ATARiS

Erkin Otles

CS 838

Overview

- RNAi
- ATARiS Method
- Proof of Concept

RNA Interference

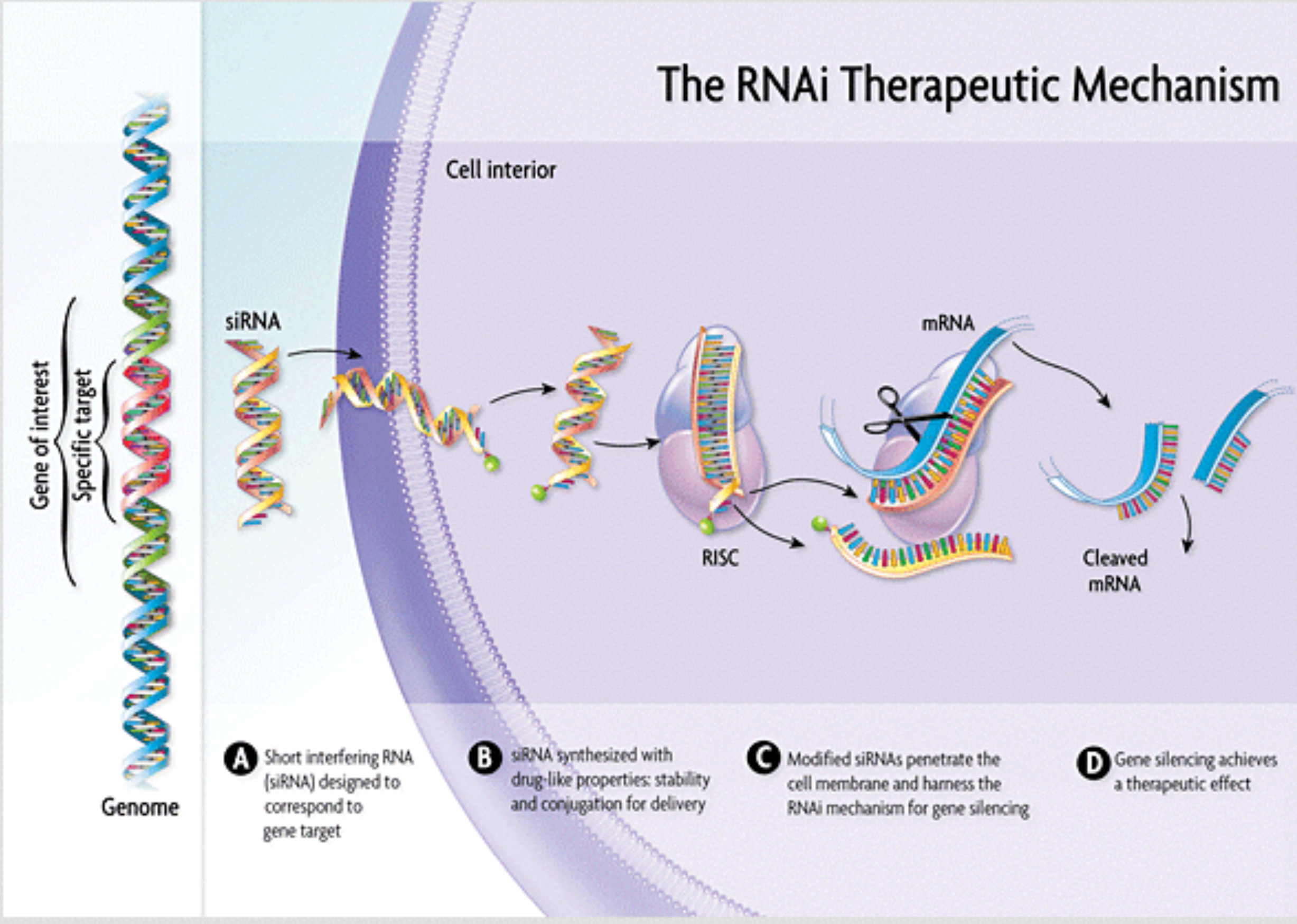
Central Dogma:

DNA → mRNA → Protein

RNAi

DNA → small RNAs (siRNA, miRNAs)

The RNAi Therapeutic Mechanism



nucleus

1 A protein called **exportin-5** transports a hairpin **primary microRNA** (pri-miRNA) out of the nucleus.

Hairpin pri-microRNA
(primary micro RNA)

Exportin-5

2½ Meanwhile, one of the strands joins a group of proteins, forming an **microRNA-protein complex**. The other strand, known as a **passenger strand** is usually discarded.
How this all happens is still not very well understood.

Protein

2 An enzyme called **dicer** (not shown) trims the pri-microRNA and removes the hairpin loop, leaving a double stranded microRNA duplex molecule.

Argonaute proteins

Base mismatch

microRNA duplex

Passenger strand

4 In animal cells, the microRNA nucleotides typically don't pair up with the mRNA nucleotides as well. Their base pairing often follows a pattern though.

microRNA

Nucleotide 1
Has an A across from it

Deadenylation

3' Poly-A tail

Base mismatches

Seed Region (Nucleotides 2–8)
Perfect base pairing

Nucleotide 9
Has an A or U across from it

Nucleotides 13–16
Good base pairing

3' UTR

3 In plant cells, the microRNA is usually perfectly complementary to its target mRNA molecule. The microRNA will bond with it, and cause the mRNA to break down.

microRNA-protein complex

microRNA
(abbreviated miRNA)
about 22 nucleotides long

5'

3' untranslated region
(3' UTR)

3' Poly-A tail

3'

