TACKLING COMPUTATIONAL CHALLENGES IN HIGH-THROUGHPUT RNA INTERFERENCE SCREENING

APPROVED BY SUPERVISORY COMMITTEE

Yang Xie, Ph.D (Mentor)

Guanghua Xiao, Ph.D (Mentor)

John Minna, MD (Chair)

Jerry Shay, Ph.D

Michael White, Ph.D

DEDICATION

I would like to thank both my advisors, Drs. Yang Xie and Guanghua Xiao. Without their mentoring and support, I couldn't have finished any of my accomplishments during graduate study. My sincere appreciation also goes to the members of my thesis committee, Drs. John Minna, Jerry Shay and Michael White. They provided me with very insightful recommendations on research as well as on presentation skills. Further, I am very grateful that I could collaborate with such excellent scientists and clinicians. Last but not least, I am really thankful that both my parents are supportive of my decision to pursue a Ph.D degree in the US since I am an only child and this means we have to be separated tens of thousands of miles away from each other.

TACKLING COMPUTATIONAL CHALLENGES IN HIGH-THROUGHPUT RNA INTERFERENCE SCREENINGS

by

RUI ZHONG

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

May, 2014

TACKLING COMPUTATIONAL CHALLENGES IN HIGH-THROUGHPUT RNA INTERFERENCE SCREENING

Rui Zhong, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2014

Supervising Professor: Yang Xie, Ph.D. & Guanghua Xiao, Ph.D.

Since the discovery of RNAi decades ago, it has been increasingly used in biomedical and biological research. The success of analyzing single genes using siRNAs has resulted in the large-scale application of RNAi for genome-wide loss-of-function phenotype screening while reducing cost and decreasing time. High-throughput RNAi screening (HTS) has been widely accepted and used in a variety of biomedical and biological research projects as the first step to identifying novel drug targets or pathway components. Huge data sets are being generated, but computational challenges remain in data analysis and hit identification, which have become

hurdles in HTS. These must be tackled before we can more accurately and precisely interpret the HTS results, since they are often blurred by spatial noise and off-target effects.

In my thesis research, I have been working on statistical modeling of high-throughput RNAi screening results. I developed SbacHTS (spatial background noise correction in high-throughput RNAi screening) to identify and remove spatially-correlated background noise from HTS, which helps enhance statistical detection power in triplicate experiments. On top of that, I also created a novel algorithm, DeciRNAi (deconvolution analysis high-throughput RNAi screening results), to quantify the strength and direction of siRNA-mimic-miRNA off-target effects in HTS projects. As a special case, image-based high-content HTS requires management of high-dimensional data analysis and visualization. I built a new R package "iScreen" (image-based high-throughput RNAi screening analysis tools) to deal with such problems.

TABLE OF CONTENTS

PRIOR PUBLICATIONS

Rui Zhong, Guanghua Xiao, Michael White and Yang Xie. Statistical Approaches for Off-Target Effects Identification and Corrections in High-throughput RNAi Screenings. In *JSM Proceedings*, Biometrics Section. 2012. 276-282.

Rui Zhong, Min Soo Kim, Yang Xie Michael White and Guanghua Xiao. SbacHTS: spatial background noise correction for high-throughput RNAi screening. Bioinformatics 2013; 29(17): 2218-2220 (Peer-reviewed).

Rui Zhong*, Ji Mi Kim*, Guanghua Xiao, Michael White and Yang Xie, et al. Computational Detection and Suppression of Sequence-specific Off-target Phenotypes from Whole Genome RNAi Screens. NAR (Revised; Peer-reviewed).

Rui Zhong*, Jeffrey D. Allen*, Guanghua Xiao and Yang Xie. Ensemble-based Network Aggregation Improves The Accuracy Of Gene Network Reconstruction. PLOS ONE (Revised; Peer-reviewed).

Mukesh Bansal, Jichen Yang, Charles Karan, Michael Patrick Menden, James C. Costello, Hao Tang, Guanghua Xiao, Yajuan Li, Jeffrey Allen, Rui Zhong, Yang Xie, Gustavo Stolovitzky, and Andrea Califanol. Predicting Activity Of Drug Combinations Through Crowdsourcing. Nature Biotechnology (Revised; Peer-reviewed).

Jichen Yang, Hao Tang, Jeffrey D. Allan, Rui Zhong, and Yang Xie, et al. DIGRE: Drug Induced Genomic Residual Effect Model for Successful Prediction of Multidrug Effects. CPT: Pharmacometrics & Systems Pharmacology (Accepted; Peer-reviewed).

Zhenyi An, Amina Tassa, Collin Thomas, Rui Zhong, Guanghua Xiao, Benjamin Tu, Daniel Klionsky and Beth Levine. Autophagy Is Required For G1/G0 Quiescence In Response To Nitrogen Starvation In Saccharomycescerevisiae. Autophagy (Under review; Peer-reviewed).

* co-first authors

LIST OF FIGURES

LIST OF TABLES

LIST OF FORMULAS

LIST OF ABBREVIATIONS

AGO	Argonaute
ANOVA	analysis of variance
AUC	area under curve
BIC	Bayesian information criterion
bp	base pair
CDF	cumulative density functions
CML	Chronic myelogenous leukaemia
CV	coefficient of variation
DecoRNAi	deconvolution analysis of high-throughput RNAi screening phenotype
DNA	deoxyribonucleic acid
dsRNA	double-stranded RNA
EM	expectation maximization
esiRNA	endonuclease-prepared siRNA
FDR	false discovery rate
GM	Gaussian Mixture
HBV	hepatitis B virus
hc	high-value
HCS	high content screening
HCV	hepatitis B virus
HIV	human immunodeficiency virus
HTS	high-throughput RNAi screen

iScreen	image-based high-throughput RNAi screening analysis tool
KS	Kolmogorov-Smirnov
LASSO	least absolute shrinkage and selection operator
lc	low-value
LOF	loss-of-function
MAD	median absolute deviation
miRNA	microRNA
mRNA	messenger RNA
nt	nucleotide
OTE	off-target effect
oligo	oligonucleotide
PKR	dsRNA-dependent protein kinase
RISC	RNA-induced silencing complex
RNAi	RNA interference
RNA	ribonucleic acid
ROC	receiver operating characteristic
RSA	redundant siRNA activity
SbacHTS	spatial background noise correction in high-throughput RNAi screening
SD	standard deviation
shRNA	short hairpin RNA
siRNA	small-interfering RNA
SNP	single nucleotide polymorphism

SSMD	strictly standardized mean difference
SVM	support vector machine

CHAPTER ONE

INTRODUCTION

The discovery of RNA interference (RNAi) has opened up a wide spectrum of biomedical and biological research, including investigation into the mechanism and function of RNAi, loss-of-function (LOF) annotation, identification of drug targets and novel therapeutics. Both RNAi itself and its applications have been of great interest to academia and industry, especially in applications, where the success of analyzing single genes using RNAi technology has led to efforts to apply it on a large scale. Therefore we can now perform genome-wide high-throughput RNAi screening (HTS) for genomic functional annotation. RNAi has been widely used in studies such as the identification of chemo-therapeutic drug sensitizers or novel component of a specific pathway. Recently, advances in microscopy have made it possible to employ high-content screening that captures multiple phenotypes after gene knock-down that increase the complexity of analysis. With the maturation of experimental technology, data analysis and modeling have become a time-limiting step in such HTS projects.

1.1 Small interfering RNA

Gene silencing caused by the injection of double-stranded RNA (dsRNA), whose sequence is complementary to that of the targeted gene, was discovered almost three decades ago (Fire, 1998; Fire, et al., 1991; Guo and Kemphues, 1995). It has been shown that the RNA interference (RNAi) pathway presents itself universally across most eukaryotes (Hannon, 2002). After entering into cells, dsRNA is processed into small interfering RNA (siRNA) of about 22

nucleotides, which is further cut into single-stranded RNA. Single-stranded siRNAs are then incorporated into the RNA-induced silencing complex (RISC) to target messenger RNA (mRNA) via a perfect complementary match (Figure 1), which we call the "on-target" effect.

