

I have a list of shRNAs

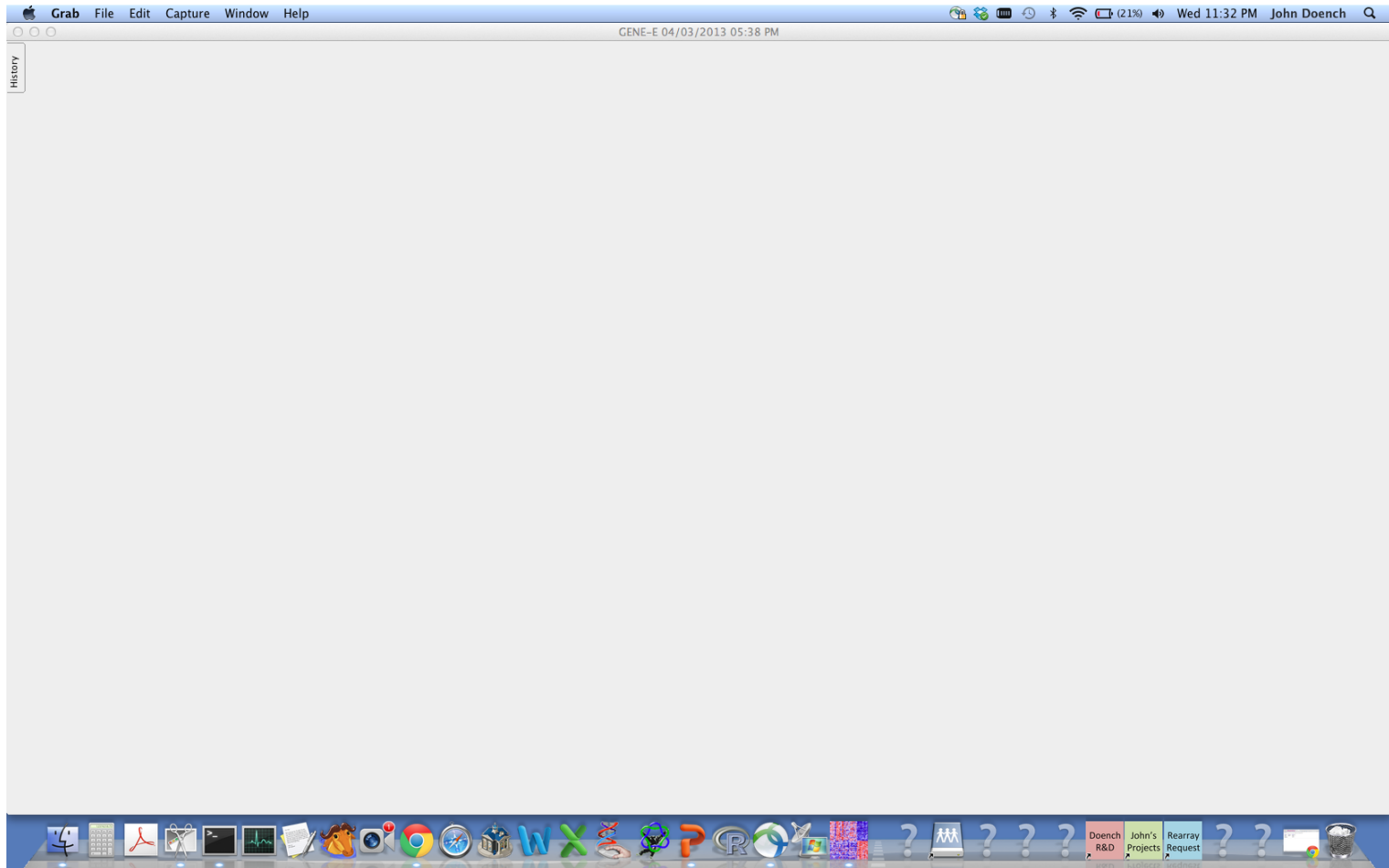


- How to appropriately map shRNAs to genes?
- To maximize on-target effect vs. off-target, require multiple shRNAs to hit the same gene

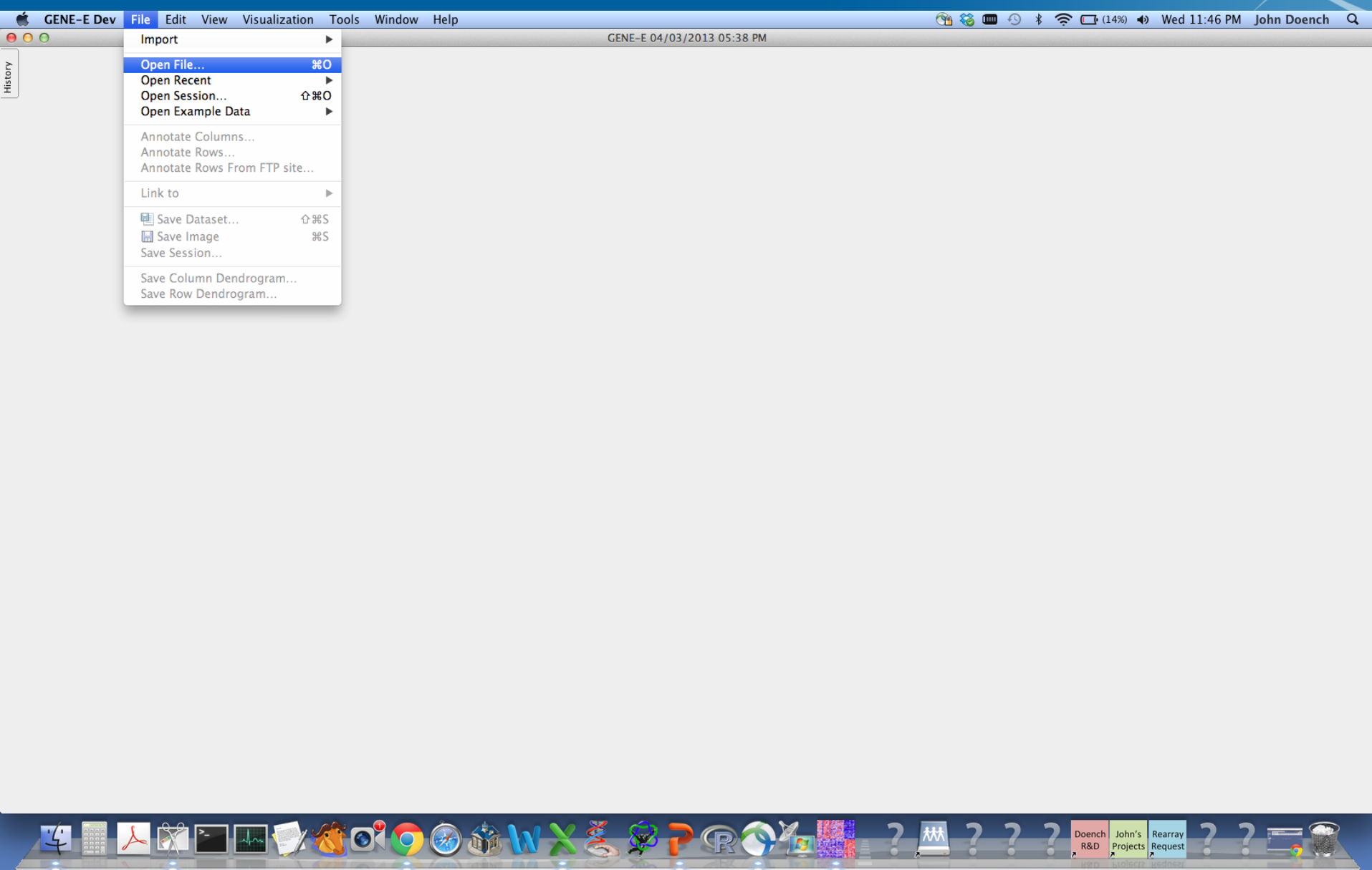
- <http://www.broadinstitute.org/~jdoench/>



- <http://www.broadinstitute.org/cancer/software/GENE-E/dev/>



Open file (no need to import)



Input File (save as .txt from Excel)



	A	B	C	D	E
1	Hairpin.Sequence	Hairpin.IDs	Gene.Symbol	Score	
2	CGGAGTGATTACGATGGCATT	TRCN0000000002	DDX3X	0.73286324	
3	CGCTTGGAACAGGAACTCTTT	TRCN0000000004	DDX3X	0.29702016	
4	AGCAGATTTAGTGGAGGGTTT	TRCN0000000005	DDX3X	-0.39430215	
5	GTTGGTTCGATAATGGCCTTT	TRCN0000000006	DHX15	0.43257301	
6	ACTGTTCTAATGAGGTCCTAT	TRCN0000000007	DHX15	-0.53199043	
7	TGGGAATACAAGGATAGGTTT	TRCN0000000008	DHX15	-0.15905649	
8	TGTAAGAGAATAAAGCGTGAA	TRCN0000000009	DHX15	0.16847474	
9	GCTTCAACAAATGCTATGCTT	TRCN0000000010	DHX15	-0.27478958	
10	GCGCCGTACATTCGAGAGTTT	TRCN0000000013	SNRNP70	0.18827007	
11	CCGGAGAGAGTTTGAGGTGTA	TRCN0000000014	SNRNP70	-0.19004955	
12	GACTATTTGGAGAGACTGATT	TRCN0000000018	PRPF18	1.85858801	
13	GAATACGTGAAGGCAAATGAT	TRCN0000000019	PRPF18	0.10936745	
14	AGCGATATTGACATGGGAGAA	TRCN0000000021	DDX3Y	-0.79597024	
15	TAGGTGCAACAGGGAGTGATT	TRCN0000000022	DDX3Y	-0.16808917	
16	AGGAGCAAGTACAGCGAGCAA	TRCN0000000023	DDX3Y	-0.34673824	
17	GACCCAACTAAGCTAAGCAA	TRCN0000000025	DHX8	0.85735113	
18	AGAGACAGATAGCGAAGGAA	TRCN0000000026	DHX8	0.00000000	

Import Data



Click the table cell containing the first data row and column.