Coding region

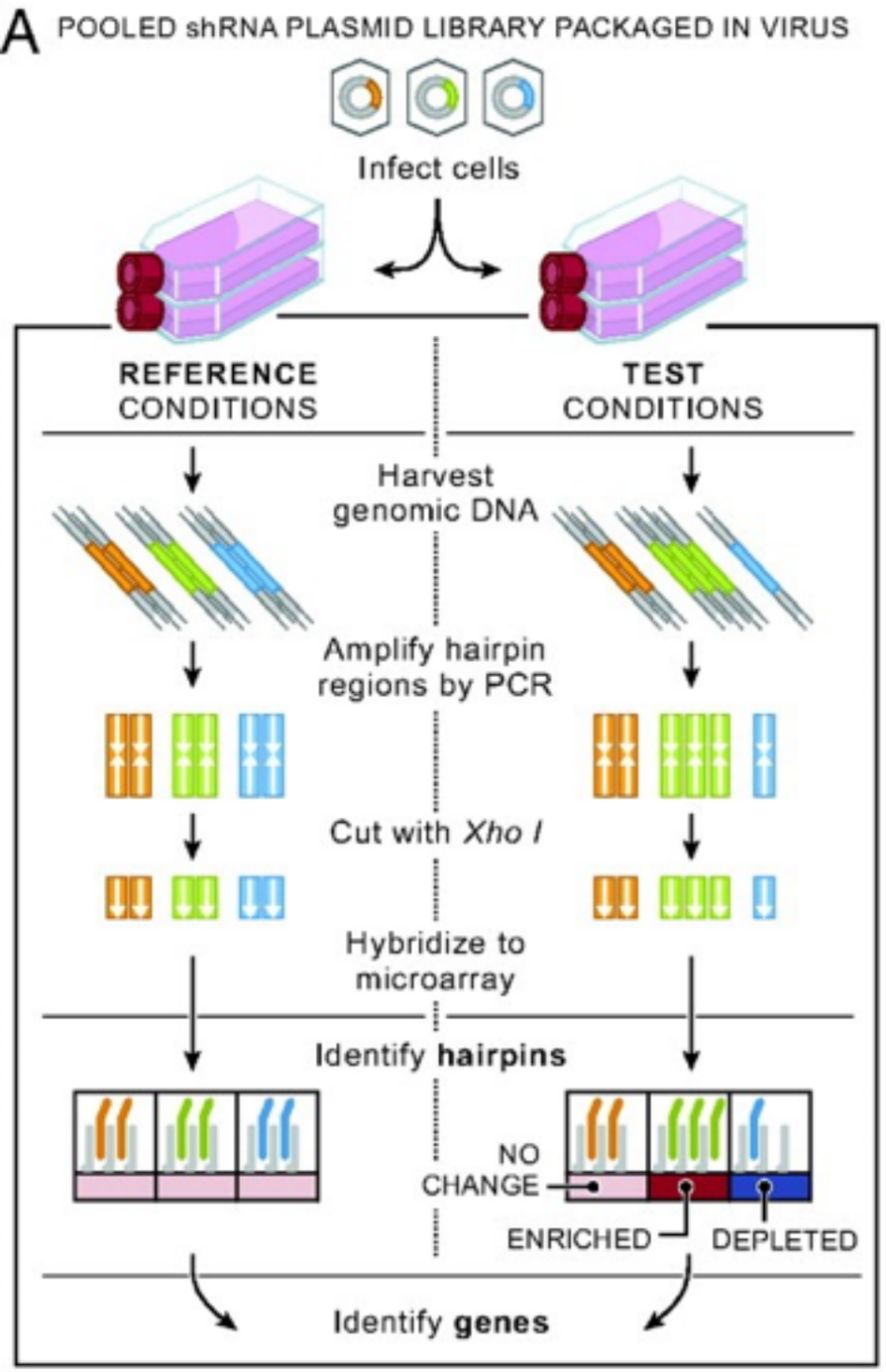
Endonucleolytic cleavage

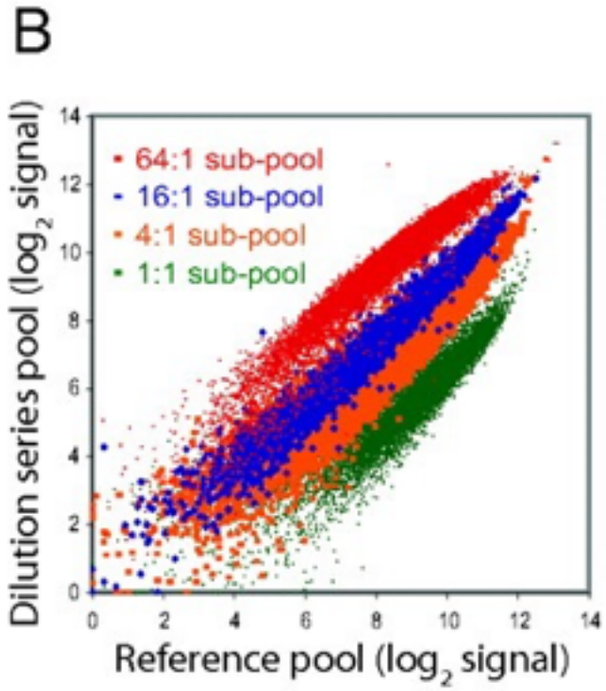
Blocked ribosome

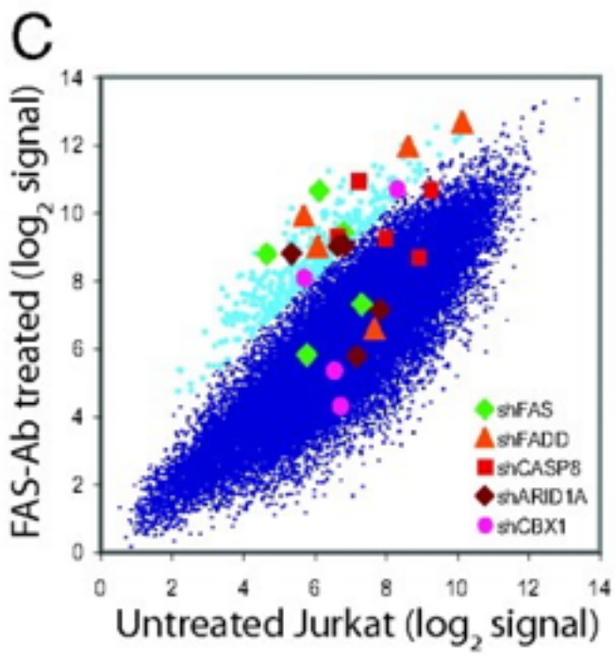
Messenger RNA strands
(mRNA)

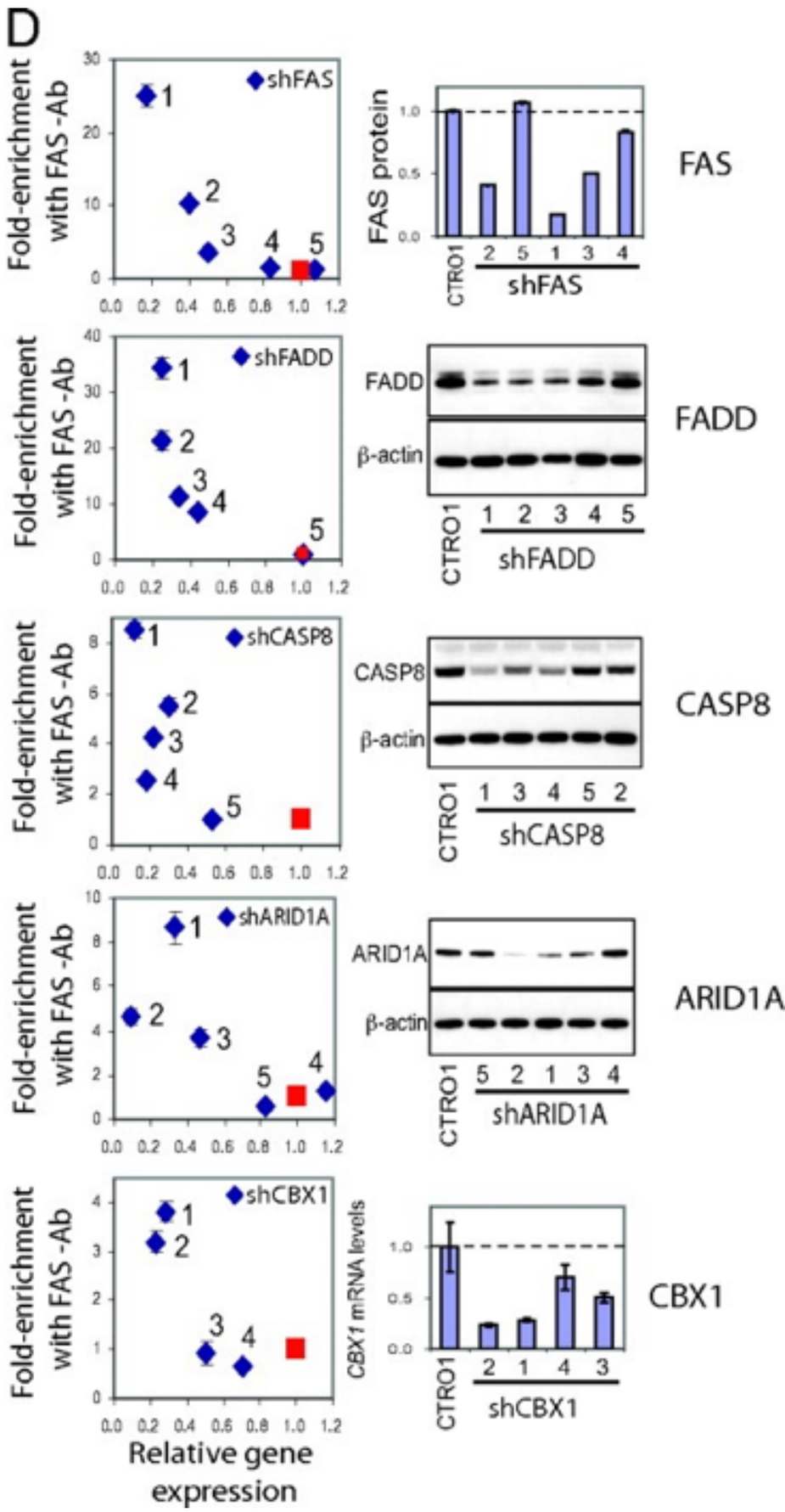
5 The microRNA-protein complex's presence blocks translation as well as speeding up **deadenylation** (breakdown of the Poly-A tail), which causes the mRNA to be degraded sooner and translated less.