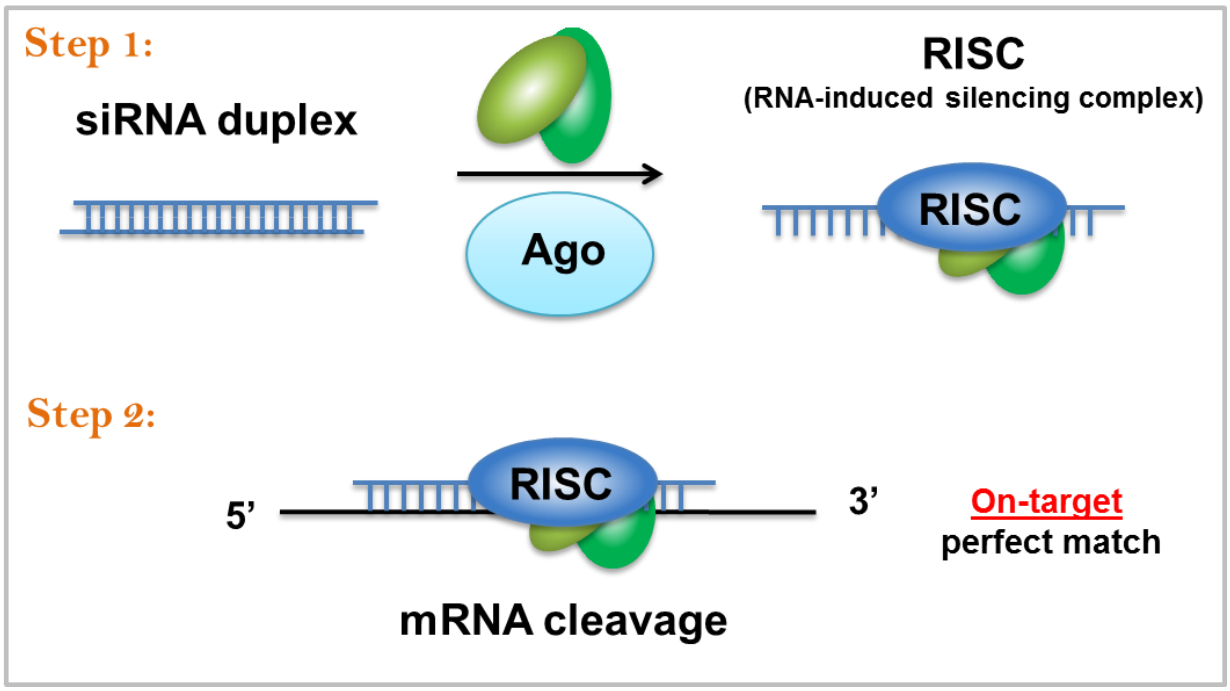


Figure 1. Gene silencing via siRNAs. After injection and transfection, double-stranded RNA (dsRNA) is cut into single-stranded small interfering RNA (siRNA) by RNase enzyme Dicer. siRNAs are then integrated with proteins like Ago to form an RNA-induced silencing complex (RISC) to target complementary messenger RNA (mRNA) via perfect sequence match, which we call "on-target" for short.

1.1.1 Applications of RNAi in mammalian cell lines

As exogenous RNA molecules, dsRNA faced considerable barriers before it could be applied to generate gene silencing in a gene-sequence-specific pattern without triggering a nonspecific response to exogenous dsRNAs, such as the antiviral mechanism. By binding to dsRNA, the enzyme dsRNA-dependent protein kinase (PKR) is activated and localized, resulting

in generalized inhibition of protein synthesis in a sequence-dependent manner (Williams, 1997). However, observations have been made that the RNA interference (RNAi) pathway is conserved across mammals, and nonspecific antiviral responses are not prevalent in mammalian cell lines (Bernstein, et al., 2001; Hammond, et al., 2001; Hannon, 2002; Williams, 1997). Developments have been made to employ RNAi as a genetic tool in mammalian cell lines and animals for functional annotation and drug target discovery.

Short dsRNAs of less than 30 base pairs (bp) have been used to produce RNAi phenotypes in sequence-specific patterns. Evidence has been observed that short dsRNAs do not effectively trigger a PKR-induced antiviral response in mammalian cell lines (Elbashir, 2001), which has led to the use of siRNAs as an RNAi tool for gene silencing in mammalian cell lines and the commercial availability of its applications as a toolkit.

A variety of standard transfection protocols and methods have been made available for siRNA introduction into mammalian cell lines, the strength and duration of which is affected by a number of factors such as efficiency of transfection, concentrations, specificity of sequence design, tissue of cell line specificity and transfection reagents (Hannon and Rossi, 2004).

1.1.2 Applications of RNAi in virus diseases treatment

As a solution to mammalian genetics tools, siRNA has been widely used in biomedical and biological research projects across different biological models such as C. elegans, mice and human cell lines (Hannon and Rossi, 2004). It was first demonstrated to induce RNAi repression in an adult animal in mouse liver via a luciferase reporter gene construction (McCaffrey, 2002).

Recently siRNAs have been proposed as potential novel therapeutics themselves based on their potency and specificity in determining which gene is knocked down, spanning from oncogenes to growth factors and single nucleotide polymorphisms (SNP) (Hannon and Rossi, 2004). For example, in defense against the human immunodeficiency virus (HIV) infection, as a new antiviral therapeutic method RNAi has been used to silence some early and late HIV-encoded RNAs, including the trans-activation response (TAR) element (Jacque, et al., 2002), tat (Coburn and Cullen, 2002; Lee, 2002; Surabhi and Gaynor, 2002), rev (Coburn and Cullen, 2002; Lee, 2002), gag (Novina, 2002; Park, 2002), and so on.

Another major health problem that could conceivably benefit from RNAi-inspired treatments is hepatitis induced by the hepatitis B virus (HBV) and hepatitis C virus (HCV), which affect millions of patients worldwide since treatment for HBV is still unsatisfactory and no vaccine is available for HCV (Hannon and Rossi, 2004). The Huh-7 human hepatoma-derived cell line has been established to study the mechanisms of liver cancer and treatment (Blight, et al., 2000; Ikeda, et al., 2002; Lohmann, 1999; Pietschmann, et al., 2001). Several groups have attempted to investigate the possibility of using siRNA to inhibit virus replication in this system and have seen some success (Kapadia, et al., 2003; Randall, et al., 2003; Wilson, 2003). In a mouse study, siRNA-treated mice survived longer than the control group. (Song, 2003).

1.1.3 Applications of RNAi in cancer treatment

Since the discovery of RNAi, scientists and researchers have been holding out hope that it would benefit treatments for cancer, which is a chronic health problem that kills tens of thousands each year. Great efforts have been made toward this goal, and a number of ongoing

projects are focused on delivering siRNA in cancer samples and cell lines for the purpose of killing cancer cells in a cell-line specific pattern (Buchele, 2003; Kittler and Buchholz, 2003; Lu, et al., 2003; Wall and Shi, 2003).

In order to achieve the highly efficient delivery of siRNA into cells, backbone modifications have been developed for synthetic siRNAs such that the half-life has been extended (Chiu and Rana, 2003; Czauderna, 2003). Chronic myelogenous leukemia (CML) is one example of a successful case where siRNA has proven useful as an anti-cancer therapeutic agent. Acquired drug resistance in CML treatments also necessitates the exploration of alternative novel therapeutics (Cowan-Jacob, 2004; Holtz and Bhatia, 2004; Marcucci, et al., 2003; Tauchi and Ohyashiki, 2004). Although no siRNA clinical trials are yet underway, based on the cell line studies siRNA may have a promising role in improving patient care and inhibiting tumor development.

1.2 MicroRNA

MicroRNA (miRNA) is an endogenous RNA molecule that can also lead to gene silencing via imperfect complementarity between miRNA and mRNA. Mature miRNAs are usually 19-25 nucleotides (nt) long and generated from hairpin-shaped transcripts (Ambros, 2003; Bartel, 2004; Cullen, 2004).

