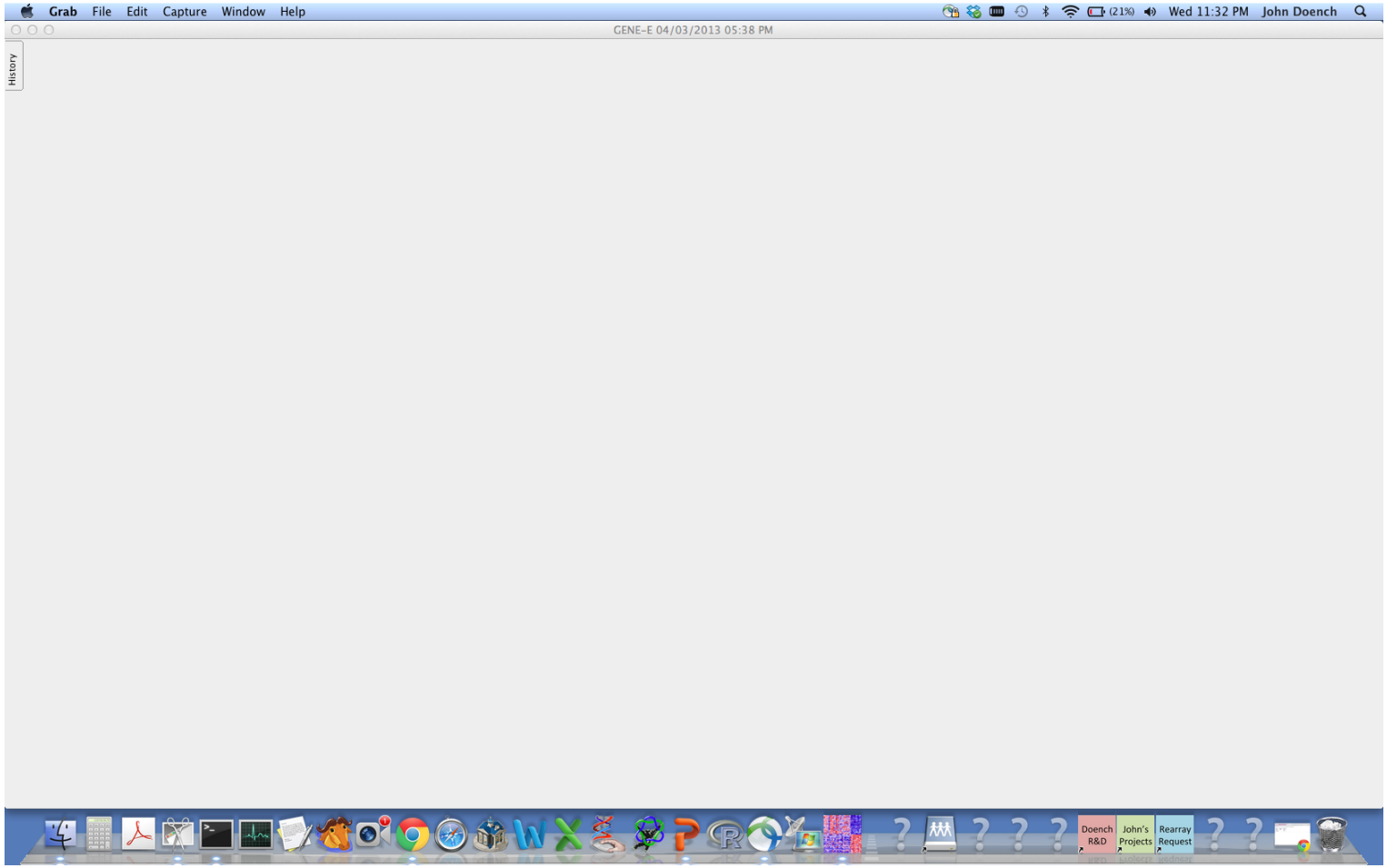
I have a list of shRNAs

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

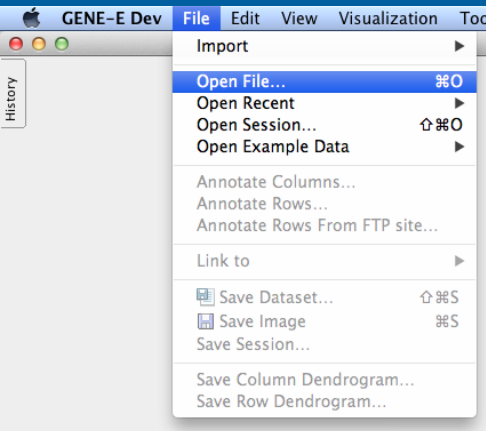
- http://www.broadinstitute.org/~jdoench/

GENE-E

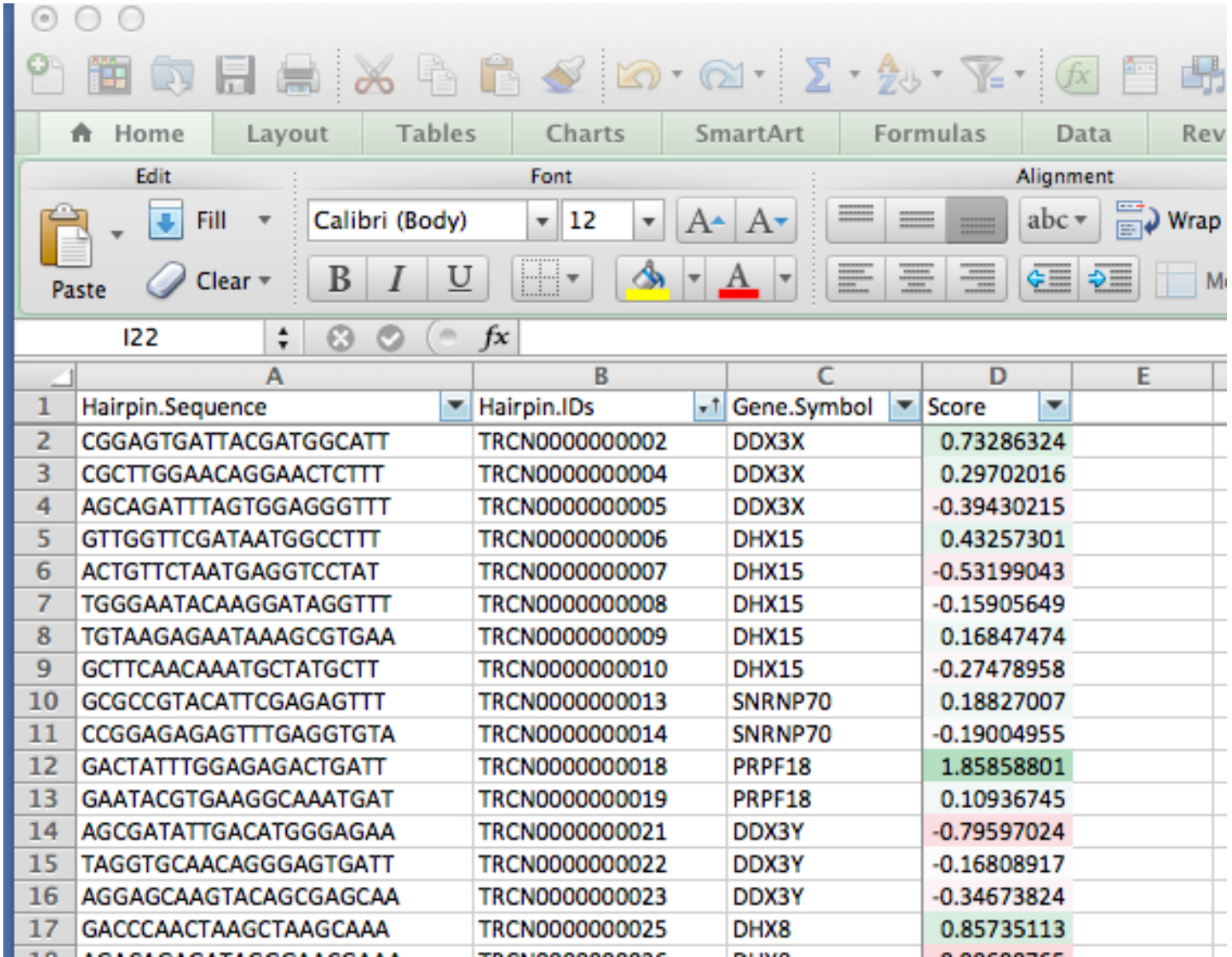
- http://www.broadinstitute.org/cancer/software/GENE-E/dev/



Open file (no need to import)

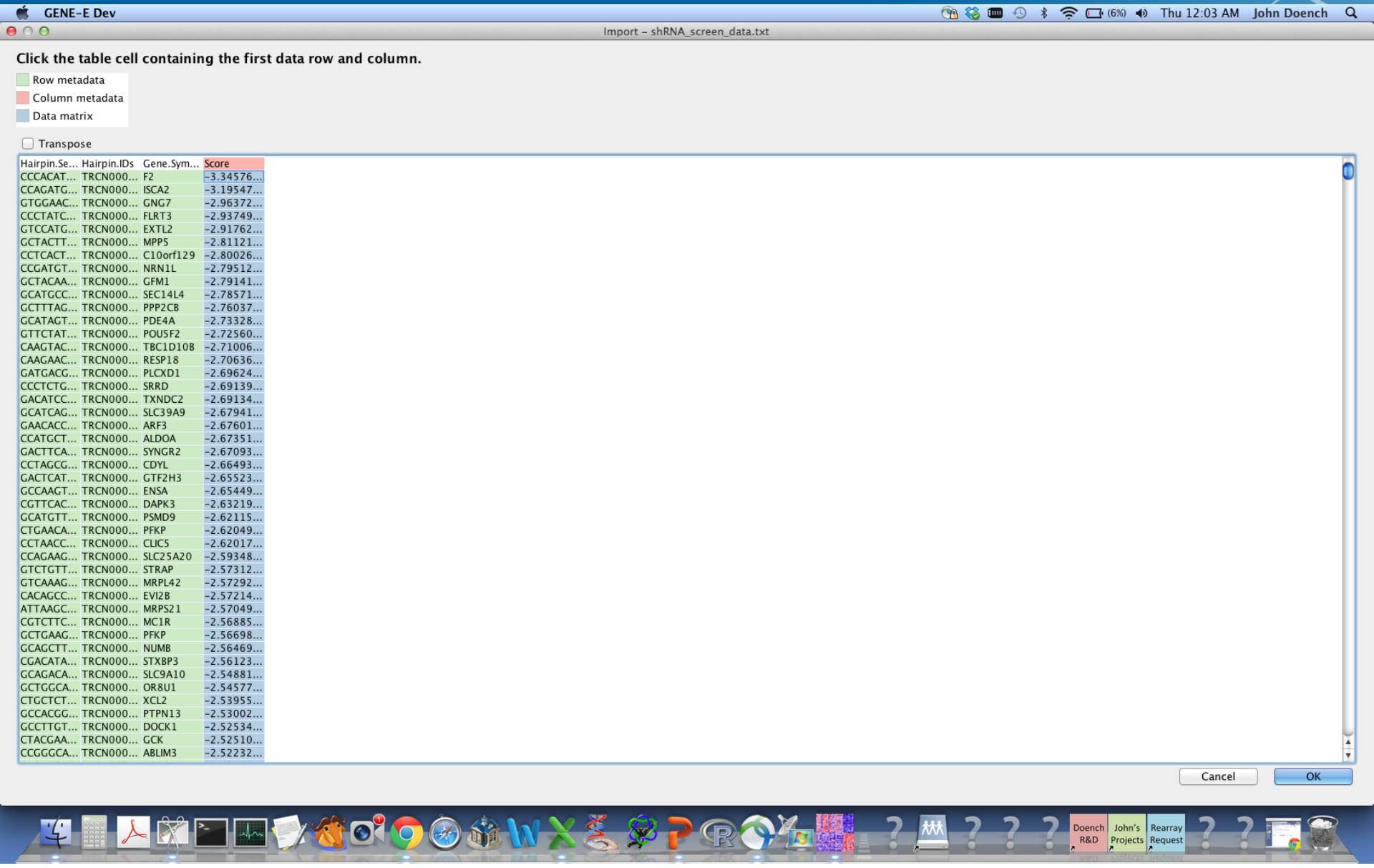


Input File (save as .txt from Excel)

