Statistical Approaches for Off-Target Effects Identification and Corrections in High-throughput RNAi Screenings

Rui Zhong[1], Guanghua Xiao[1], Michael White[3], Yang Xie[1, 2]
[1] Quantitative Biomedical Research Center, Department of Clinical Science
[2] Harold C. Simmons Comprehensive Cancer Center
[3] Department of Cell Biology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390

Abstract

High-throughput RNA-interference (RNAi) screening has been widely used in biological research, like discovery of unknown molecular machinery and identification of novel drug-targetable genes. However, off-target effects make interpretation blurry. Here we develop a novel computational approach to identify off-targeting siRNA oligos based on miRNA-mimic mechanism. Genome-wide siRNA oligos are classified into different seed families based on their seed sequences. We used KS (Kolmogorov–Smirnov) test to determined enrichment for each seed family. We modeled each Z score as a linear combination of seed families' off-target effects and on-target effects, and then estimated the off-target using penalized regression with LASSO (least absolute shrinkage and selection operator) penalty term. Using the modeling approach, we could adjust off-target effects from the original Z scores and our results showed that corrected Z scores improved accuracy in hits selection. In a real data application to identify selective autophagy factors, our method led to hits with higher confirmation rate in the secondary confirmation screening.

Key Words: High-throughput RNAi screening, off-target effects, LASSO, KS test.

1. Introduction

Genome-wide RNAi knockdown experiments reveal potential phenotype-associated genes. In recent years, high-throughput RNAi screening has been widely used as power tools in a wide spectrum of biomedical research, such as characterizing unknown molecular machinery (Orvedahl & Sumpter 2011) and identifying novel therapeutic targets (Whitehurst & Bodemann 2007, Moreau & Kumar 2011, Toyoshima & Howie 2012). However, results from RNAi screening are often hampered by off-target effects (Buehler & Khan 2012, Kulkarni & Booker 2006). Previous research showed siRNA oligos might cause off-target effect through miRNA mimics (Birmingham & Anderson 2006).

Figure 1 shows the mechanism of both on-target and off-target effects. Accordingly, after siRNA transfection, single-strand siRNA will form with factors like Ago into RISC (RNA-induced silencing complex). As designed, RISC should bind to targeted genes through perfect match; however it may also off-target other genes through seed match on 3' untranslated regions. This complicated the interpretation of high-throughput RNAi screening results, since it is difficult to distinguish between on-target and off-target effects to identify true 'hits' of the screening. Therefore, it is important to identify such off-target effects from the screening results. Here, we present a novel computational approach to identify and correct off-target effects in high-throughput RNAi screenings.

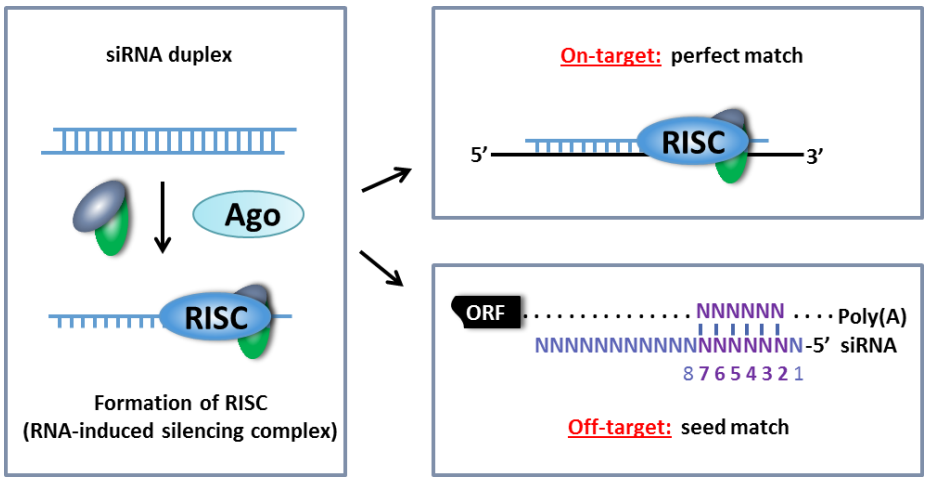

Figure 1: Graphical illustration of on-target and off-target effects: single-strand siRNA might target genes through either perfect match (on-target) or seed match (off-target).