1.2.1 MicroRNA biogenesis

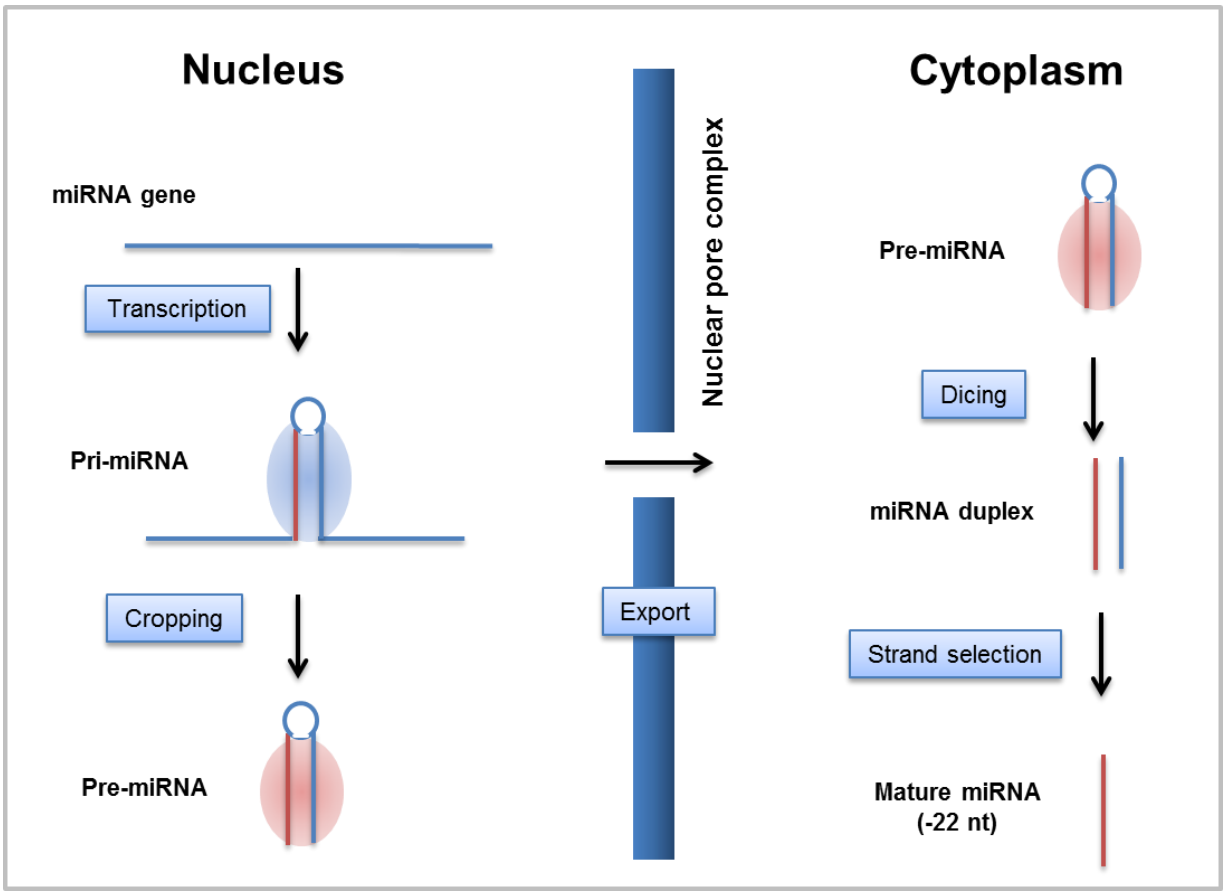


Figure 2. miRNA biogenesis. miRNA genes could be transcribed into pri-miRNA by RNA polymerase II, which is further cropped into pre-miRNA (60-100 nt stem-loop structure). Pre-miRNA is exported from the nucleus into the cytoplasm, followed by the dicing of pre-miRNA via enzyme Dicer into miRNA duplex (22 nt). Usually only one strand is left as mature miRNA while the other is degraded. Adapted from V. Narry Kim, Nature Review, 2005.

Most miRNA genes are located in intergenic regions and could be transcribed into pri-miRNA by RNA polymerase II, which is further cropped into pre-miRNA (60-100 nt stem-loop structure). Pre-miRNA is exported from the nucleus into the cytoplasm, following which it is diced via enzyme Dicer into miRNA duplex (22 nt). Usually only one strand is left as mature miRNA while the other is degraded (Figure 2) (Kim, 2005).

1.2.2 MicroRNA targeting

Targeting mRNA via miRNA is not yet well understood. At present scientists and researchers believe that miRNA regulation is quite universal in a spatial-temporal pattern. Expression of miRNA is also tissue- or cell-line-specific, which further increases the complexity of miRNA-mRNA targeting.

The mechanism of miRNA-mRNA complementarity is still under investigation. However, it is currently believed that seed match is the major determinant of miRNA-mRNA complementarity. "Seed" usually refers to 2-7 six-mer sequences on the 5' end of single-stranded mature miRNA (Figure 3) (Grimson, et al., 2007).

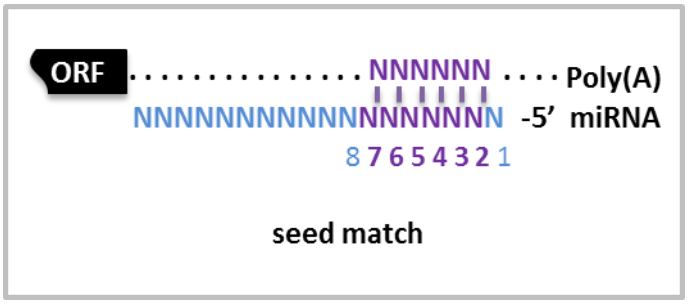


Figure 3. Seed match within miRNA/mRNA complementarity. For miRNA/mRNA complementarity, seed match plays a major role determining miRNA and mRNA base pairings. Seed sequence is conventionally defined as 2-7 six mer on the 5' end of miRNAs.

Sequences beyond seed may also help binding. Target site and AU enrichment are affecting factors, too. The detailed mechanism is still blurred, and therefore target prediction remains a computational challenge in miRNA study. One miRNA might simultaneously target multiple genes or genes within the same pathway.

1.3 Short hairpin RNA

As an exogenous approach, the application of siRNA has been hindered from the wider spectrum by the fact that the gene silencing effect is transient and not inheritable. The need for endogenous triggers of RNAi effects has been met at least partially, if not entirely, by the successful construction of short hairpin RNA (shRNA) (Paddison, et al., 2004), inspired by the discovery of miRNA (Grishok, 2001; Hutvagner, 2001; Ketting, 2001). Varying in size and design, similar to natural miRNAs, stems of shRNA range from 19 to 29 nucleotides in length (Figure 4).

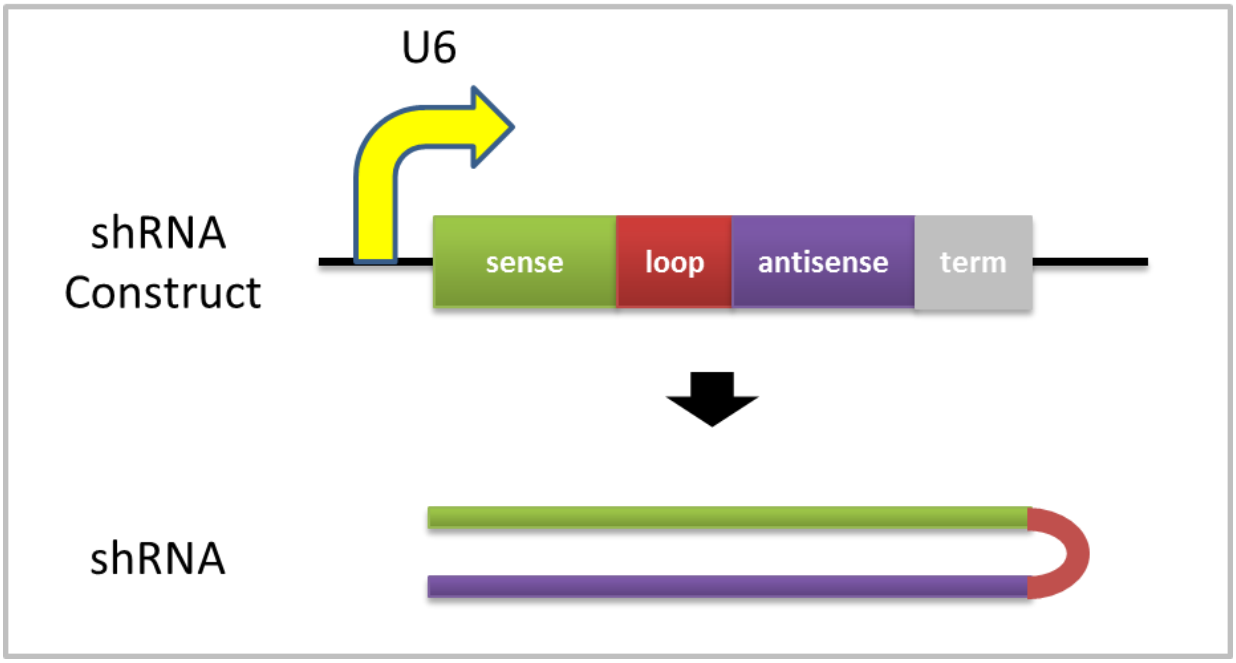


Figure 4. shRNA construction. Built from virus or promoter systems, shRNA resembles much of miRNA mechanism *in vivo,* and mature shRNA leads to gene silencing via either perfect or imperfect matching.

So far, a series of delivery systems has been made available for stable transfection of shRNA, including retroviruses, adenoviruses, constitutive or inducible promoter systems (Paddison, et al., 2004).

After transfection of shRNA into cell lines, vectors will have themselves integrated into the host genome, following which the shRNA is transcribed in the nucleus by polymerase II or III. This leads to a product that mimics pri-microRNA (pri-miRNA) and is processed by Drosha. The resulting pre-shRNA is exported from the nucleus into the cytoplasm, which is then processed by Dicer and loaded into the RNA-induced silencing complex (RISC). The sense (passenger) strand is then degraded. The antisense (guide) strand leads RISC to mRNA that has a perfect complementarity to shRNA, and RISC cleaves the mRNA. Sometimes imperfect complementarity happens, and RISC only represses translation. Either method will cause specific gene silencing.