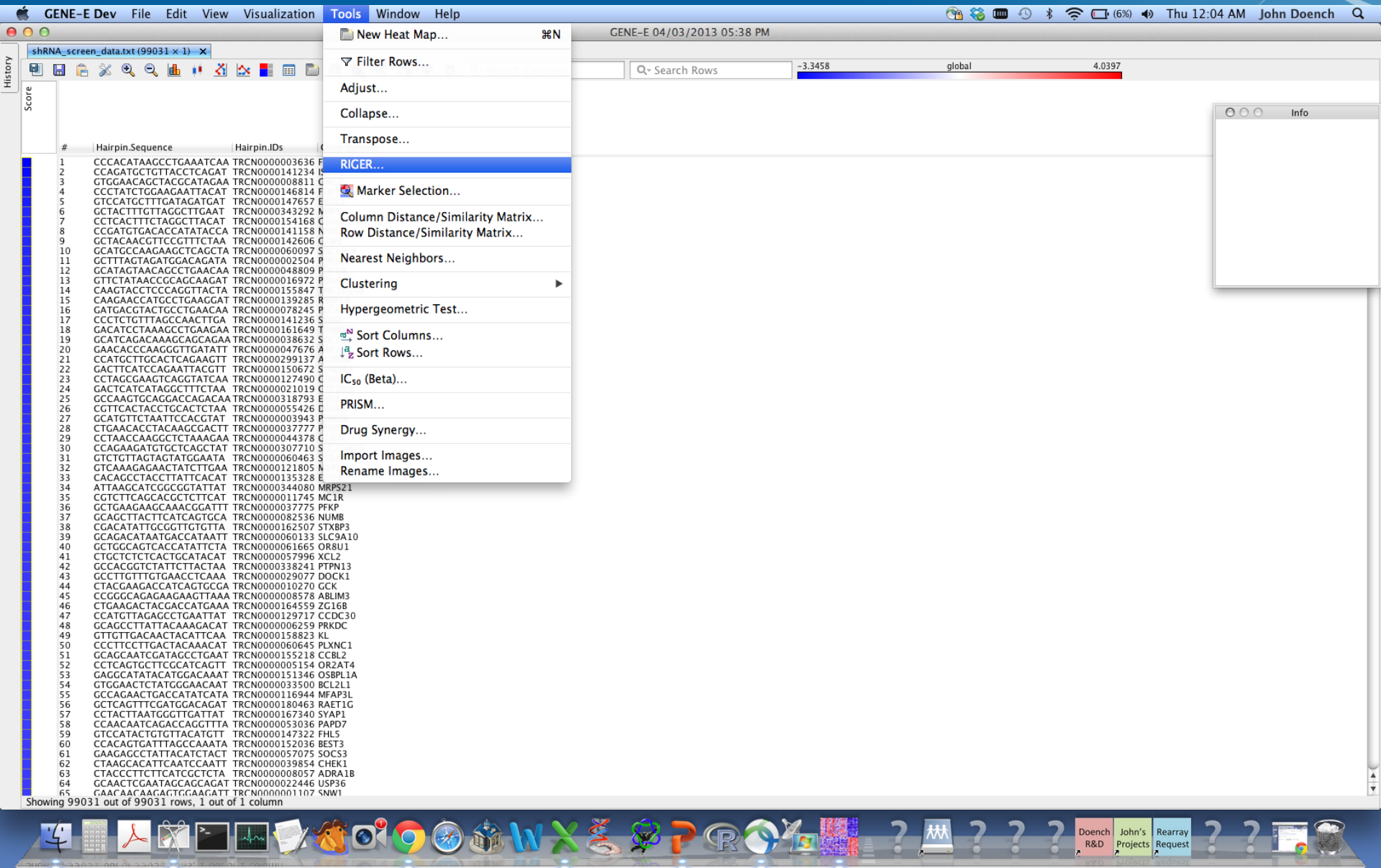


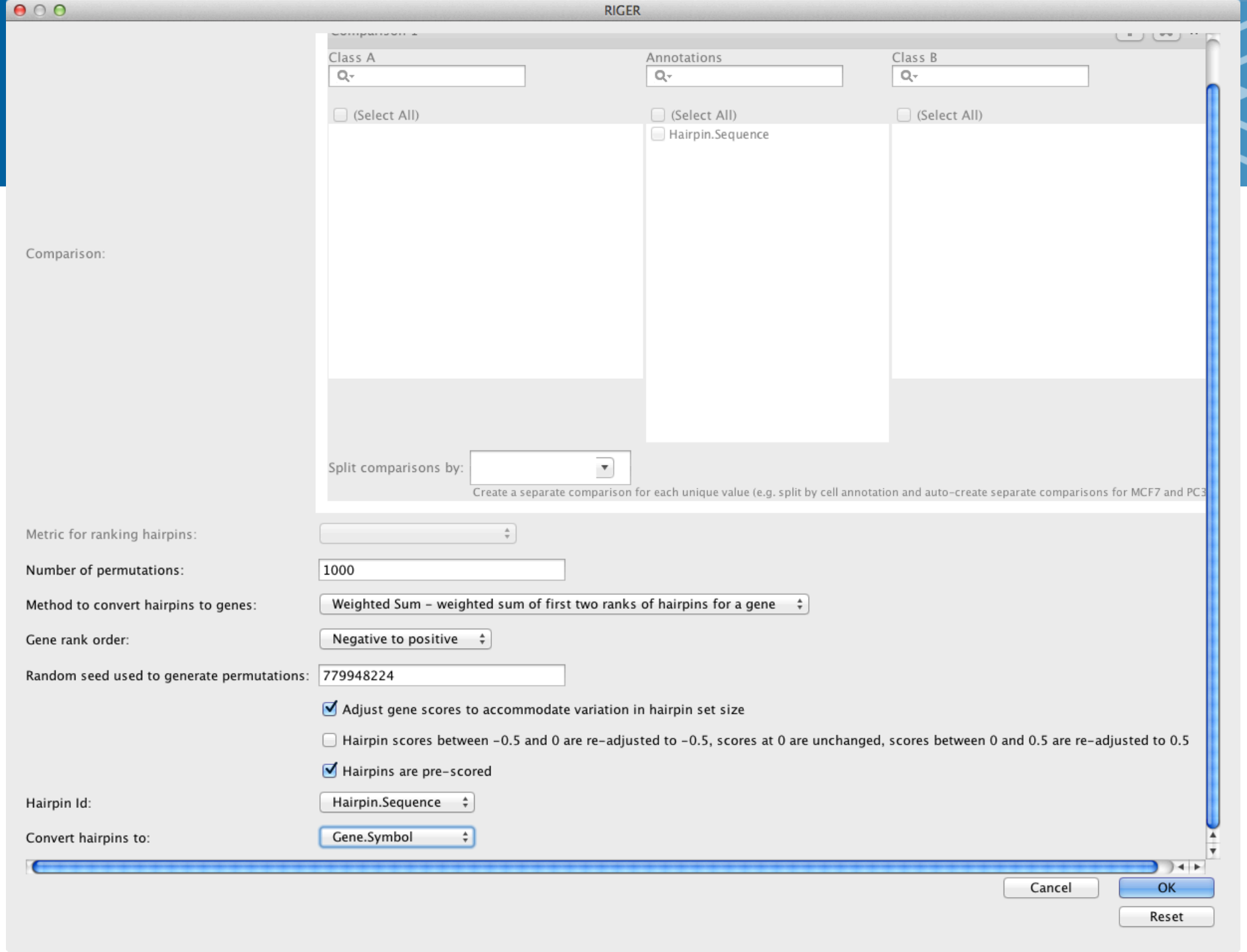
	A	B	C	D	E
1	Hairpin.Sequence	Hairpin.IDs	Gene.Symbol	Score	
2	CGGAGTGATTACGATGGCATT	TRCN0000000002	DDX3X	0.73286324	
3	CGCTTGGAACAGGAACTCTTT	TRCN0000000004	DDX3X	0.29702016	
4	AGCAGATTTAGTGGAGGGTTT	TRCN0000000005	DDX3X	-0.39430215	
5	GTTGGTTCGATAATGGCCTTT	TRCN0000000006	DHX15	0.43257301	
6	ACTGTTCTAATGAGGTCCTAT	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	GACCCAACTAAGCTAAGCAAA	TRCN0000000025	DHX8	0.85735113	

Import Data

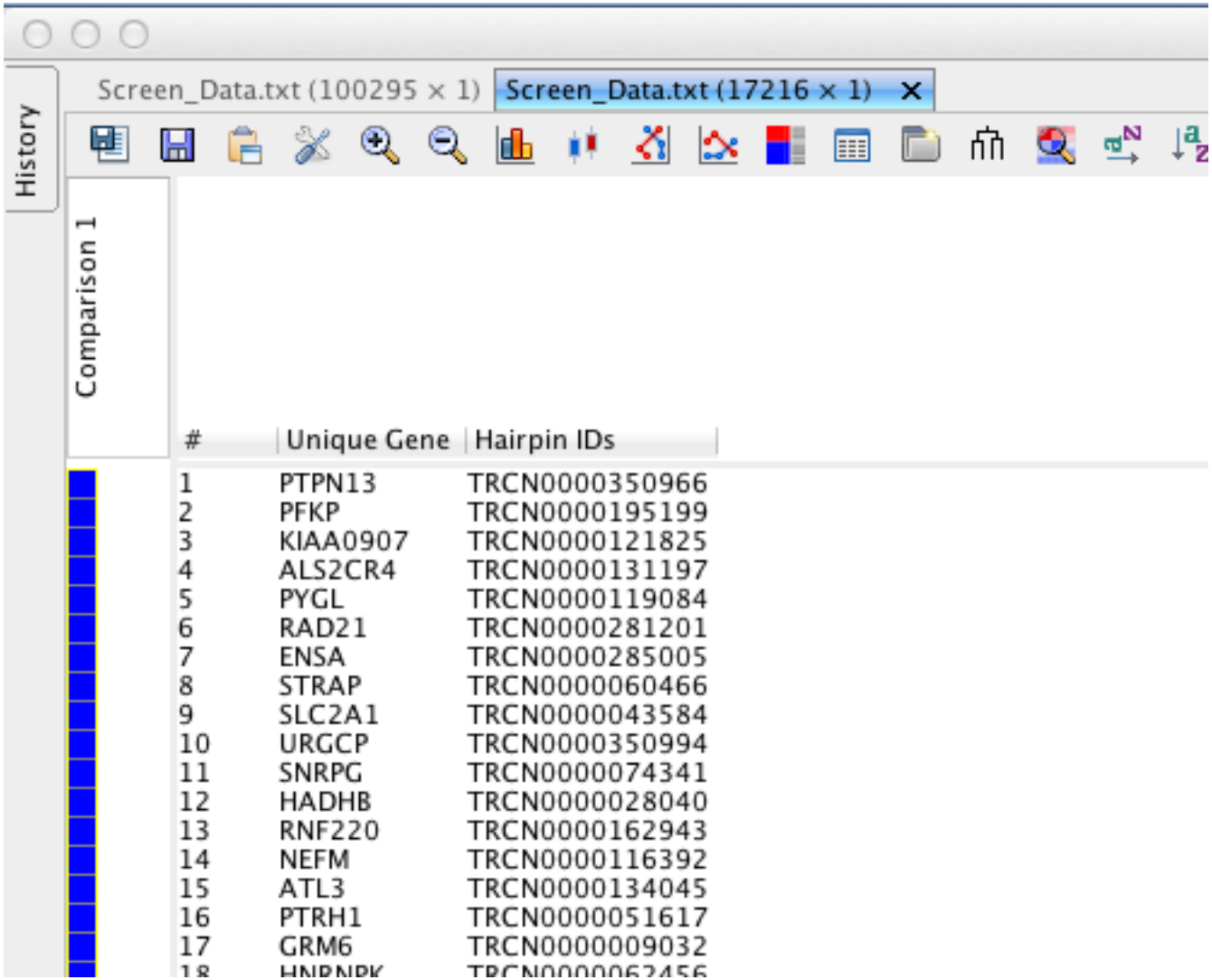

GENE-E Dev
Thu 12:03 AM
John Doench
Import – shRNA_screen_data.txt
Click the table cell containing the first data row and column.
Row metadata
Column metadata
Data matrix
Transpose
Hairpin.Se... Hairpin.IDs Gene.Sym... Score
CCCACAT... TRCN000... F2 -3.34576...
CCAGATG... TRCN000... ISCA2 -3.19547...
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...
GCCAAGT... TRCN000... ENSA -2.65449...
CGTTCAC... TRCN000... DAPK3 -2.63219...
GCATGTT... TRCN000... PSMD9 -2.62115...
CTGAACA... TRCN000... PFKP -2.62049...
CCTAACC... 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
Doench R&D
John's Projects
Rearray Request

Run RIGER from Tools menus



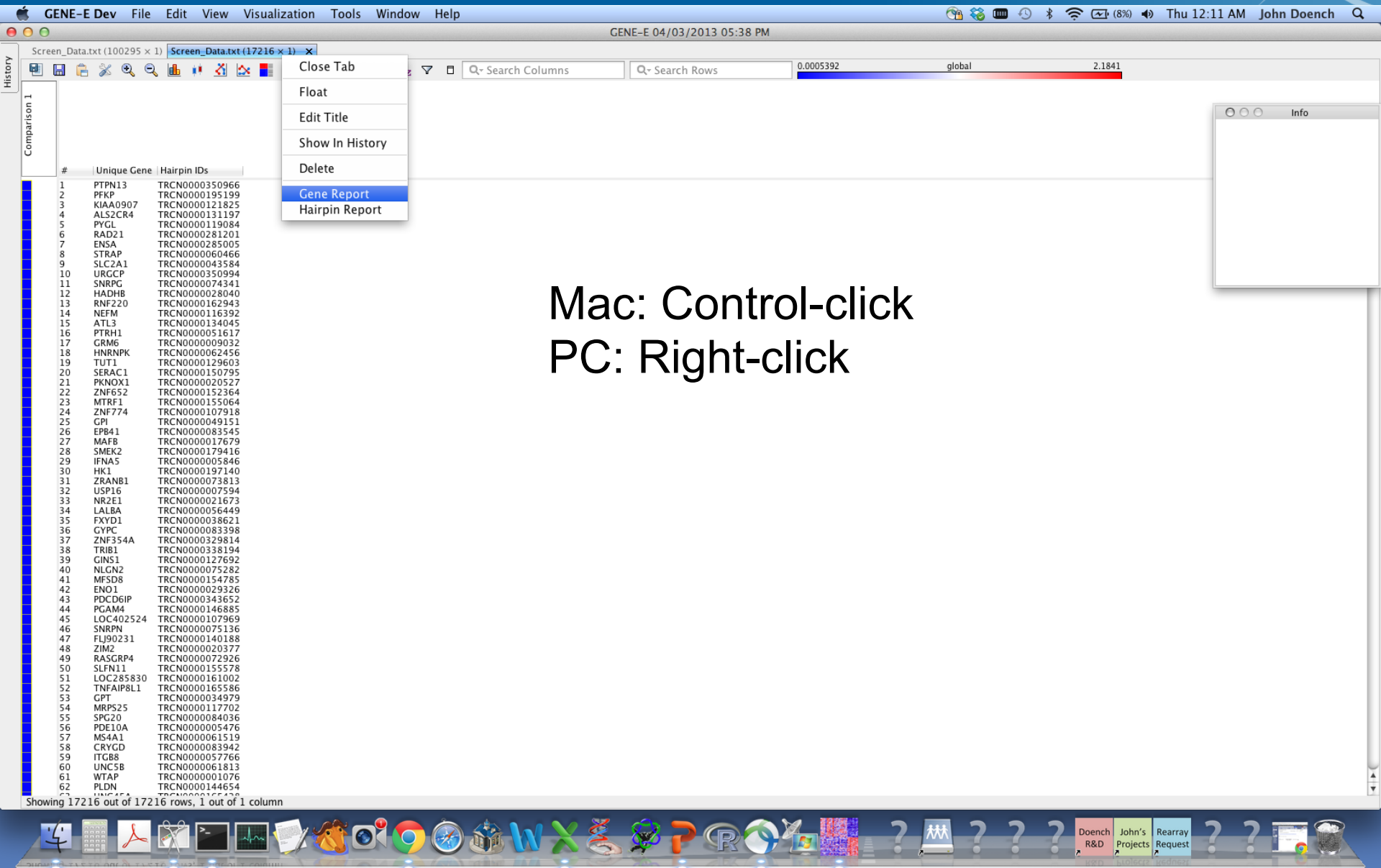

RIGER
Class A
Annotations
Class B
(Select All)
(Select All)
(Select All)
Hairpin.Sequence
Comparison:
Split comparisons by:
Create a separate comparison for each unique value (e.g. split by cell annotation and auto-create separate comparisons for MCF7 and PC3
Metric for ranking hairpins:
Number of permutations:
1000
Method to convert hairpins to genes:
Weighted Sum - weighted sum of first two ranks of hairpins for a gene
Gene rank order:
Negative to positive
Random seed used to generate permutations:
779948224
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:
Hairpin.Sequence
Convert hairpins to:
Gene.Symbol
Cancel
OK
Reset