- Row metadata
- Column metadata
- Data matrix
- Transpose

Hairpin.Se...	Hairpin.IDs	Gene.Sym...	Score
CCCACAT...	TRCN000...	F2	-3.34576...
CCAGATG...	TRCN000...	ISCA2	-3.19547...
GTGGAAC...	TRCN000...	GNG7	-2.96372...
CCCTATC...	TRCN000...	FLRT3	-2.93749...
GTCCATG...	TRCN000...	EXTL2	-2.91762...
GCTACTT...	TRCN000...	MPP5	-2.81121...
CCTCACT...	TRCN000...	C10orf129	-2.80026...
CCGATGT...	TRCN000...	NRN1L	-2.79512...
GCTACAA...	TRCN000...	GFM1	-2.79141...
GCATGCC...	TRCN000...	SEC14L4	-2.78571...
GCTTTAG...	TRCN000...	PPP2CB	-2.76037...
GCATAGT...	TRCN000...	PDE4A	-2.73328...
GTTCTAT...	TRCN000...	POU5F2	-2.72560...
CAAGTAC...	TRCN000...	TBC1D10B	-2.71006...
CAAGAAC...	TRCN000...	RESP18	-2.70636...
GATGACG...	TRCN000...	PLCXD1	-2.69624...
CCCTCTG...	TRCN000...	SRRD	-2.69139...
GACATCC...	TRCN000...	TXNDC2	-2.69134...
GCATCAG...	TRCN000...	SLC39A9	-2.67941...
GAACACC...	TRCN000...	ARF3	-2.67601...
CCATGCT...	TRCN000...	ALDOA	-2.67351...
GACTTCA...	TRCN000...	SYNGR2	-2.67093...
CCTAGCG...	TRCN000...	CDYL	-2.66493...
GACTCAT...	TRCN000...	GTF2H3	-2.65523...
GCCAACT...	TRCN000...	ENSA	-2.65449...
CGTTCAC...	TRCN000...	DAPK3	-2.63219...
GCATGTT...	TRCN000...	PSMD9	-2.62115...
CTGAACA...	TRCN000...	PFKP	-2.62049...
CCTAACCC...	TRCN000...	CLIC5	-2.62017...
CCAGAAG...	TRCN000...	SLC25A20	-2.59348...
GTCTGTT...	TRCN000...	STRAP	-2.57312...
GTCAAAG...	TRCN000...	MRPL42	-2.57292...
CACAGCC...	TRCN000...	EVI2B	-2.57214...
ATTAAGC...	TRCN000...	MRPS21	-2.57049...
CGTCTTC...	TRCN000...	MC1R	-2.56885...
GCTGAAG...	TRCN000...	PFKP	-2.56698...
GCAGCTT...	TRCN000...	NUMB	-2.56469...
CGACATA...	TRCN000...	STXBP3	-2.56123...
GCAGACA...	TRCN000...	SLC9A10	-2.54881...
GCTGGCA...	TRCN000...	OR8U1	-2.54577...
CTGCTCT...	TRCN000...	XCL2	-2.53955...
GCCACGG...	TRCN000...	PTPN13	-2.53002...
GCCTTGT...	TRCN000...	DOCK1	-2.52534...
CTACGAA...	TRCN000...	GCK	-2.52510...
CCGGGCA...	TRCN000...	ABLIM3	-2.52232...

Cancel OK

Run RIGER from Tools menu

The screenshot displays the GENE-E Dev application window. The main window shows a table of data with columns for 'Score', '#', 'Hairpin.Sequence', and 'Hairpin.IDs'. The 'Tools' menu is open, listing various analysis options. The 'RIGER...' option is highlighted in blue. The application title bar indicates the file 'shRNA_screen_data.txt (99031 x 1)'. The status bar at the bottom shows 'Showing 99031 out of 99031 rows, 1 out of 1 column'. The system tray at the bottom includes icons for various applications and system utilities.

GENE-E Dev File Edit View Visualization Tools Window Help

shRNA_screen_data.txt (99031 x 1)

Score

#	Hairpin.Sequence	Hairpin.IDs
1	CCCACATAAGCCTGAAATCAA	TRCN0000003636 F
2	CCAGATGCTGTTACCTCAGAT	TRCN0000014234 M
3	GTGGAAACGCTACGCATAGAA	TRCN0000008811 C
4	CCCTATCTGGAAAGATTACAT	TRCN0000146814 F
5	GTCATGCTTTTATAGATGAT	TRCN0000147657 E
6	GCTACTTTGTTAGGCTTGAAT	TRCN0000343292 N
7	CCTCACTTTCTAGGCTTACAT	TRCN0000154168 G
8	CCGATGTACACCATATACCA	TRCN0000141158 N
9	GCTACAACGTTCCGTTTCTAA	TRCN0000142606 G
10	GCATGCCAAGAAGCTCAGCTA	TRCN0000060097 S
11	GCTTTAGTAGATGGACAGATA	TRCN0000002504 P
12	GCATAGTAAAGCCGTAACAA	TRCN0000048809 P
13	GTTCATATAACCCGACAGAT	TRCN0000016972 P
14	CAAGTACCTCCAGGTTACTA	TRCN0000155847 T
15	CAAGAACCATGCCGTAAGGAT	TRCN0000139285 R
16	GATGACGTAAGCTGCAACAA	TRCN0000078245 P
17	CCCTCTGTTTAGCCAACCTGA	TRCN0000141236 S
18	GACATCCTAAAGCCTGAAGAA	TRCN0000161649 T
19	GCATCAGACAAGCCGACAGAA	TRCN0000038632 S
20	GAAACCCCAAGGCTGATATT	TRCN0000004679 A
21	CCATGCTTGCACTCAGAAGTT	TRCN0000299137 A
22	GACTTTCATCCAGAATTACGTT	TRCN0000150672 S
23	CCTAGCGAAGTCAGGTATCAA	TRCN0000127490 C
24	GACTCATATAGGCTTTCTAA	TRCN0000021019 C
25	GCCAAAGTGCAGGACAGACAA	TRCN00000318793 E
26	CGTTCATACCTGCAGCTTAA	TRCN0000155426 E
27	GCACTTTCTAATTCACGAT	TRCN0000003949 S
28	CTGAACACCTACAAGCGACTT	TRCN0000037777 P
29	CCTAAACAAGGCTCTAAAGAA	TRCN0000044378 C
30	CCAGAAGATGTGCTCAGCTAT	TRCN00000307710 S
31	GTCTGTTAGTAGTATGGAATA	TRCN0000060463 S
32	GTCAAAGAGAAGTATCTTGA	TRCN0000121805 M
33	CACAGCCTACCTTATTCACAT	TRCN0000135528 E
34	ATTAAGCTATCGCGGATATT	TRCN00000344080 MRPS21
35	CGTCTTCAGCACGCTCTTCAT	TRCN0000011745 MCLR
36	GCTGAAGAAGCAACGGATT	TRCN0000037775 PFKP
37	GCAGCTTACTTTCATCAGTGCA	TRCN0000082536 NUMB
38	GCACATATTGGCGTTGTGTTA	TRCN0000162507 STXBP3
39	GCCACATAATGACCATAATT	TRCN0000060133 SLC9A10
40	GCTGGCACTCAGCATAATCTA	TRCN0000061665 ORB1
41	CTGGCTCTCACTGCATAGAT	TRCN0000057996 NCL2
42	GCCACGGTCTATTCTTACTAA	TRCN0000338241 PTPN13
43	GCCTTGTGTTGAACTCAA	TRCN0000029077 DOCK1
44	CTACGAAGACCATCAGTGCGA	TRCN0000010270 GCK
45	CCGGGACAGAAAGAAAGTTAA	TRCN0000008578 ABLIM3
46	CTGAAGACTACGACCATGAAA	TRCN0000164559 ZG16B
47	CCATGTTAGAGCCCTAATTAT	TRCN0000129717 CCDC30
48	GCAAGCTATTACAAAGACAT	TRCN0000006259 PRKDC
49	GTTGTTGACAACTACATTCAA	TRCN0000158823 KL
50	CCCTTCTTGACTACAAACAT	TRCN0000060645 PLXNC1
51	GCAGCAATCGATAGCCTGAAT	TRCN0000155218 CBL2
52	CCTCAGTGCTTCGATCAGTT	TRCN0000005154 OR2A14
53	GAGGCATATACATGGACAAAT	TRCN0000151346 OSBP1A
54	GTGGAACTATGCGAACAAAT	TRCN0000033500 BCL2L1
55	GCCAGACTGACCAATACATA	TRCN0000116944 MFAP3L
56	GCTCAGTTTCGATGGACAGAT	TRCN0000180463 RAET1G
57	CCTACTTAAATGGGTTGATTAT	TRCN0000167340 SYAP1
58	CCAACAATCAGACCAGGTTTA	TRCN0000053036 PAPP7
59	GTCCATACTGTTACATGTT	TRCN0000147322 FHL5
60	CCACAGTATTAGCCAAATA	TRCN0000152036 BEST3
61	GAAGAGCTTATGACTTACT	TRCN0000057075 SOCS3
62	CTAAGCACATTAACATGCAAT	TRCN0000039854 CHEK1
63	CTACCCTTCTCATCGCTCTA	TRCN0000008057 ADRA1B
64	GCAACTGCATAAGCAGCAGAT	TRCN0000022446 USP36
65	GAACAAACAAAGTGCAGGATT	TRCN0000001107 SNW1

Showing 99031 out of 99031 rows, 1 out of 1 column

Info

Doench R&D John's Projects Rearray Request

Comparison:

Comparison 1

Class A	Annotations	Class B
<input type="text" value="Q-"/>	<input type="text" value="Q-"/>	<input type="text" value="Q-"/>
<input type="checkbox"/> (Select All)	<input type="checkbox"/> (Select All) <input type="checkbox"/> Hairpin.Sequence	<input type="checkbox"/> (Select All)

Split comparisons by:

Create a separate comparison for each unique value (e.g. split by cell annotation and auto-create separate comparisons for MCF7 and PC3)

Metric for ranking hairpins:

Number of permutations:

Method to convert hairpins to genes:

Gene rank order:

Random seed used to generate permutations:

Adjust gene scores to accommodate variation in hairpin set size

Hairpin scores between -0.5 and 0 are re-adjusted to -0.5, scores at 0 are unchanged, scores between 0 and 0.5 are re-adjusted to 0.5

Hairpins are pre-scored

Hairpin Id:

Convert hairpins to:

RIGER output – notice new tab



#	Unique Gene	Hairpin IDs
1	PTPN13	TRCN0000350966
2	PFKP	TRCN0000195199
3	KIAA0907	TRCN0000121825
4	ALS2CR4	TRCN0000131197
5	PYGL	TRCN0000119084
6	RAD21	TRCN0000281201
7	ENSA	TRCN0000285005
8	STRAP	TRCN0000060466
9	SLC2A1	TRCN0000043584
10	URGCP	TRCN0000350994
11	SNRPG	TRCN0000074341
12	HADHB	TRCN0000028040
13	RNF220	TRCN0000162943
14	NEFM	TRCN0000116392
15	ATL3	TRCN0000134045
16	PTRH1	TRCN0000051617
17	GRM6	TRCN0000009032
18	HNDNDK	TRCN0000062456

To download Gene Report



Mac: Control-click
PC: Right-click

#	Unique Gene	Hairpin IDs
1	PTPN13	TRCN0000350966
2	PFKP	TRCN0000195199
3	KIAA0907	TRCN0000121825
4	ALS2CR4	TRCN0000131197
5	PYGL	TRCN0000119084
6	RAD21	TRCN0000281201
7	ENSA	TRCN0000285005
8	STRAP	TRCN0000060466
9	SLC2A1	TRCN0000043584
10	URGCP	TRCN0000350994
11	SNRPG	TRCN0000074341
12	HADHB	TRCN0000028040
13	RNF220	TRCN0000162943
14	NEFM	TRCN0000116392
15	ATL3	TRCN0000134045
16	PTRH1	TRCN0000051617
17	GRM6	TRCN0000009032
18	HNRNPK	TRCN00000062456
19	TUT1	TRCN0000129603
20	SERAC1	TRCN0000150795
21	PKNOX1	TRCN0000020527
22	ZNF652	TRCN0000152364
23	MTRF1	TRCN0000155064
24	ZNF774	TRCN0000107918
25	GPI	TRCN0000049151
26	EPB41	TRCN0000083545
27	MAFB	TRCN0000017679
28	SMEK2	TRCN0000179416
29	IFNA5	TRCN0000005846
30	HK1	TRCN0000197140
31	ZRANB1	TRCN0000073813
32	USP16	TRCN0000007594
33	NR2E1	TRCN0000021673
34	LALBA	TRCN0000055449
35	FXD1	TRCN0000038621
36	GYPC	TRCN0000083398
37	ZNF354A	TRCN0000329814
38	TRIB1	TRCN0000338194
39	GINS1	TRCN0000127692
40	NLCN2	TRCN0000075282
41	MFSB8	TRCN0000154785
42	ENO1	TRCN0000029326
43	PDCD6IP	TRCN0000343652
44	PGAM4	TRCN0000146885
45	LOC402524	TRCN0000107969
46	SNRPN	TRCN0000075136
47	FLJ9231	TRCN0000140188
48	ZIM2	TRCN0000020377
49	RASGRP4	TRCN0000072926
50	SLFN11	TRCN0000155578
51	LOC285830	TRCN0000161002
52	TNFAIP8L1	TRCN0000165586
53	GPT	TRCN0000034979
54	MRPS25	TRCN0000117702
55	SPG20	TRCN0000084036
56	PDE10A	TRCN0000005476
57	MS4A1	TRCN0000061519
58	CRYGD	TRCN0000083942
59	ITGB8	TRCN0000057766
60	UNC5B	TRCN0000061813
61	WTAP	TRCN0000010776
62	PLDN	TRCN0000144654