We performed a cell-based image RNAi screening for genes required for the co-localization of virus capsid protein with autophagolysosomes(Orvedahl & Sumpter 2011) to expand our knowledge of autophagy regulation. In our screening, each pool contained 4 double-strand oligos designed to target one single gene, and had an associated Z score, derived from statistical analysis. A lower Z score is a phenotype of interest, indicating knockdown of the gene by siRNA could affect autophagy regulation. Based on the miRNA-mimic hypothesis, we defined the 6-mer from 2 to 7 on the 5' region of oligo sequences as seed sequence. siRNA oligos that share the common seed sequence are grouped into a seed family. In our study, we used the sequences of both strands of a siRNA oligo. Table 1 demonstrates a seed family GCAUGG. In the table, symbols indicate designed targets for each individual siRNA oliog which is assigned with a Z score. siRNA sequences showed that they all shared the common seed sequence GCAUGG on defined seed region. Figure 2 plotted cumulative Z scores of this seed family against background and illustrated that seed family Z scores dramatically deviate from the background, which indicated this seed family might cause off-target effects in our screen.

Table 1: Detailed example of seed family GCAUGG

Symbol	*siRNA sequence*	*Z score*
CXCR7	U**GCAUGG**CCAGCUGAUGUC	-6.07
C1orf210	G**GCAUGG**CCUGUGCUGGUA	-6.85
NMT1	U**GCAUGG**UCAUAUUUCUGC	-5.83
LOC387721	G**GCAUGG**AGUCCUAGGAAA	-3.98
MAPRE2	U**GCAUGG**GUUAAUGACAUA	-3.00
C9orf43	C**GCAUGG**UCAUAGUAGUUC	-1.70
FBXO11	G**GCAUGG**GUUACUUUGAAA	-1.70
CYP26B1	U**GCAUGG**UCAUCUCCUUCC	-1.61
JPH2	G**GCAUGG**GCUGGGCAUAGA	-1.28
BAHD1	A**GCAUGG**GAAGGGACUUUC	-0.05

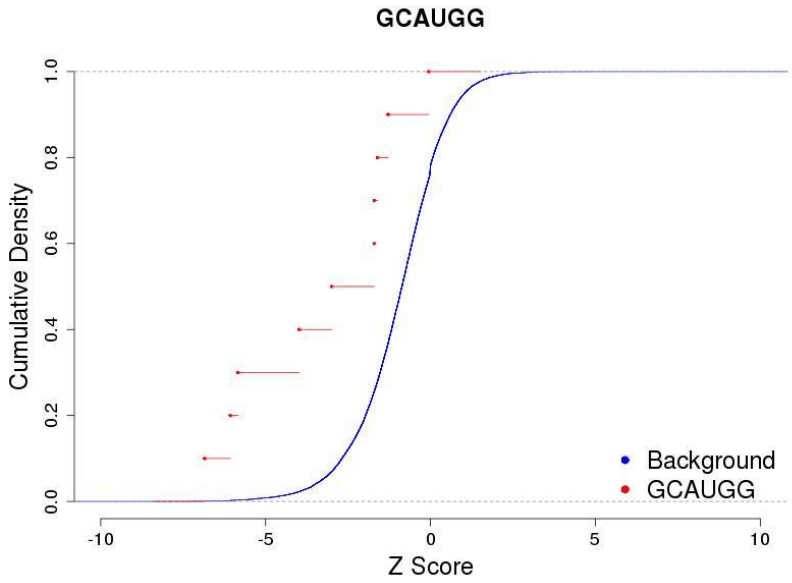

Figure 2: Comparison of cumulative Z scores between seed family GCAUGG and background.

2. Models and methods

2.1 LASSO

In our analysis, we model each Z score as a linear combination of on-target effect and off-target effects due to seeds matching, and perform LASSO regression as below:

$$\vec{Z} = X\vec{\beta} + \vec{\varepsilon}\ ,\ \text{subject to}\ \left|\vec{\beta}\right| < s$$

where Z_i is the i^{th} original Z score, β_j is the estimated off-target effect of the j^{th} seed family, ε_i is the i^{th} corrected Z score (on-target effect) and λ is the penalty parameter. X is denoted as below:

$$X = \left[x_{ij}\right],\ x_{ij} = \begin{cases} 1, & \text{if}\quad \text{i} \in \text{j} \\ 0, & \text{otherwise} \end{cases}$$

And the solution is given:

$$\beta = \arg\min_{\beta}\left[\left\|\vec{Z} - X\vec{\beta}\right\|^2 + \lambda\left|\vec{\beta}\right|\right]$$

After modeling using LASSO, for each seed family, we can estimate the coefficient which indicates the strength and direction of off-target effects. In our analysis, based on empirical experience, λ is set to 0.001, and for cut-off of coefficient, we determine those whose absolute value is bigger than 1 represent strong enough off-target effects. A negative coefficient means the seed family tends to lower Z score and vice versa.