RIGER output – notice new tab



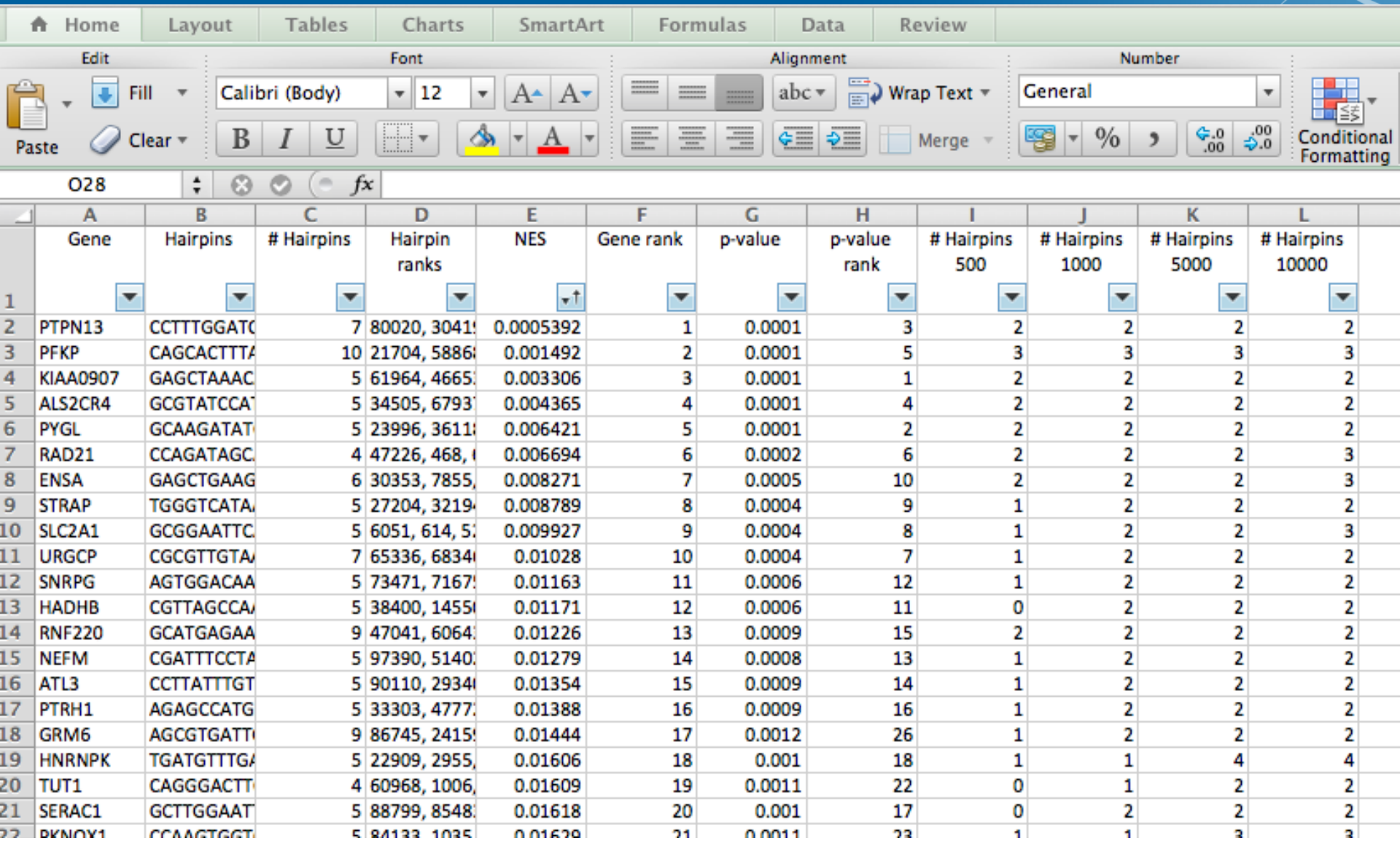
#	Unique Gene	Hairpin IDs
1	PTPN13	TRCN0000350966
2	PFKP	TRCN0000195199
3	KIAA0907	TRCN0000121825
4	ALS2CR4	TRCN0000131197
5	PYGL	TRCN0000119084
6	RAD21	TRCN0000281201
7	ENSA	TRCN0000285005
8	STRAP	TRCN0000060466
9	SLC2A1	TRCN0000043584
10	URGCP	TRCN0000350994
11	SNRPG	TRCN0000074341
12	HADHB	TRCN0000028040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

To download Gene Report



Mac: Control-click
PC: Right-click

Gene report, opened in Excel



Home | Layout | Tables | Charts | SmartArt | Formulas | Data | Review

Edit: Paste, Fill, Clear
Font: Calibri (Body), 12, B, I, U
Alignment: Wrap Text, Merge
Number: General
Conditional Formatting

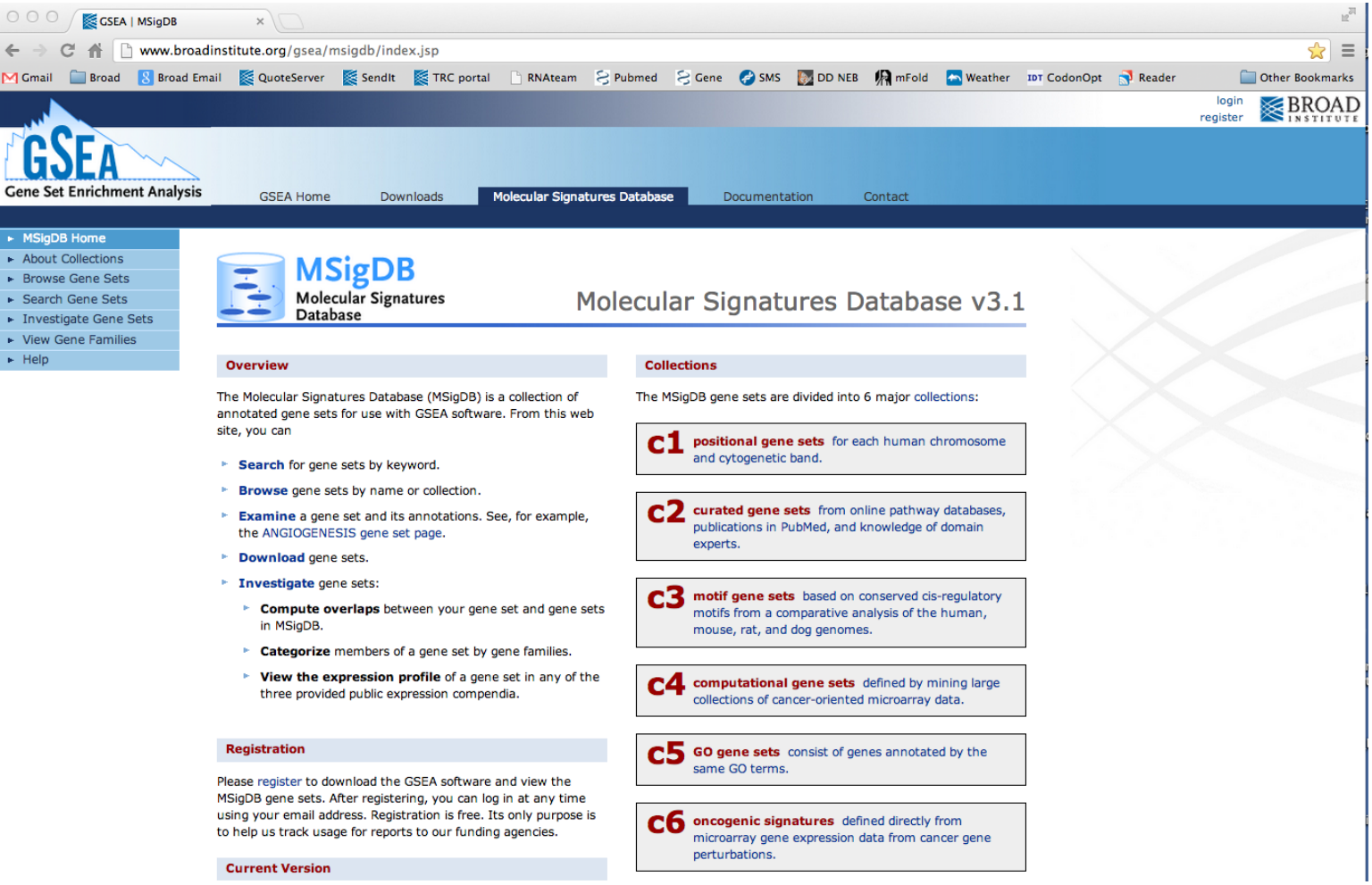
O28

	A	B	C	D	E	F	G	H	I	J	K	L
1	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	CAGCACTTTA	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	CGATTTCCTA	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	PKNOX1	CCAAGTGGT	5	84133, 1035	0.01629	21	0.0011	23	1	1	3	3

Congratulations, now you have another list!

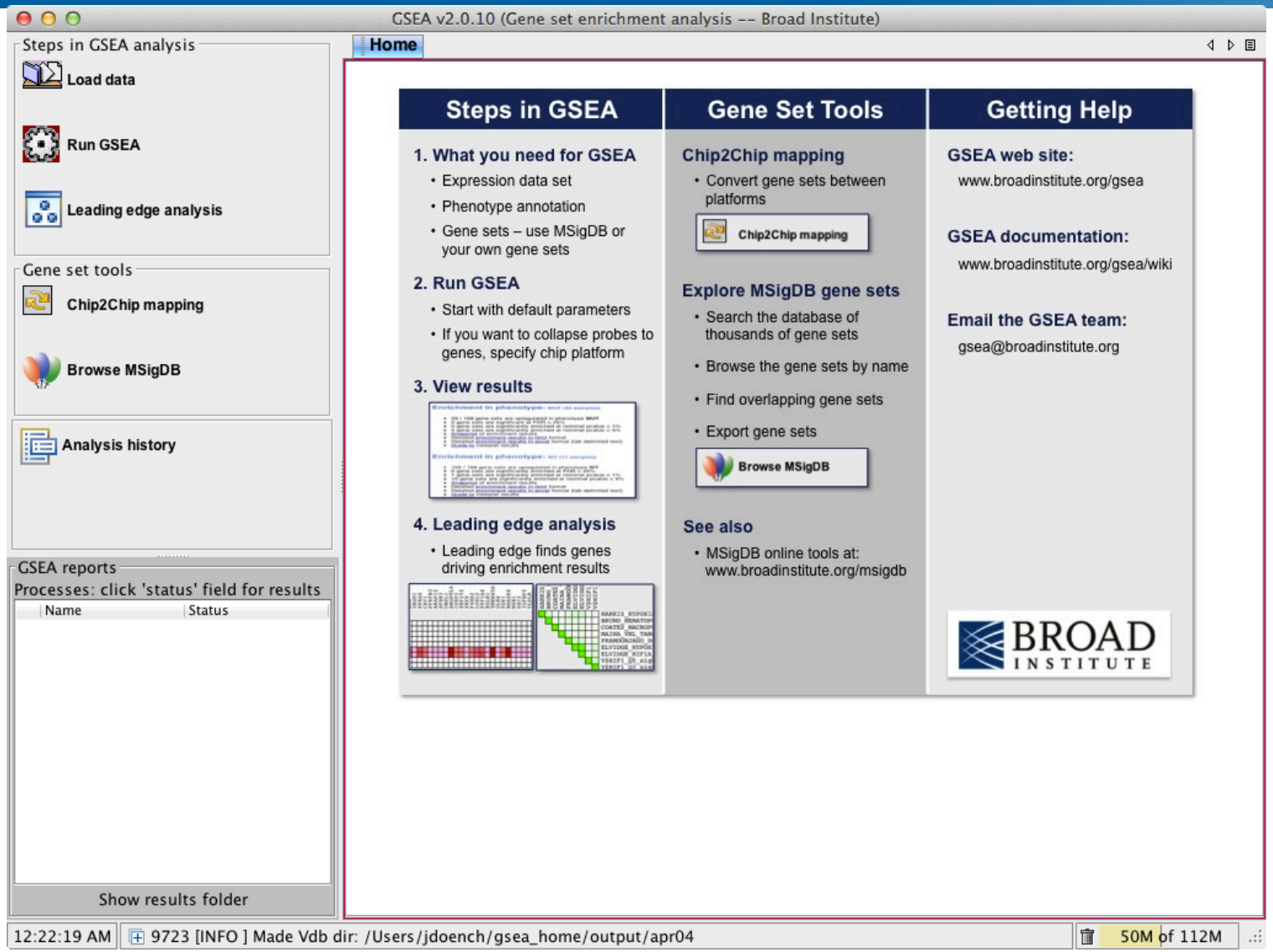
- Unbiased means of determining if your list makes sense?
- GSEA – Gene Set Enrichment Analysis
  - Originally used to analyze microarray data
  - MSigDB is a collection of curated gene sets, curated from pre-existing data
- DAPPLE – Protein-protein interaction data