Gene report, opened in Excel

	A	B	C	D	E	F	G	H	I	J	K	L
	Gene	Hairpins	# Hairpins	Hairpin ranks	NES	Gene rank	p-value	p-value rank	# Hairpins 500	# Hairpins 1000	# Hairpins 5000	# Hairpins 10000
2	PTPN13	CCTTTGGATC	7	80020, 3041	0.0005392	1	0.0001	3	2	2	2	2
3	PFKP	CAGCACTTA	10	21704, 5886	0.001492	2	0.0001	5	3	3	3	3
4	KIAA0907	GAGCTAAAC	5	61964, 4665	0.003306	3	0.0001	1	2	2	2	2
5	ALS2CR4	GCGTATCCA	5	34505, 6793	0.004365	4	0.0001	4	2	2	2	2
6	PYGL	GCAAGATAT	5	23996, 3611	0.006421	5	0.0001	2	2	2	2	2
7	RAD21	CCAGATAGC	4	47226, 468	0.006694	6	0.0002	6	2	2	2	3
8	ENSA	GAGCTGAAG	6	30353, 7855	0.008271	7	0.0005	10	2	2	2	3
9	STRAP	TGGGTCATA	5	27204, 3219	0.008789	8	0.0004	9	1	2	2	2
10	SLC2A1	GCGGAATTC	5	6051, 614, 5	0.009927	9	0.0004	8	1	2	2	3
11	URGCP	CGCGTTGTA	7	65336, 6834	0.01028	10	0.0004	7	1	2	2	2
12	SNRPG	AGTGGACAA	5	73471, 7167	0.01163	11	0.0006	12	1	2	2	2
13	HADHB	CGTTAGCCA	5	38400, 1455	0.01171	12	0.0006	11	0	2	2	2
14	RNF220	GCATGAGAA	9	47041, 6064	0.01226	13	0.0009	15	2	2	2	2
15	NEFM	CGATTCCTA	5	97390, 5140	0.01279	14	0.0008	13	1	2	2	2
16	ATL3	CCTTATTTGT	5	90110, 2934	0.01354	15	0.0009	14	1	2	2	2
17	PTRH1	AGAGCCATG	5	33303, 4777	0.01388	16	0.0009	16	1	2	2	2
18	GRM6	AGCGTGATT	9	86745, 2415	0.01444	17	0.0012	26	1	2	2	2
19	HNRNPK	TGATGTTTGA	5	22909, 2955	0.01606	18	0.001	18	1	1	4	4
20	TUT1	CAGGGACTT	4	60968, 1006	0.01609	19	0.0011	22	0	1	2	2
21	SERAC1	GCTTGGAAT	5	88799, 8548	0.01618	20	0.001	17	0	2	2	2
22	DKNOY1	CCAACTGCT	5	84133, 1035	0.01620	21	0.0011	23	1	1	2	2

Congratulations, now you have another list!



- Unbiased means of determining if your list makes sense?
- GSEA – Gene Set Enrichment Analysis
 - Originally used to analyze microarray data
 - MSigDB is a collection of curated gene sets, curated from pre-existing data
- DAPPLE – Protein-protein interaction data

GSEA & MSigDB

The screenshot shows the MSigDB website interface. At the top, there is a navigation bar with links for GSEA Home, Downloads, Molecular Signatures Database (selected), Documentation, and Contact. A sidebar on the left contains a list of links: MSigDB Home, About Collections, Browse Gene Sets, Search Gene Sets, Investigate Gene Sets, View Gene Families, and Help. The main content area is titled "Molecular Signatures Database v3.1" and is divided into four sections: Overview, Collections, Registration, and Current Version. The Overview section describes the database and lists key actions: Search, Browse, Examine, Download, and Investigate. The Collections section lists six major collections (c1-c6) with brief descriptions. The Registration section explains how to register and use the database. The Current Version section is currently empty.

MSigDB
Molecular Signatures Database

Molecular Signatures Database v3.1

Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- ▶ **Search** for gene sets by keyword.
- ▶ **Browse** gene sets by name or collection.
- ▶ **Examine** a gene set and its annotations. See, for example, the [ANGIOGENESIS gene set page](#).
- ▶ **Download** gene sets.
- ▶ **Investigate** gene sets:
 - ▶ **Compute overlaps** between your gene set and gene sets in MSigDB.
 - ▶ **Categorize** members of a gene set by gene families.
 - ▶ **View the expression profile** of a gene set in any of the three provided public expression compendia.

Registration

Please [register](#) to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Current Version

Collections

The MSigDB gene sets are divided into 6 major collections:

- c1 positional gene sets** for each human chromosome and cytogenetic band.
- c2 curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
- c3 motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
- c4 computational gene sets** defined by mining large collections of cancer-oriented microarray data.
- c5 GO gene sets** consist of genes annotated by the same GO terms.
- c6 oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.

<http://www.broadinstitute.org/gsea/msigdb/index.jsp>

Launch Java version



GSEA v2.0.10 (Gene set enrichment analysis -- Broad Institute)

Home

Steps in GSEA analysis

- Load data
- Run GSEA
- Leading edge analysis

Gene set tools

- Chip2Chip mapping
- Browse MSigDB

Analysis history


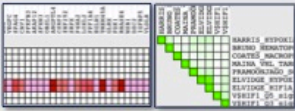
GSEA reports

Processes: click 'status' field for results

Name	Status

Show results folder

Steps in GSEA

- 1. What you need for GSEA**
 - Expression data set
 - Phenotype annotation
 - Gene sets – use MSigDB or your own gene sets
- 2. Run GSEA**
 - Start with default parameters
 - If you want to collapse probes to genes, specify chip platform
- 3. View results**
- 4. Leading edge analysis**
 - Leading edge finds genes driving enrichment results

Gene Set Tools

Chip2Chip mapping

- Convert gene sets between platforms

Chip2Chip mapping

Explore MSigDB gene sets

- Search the database of thousands of gene sets
- Browse the gene sets by name
- Find overlapping gene sets
- Export gene sets

Browse MSigDB

See also


- MSigDB online tools at: www.broadinstitute.org/msigdb

Getting Help

GSEA web site:
www.broadinstitute.org/gsea

GSEA documentation:
www.broadinstitute.org/gsea/wiki

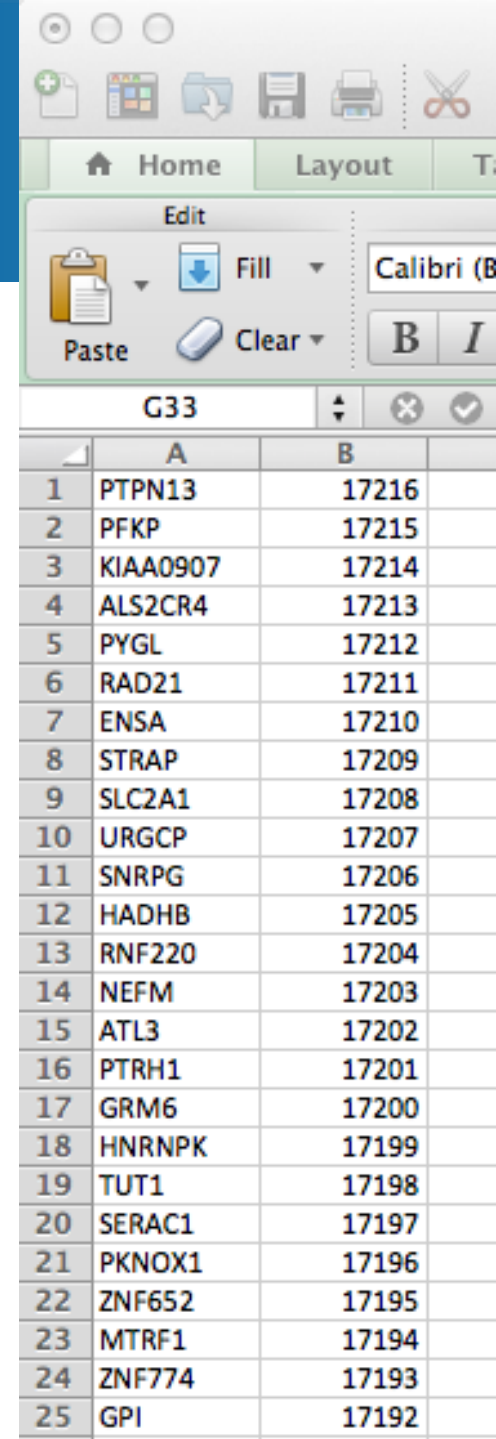
Email the GSEA team:
gsea@broadinstitute.org



12:22:19 AM 9723 [INFO] Made Vdb dir: /Users/jdoench/gsea_home/output/apr04 50M of 112M

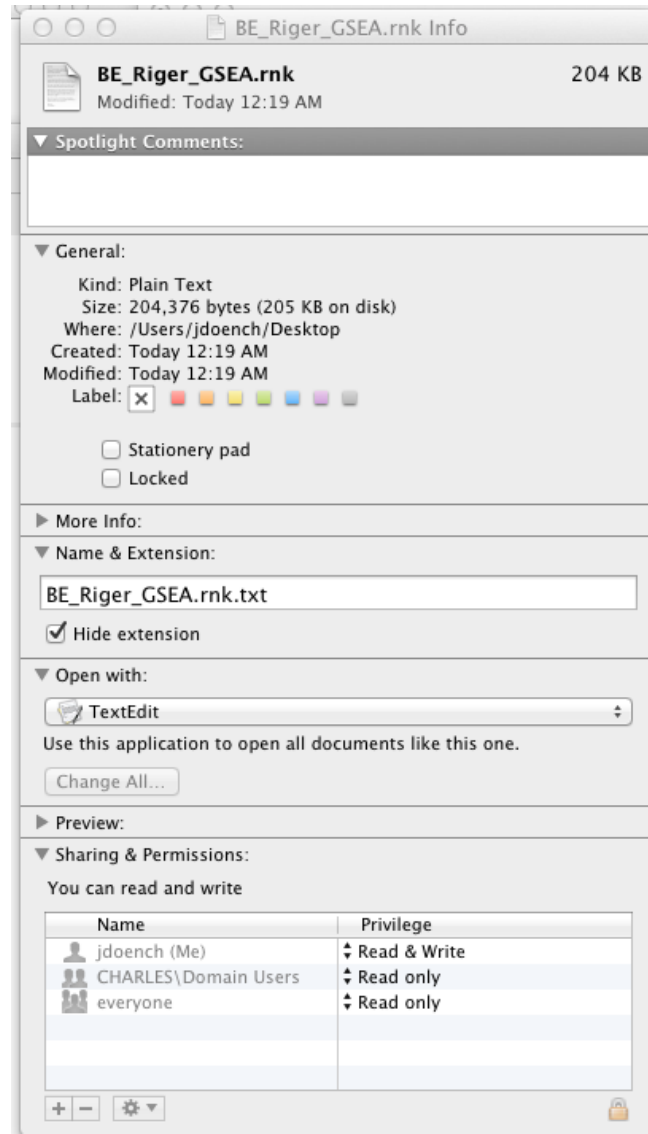
Requirements

- Each gene symbol should appear only once
- Higher value is best (i.e. rank #1 would be *least* enriched)
 - Need to invert output of RIGER