the formation and function of microRNAs

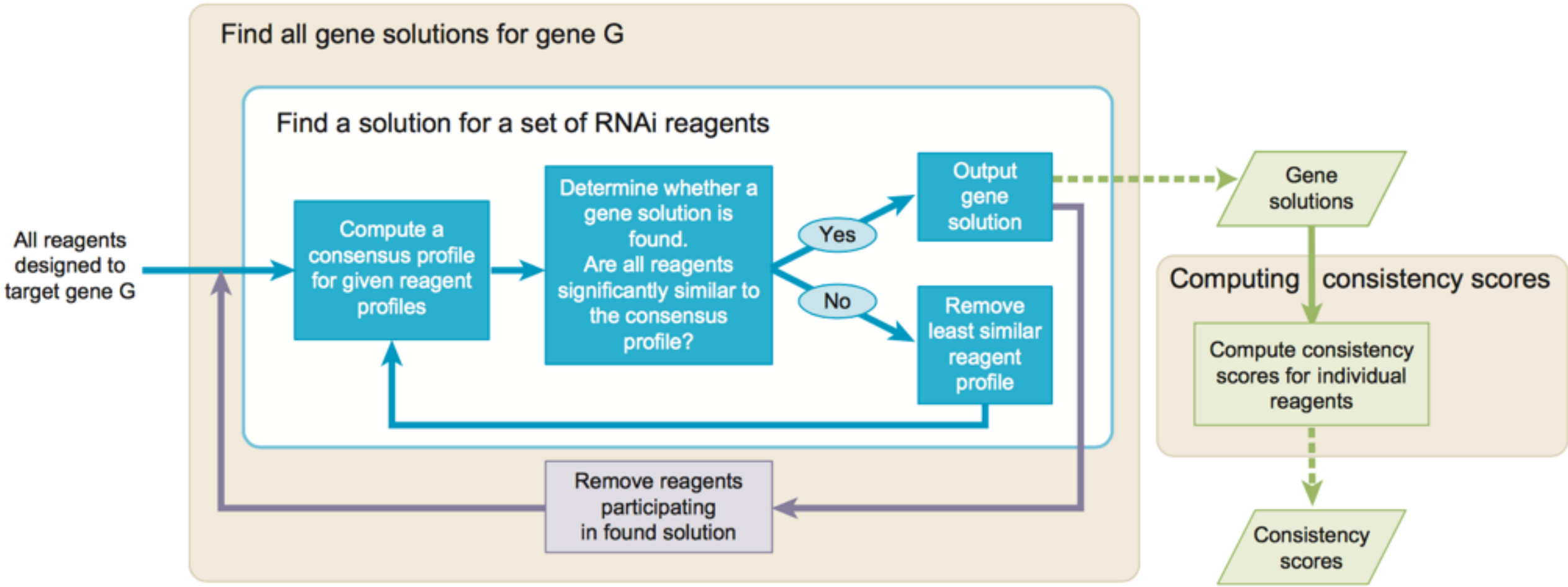

A
POOLED shRNA PLASMID LIBRARY PACKAGED IN VIRUS
Infect cells
REFERENCE CONDITIONS
TEST CONDITIONS
Harvest genomic DNA
Amplify hairpin regions by PCR
Cut with Xho I
Hybridize to microarray
Identify hairpins
NO CHANGE
ENRICHED
DEPLETED
Identify genes


B
64:1 sub-pool
16:1 sub-pool
4:1 sub-pool
1:1 sub-pool
Dilution series pool (log2 signal)
Reference pool (log2 signal)


C
FAS-Ab treated (log2 signal)
Untreated Jurkat (log2 signal)
shFAS
shFADD
shCASP8
shARID1A
shCBX1


D
Fold-enrichment with FAS -Ab
shFAS
FAS protein
CTRO1
shFAS
FAS
shFADD
FADD
β-actin
shFADD
shCASP8
CASP8
shCASP8
shARID1A
ARID1A
shARID1A
shCBX1
CBX1 mRNA levels
shCBX1
Relative gene expression

ATARiS:
Computational quantification of gene suppression phenotypes from multisample RNAi screens



Supplementary Figure 1. A schematic diagram of the ATARiS algorithm.

Statistical Method

For a given gene (G) let:

n - denote the number of screened samples

p - denote the number of reagents targeting

Let $\mathbf{X}$ denote a $p \times n$ matrix with each element $\mathbf{X}_{ij}$ representing the observed phenotypic effect produced by reagent i in sample j.

Statistical Method

Because we are only interested in finding the relative effects of gene suppression, we median-center each row of **X**.

$$\mathbf{X^*} = \mathbf{X} - \mu\ 1_n^T$$

μ is a vector of length p such that $\mu_i = \text{median}(x_{i\bullet})$ and 1_n is a vector of 1's of length n.

Statistical Method

Let:
c - denote a vector of length n representing the consensus profile for **X***
e - denote a vector of length p consisting of a relative effect size for each RNAi reagent.

ATARiS models each measurement X^*_{ij} as a product of its corresponding relative effect size $\mathbf{e}_i$ and phenotypic effect $\mathbf{c}_j$.

An approximation for $\mathbf{X}^*$ is given by

$$\mathbf{X}^* \sim \mathbf{e}\,\mathbf{c}^T$$

set max(**e**) =1 for identifiability.

Statistical Method

We begin by initializing ***c*** with the mean values of **X*** in each sample:

$$c_j \leftarrow \frac{1}{p}\sum_i x^*_{i,j} \quad \text{for} \quad j \in 1, \ldots, n.$$

We then update **e** and **c** repeatedly until convergence:

$$\mathbf{e} \leftarrow \arg\min_{\mathbf{e}} \left\| \mathbf{X}^* - \mathbf{e}\mathbf{c}^T \right\|_1$$

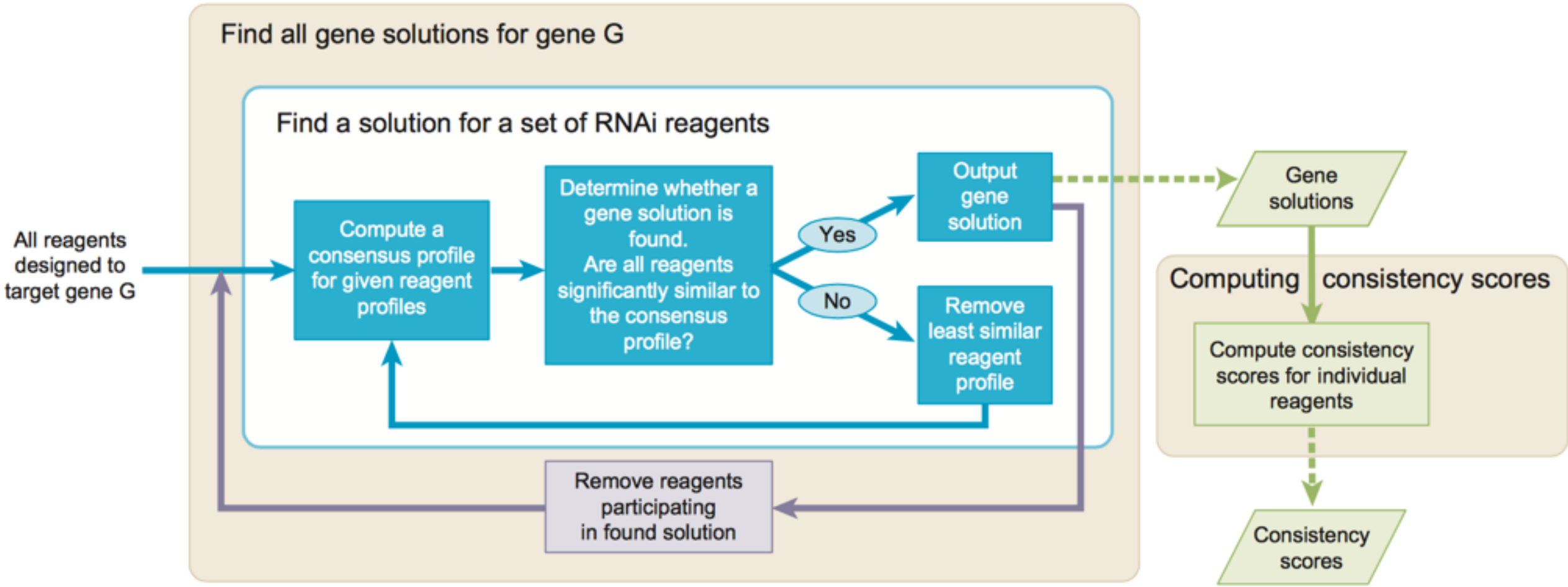
$$\mathbf{c} \leftarrow \arg\min_{\mathbf{c}} \left\| \mathbf{X}^* - \mathbf{e}\mathbf{c}^T \right\|_1.$$