GSEA & MSigDB



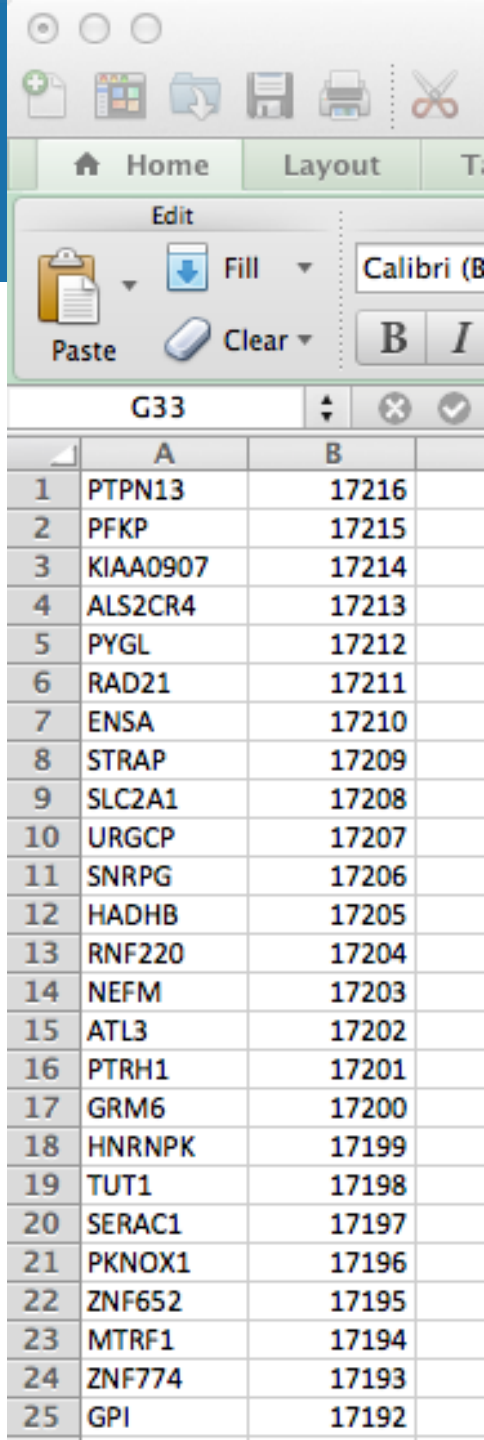
http://www.broadinstitute.org/gsea/msigdb/index.jsp

Launch Java version



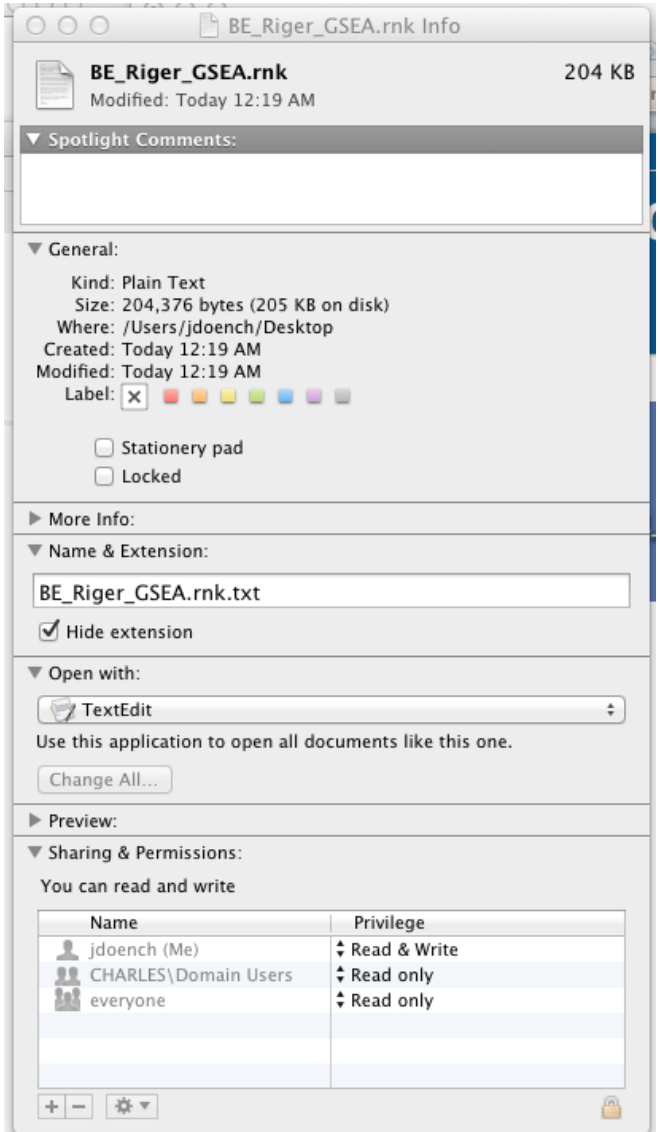
Requirements

- 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

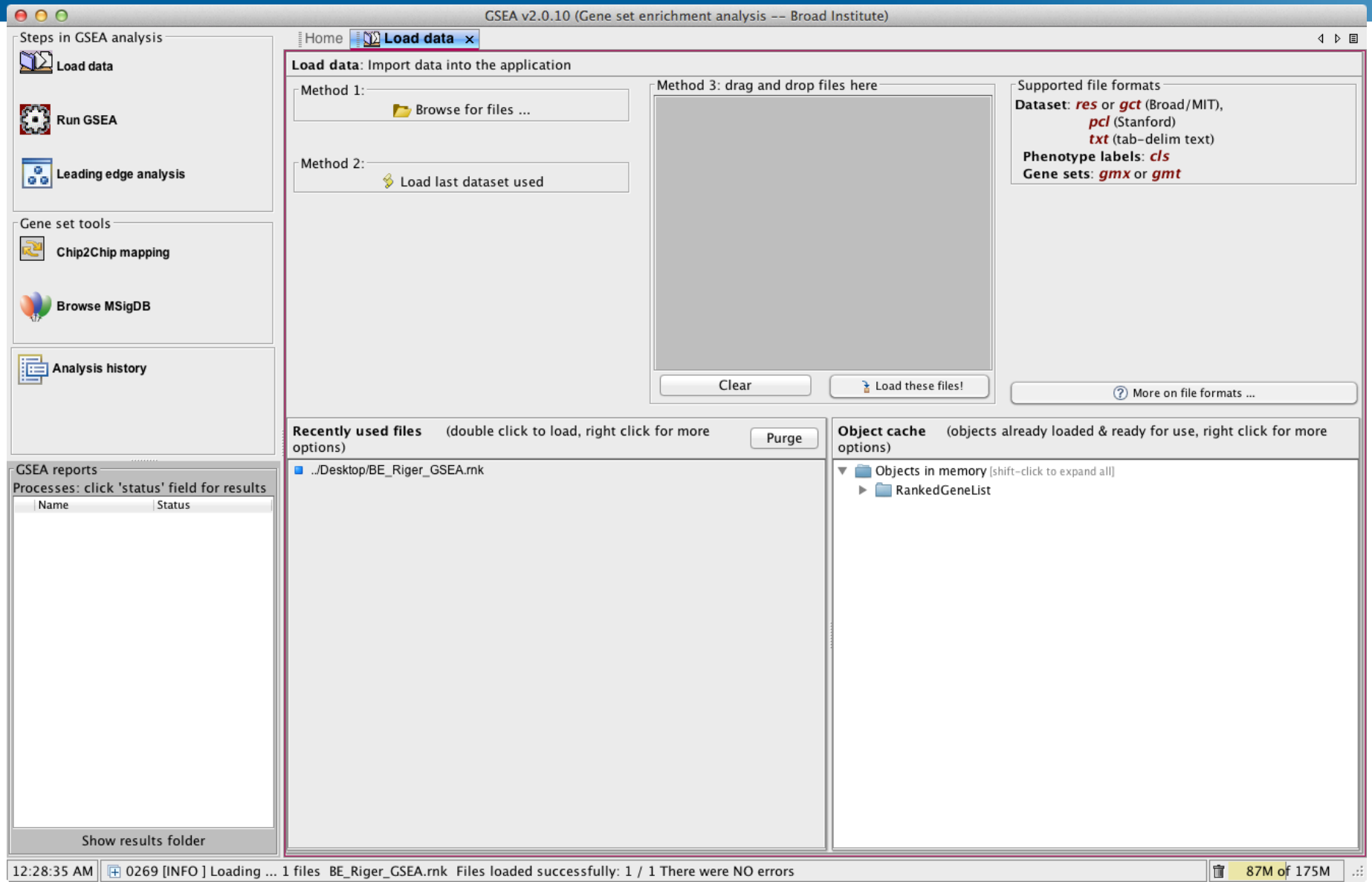


	A	B
1	PTPN13	17216
2	PFKP	17215
3	KIAA0907	17214
4	ALS2CR4	17213
5	PYGL	17212
6	RAD21	17211
7	ENSA	17210
8	STRAP	17209
9	SLC2A1	17208
10	URGCP	17207
11	SNRPG	17206
12	HADHB	17205
13	RNF220	17204
14	NEFM	17203
15	ATL3	17202
16	PTRH1	17201
17	GRM6	17200
18	HNRNPK	17199
19	TUT1	17198
20	SERAC1	17197
21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

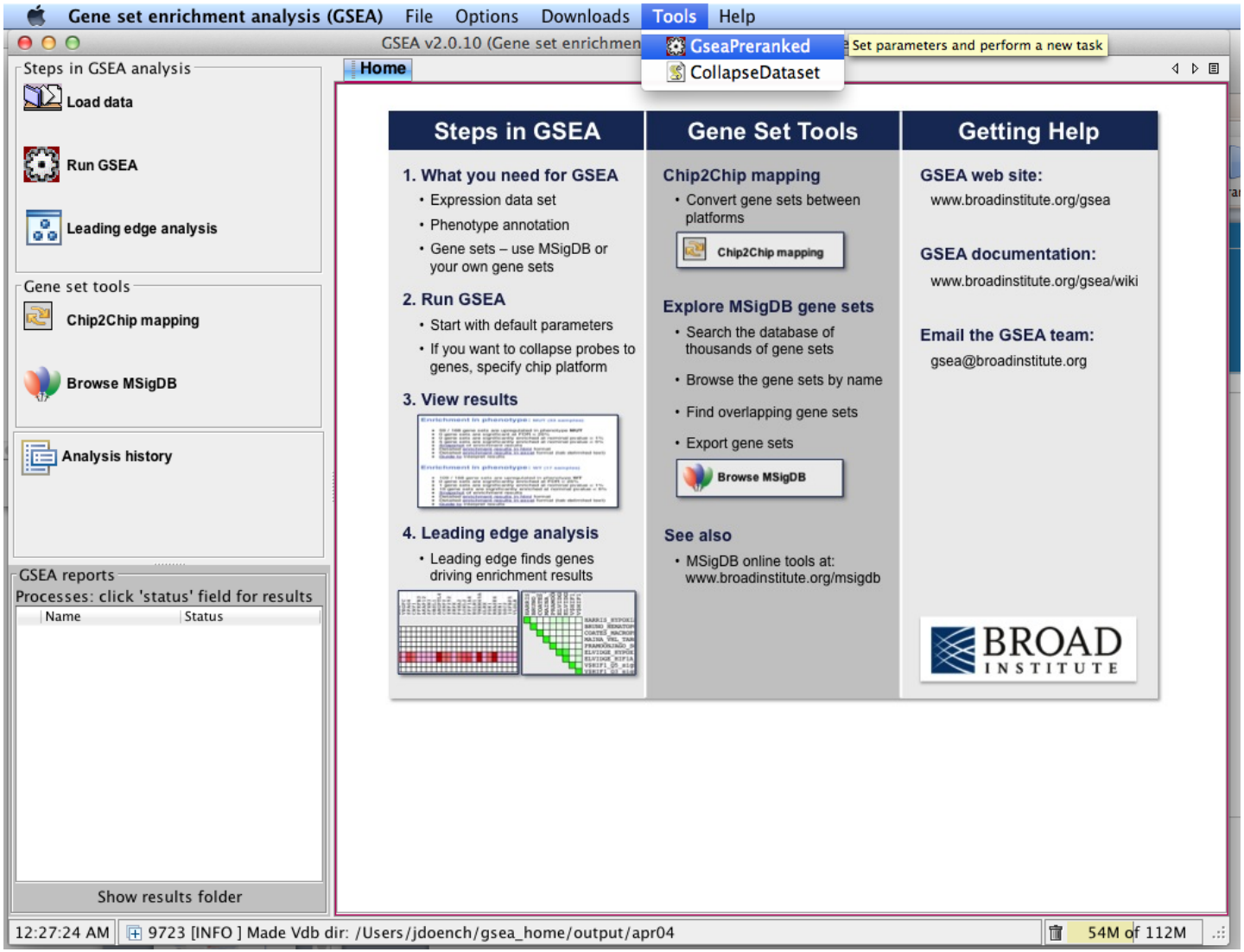
Get your extensions right!