	A	B
1	PTPN13	17216
2	PFKP	17215
3	KIAA0907	17214
4	ALS2CR4	17213
5	PYGL	17212
6	RAD21	17211
7	ENSA	17210
8	STRAP	17209
9	SLC2A1	17208
10	URGCP	17207
11	SNRPG	17206
12	HADHB	17205
13	RNF220	17204
14	NEFM	17203
15	ATL3	17202
16	PTRH1	17201
17	GRM6	17200
18	HNRNPK	17199
19	TUT1	17198
20	SERAC1	17197
21	PKNOX1	17196
22	ZNF652	17195
23	MTRF1	17194
24	ZNF774	17193
25	GPI	17192

Get your extensions right!



Load your file (.rnk)



GSEA v2.0.10 (Gene set enrichment analysis -- Broad Institute)

Steps in GSEA analysis

- Load data
- Run GSEA
- Leading edge analysis

Gene set tools

- Chip2Chip mapping
- Browse MSigDB

Analysis history

GSEA reports

Processes: click 'status' field for results

Name	Status
------	--------

Show results folder

Home Load data x

Load data: Import data into the application

Method 1:

Method 2:

Method 3: drag and drop files here

Supported file formats

Dataset: *res* or *gct* (Broad/MIT),
pcl (Stanford)
txt (tab-delim text)

Phenotype labels: *cls*

Gene sets: *gmx* or *gmt*

Recently used files (double click to load, right click for more options)

- ../Desktop/BE_Riger_GSEA.rnk

Object cache (objects already loaded & ready for use, right click for more options)

- Objects in memory (shift-click to expand all)
 - RankedGeneList

Run GSEA PreRanked



Gene set enrichment analysis (GSEA) v2.0.10

File Options Downloads **Tools** Help

Home

Steps in GSEA analysis

- Load data
- Run GSEA**
- Leading edge analysis

Gene set tools

- Chip2Chip mapping
- Browse MSigDB

Analysis history

GSEA reports

Processes: click 'status' field for results

Name	Status
------	--------

Show results folder

Steps in GSEA

- 1. What you need for GSEA**
 - Expression data set
 - Phenotype annotation
 - Gene sets – use MSigDB or your own gene sets
- 2. Run GSEA**
 - Start with default parameters
 - If you want to collapse probes to genes, specify chip platform
- 3. View results**
- 4. Leading edge analysis**
 - Leading edge finds genes driving enrichment results

Gene Set Tools

Chip2Chip mapping

- Convert gene sets between platforms

Chip2Chip mapping

Explore MSigDB gene sets

- Search the database of thousands of gene sets
- Browse the gene sets by name
- Find overlapping gene sets
- Export gene sets

Browse MSigDB

See also

- MSigDB online tools at: www.broadinstitute.org/msigdb

Getting Help

GSEA web site:
www.broadinstitute.org/gsea

GSEA documentation:
www.broadinstitute.org/gsea/wiki

Email the GSEA team:
gsea@broadinstitute.org

BROAD INSTITUTE

12:27:24 AM 9723 [INFO] Made Vdb dir: /Users/jdoench/gsea_home/output/apr04 54M of 112M

Selection of Gene Sets



The screenshot displays the GSEA v2.0.10 web interface. The main window is titled "GSEA v2.0.10 (Gene set enrichment analysis -- Broad Institute)" and has a tab for "Run Gsea on a Pre-Ranked gene list". The interface is divided into several sections:

- Steps in GSEA analysis:** Includes "Load data", "Run GSEA" (the current step), and "Leading edge analysis".
- Gene set tools:** Includes "Chip2Chip mapping" and "Browse MSigDB".
- Analysis history:** A section for tracking previous analyses.
- GSEA reports:** A table with columns for "Name" and "Status".

The main configuration area is titled "GseaPranked: Run GSEA on a pre-ranked (with external tools) gene list" and contains the following fields:

- Required fields:**
 - Gene sets database:** A text input field with a dropdown arrow.
 - Number of permutations:** A dropdown menu set to "1000".
 - Ranked List:** A dropdown menu.
 - Collapse dataset to gene symbols:** A dropdown menu set to "true".
 - Chip platform(s):** A text input field with a dropdown arrow.
- Basic fields:** A section for basic configuration options.
- Advanced fields:** A section for advanced configuration options.

An open modal window titled "Select one or more gene sets(s)" is shown in the foreground. It has two tabs: "Gene matrix (from website)" and "Gene sets (grp)". The "Gene sets (grp)" tab is active, displaying a list of gene sets with a yellow folder icon to the left of each name:

- c1.all.v3.1.symbols.gmt [Positional]
- c2.all.v3.1.symbols.gmt [Curated]
- c2.cgp.v3.1.symbols.gmt [Curated]
- c2.cp.v3.1.symbols.gmt [Curated]
- c2.cp.biocarta.v3.1.symbols.gmt [Curated]
- c2.cp.kegg.v3.1.symbols.gmt [Curated]
- c2.cp.reactome.v3.1.symbols.gmt [Curated]
- c3.all.v3.1.symbols.gmt [Motif]
- c3.mir.v3.1.symbols.gmt [Motif]
- c3.tft.v3.1.symbols.gmt [Motif]
- c4.all.v3.1.symbols.gmt [Computational]
- c4.cgn.v3.1.symbols.gmt [Computational]
- c4.cm.v3.1.symbols.gmt [Computational]
- c5.all.v3.1.symbols.gmt [Gene ontology]
- c5.bp.v3.1.symbols.gmt [Gene ontology]
- c5.cp.v3.1.symbols.gmt [Gene ontology]

The modal window includes a "Help" button with a question mark icon, and "Cancel" and "OK" buttons at the bottom right.

Settings



GSEA v2.0.10 (Gene set enrichment analysis -- Broad Institute)

Home | Load data x | **Run Gsea on a Pre-Ranked gene list** x

GseaPreranked: Run GSEA on a pre-ranked (with external tools) gene list

Required fields

Gene sets database: ...

Number of permutations:

Ranked List:

Collapse dataset to gene symbols:

Chip platform(s): ...

Basic fields

Advanced fields

Steps in GSEA analysis

- Load data
- Run GSEA**
- Leading edge analysis

Gene set tools

- Chip2Chip mapping
- Browse MSigDB

Analysis history

GSEA reports

Processes: click 'status' field for results

	Name	Status
1	GseaPreranked	Success 5
2	GseaPreranked	Success 5

Permutations: Run with 10 first, but then 1000

Show results folder

? Reset Last Command Normal (cpu usage) Run

Output, opens in browser



GSEA Report for Dataset VW_GSEA_in

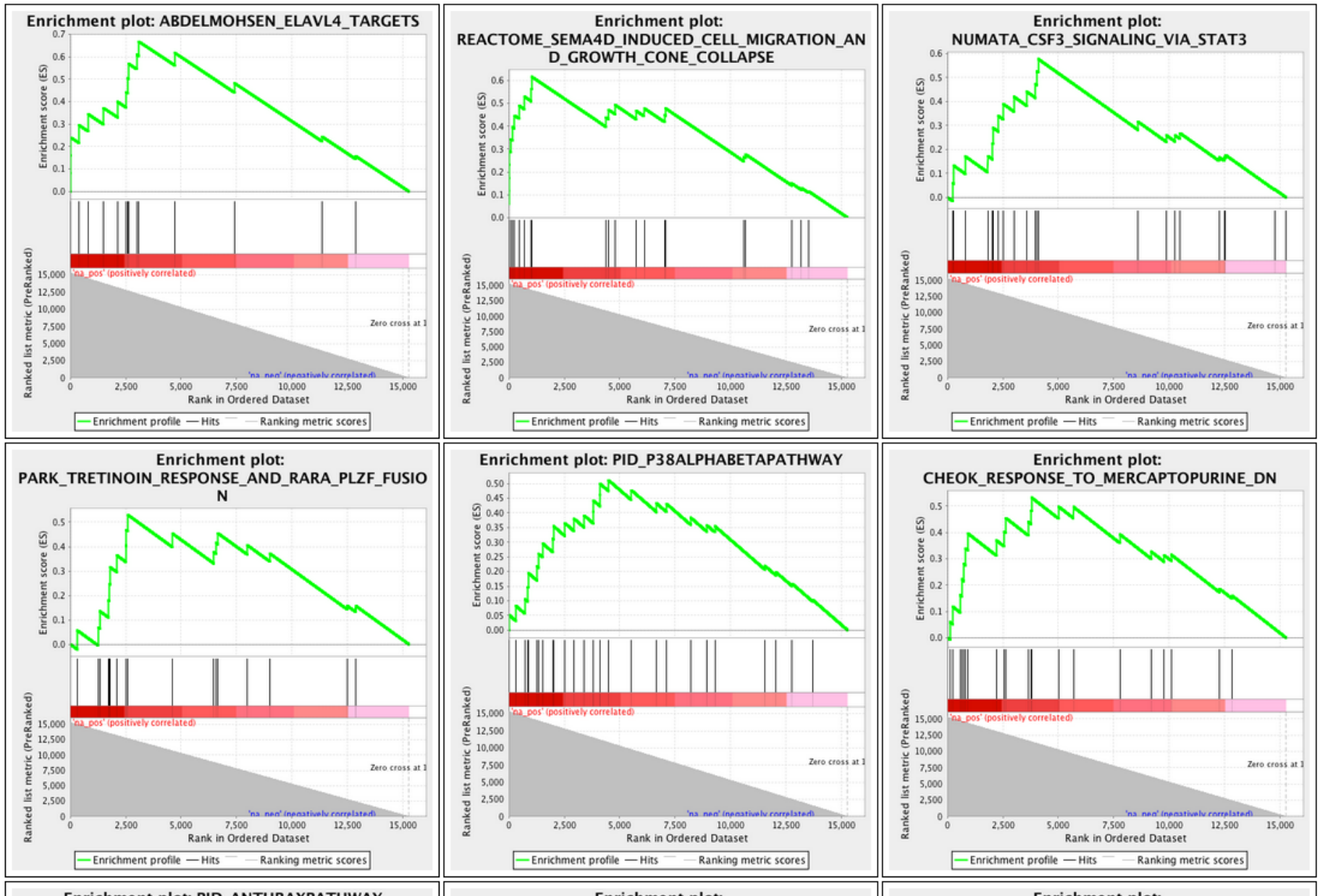
Enrichment in phenotype: **na**

- 3656 / 3687 gene sets are upregulated in phenotype **na_pos**
- 37 gene sets are significant at FDR < 25%
- 413 gene sets are significantly enriched at nominal pvalue < 1%
- 413 gene sets are significantly enriched at nominal pvalue < 5%
- [Snapshot](#) of enrichment results
- Detailed [enrichment results in html](#) format
- Detailed [enrichment results in excel](#) format (tab delimited text)
- [Guide to](#) interpret results