The elements of **e** and **c** are updated in an element-wise manner, i.e.,

$$e_i \leftarrow \arg\min_{\hat{e}} \sum_j \left| x^*_{i,j} - \hat{e}c_j \right| \quad \text{for} \quad i \in 1, \ldots, p$$

and similarly

$$c_j \leftarrow \arg\min_{\hat{c}} \sum_i \left| x^*_{i,j} - e_i\hat{c} \right| \quad \text{for} \quad j \in 1, \ldots, n.$$



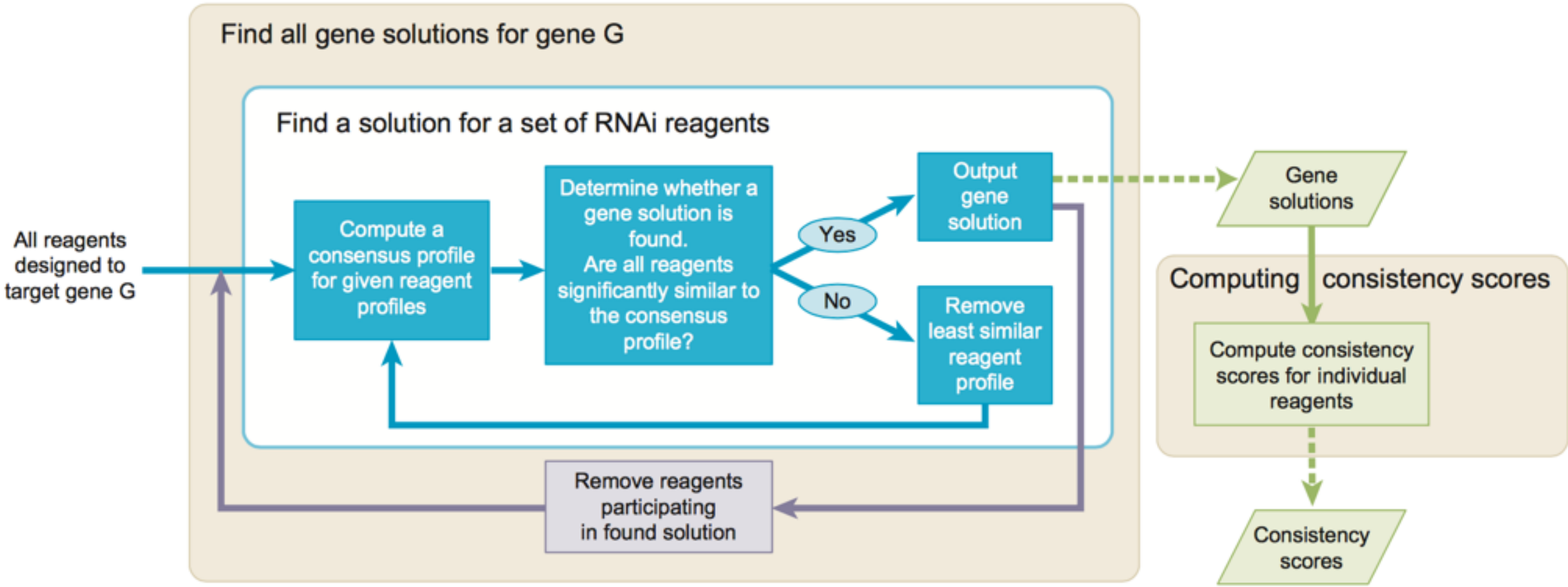
Supplementary Figure 1. A schematic diagram of the ATARiS algorithm.

Statistical Method

If either criterion is not fulfilled, we remove one reagent from the set R_G^* as follows:

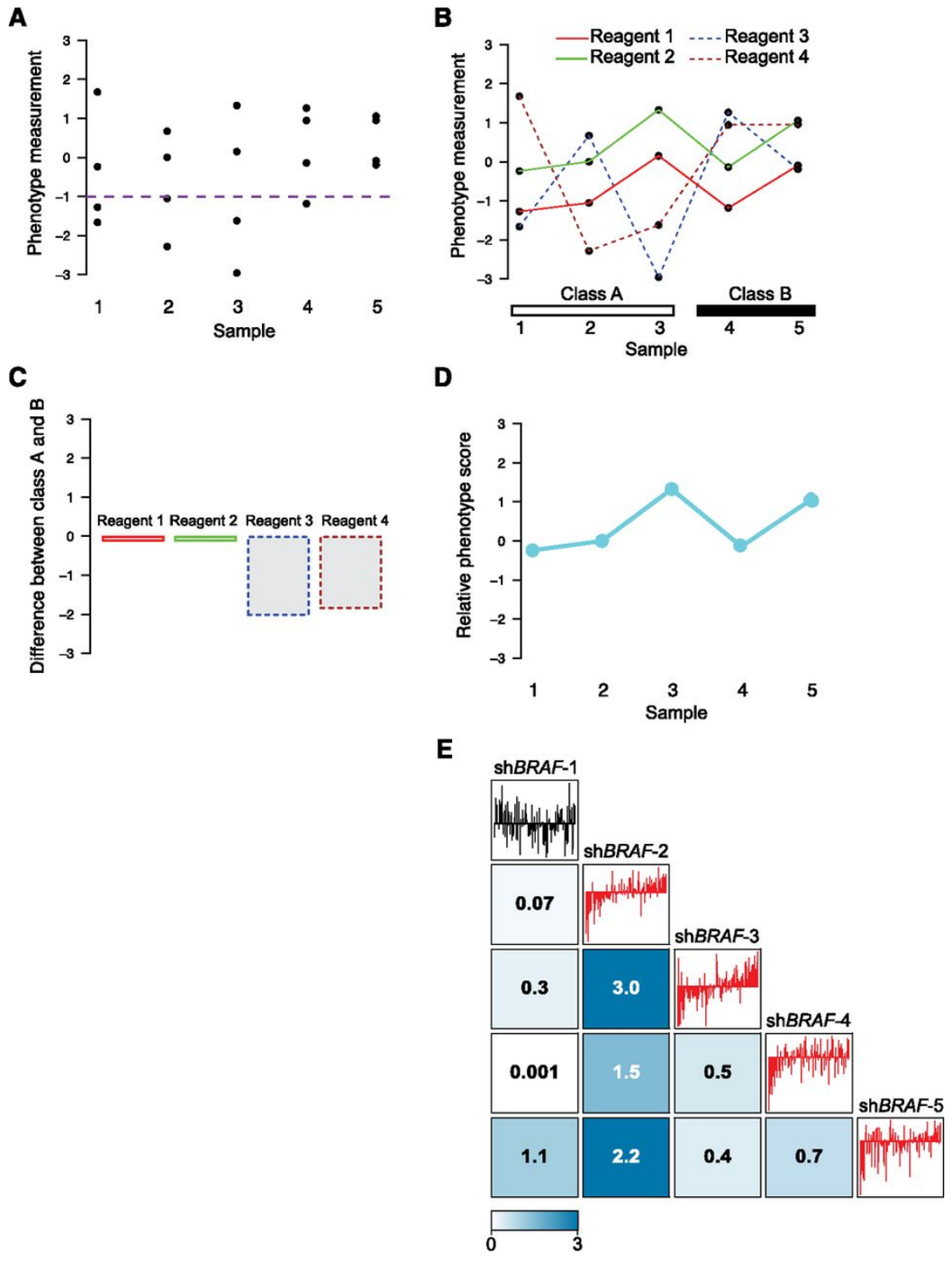
1. If any reagent $r \in R_G^*$ does not satisfy criterion (2), we discard the one with the lowest effect magnitude e_r.

2. Otherwise, we discard the reagent $r \in R_G^*$ with the lowest Spearman correlation coefficient between its profile and the consensus profile **c**.



Supplementary Figure 1. A schematic diagram of the ATARiS algorithm.