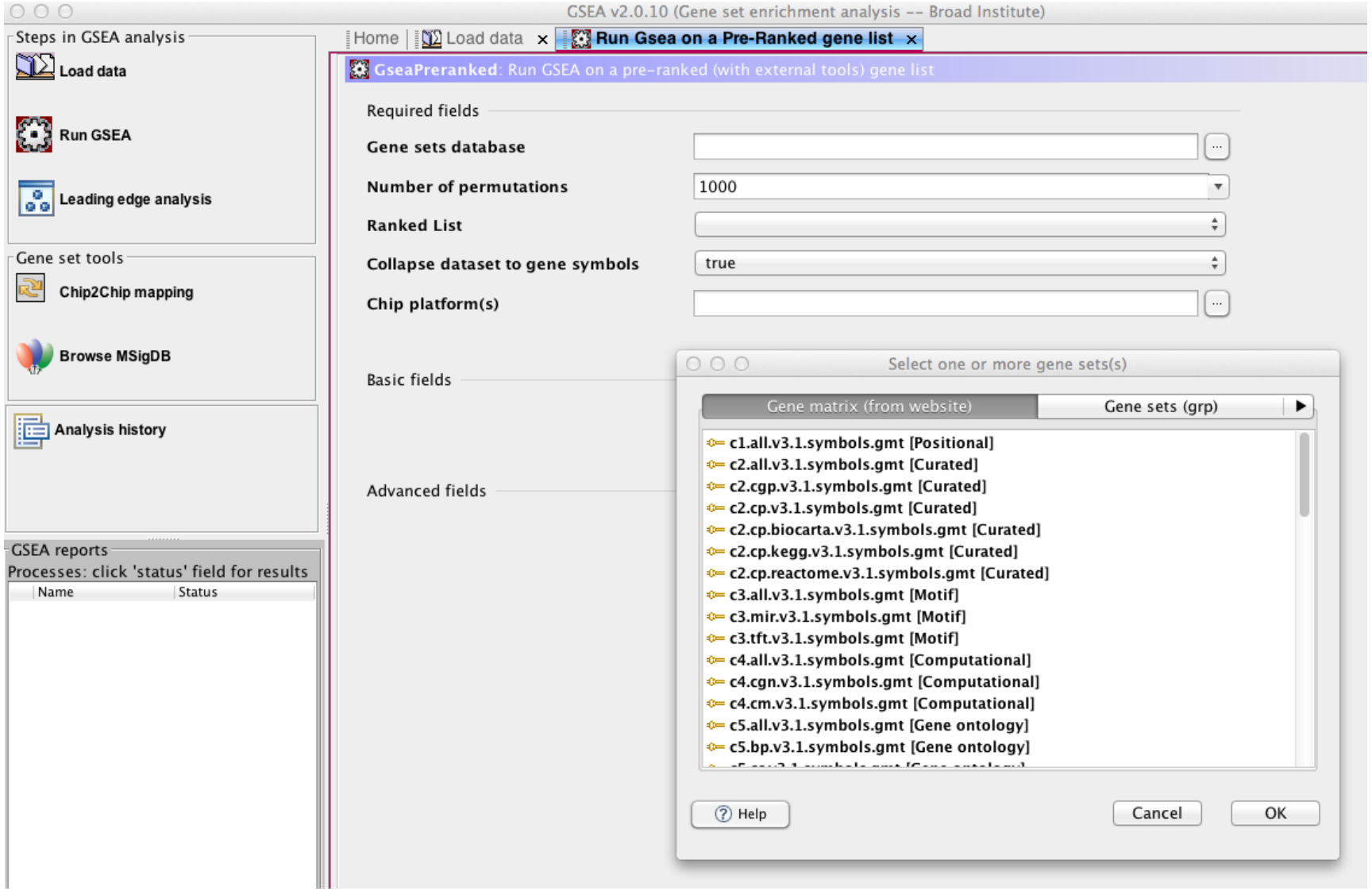
Load your file (.rnk)



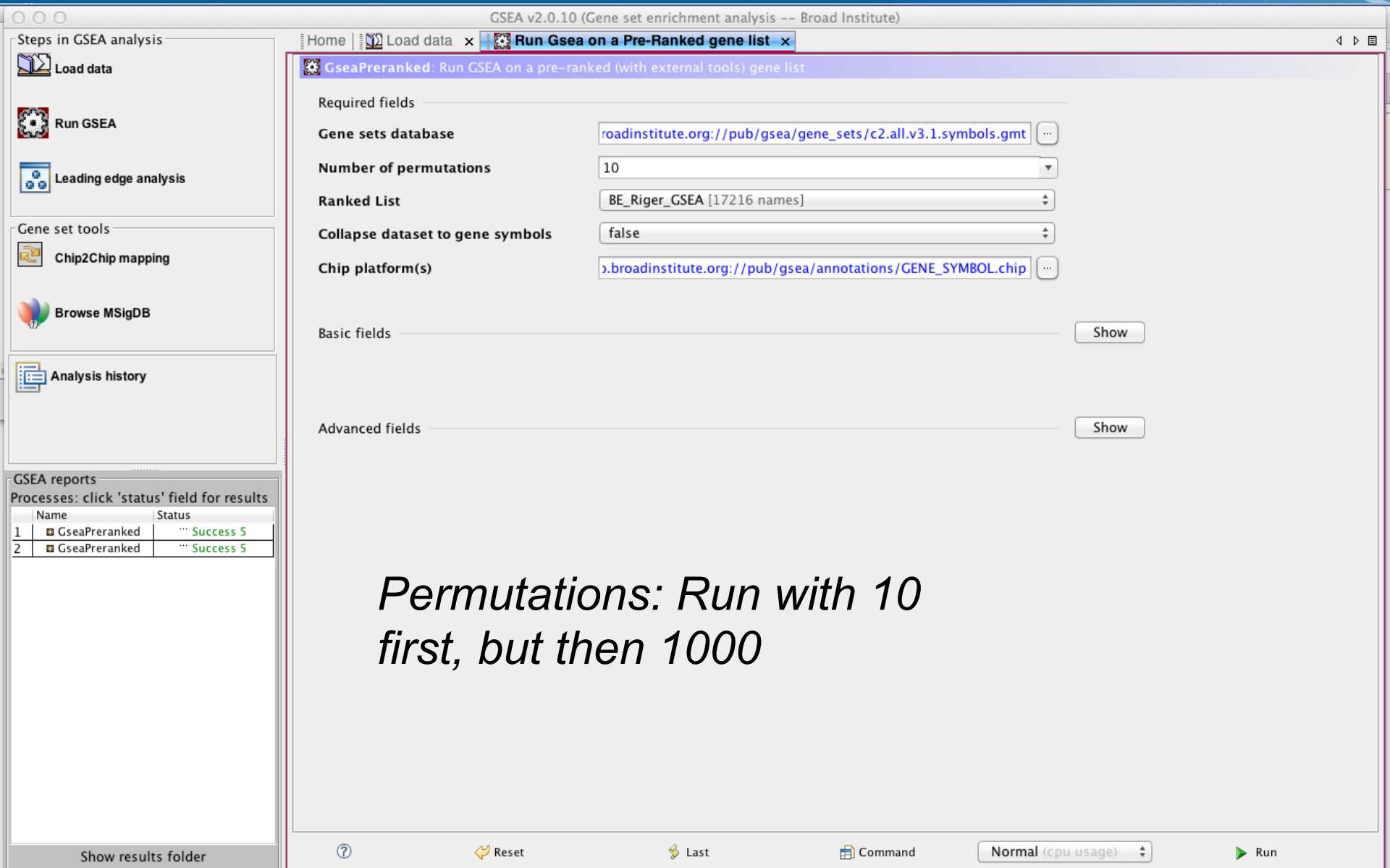
Run GSEA PreRanked



Selection of Gene Sets



Settings



Permutations: Run with 10 first, but then 1000

Output, opens in browser

GSEA Report for Dataset VW_GSEA_in

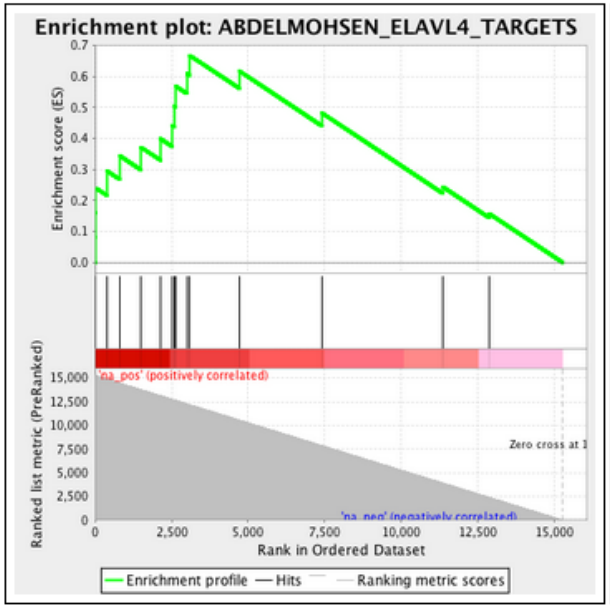
Enrichment in phenotype: na

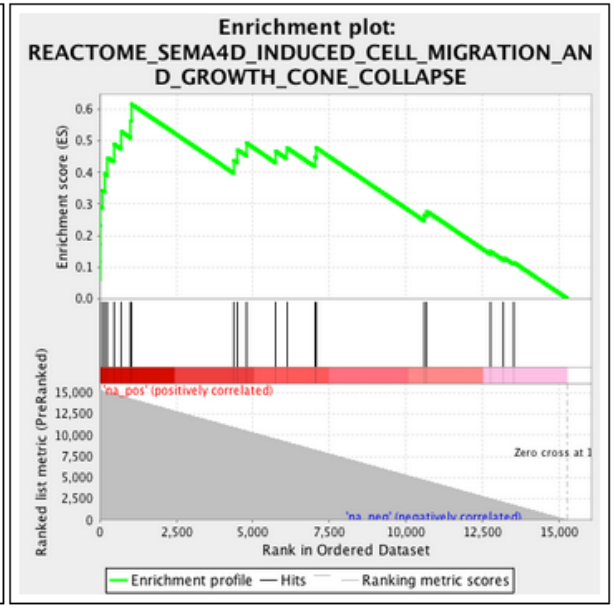
- 3656 / 3687 gene sets are upregulated in phenotype **na_pos**
- 37 gene sets are significant at FDR < 25%
- 413 gene sets are significantly enriched at nominal pvalue < 1%
- 413 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results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

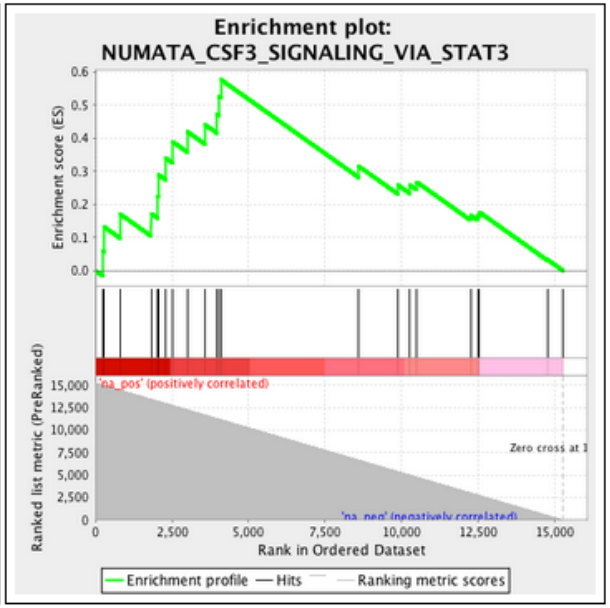
Enrichment in phenotype: na

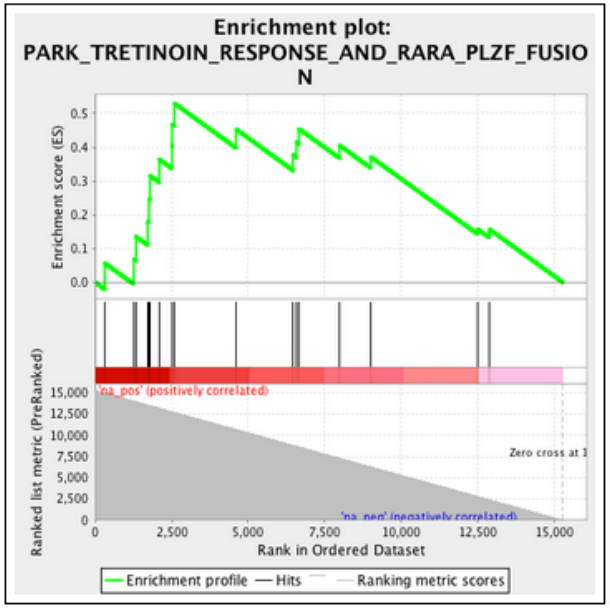
- 31 / 3687 gene sets are upregulated in phenotype **na_neg**
- 16 gene sets are significantly enriched at FDR < 25%
- 2 gene sets are significantly enriched at nominal pvalue < 1%
- 2 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results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