Enrichment in phenotype: **na**

- 31 / 3687 gene sets are upregulated in phenotype **na_neg**
- 16 gene sets are significantly enriched at FDR < 25%
- 2 gene sets are significantly enriched at nominal pvalue < 1%
- 2 gene sets are significantly enriched at nominal pvalue < 5%
- [Snapshot](#) of enrichment results
- Detailed [enrichment results in html](#) format
- Detailed [enrichment results in excel](#) format (tab delimited text)
- [Guide to](#) interpret results

Table: Snapshot of enrichment results



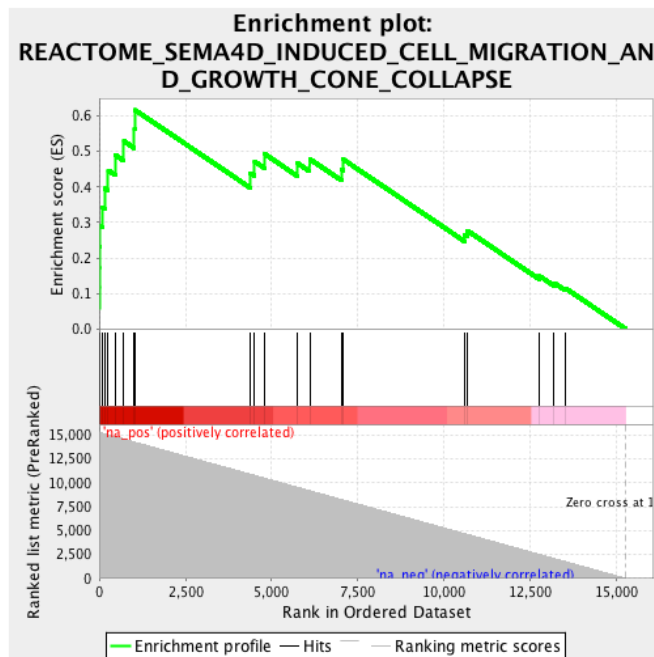


Fig 1: Enrichment plot: REACTOME_SEMA4D_INDUCED_CELL_MIGRATION_AND_GROWTH_CONE_COLLAPSE
 Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Table: GSEA details [\[plain text format\]](#)

	PROBE	GENE SYMBOL	GENE_TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	MYL12B			0	15275.000	0.0582	Yes
2	MYH9	MYH9 Entrez , Source	myosin, heavy chain 9, non-muscle	1	15274.000	0.1164	Yes
3	ARHGEF12	ARHGEF12 Entrez , Source	Rho guanine nucleotide exchange factor (GEF) 12	2	15273.000	0.1746	Yes
4	CDC42	CDC42 Entrez , Source	cell division cycle 42 (GTP binding protein, 25kDa)	5	15270.000	0.2327	Yes
5	RHOA	RHOA Entrez , Source	ras homolog gene family, member A	29	15246.000	0.2892	Yes



<http://www.broadinstitute.org/mpg/dapple/dapple.php>

that size here, in kb.

Plot:

Return a picture of your network.

Iterate:

If any gene achieves a bonferonni corrected score of $p < 0.05$, prioritize that gene and restart.

Nearest Gene:

For SNP inputs only. Select the closest gene, rather than all genes in the wingspan.

Inputs:

Inputs can be genes, SNPs or regions. Choose a file or enter inputs in the box (one line per input).

WARNING: Do not enter more than 200 snps.

See example of SNP input

See example of Region input

See example of Gene-Region input

See example of Gene input

Genes to Specify:

Only for SNP and region input. Input any genes that you would like to fix as the causal gene for an input locus, such that all other genes in that region will not be included. Each line should only contain 1 gene. Genes should be in gene symbol ID.

See example of genes to specify

Your email address (required):

Description (required):

Warning: All non-alphanumeric characters and spaces will be removed

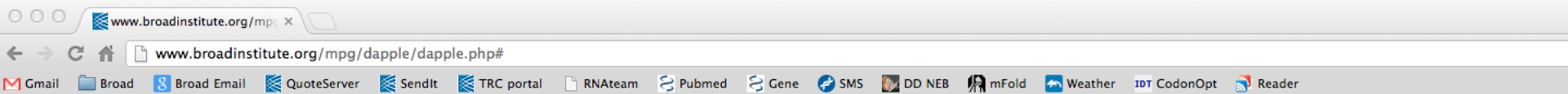
 kb downstream of TX END Plot Color by DAPPLE p-value Simplify Indirect Network? Iterate Use Nearest Gene No file chosen

```
MYL12B
MYH9
ARHGEF12
ZAK
INSL6
CDC42
```

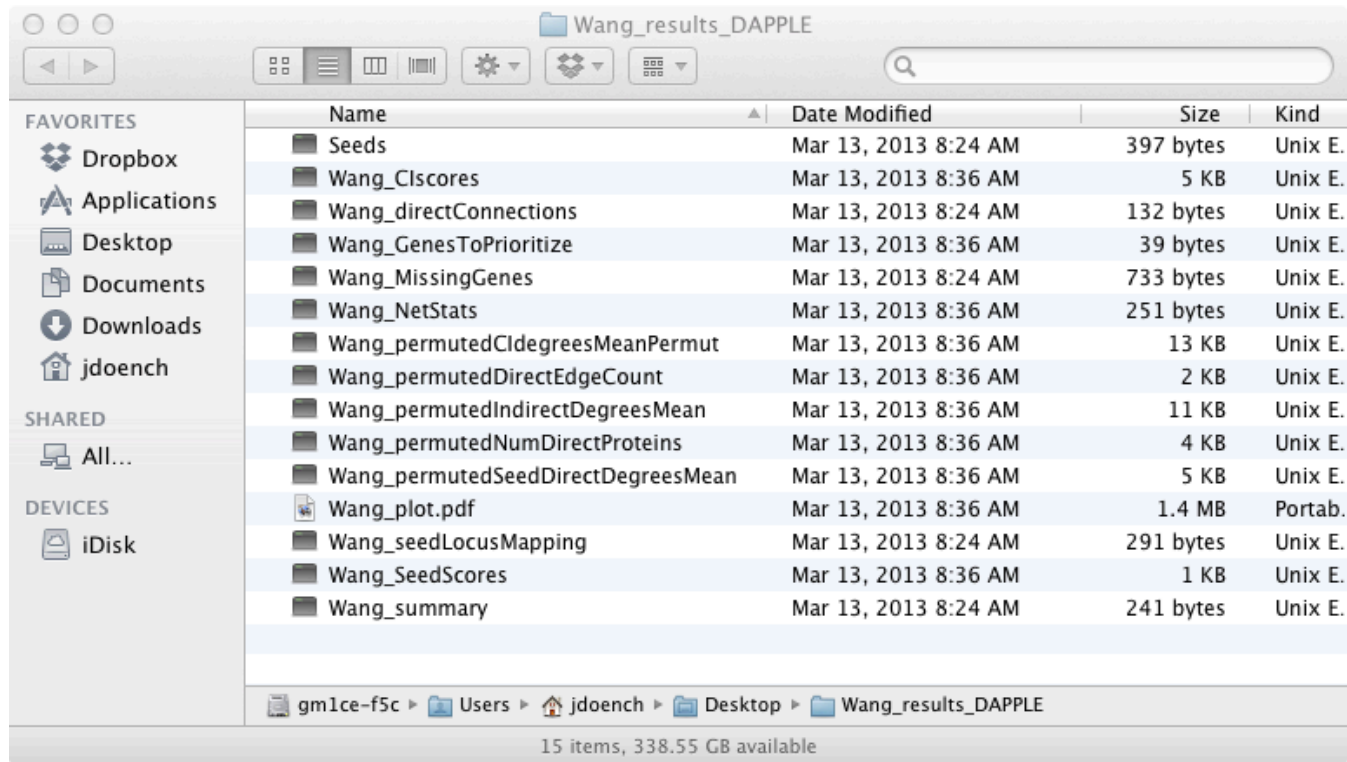
 No file chosen

If you haven't yet, click here to receive email notification of any major changes in DAPPLE

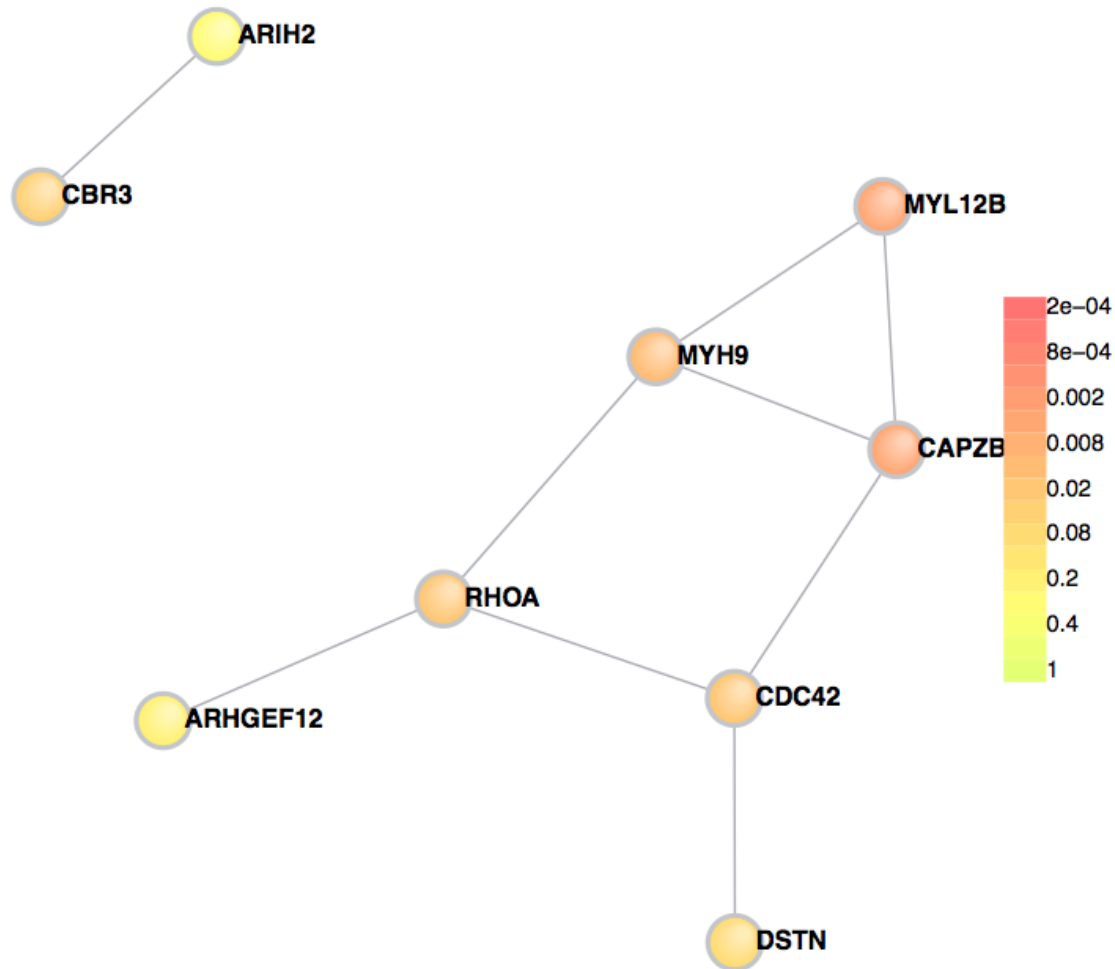
DAPPLE download when done running



DAPPLE server is currently overloaded: there are 51 DAPPLE jobs on the server currently. The maximum allowed is 50. Please try your query 6 hours later. If this issue persists, please contact dapple@broadinstitute.org. Thankyou.