2.2 KS test

After selection of strong off-target seed families, we also perform KS test to examine the statistical significance. As LASSO is already a stringent method and therefore we set KS test p value cut-off to 0.01 so that we can be confident that selected off-target seed families are both statistically and biologically significant.

3. Results

3.1 Identification of off-target seed families

When applying traditional statistical approach such as KS test, seed families are analyzed one by one individually. However, in high-throughput RNAi screenings, it is quite conventional to include 3 or 4 double-strand siRNAs in one pool, which complicates the analysis in identification of off-target seed families since a Z score might be the consequences of both on-target and multiple off-target effects. By contrast, LASSO is able to estimate those off-target effects for all seed families simultaneously, avoiding those problems caused by individual analysis. Moreover, KS test can't estimate the strength of off-target effects which is even more biologically important. In identification of off-target seed families, we set cutoff -1 as a threshold for off-target effect and p value 0.01 as discussed in models and methods shown in Figure 3.

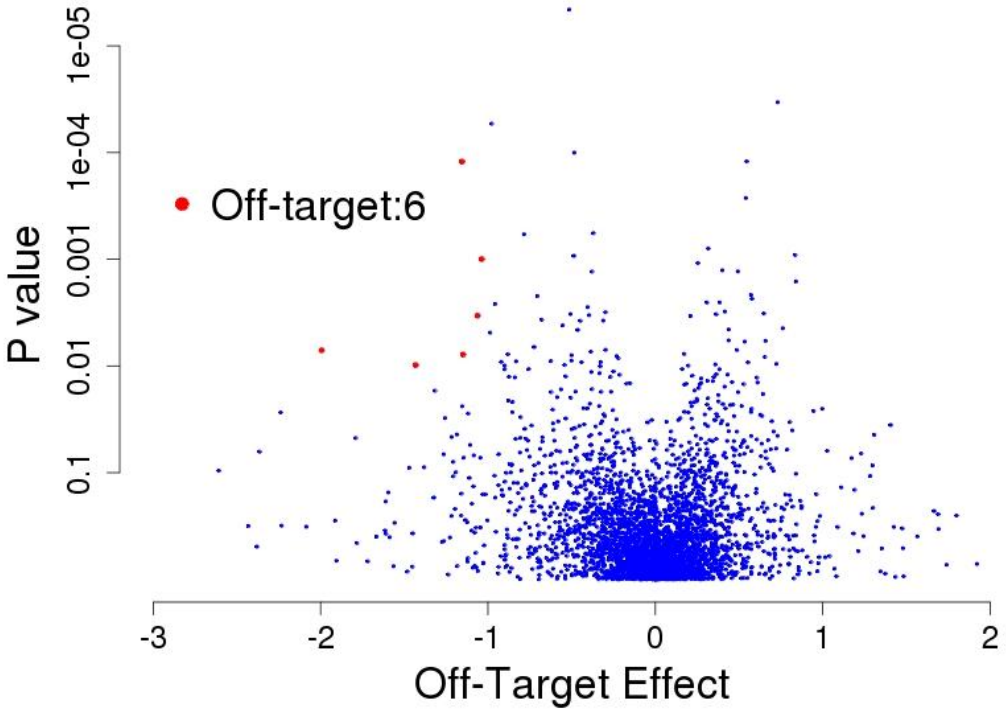

Figure 3: Volcano plot of off-target effects: each dot represents a seed family and x-axis is LASSO-estimated off-target effect and y-axis is associated KS test P values. Red dots are identified off-target seed families which are both biologically and statistically significant.

3.2 Secondary screening validation

In the primary screening, each well contains 4 oligos designed to target the same genes. In our study, we also included a secondary screening in which 195 genes were retested using 4 separate individual oligos. Thus we are able to evaluate the performance of identified off-target seed families. In the secondary screening, individual oligos were used to knock down genes. For those individual oligos belonging to identified off-target seed families, their Z scores were lower than the other oligos designed to target the same

genes by 3 (Figure 4). Thus we know the difference is due to off-target effects instead of on-target effects.