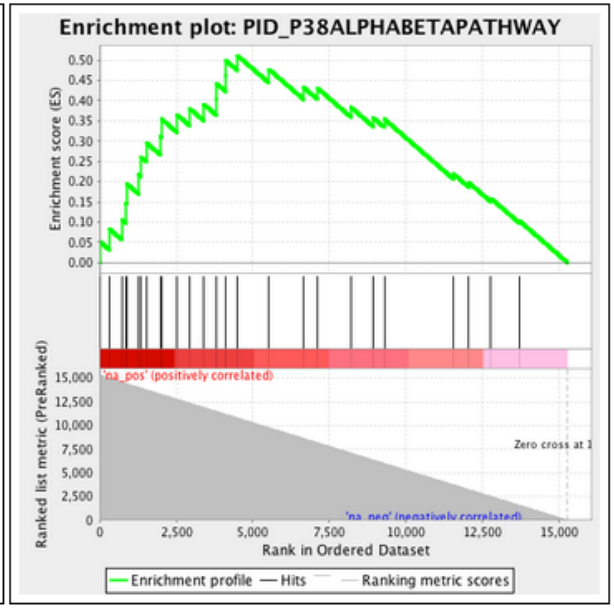
Table: Snapshot of enrichment results

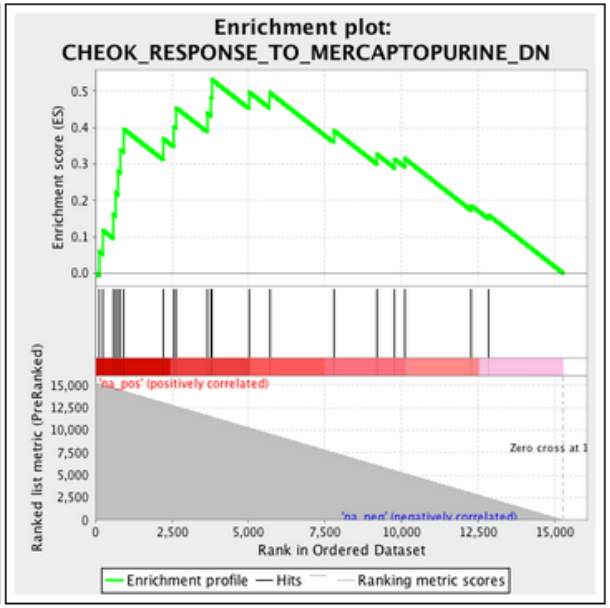












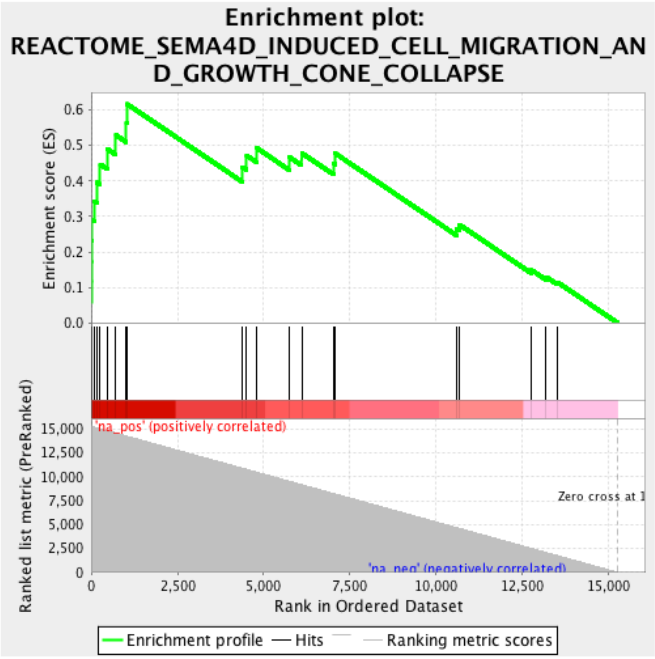


Fig 1: Enrichment plot: REACTOME_SEMA4D_INDUCED_CELL_MIGRATION_AND_GROWTH_CONE_COLLAPSE
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

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

DAPPLE

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

that size here, in kb.

50 kb downstream of Tx End

Plot:
Return a picture of your network.

☑ Plot ☑ Color by DAPPLE p-value
☐ Simplify Indirect Network?

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

☐ Iterate

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

☐ Use Nearest Gene

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

Choose File No file chosen

MYL12B
MYH9
ARHGEF12
ZAK
INSL6
CDC42

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

Choose File No file chosen

Your email address (required):

jdoench@broadinstitute.

Description (required):
Warning: All non-alphanumeric characters and spaces will be removed

BroadE

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

submit reset

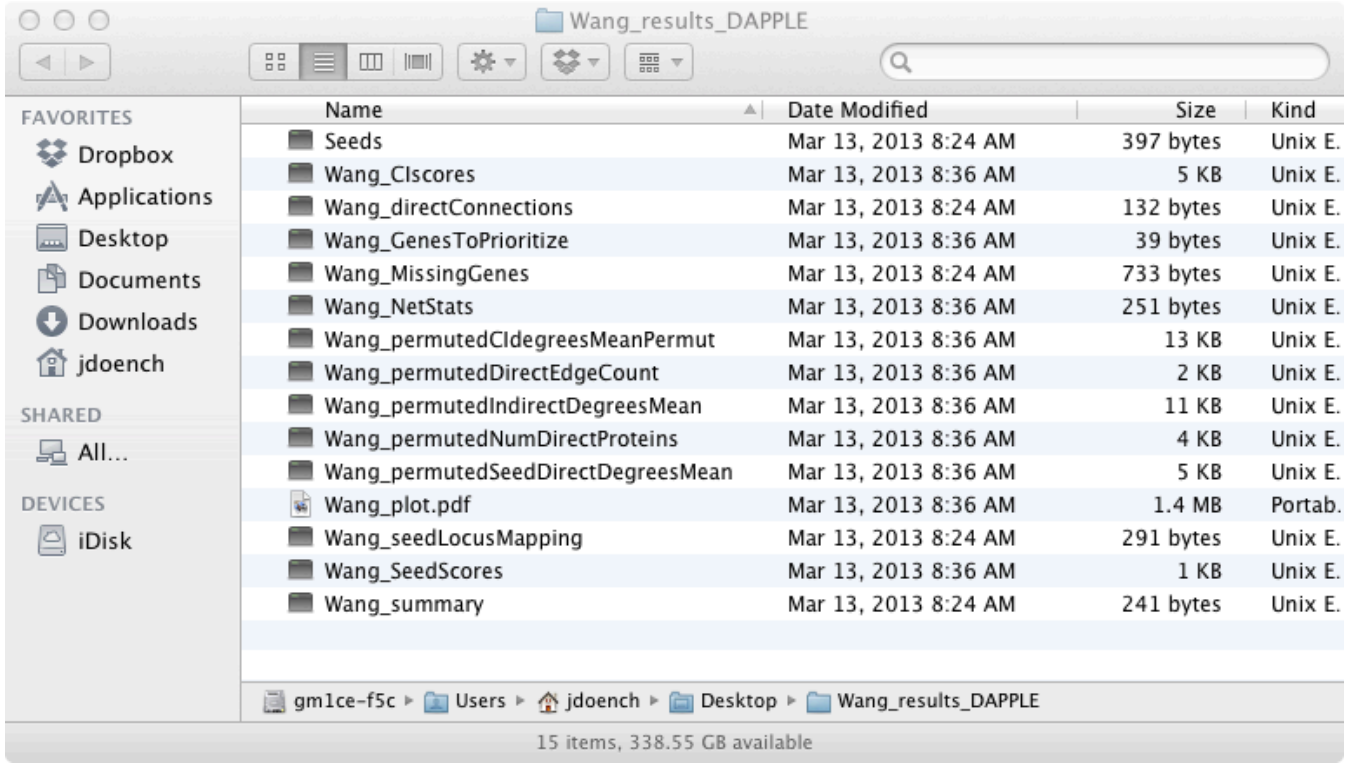
QUESTIONS? Please refer to the FAQ page before emailing dapple@broadinstitute.org with questions.

DAPPLE download when done running

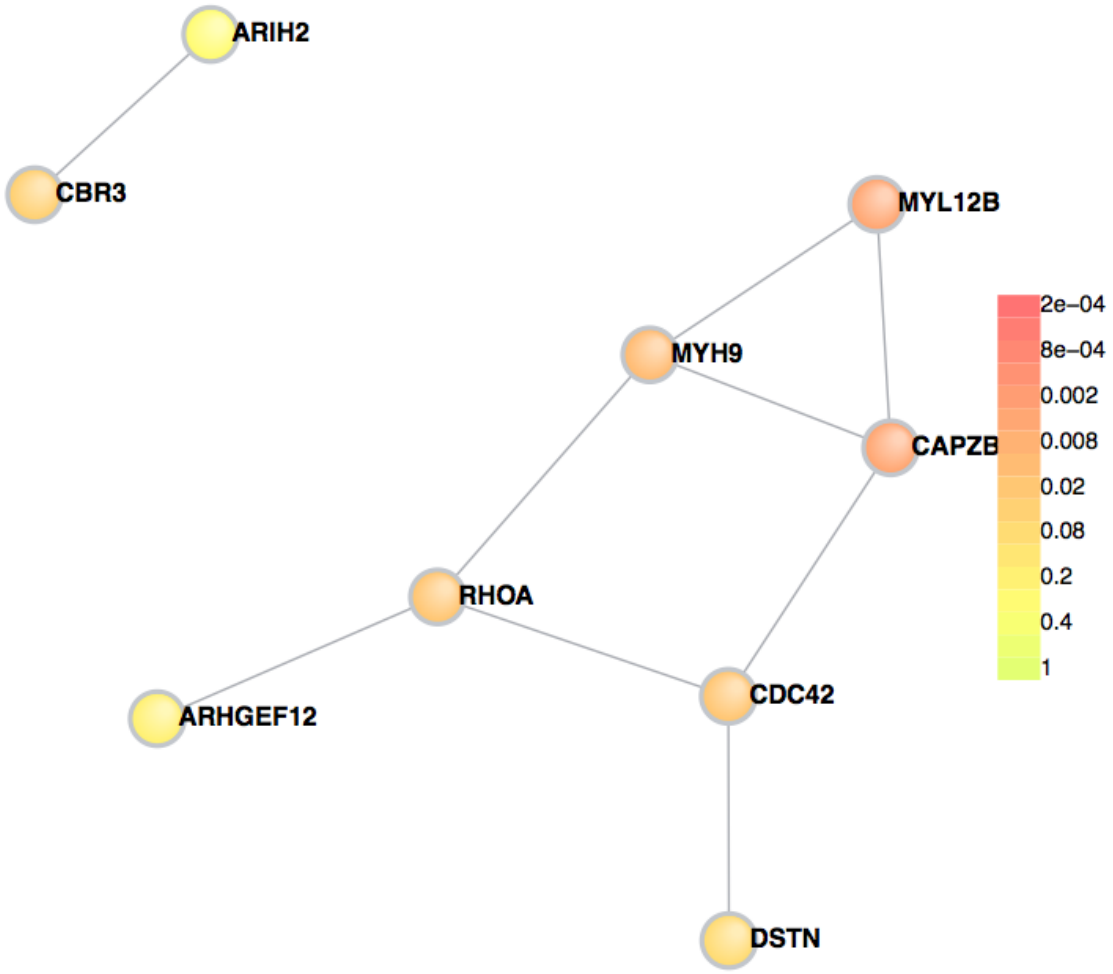
www.broadinstitute.org/mpg/dapple/dapple.php#

Gmail Broad Broad Email QuoteServer SendIt TRC portal RNAteam Pubmed Gene SMS DD NEB mFold Weather CodonOpt Reader

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

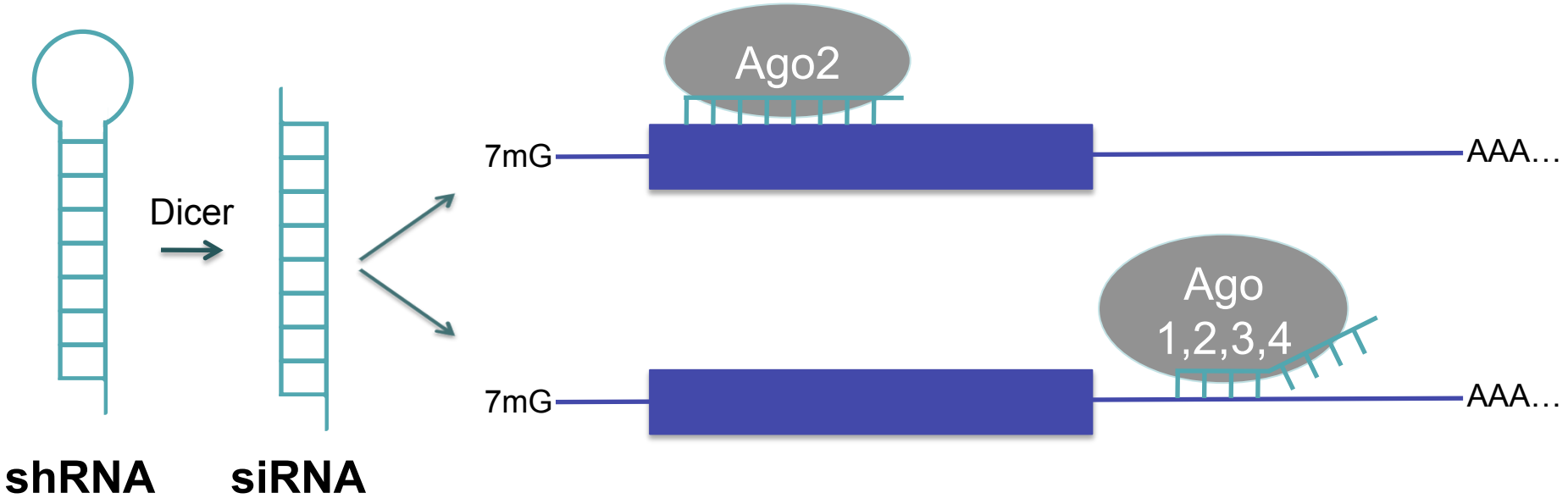
- Input is simple: siRNA/shRNA sequence and the numerical output of the screen
- Search for both
  - Enrichment of seeds to match to known microRNA sequences
  - Enrichment of UTRs preferentially targeted by multiple seeds