Scientists and researchers have demonstrated the usage of shRNA in long-term gene silencing *in vivo*, such as the xenograft model (Brummelkamp, et al., 2002; Carmell, et al., 2003; Hasuwa, et al., 2002; Hemann, 2003; Kunath, 2003; Rubinson, 2003; Tiscornia, et al., 2003). The ultimate goal is to develop an approach for the creation of inducible and tissue-specific silencing of most genes in animal models, which may shed light on the application of RNAi as novel therapeutics.

But before further application, there is an intrinsic problem with heterogenetic processing of shRNA such that the gene silencing might be blurred by off-target effects, which will be covered later. Precision in the maturation of shRNA remains a challenge, though we have observed strong evidence that shRNA expression gives rise to single siRNA, which makes it possible to use shRNA as an alternative genetic tool for employing RNAi in research and drug target discovery (Hannon and Rossi, 2004).

1.4 High-throughput RNA interference screening

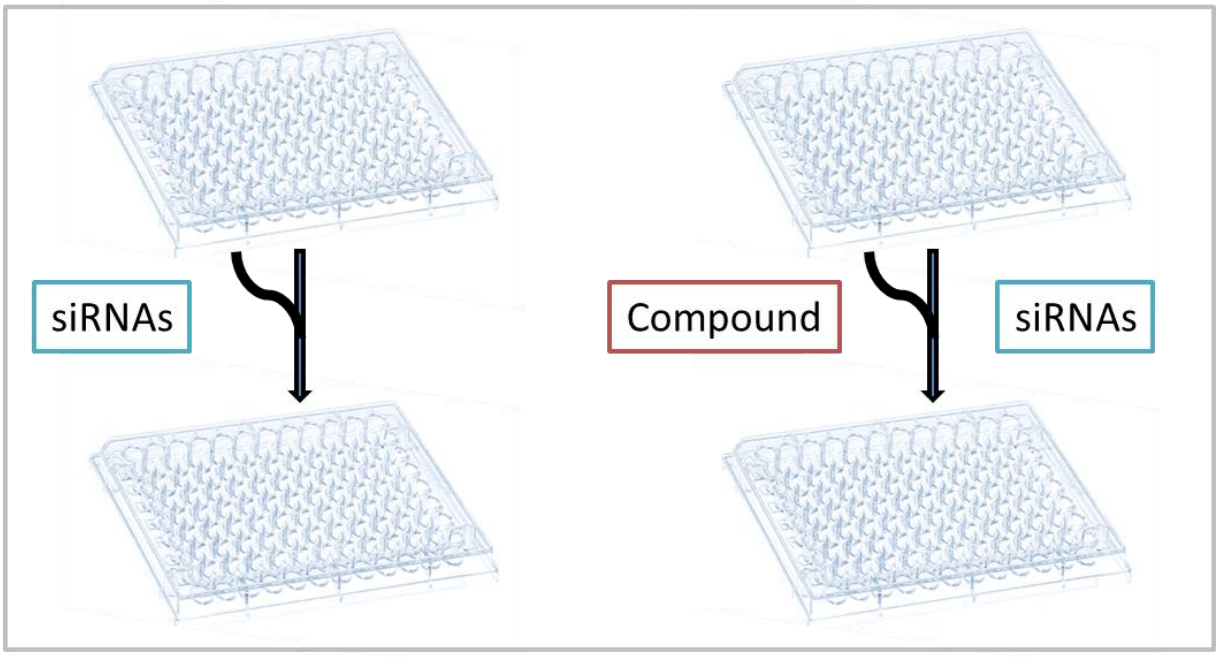


Figure 5. High-throughput RNAi screening. As the most obvious large-scale application of RNAi, loss-of function (LOF) gene knock-down provides tools to identify and functionally characterize genes of interest. As a modification of HTS, synthetic high-throughput RNAi screening helps identify genes or pathways that, when silenced, could help increase or decrease sensitivity to a specific compound or drug. Such genes could be used as chemotherapeutic sensitizers.

RNA interference revolutionized functional annotation and genetics study in a gene-specific pattern and has seen a lot of success since its discovery. Because of the success of using siRNA for analyzing single genes, high-throughput technology has made it possible to perform genome-wide high-throughput RNAi screening (HTS). This is an all-in-one technology that makes it possible to knock down all genes one at a time. It has been widely used in studies such as the identification of novel pathway components, novel drug targets and synthetic chemotherapeutic sensitizers.

1.4.1 Formats and paradigms of HTS

Similar to traditional genetic screening methods such as compound screening, some screening strategies could be readily applied to high-throughput RNAi screening. Obviously, for high-throughput RNAi screening the first large-scale application is loss-of-function (LOF) screening (Figure 5, left). Genes are knocked down individually in each well of the microplates such that they can be functionally characterized and annotated. Another modification of HTS is synthetic screening, in which compounds are used with siRNAs (Figure 5, right). In such a screening, we can identify genes that might have synergistic effects when knocked down with a sub-lethal concentration of specific compounds, like chemotherapeutic drugs (Whitehurst, et al., 2007).

As for the screening paradigms of HTS, usually two options are available: systematic screening and selection-based screening (Echeverri and Perrimon, 2006). In a systematic screening, the siRNA library is customized on a selected subset of the entire genome and planted on 96- or 384-well microplates. Each gene is silenced individually to identify relevant hits of interest. An appropriate read-out system that is both sensitive and specific has to be optimized to make sure results are reproducible. However, as the procedures suggest, a genome-wise systematic screening is considerably expensive and involves an enormous quantity of reagents, automation instruments, and computing infrastructure (Echeverri and Perrimon, 2006).

In contrast, in a selection-based HTS screening, a pooled silencing molecule library is used upon a single, large population of cells, followed by cell sorting based on a specific readout system that can be a fluorescent reporter gene, cell growth advantage, or unique bar-code (Berns, 2004; Silva, 2005; Silva, et al., 2004; Westbrook, 2005). It is faster, simpler and less expensive

than systematic screening. However, the application of selection-based HTS is expected to be more powerful in experiments where low multiplicity of infection is present, and in theory each individual cell is transfected with only one oligo (Echeverri and Perrimon, 2006). Moreover, simultaneous silencing of multiple genes might also blur the interpretation of loss-of-function phenotypes.

1.4.2 Cell line selection for HTS

Due to the simplicity of the cell culture condition, the high efficiency of transfection, and the high-resolution spatial-temporal observations, Drosophila melanogaster cell lines are excellent candidates for high-throughput RNAi screening (Clemens, 2000; Echalier, 1997; Kiger, 2003; Lum, 2003; Ui, 1994). In other cases, many mammalian cell lines have made themselves available for HTS because of features such as adherent property, fast and robust growth, efficient delivery, well-ordered monolayer and reasonable doubling time (Hannon and Rossi, 2004; Song, 2003). Sometimes, in order to model biological processes, primary cell lines are preferred in HTS since they are more likely to represent the real biological context (Ovcharenko, et al., 2005), and close attention has to be paid to reagent selection, experimental optimization and read-out system.

1.4.3 Reagents selection for HTS

In the selection of reagents for HTS, the two most popular commercially available choices are synthetic siRNAs (Elbashir, 2001) and vector-based shRNAs (Berns, 2004). For synthetic siRNAs, pooled siRNAs (3~6 siRNAs per pool) are often used in each well to target the same genes even though their sequences are different. The goal of pooling siRNAs is to increase the chance that the desired gene will be successfully knocked down and will generate loss-of-function phenotypes. However, the increased probability and decreased expense of doing so come at the cost of lower specificity due to sequence-dependent off-target effects.

In order to enhance the gene silencing efficacies of pooled siRNAs, endonuclease-prepared siRNAs (esiRNAs) have recently been developed and the pooled siRNAs effect has been taken to next stage of RNAi technology (Echeverri and Perrimon, 2006). Transcribed in vitro from DNA templates, 200~500 bp dsRNAs are cut into a cocktail of siRNA-like molecules all targeting the same gene (Kittler, 2004; Yang, 2002).

Conversely, when loss-of-function phenotypes are not detectable in readily transfectable cells or within the time frame of a transient transfection, the shRNA library has made itself an important role in high-throughput RNAi screening due to advanced technology and reduced cost (Echeverri and Perrimon, 2006; Silva, 2005).

1.5 High-content screening in HTS

Traditionally, high-throughput RNAi screening has only been associated with single-value read-out systems such as cell viability or substrate concentration or quantified pathway activity. However, the over-simplified representation of complex biological physiological

phenotype has hindered accurate interpretation of loss-of-function phenotypes. With the advent of high-content image-screening technology, image-based HTS has becoming more and more popular with the hope of capturing multiple features after gene interference (Carpenter and Sabatini, 2004).

Advanced high-throughput imaging technology and analysis pipelines have made it possible to integrate high-throughput RNAi screening into high-content screening that is derived from a small molecule screening strategy (Figure 6).

