



# siRNA off-target effects in genome-wide screens identify signaling pathway members

SUBJECT AREAS:

BIOINFORMATICS

COMPUTATIONAL BIOLOGY

RNAI

SYSTEMS BIOLOGY

Eugen Buehler<sup>1</sup>, Aly A. Khan<sup>2</sup>, Shane Marine<sup>3</sup>, Misha Rajaram<sup>4</sup>, Amit Bahl<sup>5</sup>, Julja Burchard<sup>5</sup> & Marc Ferrer<sup>1</sup>

<sup>1</sup>National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, MD 20892, <sup>2</sup>Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, <sup>3</sup>Automated Biotechnology, Merck Research Laboratories, North Wales, PA 19454, <sup>4</sup>Cardiovascular Research Institute, University of California, San Francisco, San Francisco, CA 94143, <sup>5</sup>Merck & Co., Inc., West Point, PA 19486.

Received  
28 March 2012

Accepted  
14 May 2012

Published  
29 May 2012

We introduce a method for analyzing small interfering RNA (siRNA) genetic screens based entirely on off-target effects. Using a screen for members of the Wnt pathway, we demonstrate that this method identifies known pathway components, some of which are not present in the screening library. This technique can be applied to siRNA screen results retroactively to confirm positives and identify genes missed using conventional methods for on-target gene selection.

Correspondence and requests for materials should be addressed to E.B. (eugen.buehler@nih.gov)

Large-scale siRNA screening is an increasingly popular method of interrogating signaling pathways both to identify relevant genes and to discover novel drug targets. siRNA screens typically involve transfecting small numbers of cells with a library designed to target the entire genome or some subset (such as a set of genes coding for theoretically druggable domains)<sup>1</sup>. The systematic knockdown of each gene enables high-throughput loss-of-function analysis by assessing pathway dependent phenotypes as measured through changes in the appearance of specific markers or induction of reporter constructs (Fig. 1a).

Unfortunately, siRNA screens have demonstrated a high false positive rate<sup>2</sup>. Researchers typically perform labor-intensive follow-up work on hundreds of hits to confirm a handful of relevant genes. Many false positives are likely due to off-target effects<sup>3,4</sup>, wherein partial complementarity between an siRNA and multiple transcripts, typically in the 3'UTR, results in their down-regulation, adding unintended silencing to the screen (Fig. 1b). Previous work on attenuating off-target effects has largely focused on identifying lower-risk sequences, introduction of chemically modified siRNAs, or use of multiple siRNA sequences in additional screens<sup>5</sup>.