Output: Protein Protein interactions



Off-target Effects miRkat



- Input is simple: siRNA/shRNA sequence and the numerical output of the screen
- Search for both
 - Enrichment of seeds to match to known microRNA sequences
 - Enrichment of UTRs preferentially targeted by multiple seeds



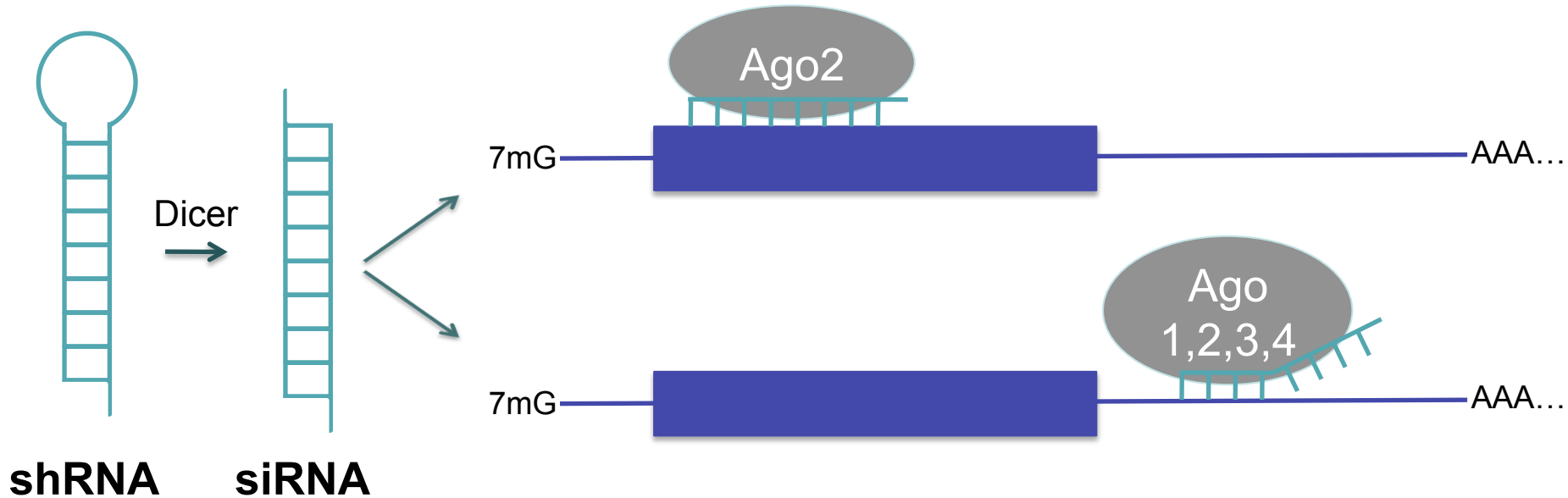
microRNA knockdown
analysis **toolkit***

Two pathways for shRNA library



RNAi

Perfect (or near-perfect) match to mRNA causes mRNA cleavage and degradation: **18 – 22 nt**



microRNA 'Seed' region binds to 3'UTR to represses translation and may destabilize mRNA: **6 – 8 nt**

LKO shRNA anatomy



- 7-mer “seed” sequence binding site, on average, corresponds to nts 11 – 17 of sense strand

0 1 2
1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 CT

GGCCGGAACTTATCGATCGATC**G** **C**

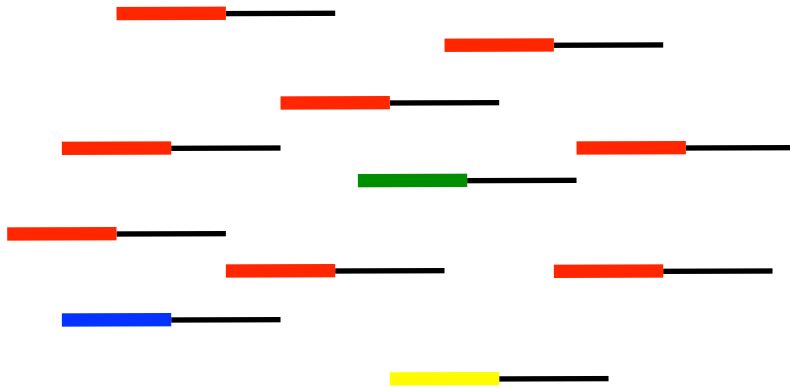
TTT**TGGCCTTGAAT****AGCTAGCT****AGC** **G**

GA

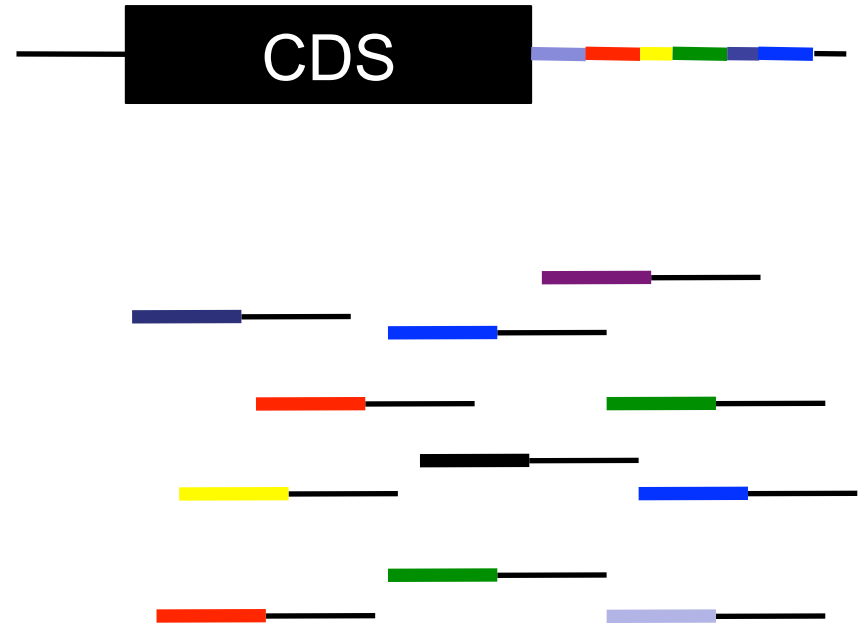
Two types of microRNA effects to search for in hit shRNAs



Hit shRNAs:

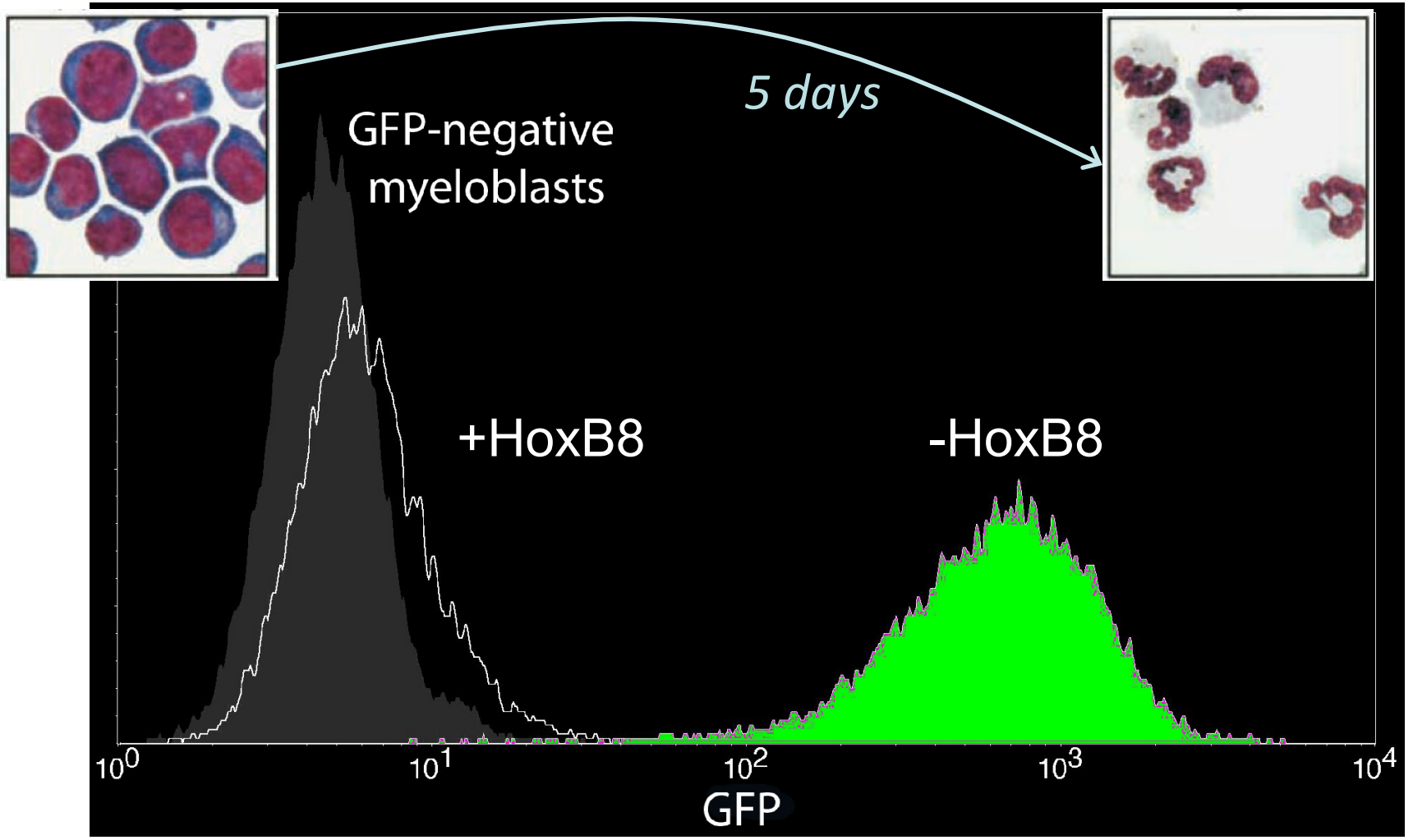


- Enrichment of seeds



- Enrichment of UTRs

AML model: HoxB8-induced differentiation block



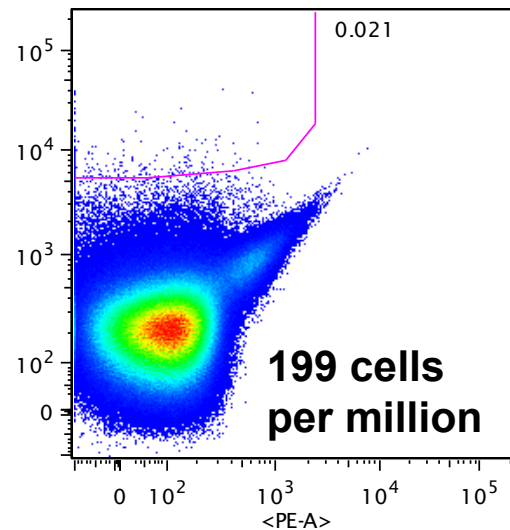
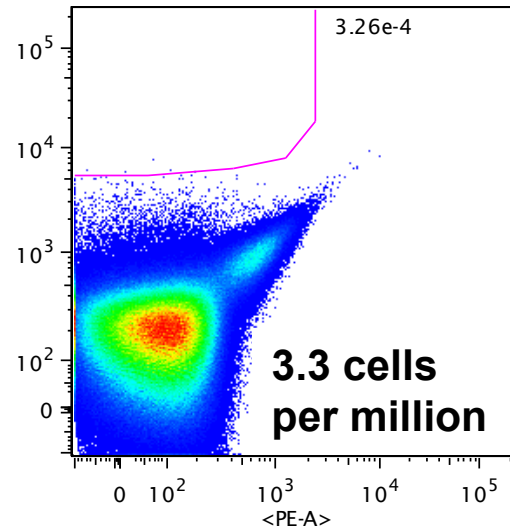
Pooled screen with flow cytometry selection of hits