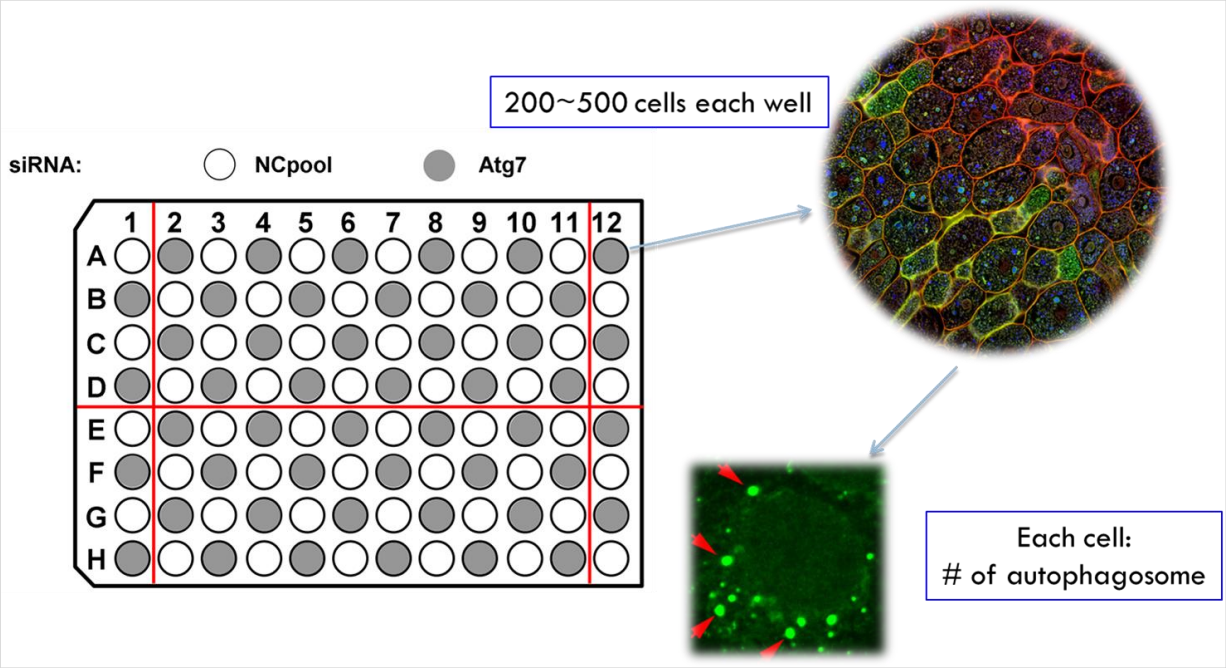


Figure 6. Schematic demonstration of high-content high-throughput RNAi screening. Data provided from Xiaonan Dong in Dr. Beth Levine's lab. In this screening, in each well, 200~500 cells are plated and transfected with RNAi reagents. For each cell, the number of autophagosomes is counted.

1.6 Summary

The discovery of RNAi has enabled genome-wide loss-of-function phenotype screening at a reduced cost in time and money, and it has been widely accepted and used in a variety of biomedical and biological research projects. Despite advancing technology in HTS, computational challenges remain in data analysis and hit identification. These computational hurdles have to be tackled so that HTS can be more accurately interpreted without confounding factors such as spatial noise and off-target effects.

In Chapter Two, I will focus on identification and correction of spatially correlated background noise from high-throughput RNAi screening. We adopted a well-established geostatistical model Kriging interpolation to fit high-throughput RNAi screening data from triplicate experiments, and experimental validation showed that reduction of spatial background noise could help enhance the signal-to-noise ratio and increase statistical detection power.

In order to identify siRNA-mimic-miRNA off-target effects from high-throughput RNAi screening, we developed a deconvolution analysis approach to model data from HTS projects in which pooled siRNAs are used to knock down genes. Data mining and statistical modeling are summarized in Chapter Three. As part of the development of a novel methodology, we tested our algorithm on multiple datasets from a variety of biological contexts across different siRNA libraries. The siRNA-mimic-miRNA off-target effect is pervasive, and identification and removal of such an off-target effect could help reduce the false positive rate from primary screening.

With the advent of advanced development of imaging instruments and technology, high-content screening has been integrated into high-throughput RNAi screening, which has

enabled multiple features from a single cell and comprehensive descriptive quantification of complex biological processes due to loss-of-function interference. Consequently, new methodologies and analysis pipelines are needed, and for this purpose we developed a new R package available within scientific community for data visualization and analysis. Details are shown in Chapter Four. We showed the high accuracy of our package in the analysis of image data.

1.7 Bibliography

Ambros, V. (2003) A uniform system for microRNA annotation, *RNA*, **9**, 277-279.

Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function, *Cell*, **116**, 281-297.

Berns, K. (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway, *Nature*, **428**, 431-437.

Bernstein, E., *et al.* (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference, *Nature*, **409**, 363-366.

Blight, K.J., Kolykhalov, A.A. and Rice, C.M. (2000) Efficient initiation of HCV RNA replication in cell culture, *Science*, **290**, 1972-1974.

Brummelkamp, T.R., Bernards, R. and Agami, R. (2002) Stable suppression of tumorigenicity by virus-mediated RNA interference, *Cancer Cell*, **2**, 243-247.

Buchele, T. (2003) Proapoptotic therapy with oblimersen (bcl-2 antisense oligonucleotide)[mdash]review of preclinical and clinical results, *Onkologie*, **26**, 60-69.

Carmell, M.A., *et al.* (2003) Germline transmission of RNAi in mice, *Nature Struct. Biol.*, **10**, 91-92.

Carpenter, A.E. and Sabatini, D.M. (2004) Systematic genome-wide screens of gene function, *Nature Rev. Genet.*, **5**, 11-22.

Chiu, Y.L. and Rana, T.M. (2003) siRNA function in RNAi: a chemical modification analysis, *RNA*, **9**, 1034-1048.

Clemens, J.C. (2000) Use of double-stranded RNA interference in Drosophila cell lines to dissect signal transduction pathways, *Proc. Natl Acad. Sci. USA*, **97**, 6499-6503.

Coburn, G.A. and Cullen, B.R. (2002) Potent and specific inhibition of human immunodeficiency virus type-1 replication by RNA interference, *J. Virol.*, **76**, 9225-9231.

Cowan-Jacob, S.W. (2004) Imatinib (STI571) resistance in chronic myelogenous leukemia: molecular basis of the underlying mechanisms and potential strategies for treatment, *Mini Rev. Med. Chem.*, **4**, 285-299.

Cullen, B.R. (2004) Transcription and processing of human microRNA precursors, *Mol. Cell*, **16**, 861-865.

Czauderna, F. (2003) Structural variations and stabilising modifications of synthetic siRNAs in mammalian cells, *Nucleic Acids Res.*, **31**, 2705-2716.

Echalier, G. (1997) Drosophila Cell in Culture.

Echeverri, C.J. and Perrimon, N. (2006) High-throughput RNAi screening in cultured cells: a user's guide, *Nat Rev Genet*, **7**, 373-384.

Elbashir, S.M. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells, *Nature*, **411**, 494-498.

Fire, A. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans, *Nature*, **391**, 806-811.

Fire, A., *et al.* (1991) Production of antisense RNA leads to effective and specific inhibition of gene expression in C. elegans muscle, *Development*, **113**, 503-514.

Grimson, A., *et al.* (2007) MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing, *Molecular Cell*, **27**, 91-105.

Grishok, A. (2001) Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing, *Cell*, **106**, 23-34.

Guo, S. and Kemphues, K.J. (1995) par-1, a gene required for establishing polarity in C. elegans embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed, *Cell*, **81**, 611-620.

Hammond, S.M., *et al.* (2001) Argonaute2, a link between genetic and biochemical analyses of RNAi, *Science*, **293**, 1146-1150.

Hannon, G.J. (2002) RNA interference, *Nature*, **418**, 244-251.

Hannon, G.J. and Rossi, J.J. (2004) Unlocking the potential of the human genome with RNA interference, *Nature*, **431**, 371-378.

Hasuwa, H., *et al.* (2002) Small interfering RNA and gene silencing in transgenic mice and rats, *FEBS Lett.*, **532**, 227-230.

Hemann, M.T. (2003) An epi-allelic series of p53 hypomorphs created by stable RNAi produces distinct tumor phenotypes in vivo, *Nature Genet.*, **33**, 396-400.

Holtz, M.S. and Bhatia, R. (2004) Effect of imatinib mesylate on chronic myelogenous leukemia hematopoietic progenitor cells, *Leuk. Lymphoma*, **45**, 237-245.

Hutvagner, G. (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA, *Science*, **293**, 834-838.

Ikeda, M., *et al.* (2002) Selectable subgenomic and genome-length dicistronic RNAs derived from an infectious molecular clone of the HCV-N strain of hepatitis C virus replicate efficiently in cultured Huh7 cells, *J. Virol.*, **76**, 2997-3006.

Jacque, J.M., Triques, K. and Stevenson, M. (2002) Modulation of HIV-1 replication by RNA interference, *Nature*, **418**, 435-438.