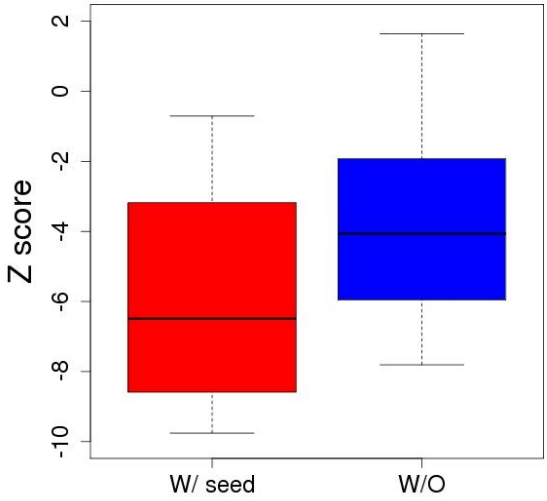

Figure 4: Boxplot showing the activity of siRNAs containing identified off-target seeds (red) by side with the activity of other siRNAs containing different seeds but targeting the seed genes (blue).

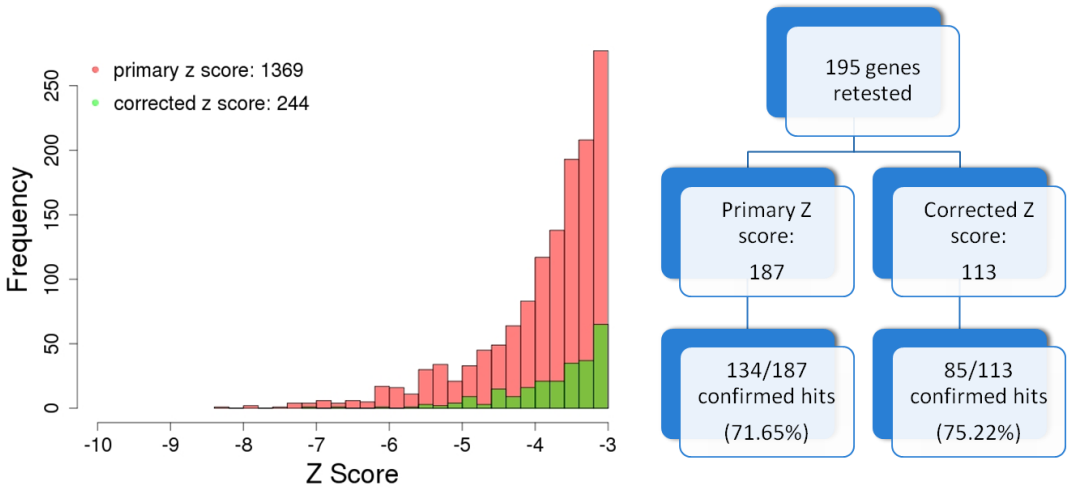

Figure 5: Corrected Z score. LASSO can estimate off-target effect of each seed family and remove them from the original Z scores. Therefore we can have off-target-free corrected Z scores for hits selection.

3.3 Corrected Z score

In identification of off-target effects in high-throughput RNAi screening, final goal is to remove such effects and improve our confirmation rate. Based on LASSO estimate, we can easily remove these off target effects. After LASSO estimation, we can remove those off-target effects, correct original Z scores and dramatically reduce number of acclaimed hits from 1369 to 244. And in the secondary screening in which 195 genes were retested, corrected Z score has an even higher confirmation rate 75.22%, compared with the original 71.65% (Figure 5).

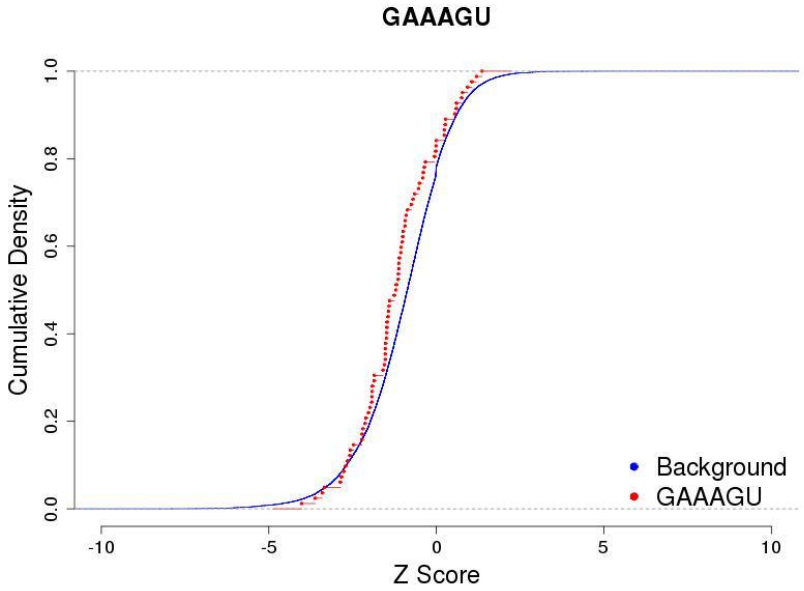

Figure 6: Cumulative density plot of seed family GAAAGU against background. Shown in this plot, seed family GAAAGU has a significant KS p value though; it is not biologically interesting because the strength of off-target effect is weak.

