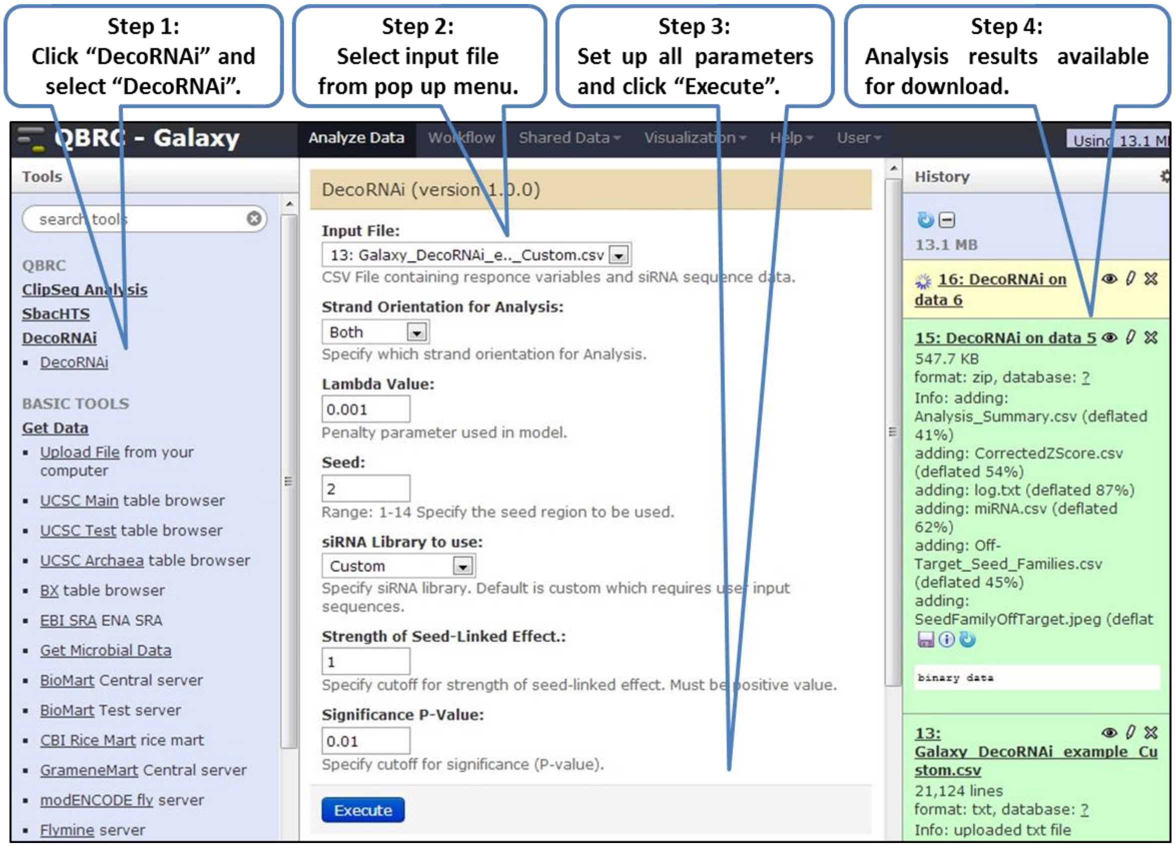
13: Galaxy_DecoRNAi_example_Custom.csv
21,124 lines
format: txt, database: ?
Info: uploaded txt file

```
AAK1,-1.647950486,GACAUGCACUGCCUAAU
AATK,-1.105522909,GUACAGAGAGGACUACU
ABL1,-1.474090144,GGAAAUCAGUGACAUAG
ABL2,0.070169561,GAAAUGGAGCGAACAGAU
TNK2,0.026905693,AAACGCAAGUCGUGGAUG
ACVR1,-0.221267053,GAACUGGGCAUUUCAC
```

12: DecoRNAi on data 6
Job is currently running

11: DecoRNAi on data 8

3) Parameters setup



Step 1:
Click “DecoRNAi” and select “DecoRNAi”.

Step 2:
Select input file from pop up menu.

Step 3:
Set up all parameters and click “Execute”.

Step 4:
Analysis results available for download.

DecoRNAi (version 1.0.0)

Input File: 13: Galaxy_DecoRNAi_e..._Custom.csv
CSV File containing responce variables and siRNA sequence data.

Strand Orientation for Analysis: Both
Specify which strand orientation for Analysis.