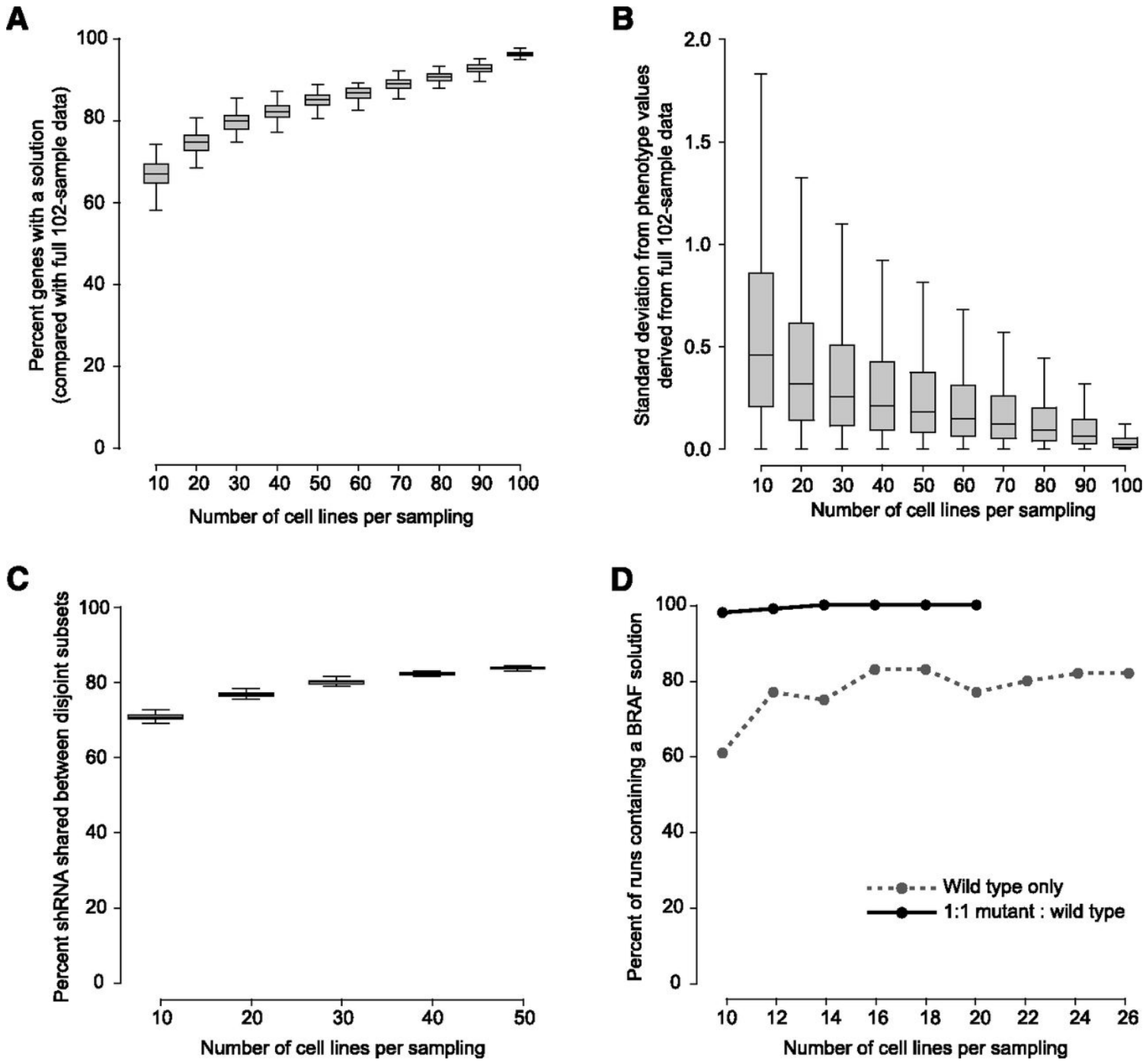
ATARiS accounts for patterns in RNAi reagent data in order to quantify the phenotypic effect of gene suppression in each sample.



Diane D. Shao et al. Genome Res. 2013;23:665-678

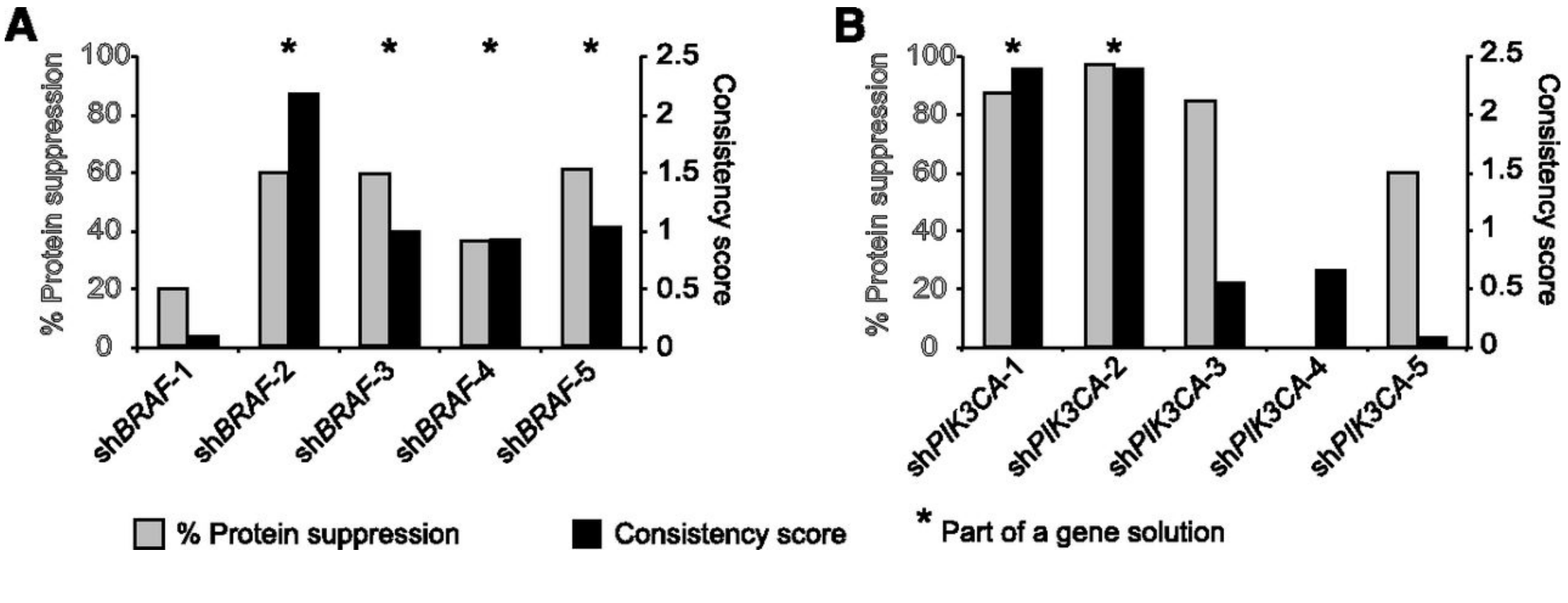
Proof of Concept

Influence of data set size and context on ATARiS results.