**Background: Introduce
negative control shRNA**

**Screen: Introduce
40,000 shRNAs targeting
8,000 mouse genes**

GFP



60-fold
enrichment of
GFP⁺ cells in
library-infected
cells



HoxB8 scores as a clear hit, but few other genes with multiple-shRNA hits



1% of shRNAs defined as hits

Rank	Symbol	Sequence	Parts per million		Fold Enrichment
			Unsorted	GFP+	
1	Hoxb8	UAGCCGUAGAAGUUGCCGUUUU	0.1	2185.1	36490
2	Traf5	AAUUCUCUCAGAGACCGGUUUU	1.1	2118.8	1981
3	LOC434093	GUGUUGACUAUACAGCCGUUUU	1.0	476.7	482
4	1810035L1	GUUCUCUCAGCUCACUCGUUUU	1.2	548.2	444
5	LOC381842	GUCUCUCUACUGGUAGGUUUU	110.8	22186.2	200
6	Itgax	UUCUCUCUGCAUGUGUGGUUUU	39.3	7362.4	188
7	Ehbp1	AUUUGGCUUUGUGAUAGCUUUU	36.3	6245.2	172
8	Eraf	AUUUGGCUAGAAACUGGCUUUU	39.6	6778.0	171
9	Oprd1	AAUUUGGUGUACCGGACGUUUU	8.2	1323.5	161
10	Slc2a8	AUUCUCUCUUCUACCUGGUUUU	11.4	1804.0	159
102	Hoxb8	ACUGCUGGGAACUUGUCUUUU	22.6	593.7	26

Recurrent sequences towards 5' end of antisense strand: miRNA seed?



1% of shRNAs defined as hits

Rank	Symbol	Sequence	Parts per million		Fold Enrichment
			Unsorted	GFP+	
1	Hoxb8	UAGCCGUAGAAGUUGCCGUUUU	0.1	2185.1	36490
2	Traf5	AAU UCUCUC AGAGACCGGUUUU	1.1	2118.8	1981
3	LOC434093	GUGUUGACUAUACAGCCGUUUU	1.0	476.7	482
4	1810035L1	GU UCUCUC AGCUCACUCGUUUU	1.2	548.2	444
5	LOC381842	G UCUCUC UUACUGGUAGGUUUU	110.8	22186.2	200
6	Itgax	U UCUCUC UGCAUGUGUGGUUUU	39.3	7362.4	188
7	Ehbp1	AUUUGG CUUUGUGAUAGCUUUU	36.3	6245.2	172
8	Eraf	AUUUGG CUAGAAACUGGCUUUU	39.6	6778.0	171
9	Oprd1	A AUUUGG UGUACCGGACGUUUU	8.2	1323.5	161
10	Slc2a8	AU UCUCUC UUCUACCUGGUUUU	11.4	1804.0	159
102	Hoxb8	ACUGCUGGGAAACUUGUCUUUU	22.6	593.7	26

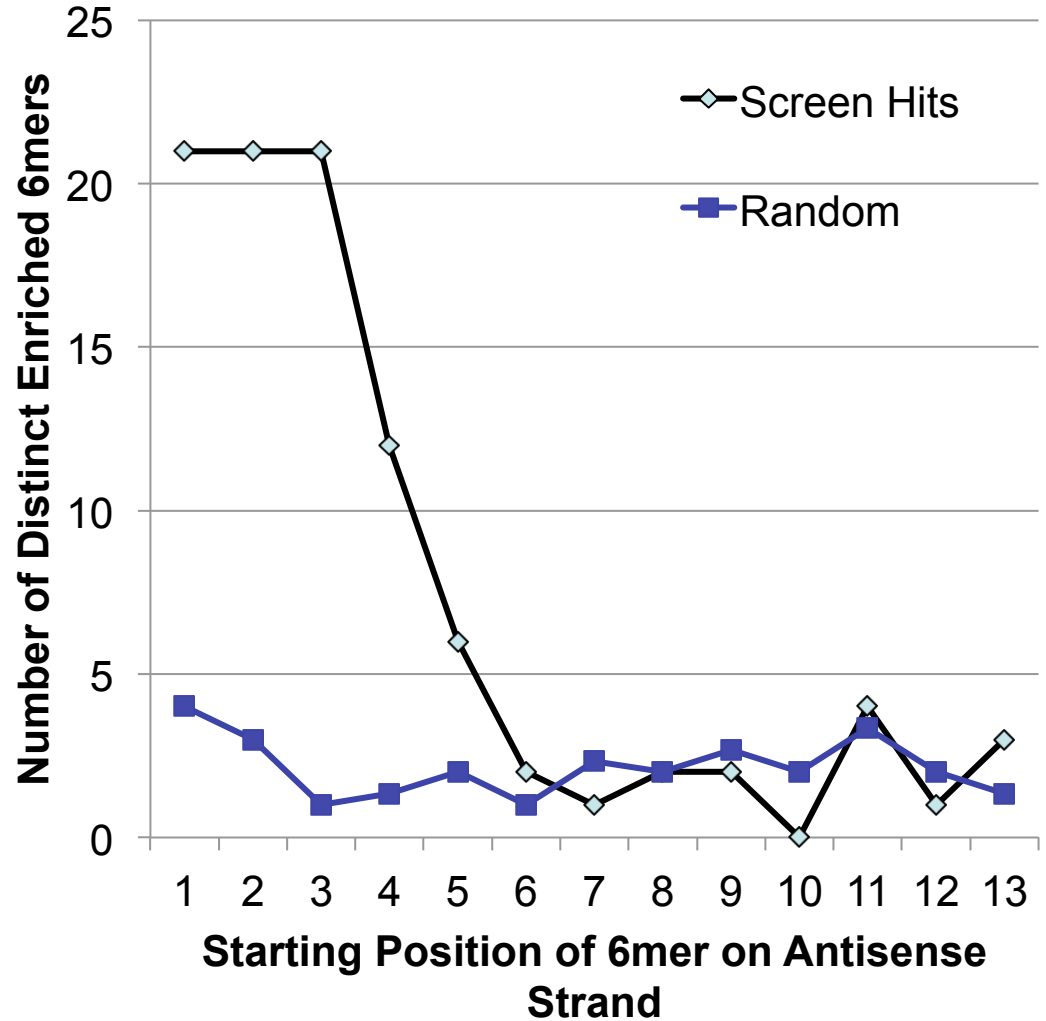
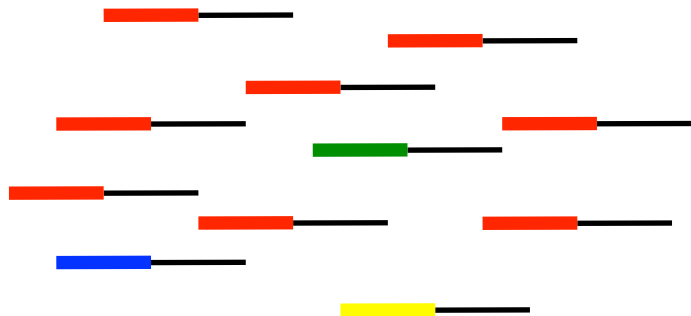
Where does this enrichment occur?



- Scan all 6mer frames of antisense strand



- Find enrichment at 5' end, i.e. miRNA seed region



Match enriched 6mers to known miRNA seed sequences



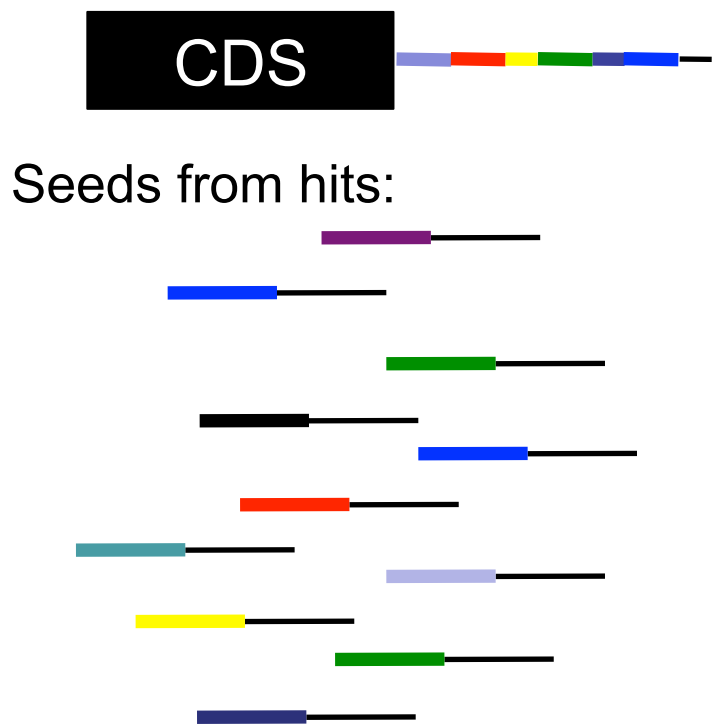
UCUCCC 13 of 19 appeared in hit list
mir-150 UCUCCCAACCCUUGUACCAGUG
mir-343 UCUCCCUUCAUGUGCCCAGA