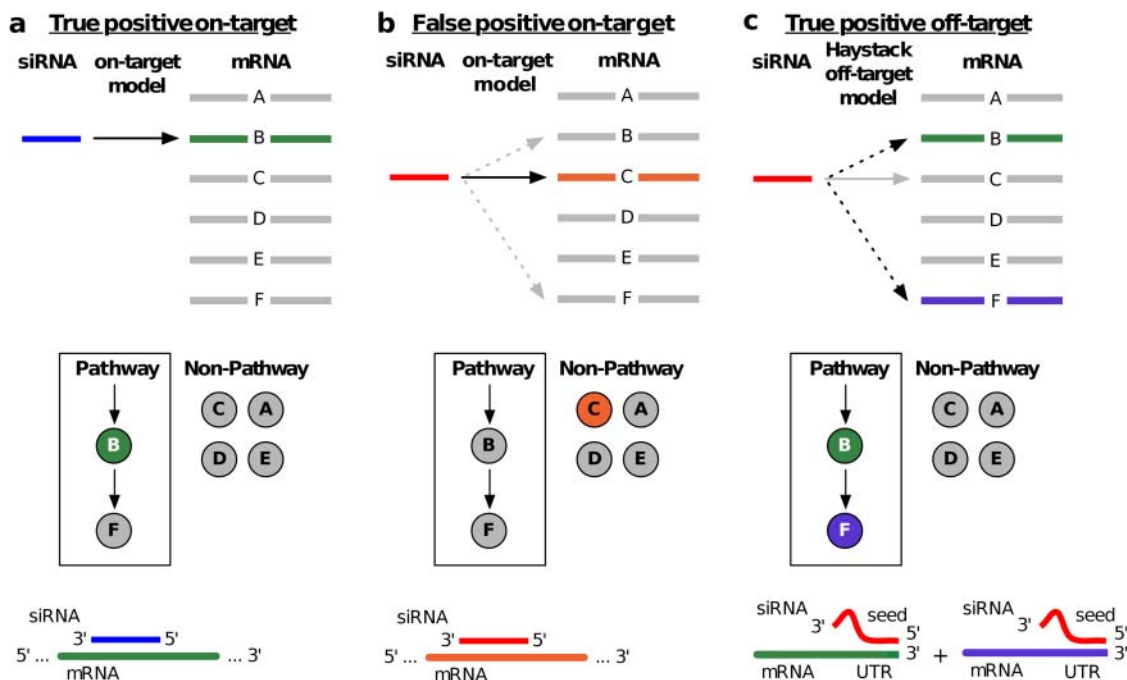
## Results

In order to understand and exploit the off-target effects present in siRNA screening data, we implemented a predictive model of down-regulation due to siRNA off-targeting. Existing predictors are microRNA-related and often use conservation or other criteria not applicable to siRNA off-targeting<sup>6</sup>. We trained a simple linear model specific for siRNAs using published gene expression profiles in which off-targeting mediated by the seed (positions 2–8 of the guide strand) has been detected<sup>7</sup>. Our model for off-target seed-based down-regulation is:

$$O_{n,k} = \beta_1 \cdot PM + \beta_2 \cdot M1 + \beta_3 \cdot M8 + \beta_4 \cdot M18 + \beta_5 \cdot 3'UTRLength$$

The model includes four types of seed matches, or reverse complementarity between the guide strand seed sequence and the 3'UTR of the transcript: PM, perfect match to guide bases 2–7 followed by adenine opposite base 1; M1, no adenine opposite base 1; M8, mismatch opposite base 1; and M18, with both terminal mismatches. We calculate as predictive variables the number of times a particular match-type occurs between the seed sequence of the siRNA  $n$  and the 3'UTR of the transcript  $k$ . In addition, we use the length of the 3'UTR as a predictive variable, since in our analyses longer 3'UTRs were correlated with up-regulation of transcripts in the absence of siRNA seed matches. Up-regulation of genes with long 3'UTRs and a lower density of siRNA matches has been previously observed in the literature<sup>8</sup>. Model parameters ( $\beta$ ) are derived from linear regression against the mean log ratio of mRNA expression data (see Fig. 2, **Supplementary Data Set 1**, and Methods).

We hypothesize that siRNAs off-targeting to pathway members can in part explain the phenotypes observed in genome-wide screens (Fig. 1c). To investigate this, we introduce Haystack, a new computational method for



**Figure 1 | siRNA on-targeting and off-targeting to genes in a hypothetical pathway.** (a) On-target model correctly infers gene B as a pathway member due to on-target effects, depicted by the solid arrow from the siRNA (blue) to gene B. Extensive base-pairing between the siRNA and target gene B results in silencing. (b) A false-positive result incorrectly infers non-pathway gene C as a pathway member by neglecting off-targeting effects, depicted by dashed gray arrows from the siRNA (red) to pathway genes B and F. (c) Haystack explains screen results as a linear combination of the predicted off-targeting effects, depicted by dashed gray arrows from the siRNA (red) to pathway genes B and F. Imperfect base-pairing between siRNA (red) and 3' UTR region of off-target genes results in down-regulation.

identifying the most statistically significant genes that explain screen results, based entirely on off-target effects. First, we predict all  $O_{n,k}$ , the off-target down-regulation of every transcript  $k$  by every siRNA  $n$ . Second, for each transcript we calculate the significance of correlation between the predicted down-regulation by every siRNA to the siRNAs' screen results  $z$ , using a t-test. Third, the most statistically significant transcript is selected and a linear parameter  $a_k$  is estimated, approximating  $z$  as a function  $a_k * O_k + c$ . Finally, the residual between  $z$  and the predicted values of  $z$  in this linear model is calculated and the next transcript is selected via the significance of the correlation of each remaining  $O_k$  to the residual. In this stepwise manner, the most statistically significant transcripts are selected and added iteratively as features to a linear model, until no transcript has a Bonferonni-corrected correlation p-value less than 0.01. The final model can be viewed as predicting the phenotypic score associated with an siRNA  $z_n$  as a linear combination of the predicted off-target effects of siRNA  $n$  on a set of transcripts  $K$  (with some constant intercept term  $c$ ):

$$z_n = c + \sum_{k=1}^K a_k O_{n,k}$$

We assess the transcripts selected based on  $a_k$ , the magnitude of the off-target effects on transcript  $k$  in explaining the screening results. The directionality of  $a_k$  indicates the effect (either positive or negative) that down-regulation of each transcript has on the assay readout.