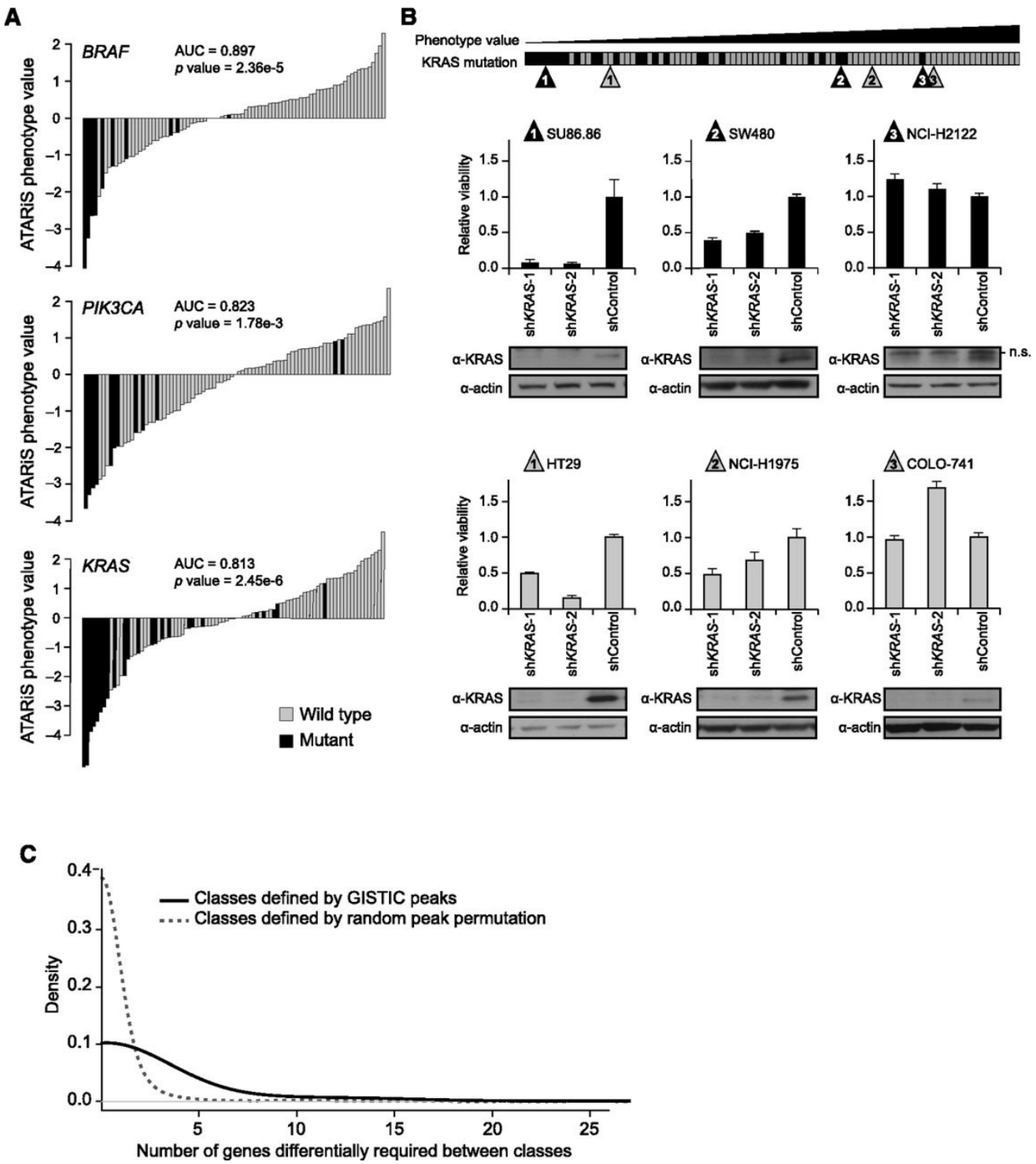
Diane D. Shao et al. Genome Res. 2013;23:665-678

ATARiS consistency scores are associated with on-target gene suppression.



Diane D. Shao et al. Genome Res. 2013;23:665-678

ATARiS gene phenotype values reflect biological dependencies.

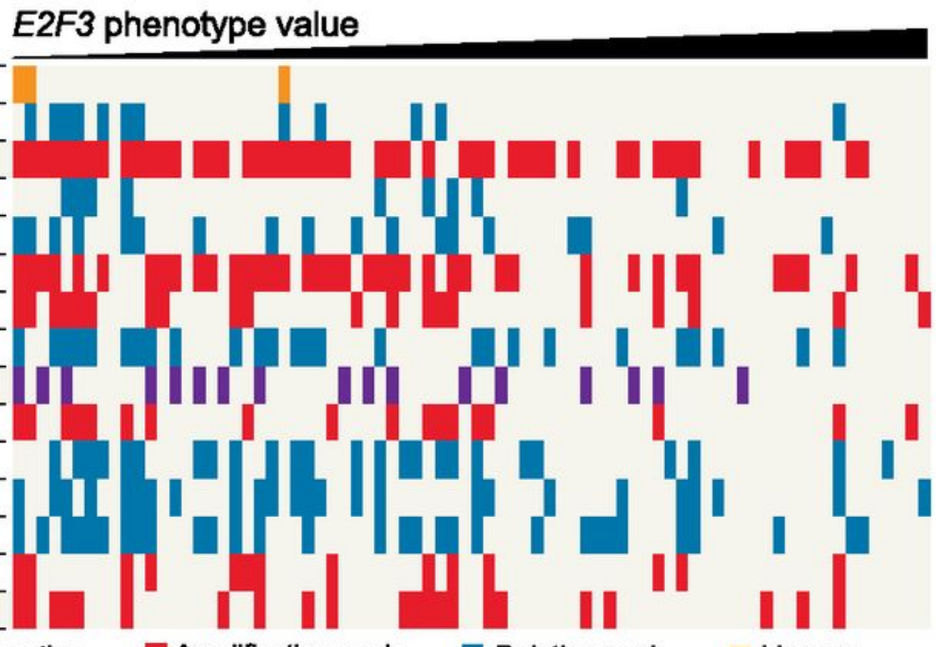


Diane D. Shao et al. Genome Res. 2013;23:665-678

ATARiS phenotype values enable phenotype-based analyses for biological discovery.

A

Rank	Feature	AUC	FDR
1	Lung Small Cell	0.909	0.148
2	Del 12p13.1	0.797	0.043
3	Amp 8q24.21	0.748	0.019
4	Del 17q11.2	0.7	0.242
5	Del 9q13	0.692	0.148
6	Amp 3q26.2	0.686	0.118
7	Amp 1p12	0.686	0.148
8	Del 22q12.3	0.684	0.148
9	Mut.CNA RB1	0.68	0.149
10	Amp 1p22.3	0.678	0.149
11	Del 13q31.3	0.667	0.148
12	Del 22q13.31	0.664	0.149
13	Del 16q23.1	0.662	0.148
14	Amp 1p34.2	0.657	0.250
15	Amp 6p22.3	0.648	0.228

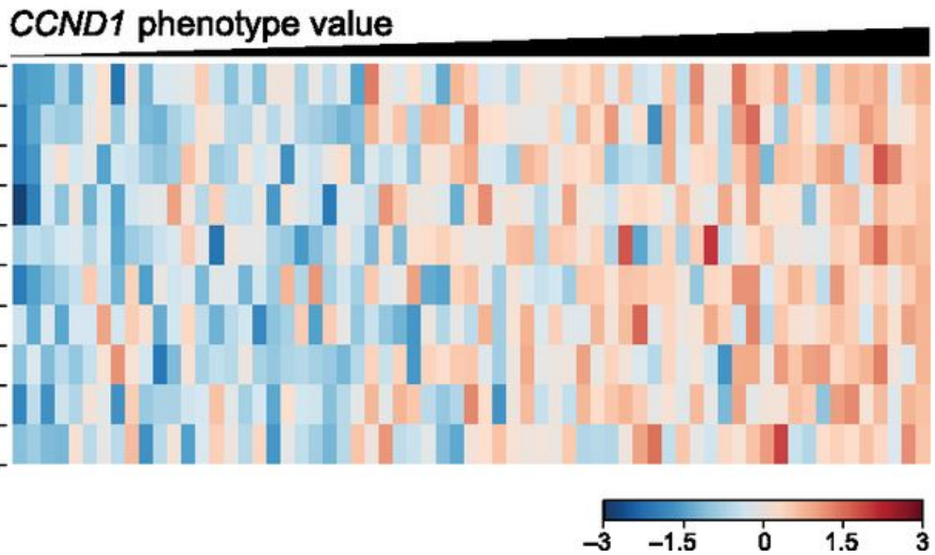


C

Rank	Solution	*p* value	FDR
1	HNF1B_1	2.00E-05	0.075
2	PAX8_1	2.00E-05	0.075
3	E2F3_1	4.00E-05	0.075
3	ELF3_1	4.00E-05	0.075
5	SOX10_1	6.00E-05	0.075
5	HIST1H4D_1	6.00E-05	0.075
7	NGEF_1	8.00E-05	0.086
8	FERMT1_1	1.00E-04	0.094
9	BCL2L1_1	1.40E-04	0.096
9	ASL_1	1.40E-04	0.096
9	POLE3_1	1.40E-04	0.096
12	MYB_1	3.40E-04	0.213
13	MPP6_1	4.00E-04	0.225
14	PITX3_1	4.20E-04	0.225
15	HNF4A_1	4.60E-04	0.230