Kapadia, S.B., Brideau-Andersen, A. and Chisari, F.V. (2003) Interference of hepatitis C virus RNA replication by short interfering RNAs, *Proc. Natl Acad. Sci. USA*, **100**, 2014-2018.

Ketting, R.F. (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans, *Genes Dev.*, **15**, 2654-2659.

Kiger, A.A. (2003) A functional genomic analysis of cell morphology using RNA interference, *J. Biol.*, **2**, 27.

Kim, V.N. (2005) MicroRNA biogenesis: coordinated cropping and dicing, *Nature reviews. Molecular cell biology*, **6**, 376-385.

Kittler, R. (2004) An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division, *Nature*, **432**, 1036-1040.

Kittler, R. and Buchholz, F. (2003) RNA interference: gene silencing in the fast lane, *Semin. Cancer Biol.*, **13**, 259-265.

Kunath, T. (2003) Transgenic RNA interference in ES cell-derived embryos recapitulates a genetic null phenotype, *Nature Biotechnol.*, **21**, 559-561.

Lee, N.S. (2002) Expression of small interfering RNAs targeted against HIV-1 rev transcripts in human cells, *Nature Biotechnol.*, **20**, 500-505.

Lohmann, V. (1999) Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line, *Science*, **285**, 110-113.

Lu, P.Y., Xie, F.Y. and Woodle, M.C. (2003) siRNA-mediated antitumorigenesis for drug target validation and therapeutics, *Curr. Opin. Mol. Ther.*, **5**, 225-234.

Lum, L. (2003) Identification of Hedgehog pathway components by RNAi in Drosophila cultured cells, *Science*, **299**, 2039-2045.

Marcucci, G., Perrotti, D. and Caligiuri, M.A. (2003) Understanding the molecular basis of imatinib mesylate therapy in chronic myelogenous leukemia and the related mechanisms of resistance. Commentary, *Clin. Cancer Res.*, **9**, 1248-1252.

McCaffrey, A.P. (2002) RNA interference in adult mice, *Nature*, **418**, 38-39.

Novina, C.D. (2002) siRNA-directed inhibition of HIV-1 infection, *Nature Med.*, **8**, 681-686.

Ovcharenko, D., *et al.* (2005) High-throughput RNAi screening in vitro: from cell lines to primary cells, *RNA*, **11**, 985-993.

Paddison, P.J., *et al.* (2004) Short hairpin activated gene silencing in mammalian cells, *Methods Mol. Biol.*, **265**, 85-100.

Park, W.S. (2002) Prevention of HIV-1 infection in human peripheral blood mononuclear cells by specific RNA interference, *Nucleic Acids Res.*, **30**, 4830-4835.

Pietschmann, T., *et al.* (2001) Characterization of cell lines carrying self-replicating hepatitis C virus RNAs, *J. Virol.*, **75**, 1252-1264.

Randall, G., Grakoui, A. and Rice, C.M. (2003) Clearance of replicating hepatitis C virus replicon RNAs in cell culture by small interfering RNAs, *Proc. Natl Acad. Sci. USA*, **100**, 235-240.

Rubinson, D.A. (2003) A lentivirus-based system to functionally silence genes in primary mammalian cells, stem cells and transgenic mice by RNA interference, *Nature Genet.*, **33**, 401-406.

Silva, J.M. (2005) Second-generation shRNA libraries covering the mouse and human genomes, *Nature Genet.*, **37**, 1281-1288.

Silva, J.M., *et al.* (2004) RNA interference microarrays: high-throughput loss-of-function genetics in mammalian cells, *Proc. Natl Acad. Sci. USA*, **101**, 6548-6552.

Song, E. (2003) RNA interference targeting Fas protects mice from fulminant hepatitis, *Nature Med.*, **9**, 347-351.

Song, E. (2003) Sustained small interfering RNA-mediated human immunodeficiency virus type 1 inhibition in primary macrophages, *J. Virol.*, **77**, 7174-7181.

Surabhi, R.M. and Gaynor, R.B. (2002) RNA interference directed against viral and cellular targets inhibits human immunodeficiency virus type-1 replication, *J. Virol.*, **76**, 12963-12973.

Tauchi, T. and Ohyashiki, K. (2004) Molecular mechanisms of resistance of leukemia to imatinib mesylate, *Leuk. Res.*, **28**, 39-45.

Tiscornia, G., *et al.* (2003) A general method for gene knockdown in mice by using lentiviral vectors expressing small interfering RNA, *Proc. Natl Acad. Sci. USA*, **100**, 1844-1848.

Ui, K. (1994) Newly established cell lines from Drosophila larval CNS express neural specific characteristics, *In Vitro Cell. Dev. Biol. Anim.*, **30A**, 209-216.

Wall, N.R. and Shi, Y. (2003) Small RNA: can RNA interference be exploited for therapy?, *Lancet*, **362**, 1401-1403.

Westbrook, T.F. (2005) A genetic screen for candidate tumor suppressors identifies REST, *Cell*, **121**, 837-848.

Whitehurst, A.W., *et al.* (2007) Synthetic lethal screen identification of chemosensitizer loci in cancer cells, *Nature*, **446**, 815-819.

Williams, B.R. (1997) Role of the double-stranded RNA-activated protein kinase (PKR) in cell regulation, *Biochem. Soc. Trans.*, **25**, 509-513.

Wilson, J.A. (2003) RNA interference blocks gene expression and RNA synthesis from hepatitis C replicons propagated in human liver cells, *Proc. Natl Acad. Sci. USA*, **100**, 2783-2788.

Yang, D. (2002) Short RNA duplexes produced by hydrolysis with Escherichia coli Rnase III mediate effective RNA interference in mammalian cells, *Proc. Natl Acad. Sci. USA*, **99**, 9942-9947.

CHAPTER TWO

SPATIAL BACKGROUND NOISE CORRECTION IN HIGH-THROUGHPUT RNA INTERFERENCE SCREENING

High-throughput cell-based RNAi screening has become an increasingly important technology that is widely used for discovering new drug targets and annotating gene functions. Screening strategies usually use hundreds of 96-well or 384-well plates in order to cover genes in a genome-wide pattern, and often collect measurements that are dampened by spatial background noise whose patterns may vary across each individual plate. Identification and correction of such position effects can substantially enhance measurement accuracy and screening success. We built SbacHTS (Spatial background noise correction for High-Throughput RNAi Screening) software for visualization, estimation and correction of spatial background noise in HT-RNAi screens. SbacHTS is available as a web-based user-friendly bioinformatics tool on the Galaxy open-source framework with open access web interface. We showed that SbacHTS software could effectively detect and correct spatial background noise, improve the signal-to-noise ratio and increase statistical detection power in high-throughput RNAi screening experiments.

2.1 Introduction

RNAi is a process in which gene expression is silenced by small RNA molecules such as siRNAs (small interfering RNAs) and shRNAs (short hairpin RNAs). High-throughput RNAi screening is a groundbreaking technology for functional genomics and for drug target discovery

and has been widely used in biological and biomedical research (Orvedahl, et al., 2011; Whitehurst, et al., 2007).

2.1.1 From small-molecule screening to HTS

A frequently employed screening strategy relies on 96-well or 384-well microwell plates, on each of which is pooled siRNAs (3~6 siRNAs per pool) designed to target a specific gene. A similar screening strategy has been used for small-molecule screening for decades (Boutros and Ahringer, 2008). They both share some elements in common with respect to analysis workflow; however, the intrinsic features and properties of RNAi exercise their own challenges for data analysis and visualization. Empirically, HTS results tend to be normally distributed but with more noise, a decreased signal-to-noise ratio and an increased coefficient of variations (Birmingham, et al., 2009). Consistently, the Z' factor of Z factor (defined as below) from high-throughput RNAi screening is prone to be lower than in small-molecule screening (Zhang, et al., 1999).

$$Z'\ factor = 1 - \left(3\sigma_{hc} + 3\sigma_{lc}\right) / \left|\mu_{hc} - \mu_{lc}\right| \quad (1)$$

$$Z\ \ factor = 1 - \left(3\sigma_{s} + 3\sigma_{c}\right) / \left|\mu_{s} - \mu_{c}\right| \quad (2)$$

where μ indicates mean, and σ represents standard deviations. Subscription "hc" and "lc" represent high-value control and low-value control, while "s" and "c" represent the sample value and control value, respectively (Birmingham, et al., 2009).

2.1.2 Data triage and quality control

Due to the higher inter-well variability of high-throughput RNAi screening, which results from a range of factors such as transfection efficiency and incubation conditions, researchers should keep in mind that analysis has to be customized from a repertoire of selections and quality control has to be performed as a work-in-progress from the very beginning in case of potential problems that might occur anytime (Birmingham, et al., 2009).