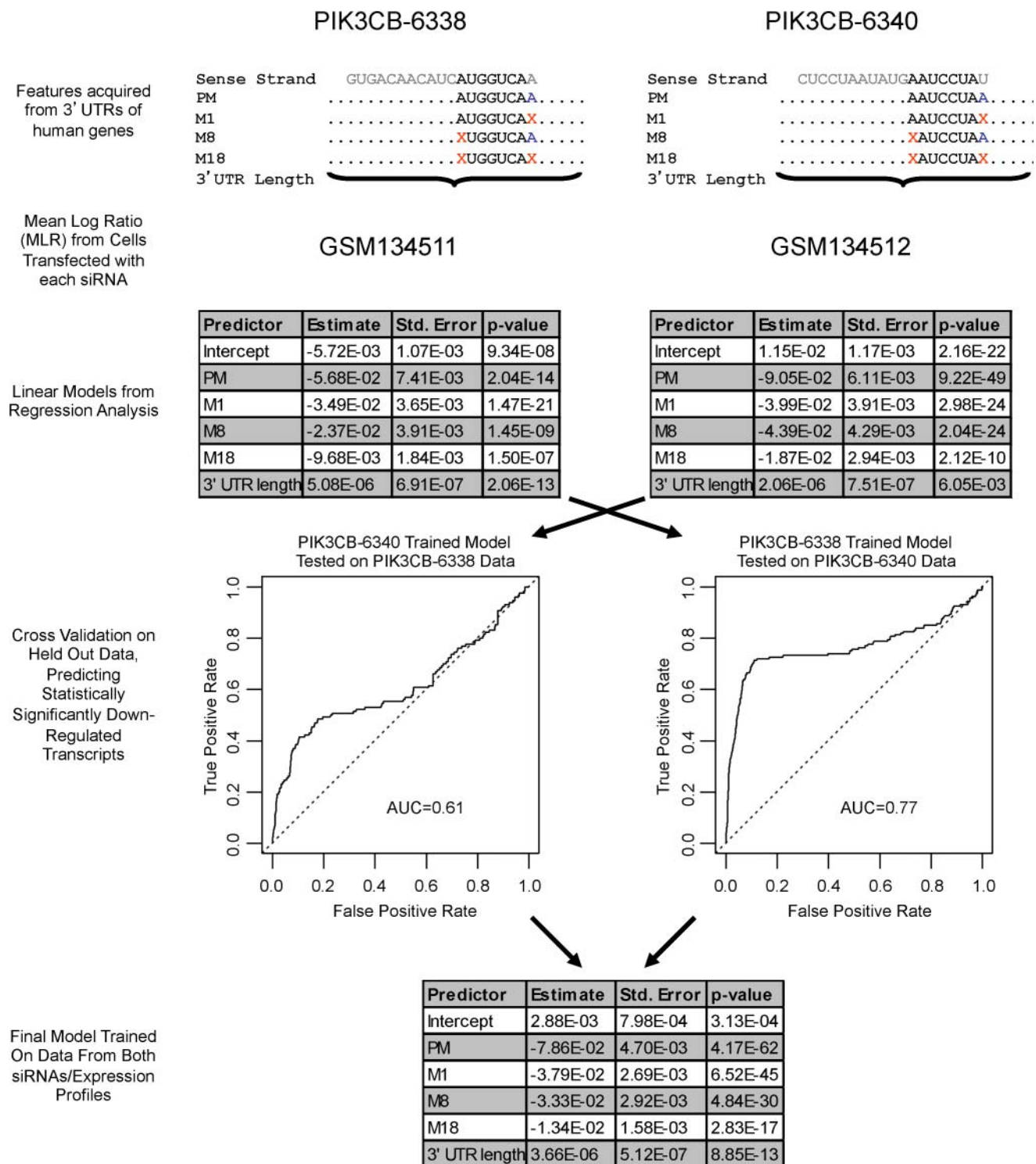
We applied Haystack to 19,815 siRNAs used in screening 6,605 theoretically druggable genes for activity in the Wnt/ $\beta$ -catenin signaling pathway (Supplementary Data Set 2). The Wnt/ $\beta$ -catenin pathway is constitutively active in many human cancers. To screen for novel factors in the Wnt/ $\beta$ -catenin pathway, HT1080 sarcoma cells were engineered to contain a firefly luciferase reporter coupled to a  $\beta$ -catenin-driven promoter, activated in the screen by

conditioned media containing Wnt-3a. A control EF1 $\alpha$ -driven Renilla luciferase reporter was used for normalization. Three siRNAs per gene were transfected individually into the reporter cell line in three separate screens. We calculated z-scores for the siRNAs from the log ratio of reporter intensities.