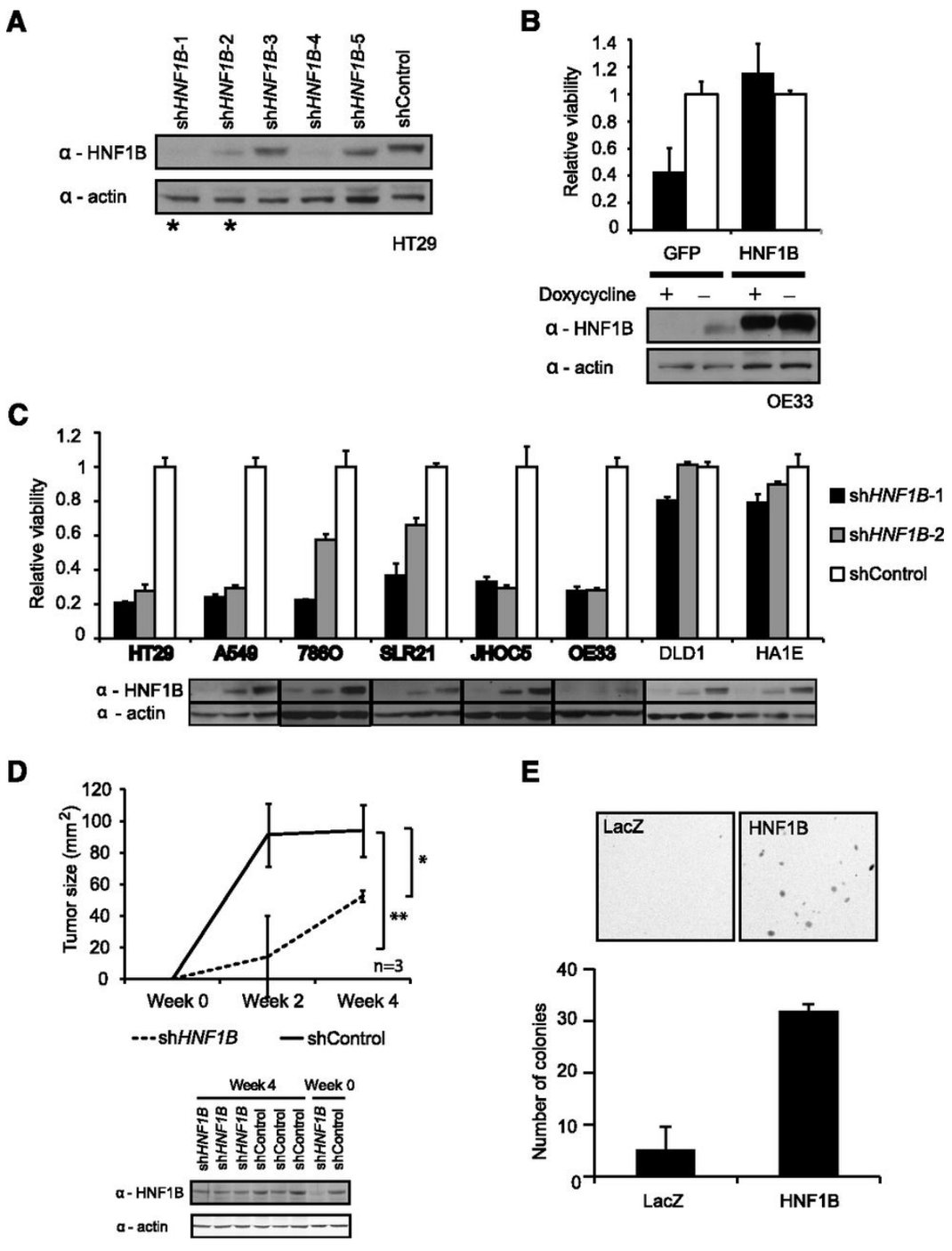
B

Rank	Solution	Correlation	FDR
1	CCL8_1	0.590	0.010
2	KRAS_2	0.566	0.010
3	64748_1	0.565	0.010
4	KRAS_1	0.557	0.010
5	CRX_1	0.543	0.010
6	RPS11_1	0.543	0.010
7	Gpr12_1	0.538	0.012
8	CDK6_2	0.530	0.012
9	RAF1_1	0.524	0.010
10	HIATL1_1	0.516	0.012



Diane D. Shao et al. Genome Res. 2013;23:665-678

Characterizing the role of HNF1B in cancer.



Diane D. Shao et al. Genome Res. 2013;23:665-678

Name	Description	786O_KIDNEY	A204_SOFT_TISSUE	A2058_SKIN	A2780_OVARY	A549_LUNG	AGS_STOMACH	ASPC1_PANCREAS	BXPC3_PANCREAS
TRCN0000072180cA_st	GFP	-0.846910266578903	-0.597219002565684	-1.36667290213915	-1.2652302300513	-1.35873485366065	-0.22486		
TRCN0000072180cB_st	GFP	-0.71236866572242	-0.540902298376377	-1.27941356651753	-1.18026578819127	-1.31092966512485	-0.18878		
TRCN0000072180cC_st	GFP	-0.893792214881376	-0.417136518239409	-1.2683064041415	-1.15433501750928	-1.12318964004888	-0.17056		
TRCN0000072180cD_st	GFP	-0.704758106365016	-0.375226412796204	-1.05492452042385	-1.08481781052728	-0.865811020008964	-0.13029		
TRCN0000072180cE_st	GFP	-0.886542719867097	-0.569060650471031	-1.5155182744791	-1.32907959138011	-1.31674236573503	-0.18961		
TRCN0000072183cA_st	GFP	-0.0627716958359524	-0.432227144441775	-0.39134337986517	0.0154526695714189	0.123194455733714	-0.54049		
TRCN0000072183cB_st	GFP	-0.133141520318345	-0.351071839157159	-0.460488734413301	-0.0992484323217213	-0.0342570306786797	-0.49226		
TRCN0000072183cC_st	GFP	-0.0554756334659765	-0.360690245709409	-0.539219874948822	-0.0900498998420714	-0.0997257750124875	-0.60686		
TRCN0000072183cD_st	GFP	-0.0496434307434803	-0.342054583014424	-0.388289043015903	0.0441334935284733	0.127691080820872	-0.56534		
TRCN0000072183cE_st	GFP	-0.118030461206006	-0.664872352924344	-0.471726795843165	-0.0148335156900851	-0.0118724382444568	-0.48773		
TRCN0000072184cA_st	GFP	-0.0177153690728369	-2.75100551510794	-1.25918855228197	0.643692406482372	-0.629272998090856	-0.81085		
TRCN0000072184cB_st	GFP	-0.064420224062705	-3.06843147549994	-1.27481673841851	0.621452376301898	-0.685656051506292	-0.89924		
TRCN0000072184cC_st	GFP	-0.0355770718949888	-2.89769088415916	-1.29207138718201	0.6703678092282	-0.546375971417545	-0.7786512451096		
TRCN0000072184cD_st	GFP	0.0603061918999312	-3.28535935950882	-1.27849951532949	0.653127105194331	-0.525243306665941	-0.74371		
TRCN0000072184cE_st	GFP	0.0167077200095679	-3.54205875534845	-1.55004431818659	0.586812196237017	-0.652227417480878	-0.82030		
TRCN0000072185cA_st	GFP	-0.629469047193879	-0.0094027522231871	0.0328583196863022	-0.400995011244995	-0.139731340412025	-0.30233		
TRCN0000072185cB_st	GFP	-0.780922258579557	0.0670729658587317	0.0682828963071088	-0.448073924496816	-0.163927600226571	-0.33967		
TRCN0000072185cC_st	GFP	-0.739261653680761	0.0664461157105195	-0.0544823661408814	-0.413776669182753	-0.0831826080525075	-0.32248		
TRCN0000072185cD_st	GFP	-0.772937257091854	0.0564165133391204	-0.0131036606052417	-0.410998216140274	-0.128738428612712	-0.32509		
TRCN0000072185cE_st	GFP	-0.725239199308351	0.0498954839898017	-0.00266183530957857	-0.451993942508002	-0.0964297834998073	-0.30106		
TRCN0000072186cA_st	GFP	1.25333107071751	0.579209536948303	1.52504591936611	0.701158040684366	1.20526151924963	0.732964		
TRCN0000072186cB_st	GFP	1.27478164899102	0.603656692728588	1.75634944411323	0.672520579472317	1.24946901434997	0.872292		

Supplementary Data 1: Achilles_102lines.gct

The Project Achilles dataset that was analyzed by ATARiS, after pre-processing and normalization as described in Methods. A tab-delimited table in gct format. Each row contains data for one shRNA out of 53341. The first column contains unique shRNA identifiers. The second column describes for each shRNA its target gene. Each following column contains data for one of the 102 screened cell lines. Values are Z-scores of log2 fold-change values. Lower values represent lower abundance of the shRNA relative to the levels detected in the DNA plasmid reference pool.