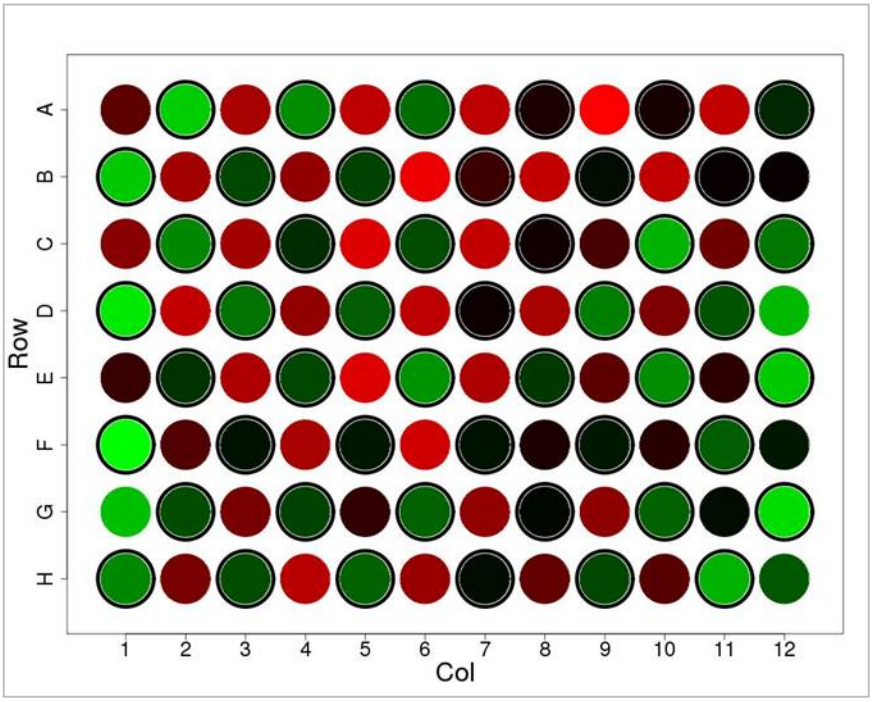


Figure 7. Visualization of raw data. Data provided from Xiaonan Dong in Dr. Beth Levine's lab. The number of autophagosomes in each cell from the same well is measured, log-transformed and plotted as in the above figure. Red means a higher number and green means a lower number.

So far, visualization is a popular approach to control quality for high-throughput RNAi screening in progress, and available approaches include a heat map of raw data (Figure 7), replicate correlation plot (Figure 8) and row-column effect visualization (Figure 9), all of which

can help to control reproducibility in progress (Ogier and Dorval, 2012; Zhang and Zhang, 2013).

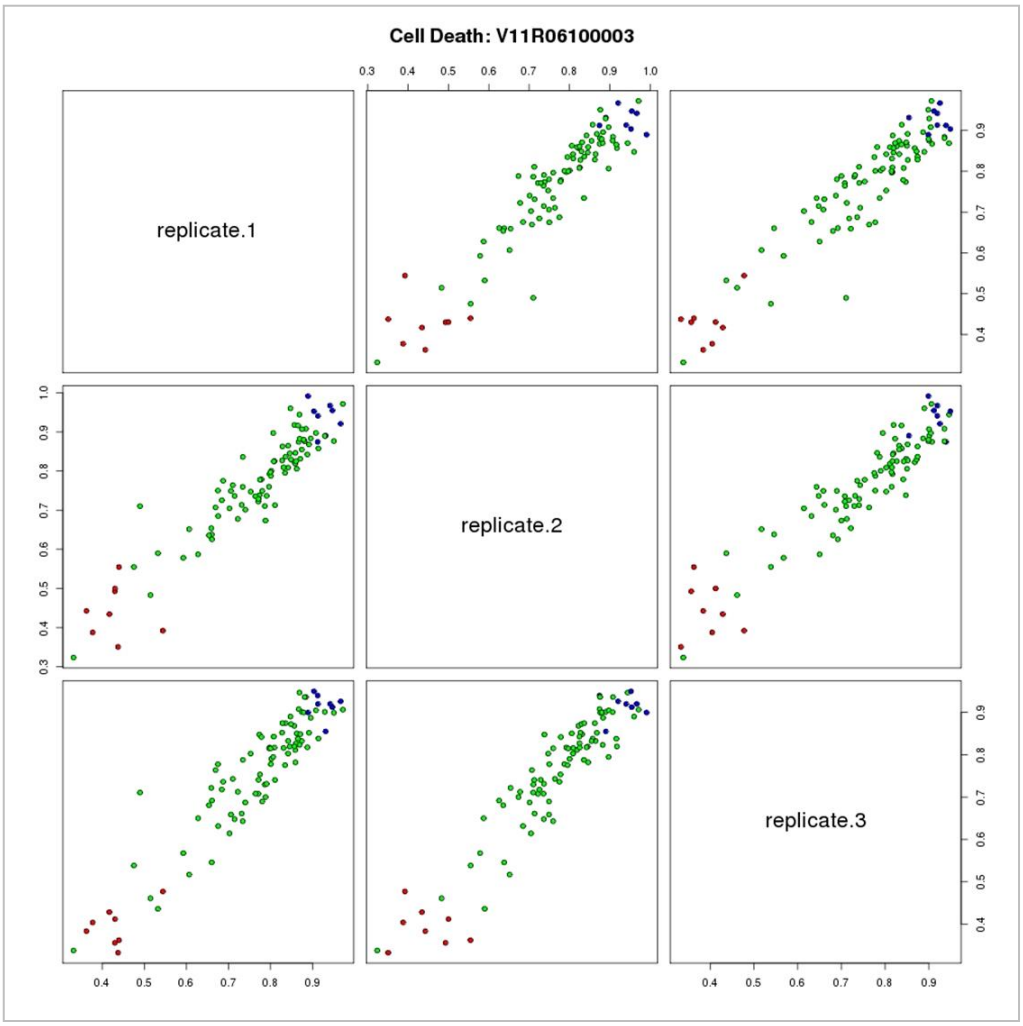


Figure 8. Replicate correlation plot. Data provided from Yang Liu in Dr. Beth Levine's lab. Cell death is measured and experiments are performed in triplicate. Thus the replicate correlation plot helps to visualize quality control. Red indicates positive control; blue indicates negative control; green indicates the sample.

A quantitative quality metrics calculation is also available to facilitate appropriate operation of experimental procedures (Birmingham, et al., 2009). We have seen the Z' factor and

Z factor. Another metric is the strictly standardized mean difference (SSMD), depicted as below (Zhang, 2007),

$$SSMD = (\mu_1 - \mu_2) / \sqrt{\sigma_1^2 + \sigma_2^2} \quad (3)$$

which measures the ratio between the mean difference and pooled standard deviation.

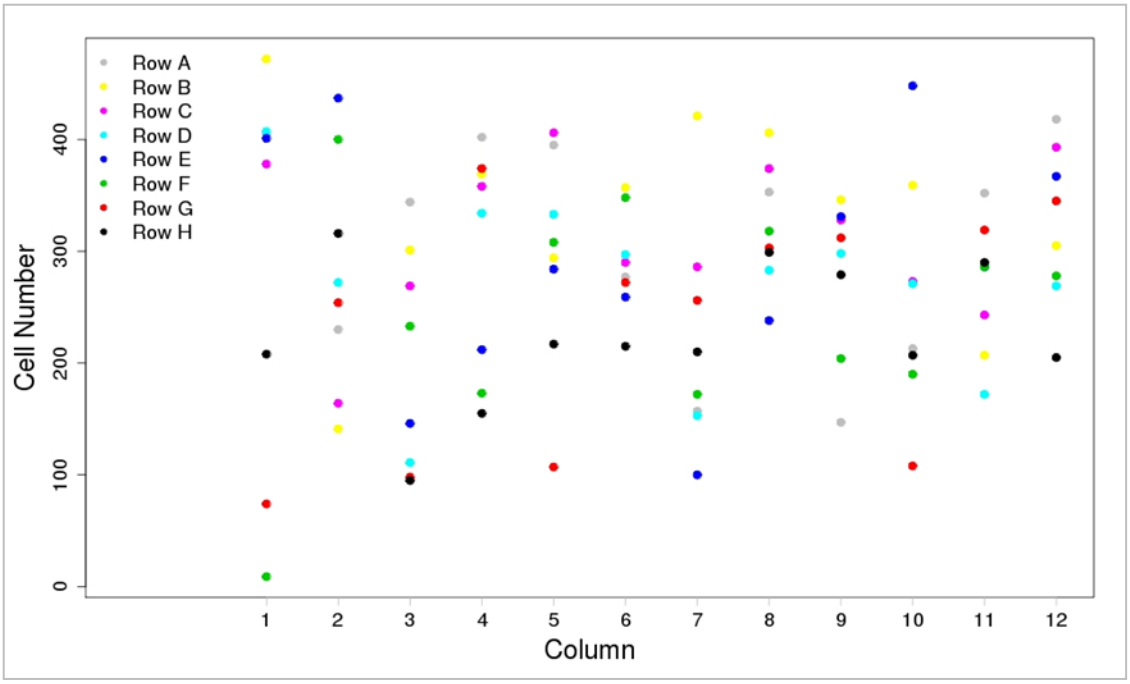


Figure 9. Row-column effect plot. Data provided from Xiaonan Dong in Dr. Beth Levine's lab. The cell number is counted and plotted in a row-column pattern such that researchers can identify the specific position effect.