Table 1 lists, ordered by p-value, the top 10 genes included in the model built via Haystack from the siRNA screens in combination. Predicted activities per gene correlated well between screens when analyzed separately (Fig. 3). Supplementary Table 1 contains all 61 hits identified. In the case of the Wnt pathway, a large number of "canonical" pathway members have been previously identified. To measure pathway enrichment in screening results, we used 158 Wnt related genes from the KEGG pathway database<sup>9</sup>. Of the top 10 most statistically significant transcripts, 6 (LEF1, AXIN2, CCND1, RAC1, CTNNB1, and LRP6) are in the KEGG list of Wnt pathway genes, a remarkable level of enrichment ( $P < 3.7e-11$ ; hypergeometric test for enrichment).

We reviewed literature on the genes present in Table 1 but not listed in KEGG to check for other associations to the Wnt pathway. SUMO2, the third most statistically significant hit, has been identified as sumoylating and repressing LEF1<sup>10</sup>. Consistent with that role, its parameter estimate in the regression indicates that down-regulation of SUMO2 increases reporter activity. Another hit (Supplementary Data Set 3), SENP7, has been shown to reverse the sumoylation of SUMO2/3<sup>11</sup>, and accordingly the regression analysis shows that its down-regulation decreases Wnt reporter activity. Finally, the transcription factor POU4F2 (BRN3B) is involved in both cell development<sup>12</sup> and cancer<sup>13</sup>, as is the Wnt pathway, making a role for POU4F2 in Wnt signaling seem plausible.

It is worth noting that three transcripts identified in the top 10 hits (RAC1, CLEC18B, and POU4F2) did not have siRNAs designed against them in the druggable library screened here. Although RAC1 is not included in the library, it is found by Haystack analysis and is a canonical member of the Wnt pathway. Since off-target



**Figure 2 | Training an siRNA off-target model.** The seed sequence for an siRNA is defined here as the reverse complement of the heptamer at the 5' end of the guide strand of the siRNA (bases 2–8), appended with an “A”. Four orthogonal match types are defined between the seed sequence of the guide (antisense) strand and a given 3'UTR: PM (octamer; perfect match), M1 (heptamer, mismatch on base 1 of the guide seed), M8 (heptamer, mismatch of base 8 of the guide seed), M18 (mismatch of bases 1 and 8 of the guide seed). The sequence of these match types are defined for two example siRNAs, PIK3CB-6338 and PIK3CB-6340. The length of a 3'UTR is also used as a predictive feature, as it has been empirically observed to be correlated with up-regulation when there are no matches of an siRNA to the 3'UTR (Figure 1). These features were then used in a linear regression against the mean log ratio of the transcript from gene expression profiles in which the siRNAs were transfected into cells (profiles GSM134511 and GSM134512 respectively, downloaded from GEO). The linear models trained from these two data sets were then cross-validated on each other, to demonstrate models derived from one siRNA can be successfully applied to another. Each model was used to predict significantly ( $p$ -value  $< .01$ ) down-regulated transcripts in the data set/siRNA that it was not derived from, and the results of this cross-validation were displayed as ROC curves. The dashed line in each graph corresponds to the expected performance of a random model (AUC=0.5). Finally, the data sets were merged to generate a final off-target model.



**Table 1 | Top 10 most statistically significant genes identified by Haystack Analysis.** Genes that are part of the canonical Wnt pathway are in bold (LEF1, AXIN2, CCND1, RAC1, CTNNB1, and LRP6). Z-scores and their respective ranks for three different siRNAs designed against these genes are displayed. Scores and ranks in the top 10% are in bold, and would be considered hits. Some z-scores are not present (--) because the library used in the screen does not contain any siRNAs designed against these genes

Gene	Parameter Estimate from Haystack	p-value	Single 1		Single 2		Single 3	
			Z-score	Rank	Z-Score	Rank	Z-Score	Rank
<b>LEF1</b>	-51.01	2.90E-62	<b>-12.38</b>	<b>749</b>	<b>-14.39</b>	<b>387</b>	-7.47	3450
<b>AXIN2</b>	26.49	3.14E-20	<b>12.85</b>	<b>633</b>	3.15	10696	0.58	17900
SUMO2	34.61	9.77E-15	7.92	3033	<b>9.68</b>	<b>1790</b>	9.24	2075
<b>CCND1</b>	-15.28	4.26E-14	-4.22	8233	<b>-12.54</b>	<b>705</b>	<b>-13.05</b>	<b>591</b>
SLC25A23	-14.72	3.76E-13	-8.88	2300	<b>-9.94</b>	<b>1638</b>	1.92	13800
<b>RAC1</b>	19.37	1.41E-11	-	-	-	-	-	-
CLEC18B	-43.02	1.96E-11	-	-	-	-	-	-
<b>CTNNB1</b>	-27.45	5.02E-11	<b>-45.4</b>	<b>1</b>	-0.15	19344	<b>-34.27</b>	<b>2</b>
POU4F2	17.21	2.59E-10	-	-	-	-	-	-
<b>LRP6</b>	-8.65	6.56E-09	<b>-19.75</b>	<b>68</b>	<b>-25.61</b>	<b>11</b>	<b>-11.65</b>	<b>973</b>

effects span the transcriptome, Haystack can detect biologically relevant genes for which no siRNA exists in our screening library.