Name	Description	X786O_KIDNEY	A204_SOFT_TISSUE	A2058_SKIN	A2780_OVARY	A549_LUNG	AGS_STOMACH	ASPC1_PANCREAS	BXPC3_PANCREAS
DDX3X_1_11110	DDX3X	0.211590453973692	0.518631377437002	-0.277199966616749	0.500443850895471	-0.098664262150512	1.55516402234618		
SNRNP70_1_11101	SNRNP70	0.0129140297330312	0.316551728762962	0.644667180493832	-0.172106200951815	0.629822365509886	0.42865156092157		
DDX3Y_1_1001	DDX3Y	-0.662606288793878	-0.940198773929798	0.173424611732904	-0.00181195640802421	0.85246711217165	0.46204001133245		
DHX8_1_1101	DHX8	-0.329424086861835	0.258580903395019	-0.515050362787001	0.538438209262556	0.442021030622888	1.73286355442932		
THOC1_1_01110	THOC1	-0.592319496300459	-0.516995175410268	-0.181493137950646	0.438929942941496	-0.542903987362922	0.31740187964502		
SF3B4_1_1111	SF3B4	0.288752225663586	0.596179939917817	-0.52009325457307	0.209277426695433	0.0965247701608617	0.31322828831008		
KHDRBS1_1_10111	KHDRBS1	-0.4333427406946 1	-0.664626699852022	0.0196730364794974	-0.194994779005351	-1.575182664051	-0.280940505081553		
RNPS1_1_01110	RNPS1	0.0947957167467381	0.015782053218945	0.352644462105955	0.0362729184867278	-0.0943501666687559	1.08689888825479		
SF3A3_1_01111	SF3A3	0.0267765672038638	0.268815108350435	-0.316576525682795	0.105943597584211	0.134241883172188	-0.0481068792687		
SF3A2_1_1111	SF3A2	0.563792102817082	0.960308522096028	-0.444272215807095	-0.284559350958382	-1.2888451164686	-0.4530168548563		
SF3B3_1_11010	SF3B3	-0.590085912450422	-0.420185919888604	-1.69970554466075	1.12197947014214	-0.594168189591323	-1.0331009167584		
SF3B1_1_11110	SF3B1	-0.097132373435564	1.65083198465813	-1.12376695670118	0.395598826532287	0.669396209997202	-1.8299543185687		
SFRS2_1_11101110101011	SFRS2	-0.106315348179639	1.3246340607673	-0.521770298688503	-0.591277809018098	0.245828946690271	0.53252969199909		
PABPN1_1_1100	PABPN1	-0.338626897973457	0.477271697114326	-1.46012297969965	-0.196571825524862	-0.243415510393393	0.65151079452029		
SFRS11_1_0011	SFRS11	-0.0349249112837136	0.339773203802342	0.550597367063392	0.799874690970715	0.566453840342808	0.50585378200152		
CWC27_1_01010	CWC27	-0.75339675553671	-0.067133768960208	-1.44843211685619	1.42365829437568	-1.1764979423836	-0.4115140970582		
NUDT21_1_10111	NUDT21	-0.10383564711311	0.264766760304623	-0.111217790733854	0.0701036150088541	0.0641259790132206	-0.0096189710900		
CPSF1_1_01110	CPSF1	0.129107334947478	0.71875494957015	-0.555860793100707	0.394530320779611	-0.715165269307273	-0.2850143099454		
PPWD1_1_11001	PPWD1	-1.16927993803962	0.0220536437942604	-1.43371565543742	1.05182176281163	-0.0549539508361078	-1.20293385861 53		
PPIL1_1_0110	PPIL1	-0.217216223924796	-0.176062555227138	-0.0967677858097028	-0.434425484020418	-0.252406076292408	-0.2940807488016		
C21orf66_1_00111	C21orf66	0.534543078141589	-0.70115536354154	0.2902193187822	0.738212421898294	0.0086449645006627	0.203022		
HGD_1_10110	HGD	-0.233645794222247	-0.0566287114975277	-0.845734099337523	-0.198504851642361	0.660043020577863	0.84498883987648		

Supplementary Data 2: Achilles_102lines_gene_solutions.gct

Tab-delimited table in gct format. Rows are 8280 gene solutions. The first column is a unique gene solution identifier. The second column is the gene symbol of the targeted gene. The next 102 columns are cell lines. Lower values represent higher relative dependency of a cell line on the targeted gene.

shRNA	gene.symbol	isUsed	sol.number	sol.name	sol.id	cscore	pval	qval
TRCN0000073948m_st	10404	FALSE	NA	NA	NA	0.212	0.614	0.811
TRCN0000073949m_st	10404	TRUE	1	10404_1_01111	01111	1.07	0.0855	0.374
TRCN0000073950m_st	10404	TRUE	1	10404_1_01111	01111	0.333	0.465	0.722
TRCN0000073951m_st	10404	TRUE	1	10404_1_01111	01111	0.652	0.223	0.533
TRCN0000073952m_st	10404	TRUE	1	10404_1_01111	01111	0.33	0.468	0.724
TRCN0000082663m_st	10777	FALSE	NA	NA	NA	0.0343	0.924	0.969
TRCN0000082664m_st	10777	FALSE	NA	NA	NA	0.0311	0.931	0.972
TRCN0000082665m_st	10777	FALSE	NA	NA	NA	0.138	0.727	0.872
TRCN0000082666m_st	10777	FALSE	NA	NA	NA	0.481	0.33	0.626
TRCN0000082667m_st	10777	FALSE	NA	NA	NA	0.693	0.203	0.516
TRCN0000063783m_st	11039	TRUE	1	11039_1_1110011000	1110011000	1.05	0.0891	0.213
TRCN0000063784m_st	11039	TRUE	1	11039_1_1110011000	1110011000	1.53	0.0295	0.115
TRCN0000063785m_st	11039	TRUE	1	11039_1_1110011000	1110011000	1.05	0.0897	0.213
TRCN0000063786m_st	11039	TRUE	2	11039_2_0001100111	0001100111	1.08	0.0836	0.206
TRCN0000063787m_st	11039	TRUE	2	11039_2_0001100111	0001100111	2.31	0.00487	0.039
TRCN0000083208m_st	11039	TRUE	1	11039_1_1110011000	1110011000	1.46	0.0344	0.123
TRCN0000083209m_st	11039	TRUE	1	11039_1_1110011000	1110011000	1.13	0.0738	0.191
TRCN0000083210m_st	11039	TRUE	2	11039_2_0001100111	0001100111	0.209	0.618	0.746
TRCN0000083211m_st	11039	TRUE	2	11039_2_0001100111	0001100111	1.2	0.0633	0.175
TRCN0000083212m_st	11039	TRUE	2	11039_2_0001100111	0001100111	0.462	0.345	0.51
TRCN0000053919m_st	11261	TRUE	1	11261_1_1010	1010	0.652	0.223	0.548
TRCN0000053920m_st	11261	FALSE	NA	NA	NA	0.263	0.546	0.781

Supplementary Data 3: Achilles_102lines_shRNA_table.txt

Tab-delimited table. Each row contains information for one screened shRNA reagent. The field ‘isUsed’ signifies whether the data of the shRNA was used to generate any gene solution. ‘sol.number’ identifies which solution of the targeted gene used data from the shRNA. ‘sol.name’ is a unique identifier for the solution generated using the shRNA data. ‘sol.id’ is a binary string with each digit representing one shRNA targeting the gene, according to the order of appearance in this file. A ‘1’ means the shRNA was used to generate the current gene solution and ‘0’ means it was not used. The ‘cscore’, ‘pval’ and ‘qval’ columns hold the consistency score and its corresponding p-value and q-value for each shRNA.