micro**R**NA **k**nockdown
analysis **t**oolkit*

Two pathways for shRNA library

RNAi

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

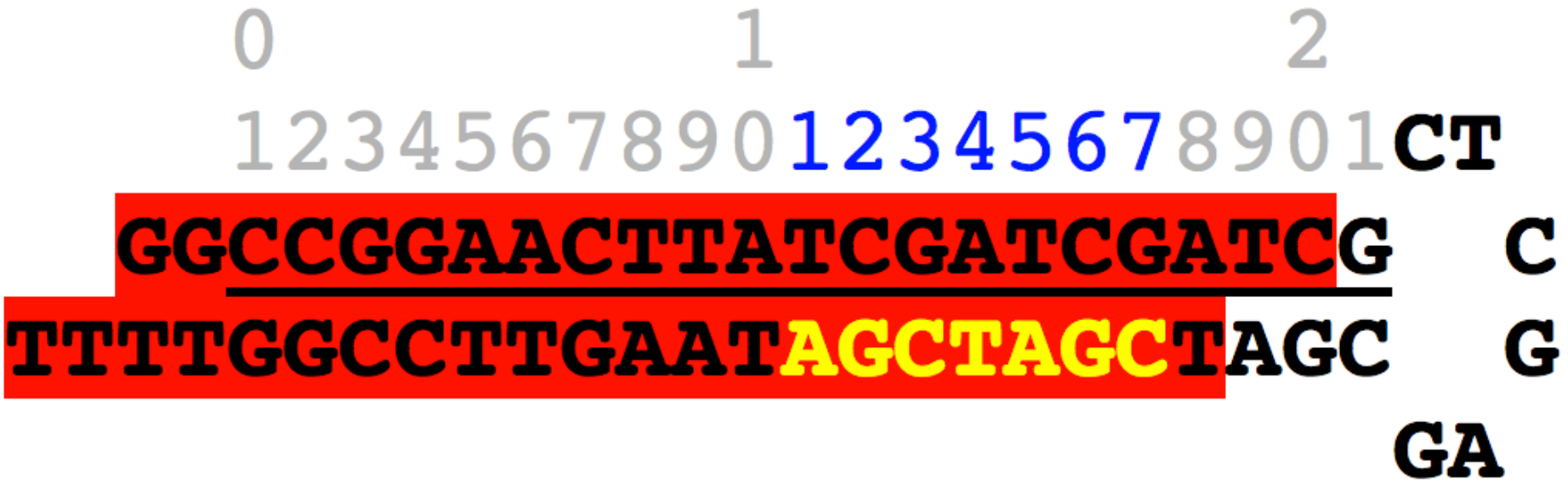


microRNA

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

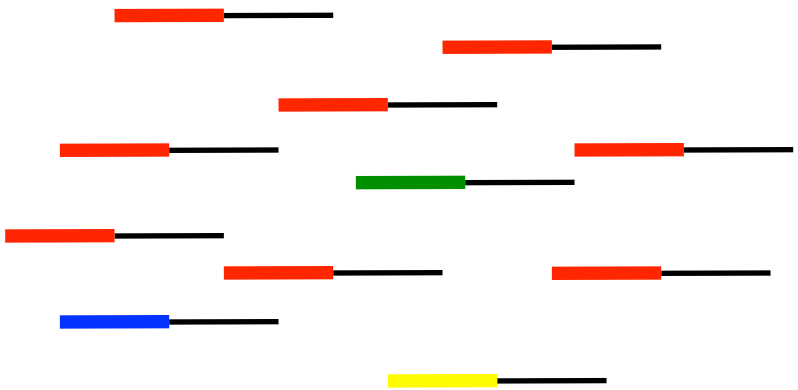
LKO shRNA anatomy

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

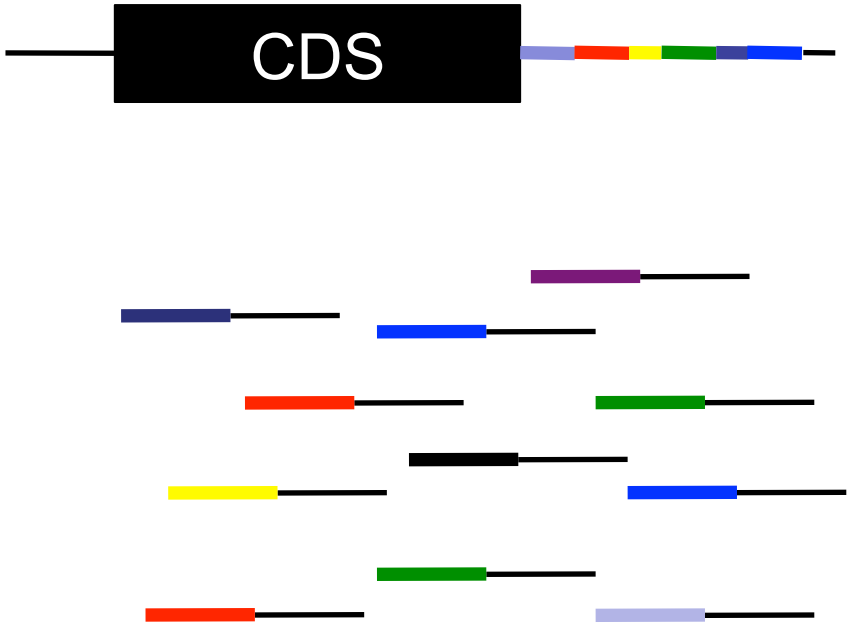


Two types of microRNA effects to search for in hit shRNAs

Hit shRNAs:

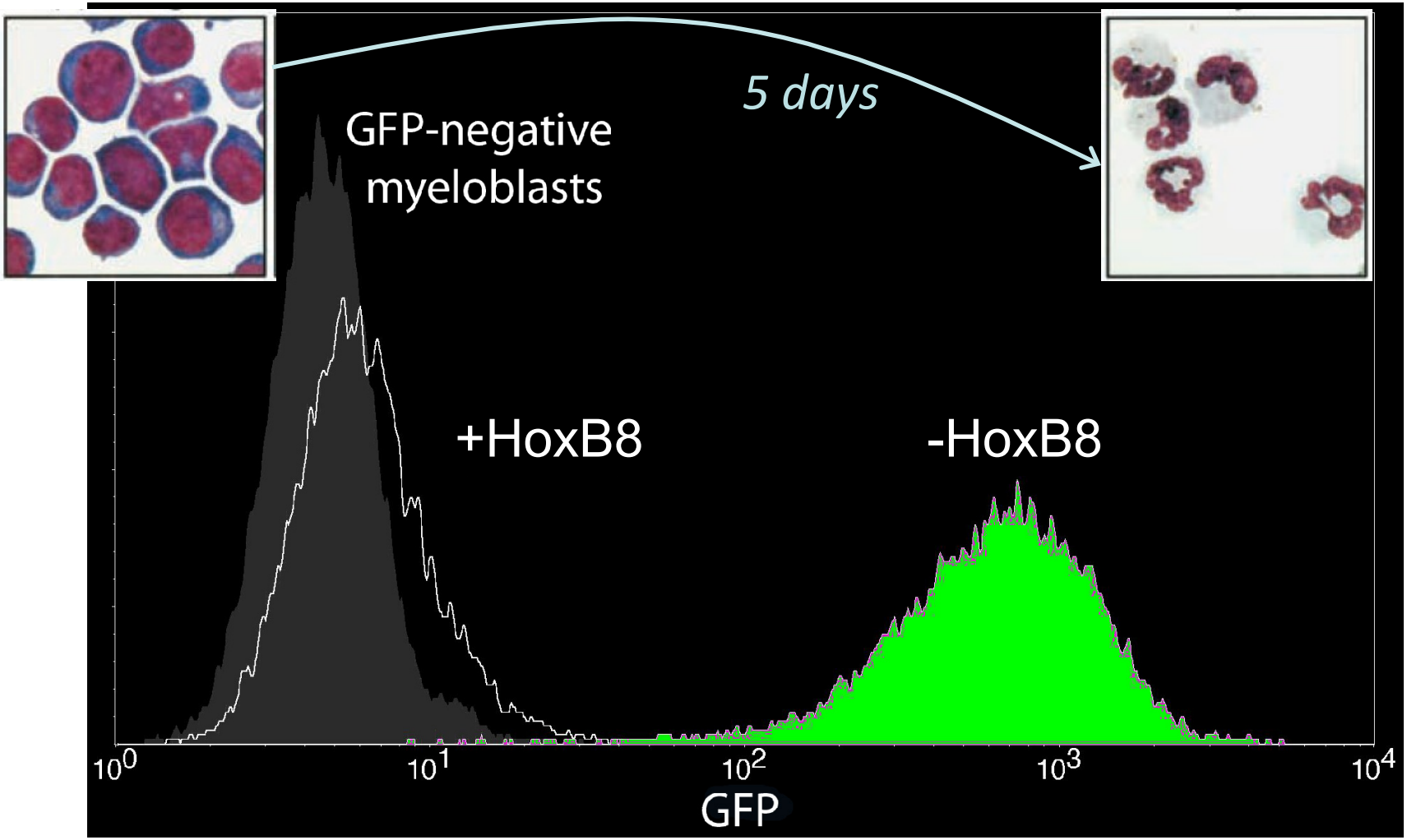


- Enrichment of seeds



- Enrichment of UTRs

AML model: HoxB8-induced differentiation block

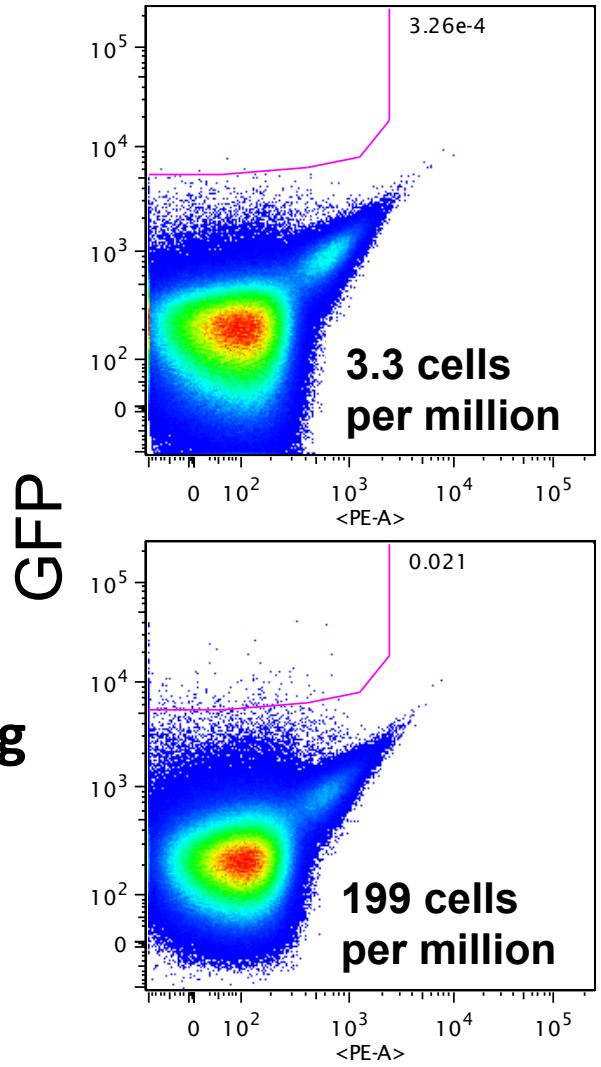


David Sykes, Scadden Lab

Pooled screen with flow cytometry selection of hits

Background: Introduce negative control shRNA

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



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	GUCUCUCUUACUGGUAGGUUUU	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**	**ACUGCUGGGAAACUUGUCUUUU**	**22.6**	**593.7**	**26**