Lambda Value: 0.001
Penalty parameter used in model.

Seed: 2
Range: 1-14 Specify the seed region to be used.

siRNA Library to use: Custom
Specify siRNA library. Default is custom which requires user input sequences.

Strength of Seed-Linked Effect.: 1
Specify cutoff for strength of seed-linked effect. Must be positive value.

Significance P-Value: 0.01
Specify cutoff for significance (P-value).

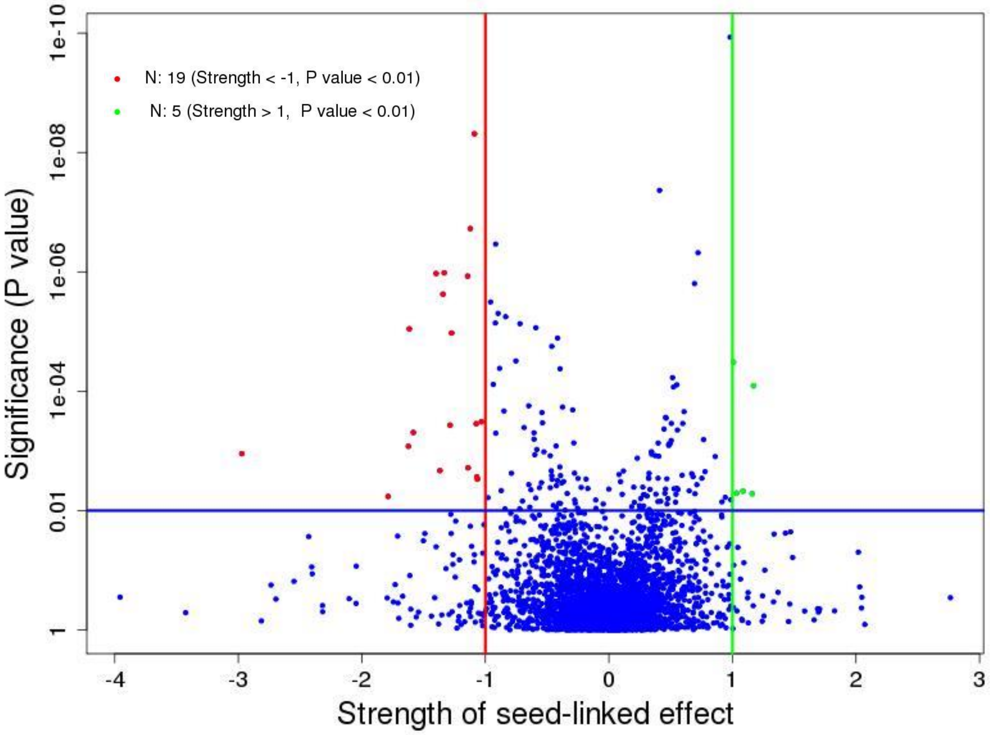
16: DecoRNAi on data 6

15: DecoRNAi on data 5
547.7 KB
format: zip, database: ?
Info: adding: Analysis_Summary.csv (deflated 41%)
adding: CorrectedZScore.csv (deflated 54%)
adding: log.txt (deflated 87%)
adding: miRNA.csv (deflated 62%)
adding: Off-Target_Seed_Families.csv (deflated 45%)
adding: SeedFamilyOffTarget.jpeg (deflat

binary data

13: Galaxy_DecoRNAi_example_Custom.csv
21,124 lines
format: txt, database: ?
Info: uploaded txt file

4) Global visualization of seed-linked off-target effects (“SeedFamilyOffTarget.jpeg” and “SeedFamilyOffTargetLegend.jpeg”): this figure provides visualization of both strength of seed-lined effect and significance (P values). On the plot, each dot represents a seed family and x axis indicates strength of seed-lined effect and y axis statistical significance (P value). Red dots represent negative off-target seed families and green dots represent positive off-target seed families. Criterion is defined by users and false discovery rate is reported in analysis summary (see below demonstration).



5) Seed family summary table ("Seed_Famliy_Summary.csv"): In a genome-wide siRNA screening, about 4000 seed families are present in an analysis and we summarize for each of them as below. "Seed family" represents 6 nt seed sequence, "strength of seed-linked effect" and "Significance (P value)" are used for selection of off-target seed families and for above visualization.