## Discussion

Haystack analysis can be applied to any previous siRNA screen of sufficient scale, preferably with normally distributed scores, to confirm hits via an orthogonal metric or to identify genes originally missed due to library composition. Unfortunately, it is not yet standard for raw data from siRNA screens to be published, as microarray expression profiling data are. Consequently, Haystack is available as a download (<http://rna.nih.gov/haystack/Software.zip>) or web-based service (<http://rna.nih.gov/haystack/>) for researchers who have performed screens to analyze their data and confirm hits or find new hits potentially missed by “on-target” analysis.

Finally, although a genome-specific siRNA library would be expected to have relatively poor on-target performance against cells from another organism, a large-scale library would likely have the seed sequence diversity required to perform Haystack analysis on any organism having a RNA interference pathway. Thus Haystack also presents a novel solution for siRNA screens in cells from organisms for which no library is yet available.

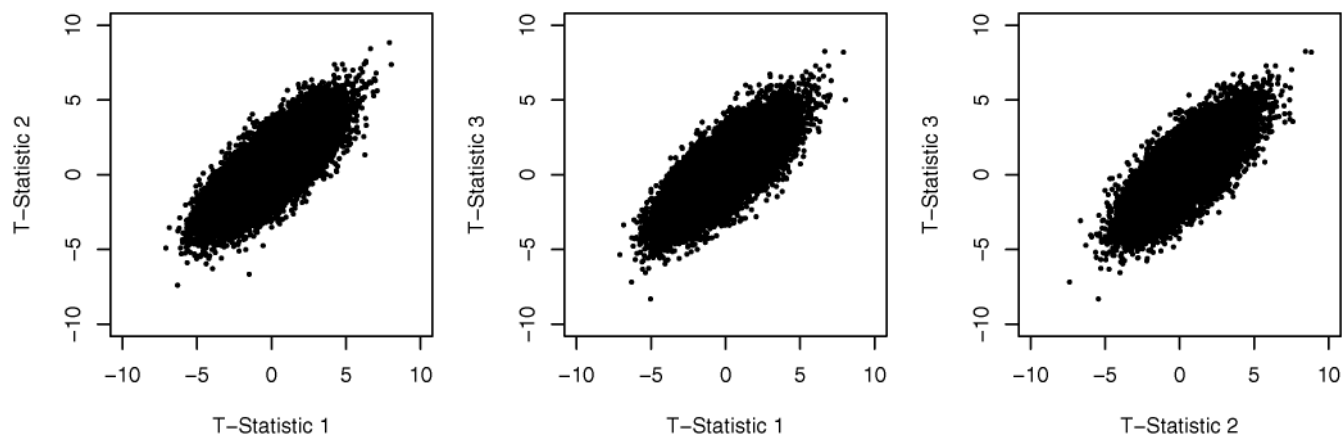
## Methods

**Software used.** Statistical analysis and graphs were generated using R<sup>14</sup>. ROC curves were generated using the ROCR package<sup>15</sup>. The kernel density plot of 3'UTR length distributions was generated using the sm package<sup>16</sup>.