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

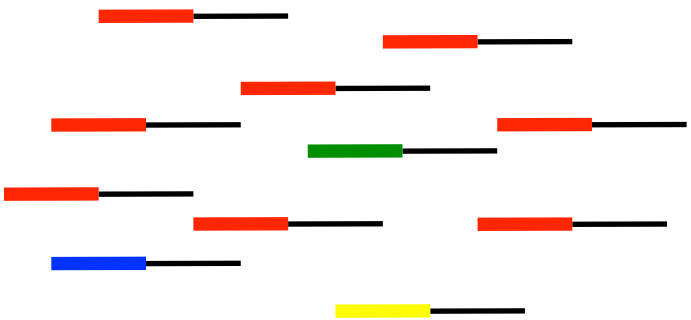
1% of shRNAs defined as hits

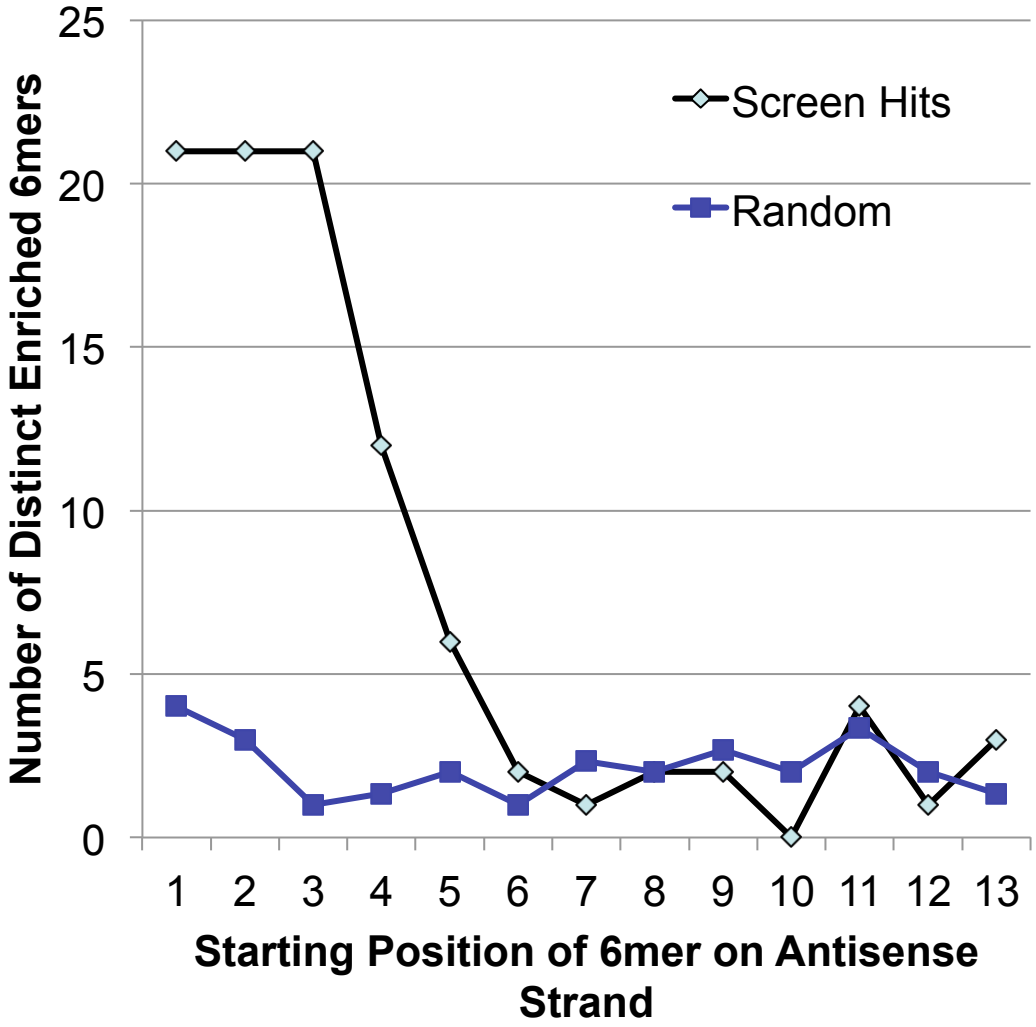
Rank	Symbol	Sequence	Parts per million		Fold Enrichment
			Unsorted	GFP+	
1	Hoxb8	UAGCCGUAGAAGUUGCCGUUUU	0.1	2185.1	**36490**
2	Traf5	AAU**UCUCUC**AGAGACCGGUUUU	1.1	2118.8	**1981**
3	LOC434093	GUGUUGACUAUACAGCCGUUUU	1.0	476.7	**482**
4	1810035L1	GU**UCUCUC**AGCUCACUCGUUUU	1.2	548.2	**444**
5	LOC381842	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

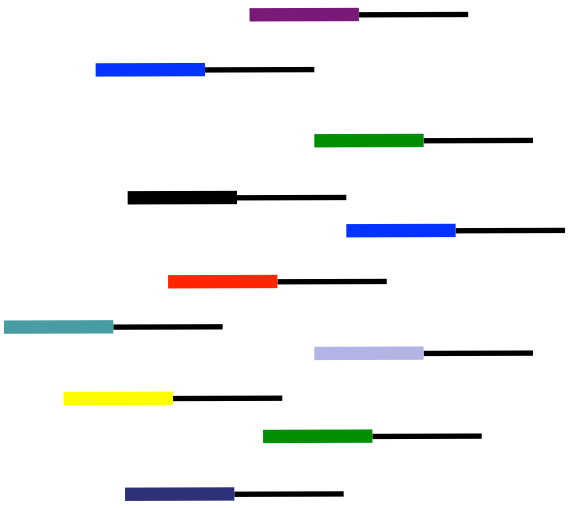
mir-130 CAGUGCAAUAGUAUUGUCAAAGC

mir-301 CAGUGCAAUGUUAAAAGGGCAU

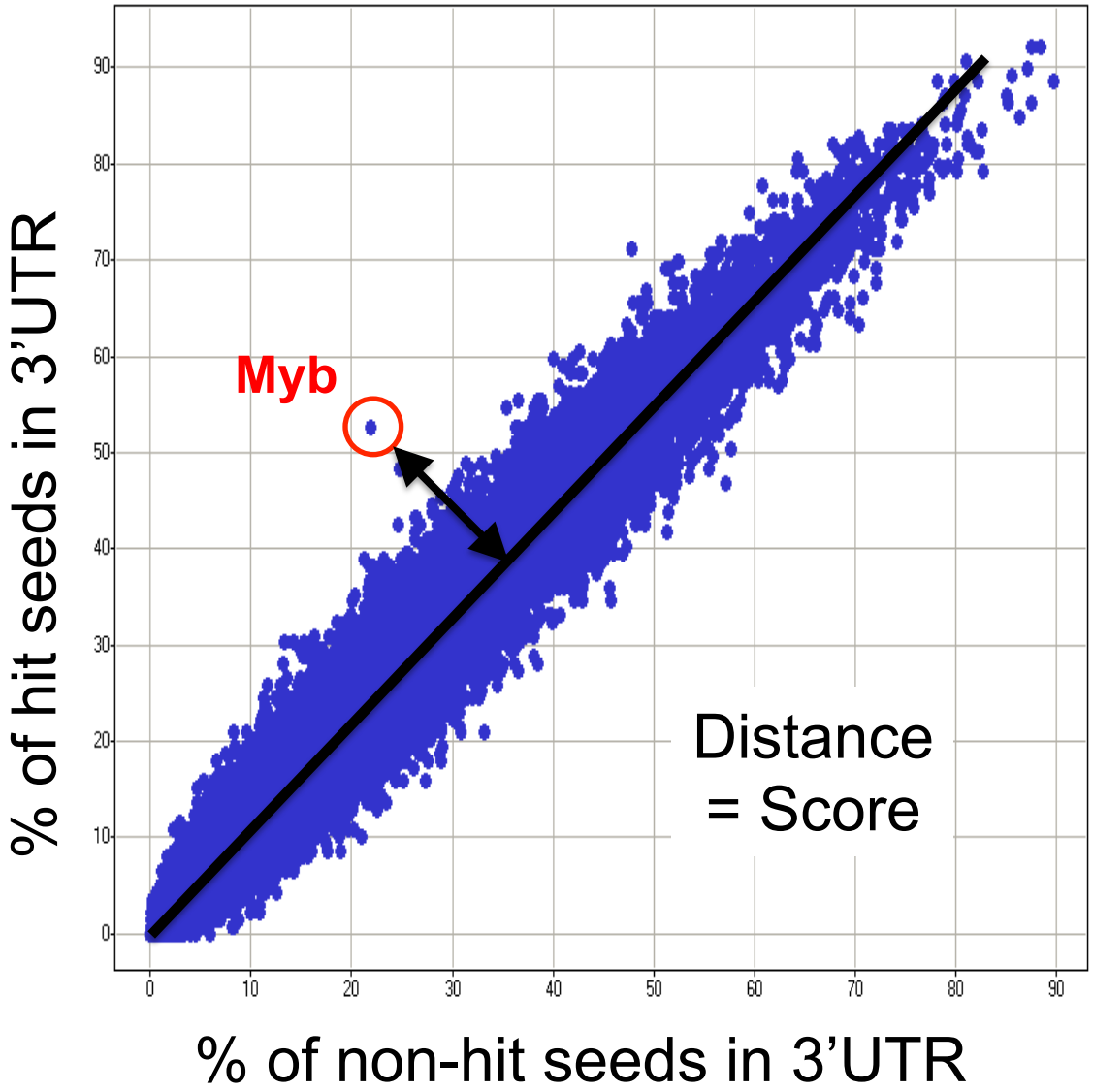
Discover microRNAs involved in biological phenotype

Enrichment of UTRs

Seeds from hits:



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,*]
[1]The CBR Institute for Biomedical Research, Harvard Medical School, Boston, MA 02115, USA
[2]Center for Comparative Functional Genomics, Department of Biology, New York University, New York, NY 10003, USA
[3]Department of Microbiology, University of Virginia Health System, Charlottesville, VA 22908, USA
[4]These authors contributed equally to this work.
[5]Present 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

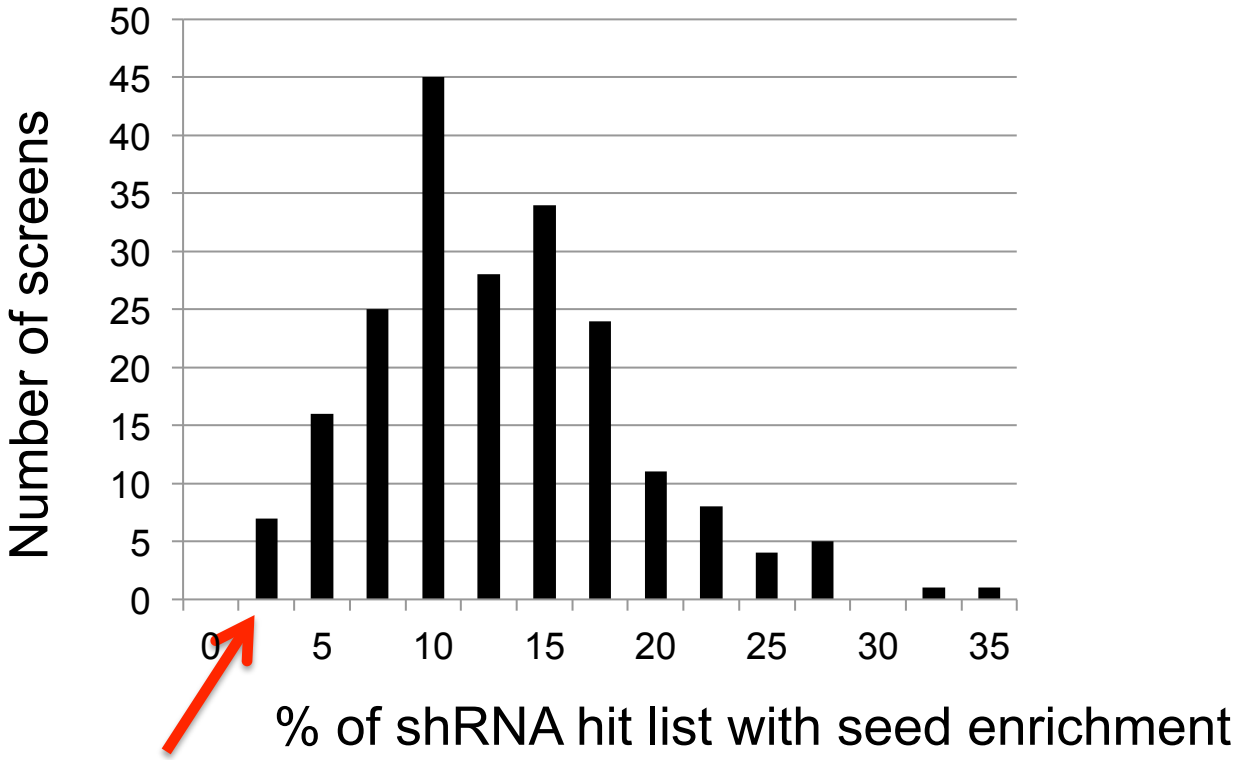
Cell lines with known cancer driver mutations depend on those drivers

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?

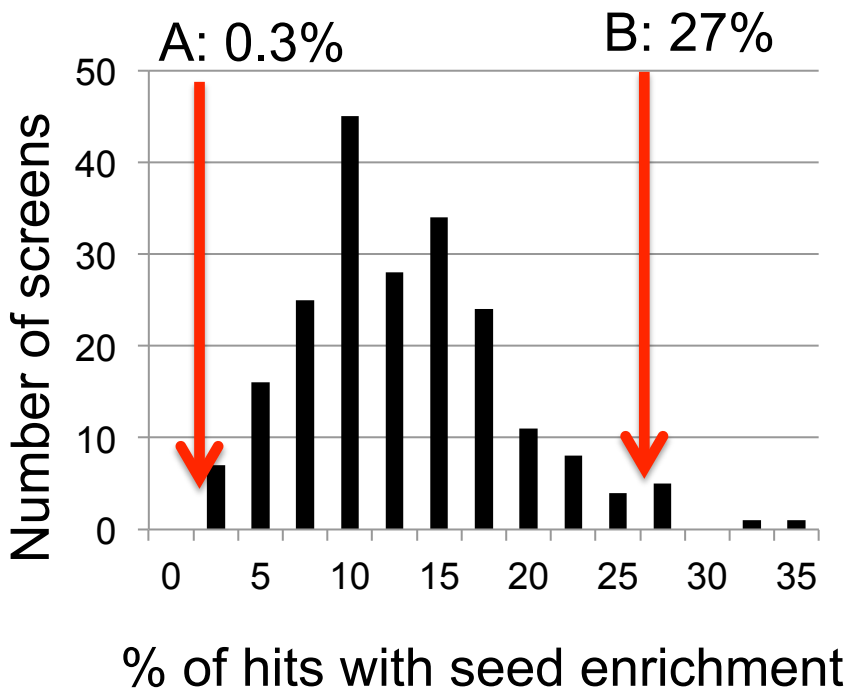
209 proliferation screens
Typical screen: ~10% of hits clearly from miRNA seed effects



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