4. Discussion

Our analysis showed that LASSO can be used as a method to identify off-target seed families in high-throughput RNAi screenings. In our analysis, it is not necessary to make any pre-cutoff for Z score like in hypergeometric test, and therefore we can evaluate the overall off-target effect of a seed family as a whole.

To our knowledge, we first apply LASSO to identify off-target effects in high-throughput RNAi screening and prove that it is better than traditional statistical approaches (Sudbery & Enright 2010, Marine & Bahl 2012). Figure 6 showed a seed family GAAAGU. It has a significant KS test P value; however cumulative density plot indicated this is not a biologically interesting off-target seed family since the difference between seed family Z scores and background are trivial. That is because it has a big family size and makes itself easier detected by KS test. Therefore, LASSO-identified off-target seed families proved better accuracy. Off-target seed families identified by LASSO were confirmed in a secondary screening where individual oligos were used and showed manifest off-target effects. Since LASSO is able to estimate off-target effects globally

and therefore is more powerful to correct Z scores while maintaining an even better confirmation rate.

Acknowledgements

This work was supported by NIH 5R01CA152301, 1R33DA027592 and NSF DMS-0907562.

References

Birmingham, A., E. M. Anderson, A. Reynolds, D. Ilsley-Tyree, D. Leake, Y. Fedorov, S. Baskerville, E. Maksimova, K. Robinson, J. Karpilow, W. S. Marshall&A. Khvorova (2006). "3' UTR seed matches, but not overall identity, are associated with RNAi off-targets." *Nat Methods* **3**(3): 199-204.

Buehler, E., A. A. Khan, S. Marine, M. Rajaram, A. Bahl, J. Burchard&M. Ferrer (2012). "siRNA off-target effects in genome-wide screens identify signaling pathway members." *Sci Rep* **2**: 428.

Kulkarni, M. M., M. Booker, S. J. Silver, A. Friedman, P. Hong, N. Perrimon&B. Mathey-Prevot (2006). "Evidence of off-target effects associated with long dsRNAs in Drosophila melanogaster cell-based assays." Nat Methods 3(10): 833-838.

Marine, S., A. Bahl, M. Ferrer&E. Buehler (2012). "Common seed analysis to identify off-target effects in siRNA screens." J Biomol Screen 17(3): 370-378.

Moreau, D., P. Kumar, S. C. Wang, A. Chaumet, S. Y. Chew, H. Chevalley&F. Bard (2011). "Genome-wide RNAi screens identify genes required for Ricin and PE intoxications." Dev Cell 21(2): 231-244.

Orvedahl, A., R. Sumpter, Jr., G. Xiao, A. Ng, Z. Zou, Y. Tang, M. Narimatsu, C. Gilpin, Q. Sun, M. Roth, C. V. Forst, J. L. Wrana, Y. E. Zhang, K. Luby-Phelps, R. J. Xavier, Y. Xie&B. Levine (2011). "Image-based genome-wide siRNA screen identifies selective autophagy factors." Nature 480(7375): 113-117.

Sudbery, I., A. J. Enright, A. G. Fraser&I. Dunham (2010). "Systematic analysis of off-target effects in an RNAi screen reveals microRNAs affecting sensitivity to TRAIL-induced apoptosis." BMC Genomics 11: 175.

Toyoshima, M., H. L. Howie, M. Imakura, R. M. Walsh, J. E. Annis, A. N. Chang, J. Frazier, B. N. Chau, A. Loboda, P. S. Linsley, M. A. Cleary, J. R. Park&C. Grandori (2012). "Functional genomics identifies therapeutic targets for MYC-driven cancer." Proc Natl Acad Sci U S A.

Whitehurst, A. W., B. O. Bodemann, J. Cardenas, D. Ferguson, L. Girard, M. Peyton, J. D. Minna, C. Michnoff, W. Hao, M. G. Roth, X. J. Xie&M. A. White (2007). "Synthetic lethal screen identification of chemosensitizer loci in cancer cells." Nature 446(7137): 815-819.