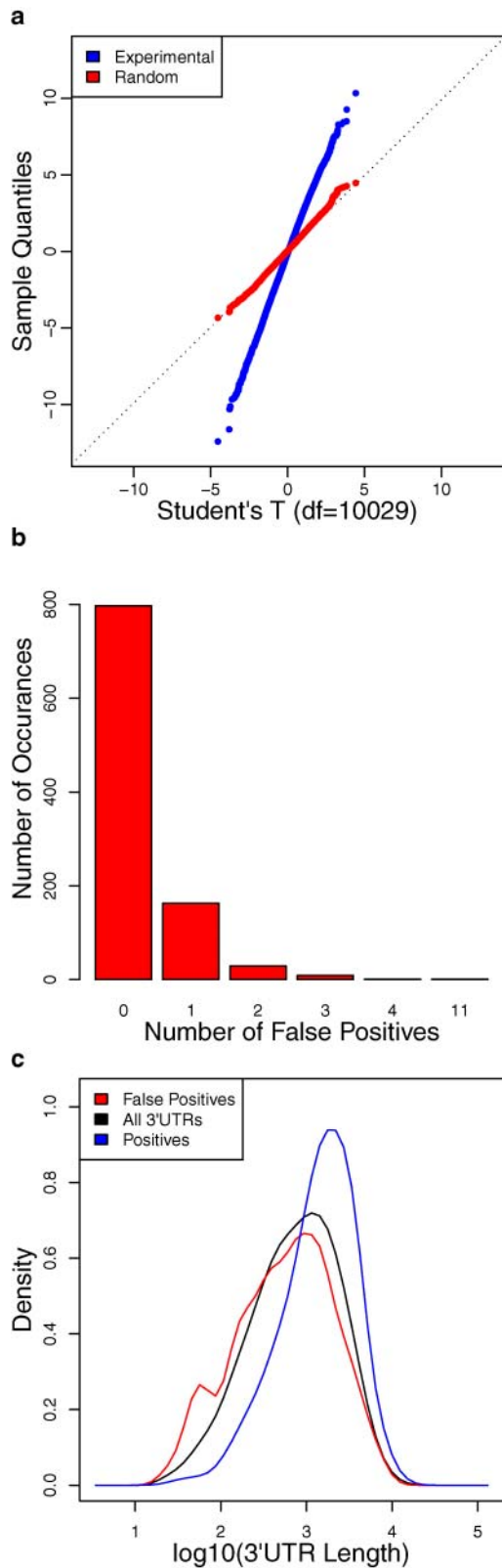
**3'UTR sequences.** Human 3'UTR sequences (a total of 34268 sequences) from the Refseq<sup>17</sup> collection were downloaded using the UCSC table browser<sup>18</sup>. These sequences were matched against the IDs for the gene expression data sets for purposes of training the off-target model. To create the matrix of predicted off-target effects, 170 sequences that were shorter than 20 nucleotides and 2554 sequences that were non-coding (NR prefix in accession) were removed. In addition, 10852 sequences were removed because they were identical to a 3'UTR already included. The remaining 20692 3'UTR sequences were used to create the off-target matrix.

**Gene expression data.** Expression data previously shown to be enriched with transcripts that were down-regulated and contained seed matches to transfected siRNAs in their 3' UTRs<sup>7</sup> were downloaded from the GEO website<sup>19</sup> and used in training and testing the siRNA off-target model.

**Training and testing the off-target model.** Features for predicting siRNA off-target down regulation were selected based on the previously observed importance of the seed sequence<sup>7,20</sup> (bases 2–7 of the guide strand) and the observed preference for adenine opposite from base 1 of the guide strand<sup>21</sup>. Although all possible match types to the seed including one or two base-mismatch were considered, only four match types (PM, M1, M8, M18) were statistically significant predictors of down-regulation. The reverse complement of the guide strand seed sequence followed by an adenine forms a perfect match (PM) and mismatches to base 1 (M1), base 8 (M8), or both (M18) form the four different seed match types selected. In addition, we used the length of the 3'UTR as a predictive variable, since in our analyses longer 3'UTRs were correlated with up-regulation of transcripts in the absence of siRNA seed matches. Linear models predicting the down-regulation of a transcript based on the number of each match type observed in the 3'UTR and the total length of the 3'UTR were trained separately on gene expression data from the transfection of two different siRNAs. These models were then cross-validated on the data not used to train them, and finally the data sets (**Supplementary Data Set 1**) were combined to train a final model (Fig. 2).



**Figure 3 | Reproducibility of t-statistics.** T-statistics calculated in the first step of the method, used to evaluate the null-hypothesis that the predicted off-target effects of the library on a given transcript are not correlated with the screening results. Large t-statistics in either the positive or negative direction indicate that there is significant correlation or anti-correlation respectively between the off-target effects on the transcript and the screening results. These correlations are reproducible across subsets of the data, in this case, across the data from each of the three single siRNAs designed against each gene.



**Figure 4 | Validation of Statistical Framework.** (a) A q-q plot of t-statistics generated by random permutation of the druggable singles data (in red) and non-randomized data (in blue). (b) Incidences and number of false positives in 1000 random permutations the druggable singles data. (c) 3'UTR length distribution for false positives (red) and positives (blue) compared to the overall distribution of 3'UTR lengths (black).

**siRNA libraries.** The custom siRNA libraries used in screening were synthesized by Sigma–Proligo. siRNAs were designed to have sequence asymmetry to increase efficacy and had less than 17 bp of complementarity to other genes<sup>22</sup>. The library contains 6605 sets of 3 siRNAs each designed to theoretically druggable genes.

**siRNA screening data.** Data used to evaluate these methods came from a genome-scale siRNA screen<sup>23,24</sup> in the HT1080 sarcoma cell line for genes involved in the Wnt/ $\beta$ -catenin pathway. Cells were engineered to contain a luciferase reporter coupled to a  $\beta$ -catenin-driven promoter, which was subsequently activated with conditioned media containing Wnt-3a. In primary screening, both a whole genome and a druggable library were screened in pools with at least three replicates. In addition, all single siRNAs contained within the druggable pools were assayed separately. All assay results were normalized to a plate-based non-targeted control to yield a z-score.