CUUCUC 10 of 47 appeared in hit list
mir-207 GCUUCUCCUGGCUCUCCUCCCUC

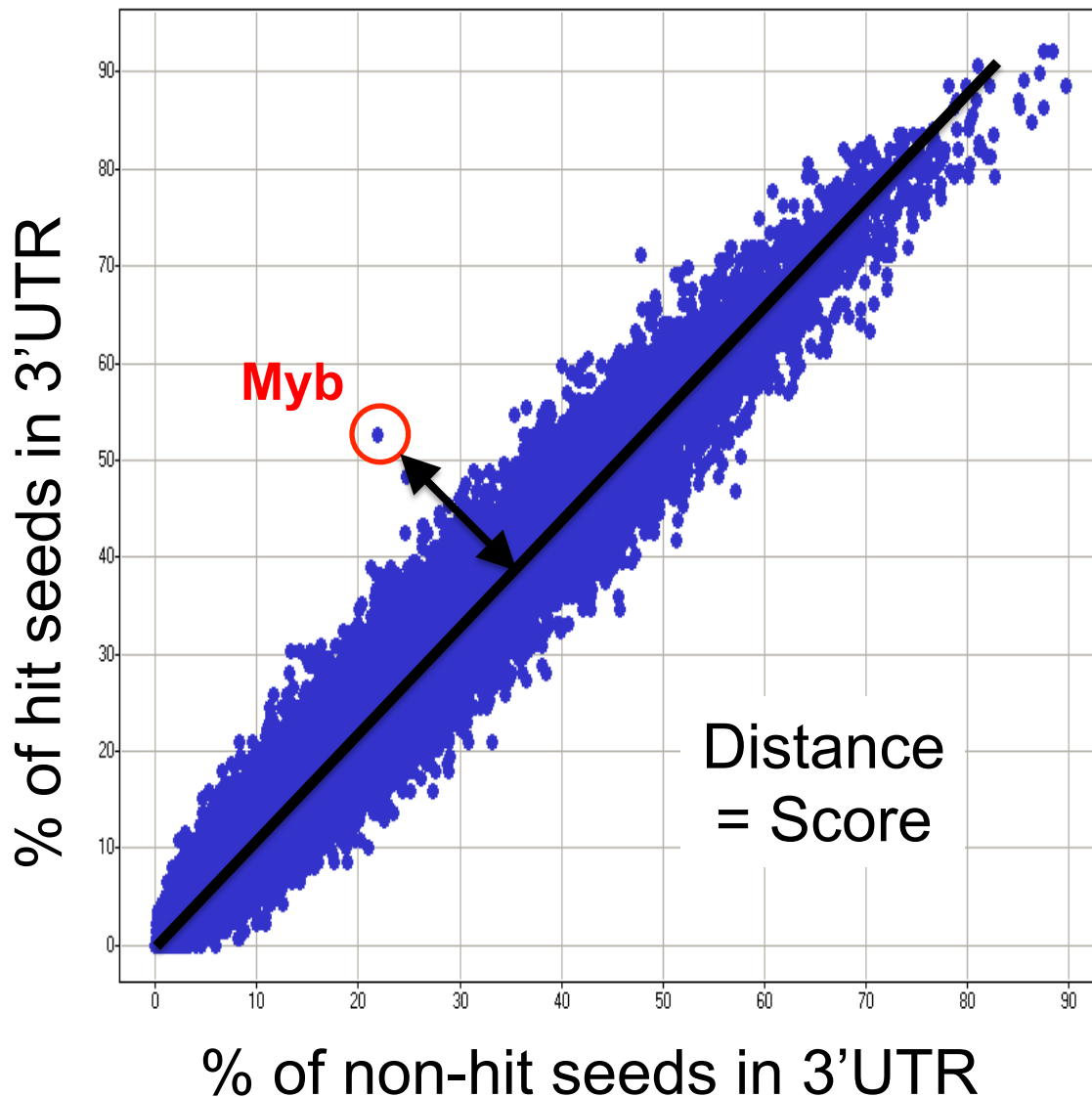
AGUGCA 8 of 25 appeared in hit list
mir-130 CAGUGCAAUAGUAUUGUCAAGC
mir-301 CAGUGCAAUGUUAAAAGGGCAU

Discover microRNAs involved in biological phenotype

Enrichment of UTRs



52% of hit seeds
match Myb 3'UTR
21% of non-hit seeds
match



Summary of Sykes AML/HoxB8 screen



- Primary screen properly identified HoxB8 but few other multiple-shRNA hits
- Enrichment for shRNAs matching microRNA seeds
 - mir-150
- Enrichment for target UTRs
 - Myb
 - GSEA to generate hypotheses about additional genes and pathways

MiR-150 Controls B Cell Differentiation by Targeting the Transcription Factor c-Myb

Changchun Xiao,¹ Dinis Pedro Calado,^{1,4} Gunther Galler,^{1,4} To-Ha Thai,¹ Heide Christine Patterson,¹ Jing Wang,¹ Nikolaus Rajewsky,^{2,5} Timothy P. Bender,³ and Klaus Rajewsky^{1,*}

¹The CBR Institute for Biomedical Research, Harvard Medical School, Boston, MA 02115, USA

²Center for Comparative Functional Genomics, Department of Biology, New York University, New York, NY 10003, USA

³Department of Microbiology, University of Virginia Health System, Charlottesville, VA 22908, USA

⁴These authors contributed equally to this work.

⁵Present address: Systems Biology, Max Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse 10, 13092 Berlin, Germany.

*Correspondence: rajewsky@cbr.med.harvard.edu

DOI 10.1016/j.cell.2007.07.021

How common is this effect?



Project “Achilles”

Genome-scale shRNA screens for proliferation-essential genes in 100s of cancer cell lines

(To identify vulnerabilities of particular tumor types based on oncogenes, tumor suppressors, tissue or origin, etc.)

Survey miRNA effects in 209 genome-scale proliferation screens
55,000 shRNAs

These proliferation screens produce strong, expected on-target hits



Cell lines with known cancer driver mutations depend on those drivers

Known
Paradigms

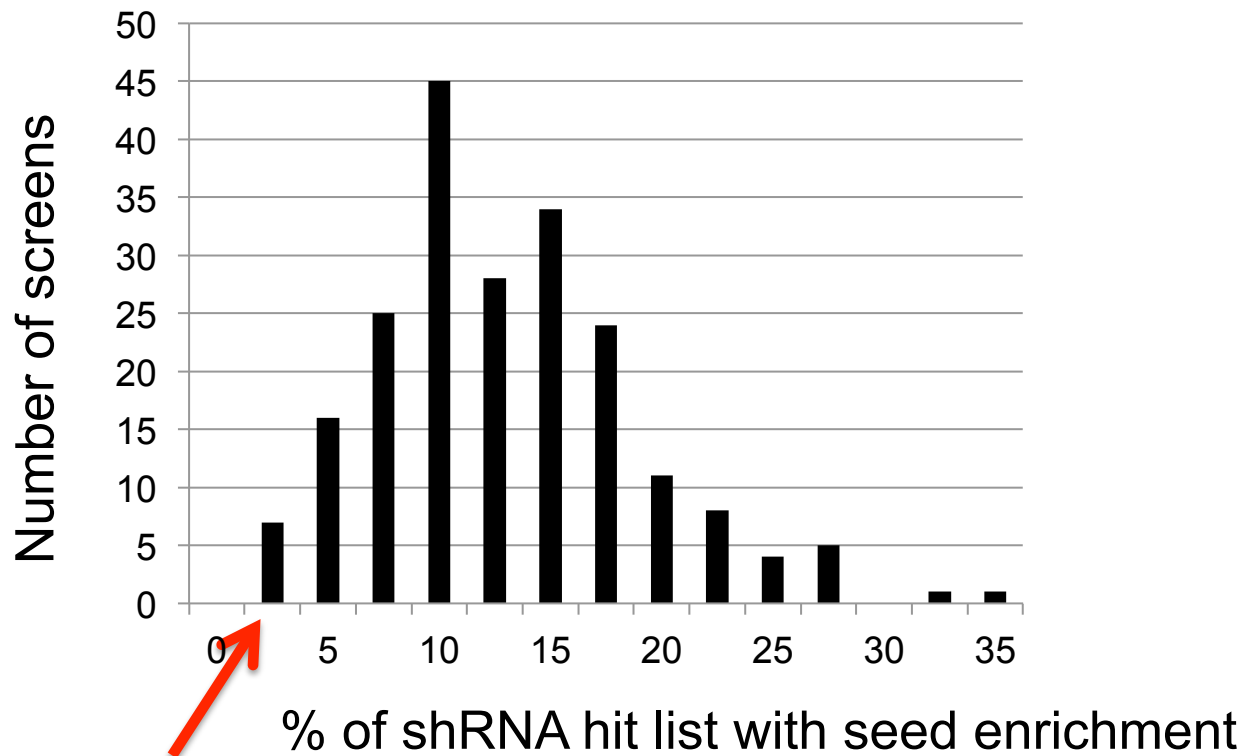
Dependency	Most significant correlate	Q-value
PIK3CA	PIK3CA mut	0
KRAS	KRAS mut	0
BRAF	BRAF mut	0
NRAS	NRAS mut	0
mTOR	PIK3CA mut	0
CTNNB1	APC mut	0.02
MDM4	TP53 wt	0.48

How common are miRNA-based effects in these screens?



209 proliferation screens

Typical screen: ~10% of hits clearly from miRNA seed effects



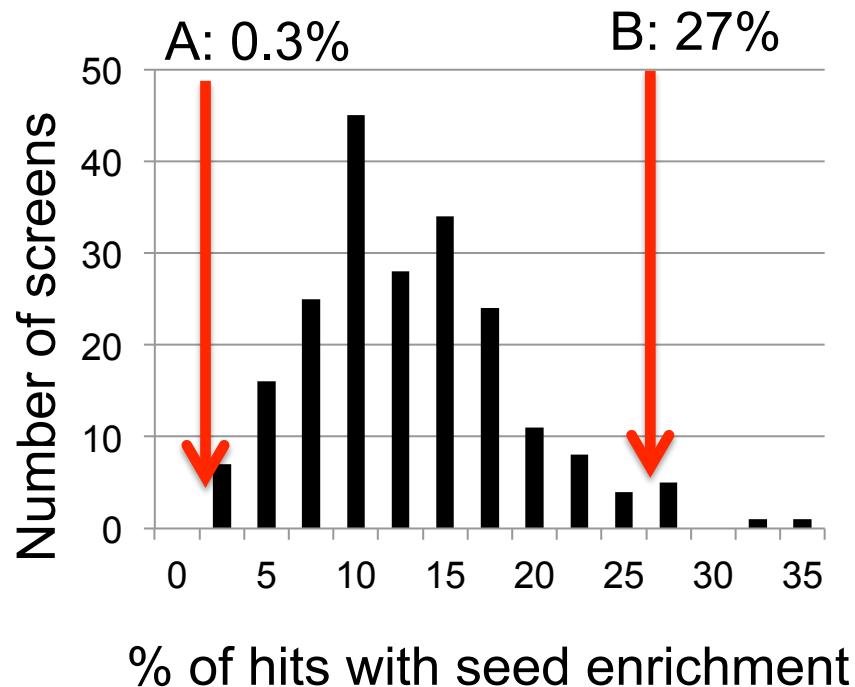
0.2% for randomized hit list

How common are the miRNA effects in other types of screens?



2 modifier screens:

- A. Rescue of sensitive cells from a chemotherapeutic
- B. Rescue from hormone-deprivation



So, what to do?



- miRNA effects can be found in screens
- Always look for them, subtract them out of your on-target analysis
- Use them to discover miRNA-related biology and genes of interest

Where is analysis headed?



- Annotate each shRNA to transcript, not to gene
- Use Ataris/Achilles, qPCR, L1000 to annotate effective shRNAs