Seed family	Strength of seed-linked effect	Family size	Significance (P value)
UACUGU	0.98	63	1.17E-10
UAUUGG	-1.09	68	4.87E-09
AUUCCU	0.41	113	4.32E-08
UACCAG	-1.12	48	1.89E-07
AGUUGA	-0.92	69	3.44E-07
AAUACU	0.72	132	4.79E-07
UGUUGU	-1.33	46	1.04E-06
ACAUGU	-1.40	32	1.07E-06
ACCAGG	-1.14	68	1.18E-06
… …	… …	…	… …

6) Identified off-target seed families (“Off-Target_Seed_Families.csv”):Identified off-target seed families. Criterion defined by users.

Seed family	Strength of seed-linked effect	Family size	Significance (P value)
ACAUGU	-1.40	32	1.07E-06
ACCAGG	-1.14	68	1.18E-06
ACCAGU	-1.06	43	2.98E-03
ACUAGG	-1.62	23	8.37E-04
ACUAGU	-1.62	23	9.03E-06
AGCUCC	1.17	19	8.13E-05
AGUACC	-1.29	18	3.71E-04
GAUGCU	-1.27	37	1.06E-05
GCCUAA	1.16	6	5.22E-03
GUUACU	-1.03	19	3.25E-04
GUUUGG	1.09	34	4.72E-03
UAAGUC	-1.58	10	4.91E-04
UACCAG	-1.12	48	1.89E-07
UAUGAU	-1.07	39	2.73E-03
UAUUGG	-1.09	68	4.87E-09
UCCUUU	-1.07	28	3.51E-04
UCUCAG	1.01	27	3.25E-05
UCUCCC	-1.37	10	2.14E-03
UCUGAC	-2.97	6	1.11E-03
UCUGCA	-1.14	44	1.92E-03
UGUCCC	-1.79	9	5.80E-03
UGUUGU	-1.33	46	1.04E-06
UUCUCC	-1.34	34	2.37E-06
UUGGGU	1.03	26	5.09E-03

7) siRNAs pools with off-target seed family (“siRNAsPool_OffTargetSeedFamily_ACUAGG.csv”):

ID	Z Score	off-target seed
5993	-5.08	ACUAGG
493	-3.55	ACUAGG
55014	-3.26	ACUAGG
6530	-3.21	ACUAGG
51320	-3.21	ACUAGG
79917	-3.09	ACUAGG
79780	-2.91	ACUAGG
1241	-2.69	ACUAGG
4299	-2.02	ACUAGG
387837	-1.76	ACUAGG
9814	-1.30	ACUAGG
124842	-1.23	ACUAGG
5970	-1.15	ACUAGG
9786	-0.91	ACUAGG
10725	-0.72	ACUAGG
56961	-0.46	ACUAGG
390195	-0.15	ACUAGG
22875	0.00	ACUAGG
10181	0.17	ACUAGG
83849	0.20	ACUAGG
10957	0.62	ACUAGG
2263	1.01	ACUAGG
7780	1.38	ACUAGG

8) Annotated miRNAs with phenotype of interest (“miRNA.csv”): based identified off-target seed families, we could annotate known miRNAs with potential phenotype of interest.

MiRBase ID	MiRBase Accession	Mature sequence	Seed.sequence
hsa-miR-3923	MIMAT0018198	AACUAGUAAUGUUGGAUUAGGG	ACUAGU
hsa-miR-4713-5p	MIMAT0019820	UUCUCCCACUACCAGGCUCCCA	UCUCCC
hsa-miR-4256	MIMAT0016877	AUCUGACCUGAUGAAGGU	UCUGAC

9) Corrected Z score (“CorrectedZScore.csv): based on our algorithm, after removing off-target effect, we provide a list of corrected Z score which has a lower false positive rate.

ID	Z Score	Corrected Z Score
7314	-23.60	-13.78
27243	-22.32	-16.79
8837	-19.94	-17.51
10482	-17.53	-12.10
7316	-14.07	-11.24
26137	-11.77	-9.26
285877	-10.35	-7.77
128866	-9.66	-8.07
5430	-9.53	-7.94
29080	-9.26	-7.91
… …	… …	… …

10) Analysis summary (“Analysis_Summary.csv”): summary of whole analysis.

Analysis is successfully done!	
Parameters set up	
Strand Orientation for Analysis:	both
Lambda value:	0.001
Seed:	2
siRNA Library:	new_dharmacon
Strength of seed-linked effect:	1
P value:	0.01
Analysis summary	
Number of negative off-target seed families:	19
Number of positive off-target seed families:	5
FDR:	0.026
Identify miRNA with phenotypic effects:	Yes