**Correlation and statistical significance calculations.** Correlation between the screen results  $z$  (or later the residual) and the predicted off-target effects of the library on a given transcript were calculated using the formula:

$$r_k = \frac{\sum_{n=1}^N z_n O_{n,k} - n\bar{z}\bar{O}_k}{(n-1)s_z s_{O_k}}$$

where  $O_{n,k}$  is our prediction of the off-target down-regulation of each mRNA transcript  $k$  by each siRNA  $n$ . The t-statistic of the null hypothesis that the correlation is actually zero is then calculated as:

$$t_k = \frac{r_k}{\sqrt{(1-r_k^2)/(n-2)}}$$

with the associated p-value of the null hypothesis being derived from the Student's t-distribution with  $n-2$  degrees of freedom, where  $n$  is the number of siRNA screening results.

Since the off-target model makes the same predictions for every siRNA with the same heptamer seed sequence (bases 2–8 of the guide strand) and there are only 16384 possible heptamers, we can pre-compute the predictions for each 3'UTR. Using this simplification, we do not need a separate set of predictions for each siRNA library. To use this pre-computed matrix to calculate Haystack t-statistics for a set of single siRNAs used in a screen, we first calculate the mean assay response for all siRNAs that contain the same seed sequence. We then calculate the weighted correlation coefficient and corresponding t-statistic between the predicted down-regulation caused by each seed sequence and the mean observed response of that seed sequence in the assay (or the residual in later steps), where the weights are the number of occurrences of each seed sequence in the library. A weighted linear regression is performed at each iteration between the selected  $O_k$  and the mean assay response for each seed. Likewise, it is possible to use Haystack analysis on siRNA screens performed using pools by calculating the mean response of the assay for all pools in which one of the component siRNAs has a given seed sequence and then weighting the correlation coefficient and linear regression by the number of pools in the screen containing that seed sequence. PERL and R code for pre-computing off-target predictions from a FASTA formatted file of 3'UTRs and calculating these t-statistics from mean seed sequence responses are supplied from the Haystack website (<http://rnaai.nih.gov/haystack/Software.zip>).

**Validation of statistical techniques.** The validity of using Pearson's correlation coefficient, the t-statistic and corresponding p-value to estimate the statistical significance of non-normally distributed vectors has been a subject of some debate<sup>25</sup>, and seems to depend on the size of the vectors and the particular characteristics of the distributions considered. Although the z-scores generated from this set of screening data are fairly normal in their distribution, the predictions of off-target effects to which we are comparing them are non-normal. Therefore, we tested how well the Student's T distribution was approximating the distribution expected at random.

Z-scores from the druggable singles collection were randomly permuted against their original heptamers. Fig. 4a shows a q-q plot of the t-statistics resulting from the random permutation (in red) and the distribution observed for the original non-permuted data (in blue) compared to the Student's T-distribution with 10029 degrees of freedom (which is the length of the vectors for which the null hypothesis of zero correlation is being tested, minus 2). As expected, the quantiles from our randomly permuted data track well against the Student's t-distribution, while the non-permuted data does not.

We next performed 1000 random permutations of the same data set and noted transcripts for which the Bonferroni corrected p-value (calculated from the t-statistic) was less than 0.01. Fig. 4b shows the distribution of false positives observed in 1000 trials. Although the majority of permutations did not result in any false positives, 20.3% of the permutations resulted in one or more false positives, which is significantly greater than the 1% of false positives that we would expect if Student's t-distribution perfectly modeled the null hypothesis we are testing. The mean number of false positives per random permutation was 0.263. In comparison to the 1217 putative positives observed for the original, non-permuted data, we judge the risk of false positives to be small and error in this approximation to be acceptable, given that the alternative would be to use non-parametric statistics or random permutation to estimate p-values, either of which would significantly reduce our sensitivity and increase the already significant computation time required.



To further investigate the source of false positives, we plotted the distribution of 3'UTR length for the false positives identified in the 1000 random permutations versus all transcripts analyzed and those transcripts identified as positives in the non-permuted data (Fig. 4c). There is a significant shift towards smaller 3'UTRs in the set of false positives compared to the general population. This may indicate that the approximation of the Student's T distribution is slightly less valid for shorter transcripts. Similarly, transcripts identified as positives in the non-permuted set are shifted towards longer 3'UTRs than the set of all 3'UTRs. We believe this may be the case because true positives with short 3'UTRs have a smaller "off-target signature", being down-regulated by fewer siRNAs off-target and thus presenting less signal for detection by the method. It is also possible that short 3'UTRs represent incorrect or incomplete sequences. Thus we conclude that transcripts with short 3'UTRs may be more likely to be both false positives and false negatives in Haystack analysis.

- Mohr, S., Bakal, C. & Perrimon, N. Genomic Screening with RNAi: Results and Challenges. *Annual Review of Biochemistry* **79**, 37–64 (2010).
- Echeverri, C. J. *et al.* Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nat Methods* **3**, 777–779 (2006).
- Birmingham, A. *et al.* 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nat Methods* **3**, 199–204 (2006).
- Ma, Y., Creanga, A., Lum, L. & Beachy, P. A. Prevalence of off-target effects in *Drosophila* RNA interference screens. *Nature* **443**, 359–363 (2006).
- Jackson, A. L. & Linsley, P. S. Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nature reviews* **9**, 57–67 (2010).
- Hammell, M. Computational methods to identify miRNA targets. *Seminars in cell & developmental biology* **21**, 738–744 (2010).
- Jackson, A. L. *et al.* Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. *Rna* **12**, 1179–1187 (2006).
- Khan, A. A. *et al.* Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. *Nat Biotechnol* **27**, 549–555 (2009).
- Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research* **38**, D355–D360 (2010).
- Sachdev, S. *et al.* PIASy, a nuclear matrix-associated SUMO E3 ligase, represses LEF1 activity by sequestration into nuclear bodies. *Genes Dev* **15**, 3088–3103 (2001).
- Shen, L. N., Geoffroy, M. C., Jaffray, E. G. & Hay, R. T. Characterization of SENP7, a SUMO-2/3-specific isopeptidase. *Biochem J* **421**, 223–230 (2009).
- Xiang, M. *et al.* Role of the Brn-3 family of POU-domain genes in the development of the auditory/vestibular, somatosensory, and visual systems. *Cold Spring Harb Symp Quant Biol* **62**, 325–336 (1997).
- Budhram-Mahadeo, V. S. & Latchman, D. S. Targeting Brn-3b in breast cancer therapy. *Expert Opin Ther Targets* **10**, 15–25 (2006).
- R Development Core Team, R. F. F. S. C. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* **1**, 2673 (2008).
- Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCr: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
- Bowman, A. W. & Azzalini, A. Computational aspects of nonparametric smoothing with illustrations from the sm library. *Computational Statistics & Data Analysis* **42**, 545–560 (2003).
- Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61–65 (2007).
- Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493–496 (2004).
- Barrett, T. *et al.* NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Research* **37**, D885–D890 (2009).
- Lin, X. *et al.* siRNA-mediated off-target gene silencing triggered by a 7 nt complementation. *Nucleic Acids Research* **33**, 4527–4535, doi:10.1093/nar/gki762 (2005).
- Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
- Majercak, J. *et al.* LRRMT3 promotes processing of amyloid-precursor protein by BACE1 and is a positional candidate gene for late-onset Alzheimer's disease. *Proc Natl Acad Sci U S A* **103**, 17967–17972 (2006).
- Major, M. B. *et al.* New regulators of Wnt/beta-catenin signaling revealed by integrative molecular screening. *Sci Signal* **1**, ra12 (2008).
- Tang, W. *et al.* A genome-wide RNAi screen for Wnt/beta-catenin pathway components identifies unexpected roles for TCF transcription factors in cancer. *Proc Natl Acad Sci U S A* **105**, 9697–9702 (2008).
- Kowalski, C. J. On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **21**, 1–12 (1972).

## Acknowledgements

Thanks to Christina Leslie of Sloan-Kettering Institute, Steven Bartz, Peter Linsley and Alan Sachs of Merck & Co., Inc. for their guidance, encouragement and support.

## Author contributions

EB conceived, developed and tested the algorithm, software, and related statistics, wrote the original manuscript and produced all the supplementary materials. AAK helped significantly revise the manuscript and made Figure 1, as well as testing alternative methods of feature selection. SM generated all of the screening data used. EB and JB developed the mathematical model of siRNA off-targeting. SM and MF provided feedback on biological aspects of the manuscript. MF supervised and guided the siRNA screen. MR explored alternative methods of deriving parameters for the off-target model.

## Additional information

**Supplementary information** accompanies this paper at <http://www.nature.com/scientificreports>

**Competing financial interests:** The authors declare no competing financial interests.

**License:** This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/>

**How to cite this article:** Buehler, E. *et al.* siRNA off-target effects in genome-wide screens identify signaling pathway members. *Sci. Rep.* **2**, 428; DOI:10.1038/srep00428 (2012).