Receiver operating characteristic (ROC) curves, in which both the true positive rate and false positive rate are calculated and plotted, have been widely applied in statistical analysis and are also applicable for high-throughput RNAi screening as a simple and intuitive quantitative metric and visualization (Fawcett, 2006; Forster, et al., 2003; Wagner, 2002; Wiles, et al., 2008).

The area under the curve (AUC) is a desirable measurement of quality, with 1 representing a perfect predictor and 0.5 as random chance.

$$true\ positive\ rate = \frac{true\ positive}{true\ positive + false\ negative} \tag{4}$$

$$false\ positive\ rate = \frac{false\ positive}{true\ negative + false\ positive} \tag{5}$$

2.1.3 Data normalization

Before we talk about the details of data normalization strategies in high-throughput RNAi screening, we need to be cognizant of the two main streams, control-based and sample-based normalization (Birmingham, et al., 2009). When applicable, both negative control and positive control are preferred in high-throughput RNAi screening to facilitate the calculation of quality metrics and data normalization. In such instances, either is unavailable and one is also working well most of the time. However, in some cases when controls are not working well, given the batch effects and position effects, sample-based normalization can be useful. However, the assumption for sample-based normalized is that on a plate (Whitehurst, et al., 2007), most genes have little or no phenotypic effect, and researchers have to keep that in mind in case the assumption is violated for some high-throughput RNAi screening projects.

Normalization helps remove systematic errors such as batch effects, and therefore lends a hand to the comparison of data from different plates or even experimental dates. A fraction or percentage of either the controls or the samples are the two most obvious normalization methods, such as relative cell viability or cell death rate. A z score representing the number of standard

deviation from sample mean is also applicable, though it is quite sensitive to outliers. Therefore a robust z score is used in which the median replaces the mean and the median absolute deviation (MAD) replaces the standard deviation.

$$z\ score = \frac{sample\ value - sample\ mean}{sample\ deviation} \tag{6}$$

$$robust\ z\ score = \frac{sample\ value - sample\ median}{sample\ median\ absolute\ deviation(MAD)} \tag{7}$$

$$MAD = median\left(\left|sample - median\left(sample\right)\right|\right) \tag{8}$$

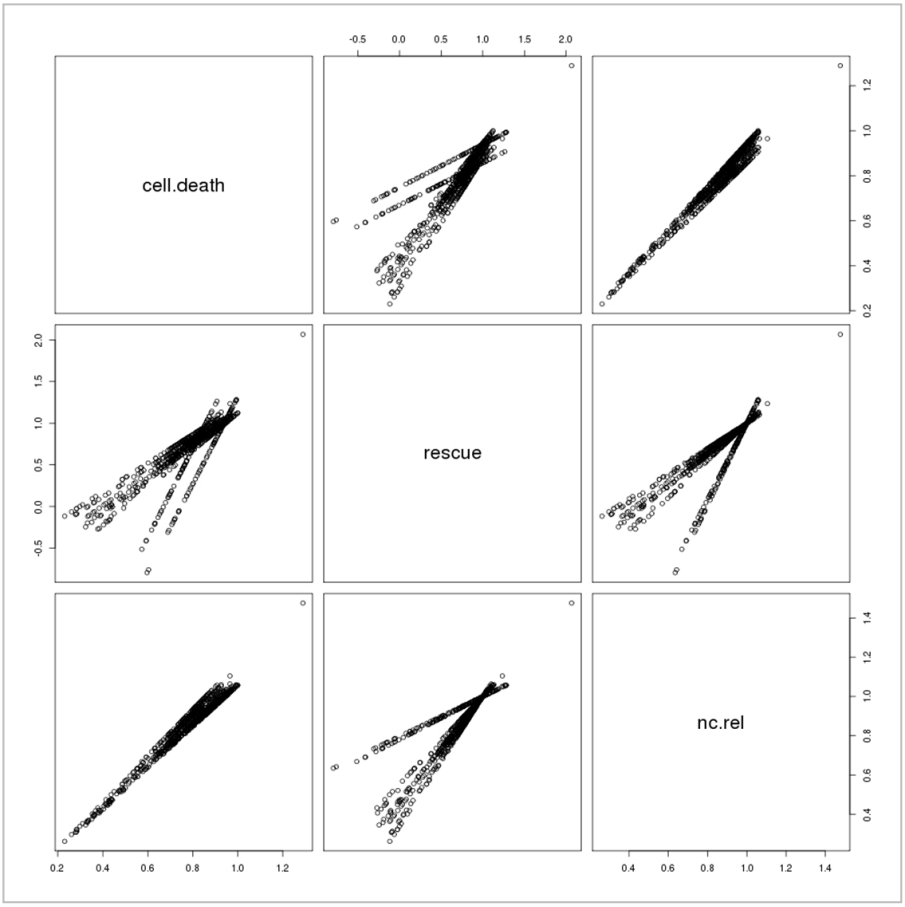


Figure 10. Correlation between normalization methods. Data provided from Yang Liu in Dr. Beth Levine's lab.

Employing the Tukey median algorithm, the B score has recently been used for HTS normalization to account for within-well variations such as the column effect and row effect (Brideau, et al., 2003), and it is easily available in the R package cell HTS2 (Boutros, et al., 2006). A recently modified t-test and goodness-of-fit test can also be generalized to normalize HTS results and show improvement (Dragiev, et al., 2012). Scores from different normalizations could be plotted against each other for double-checking (Figure 10).

2.1.4 Identifying hits from HTS

As the ultimate goal of any primary high-throughput RNAi screening project, hits identification is to select as many true positive hits as possible. Many techniques are available and should be used on a case-by-case basis.

Both the $\text{mean} \pm k$ standard deviation and $\text{median} \pm k$ MAD are derived from small-molecule screening, and quite easily used (Chung, 2008; DasGupta, et al., 2005; Muller, et al., 2005; Possemato, et al., 2011; Zhang, 2006). Both are easy to calculate, with the former easily aligned to the P value and the latter more robust to outliers.

Multiple t-tests could be used when a comparison has to be made between the case and control groups, and follow-up multiple test correction has to be carefully performed in order to balance stringency and detection power (Manly, et al., 2004; Whitehurst, et al., 2007; Zhang, 2008). It is easy to calculate, but it requires triplicates and is inappropriate when data is not normally distributed (Birmingham, et al., 2009).

Robust to outliers, a quantile-based hit selection criterion, in which hits are defined as bigger than the third quantile or lower than the first quantile, is good for nonsymmetrical data

distribution after researchers and scientists evaluate their data distribution. However, it comes with limited additional power and is not linked with P values (Zhang, 2006).

Though introduced as a quality control metric, SSMD is also applicable in the identification of hits selection, depending on whether the goal is to control rates of false negatives, false positives or both (Birmingham, et al., 2009; Zhang, 2007; Zhang, 2007; Zhang, et al., 2009). It is dependent on sample size and linked to rigorous probability interpretation. However, it is not available in most analysis software and is not very intuitive for biologists.

In some cases, multiple reagents are used to target the same gene in a high-throughput RNAi screening projects, and a newly developed method, redundant siRNA activity (RSA), has been employed to deal with such cases (Konig, 2007). It is robust to outliers and helps identify weaker hits as well as reduce false positives from off-target effect. However, it is very difficult to calculate and not applicable to pooled siRNA HTS (Birmingham, et al., 2009).

When it comes to merging multiple datasets from different HTS projects, the analysis may consider algorithms such as rank product given the assumption that a hit should be consistent across different biological context and background (Breitling, et al., 2004). A simulation is needed to generate the null distribution to calculate P values, and many replicates are required through this approach, which is more robust to outliers and can identify weaker hits (Birmingham, et al., 2009).

A Bayesian approach is also available for hits identification in an HTS project (Zhang, 2008); it is not sensitive to outliers, provides P values, and allows for calculation of the false discovery rate (FDR). Besides these it uses both negative controls and samples and includes both

experimental-wide and plate-wide information. However, it is quite difficult to calculate and interpret for a non-statistician audience (Birmingham, et al., 2009).

2.1.5 Spatial noise in HTS

An experimental and computational challenge confronting the collection of precision measurements during HTS is that of the required experimental steps, whose procedures (including cell culture, transfection, reagent delivery, incubation and HTS-plate scanning) may introduce spatially correlated background noise (Figure 11) varying across experiments, batches and plates (Birmingham, et al., 2009; Carralot, et al., 2011; Malo, et al., 2006).

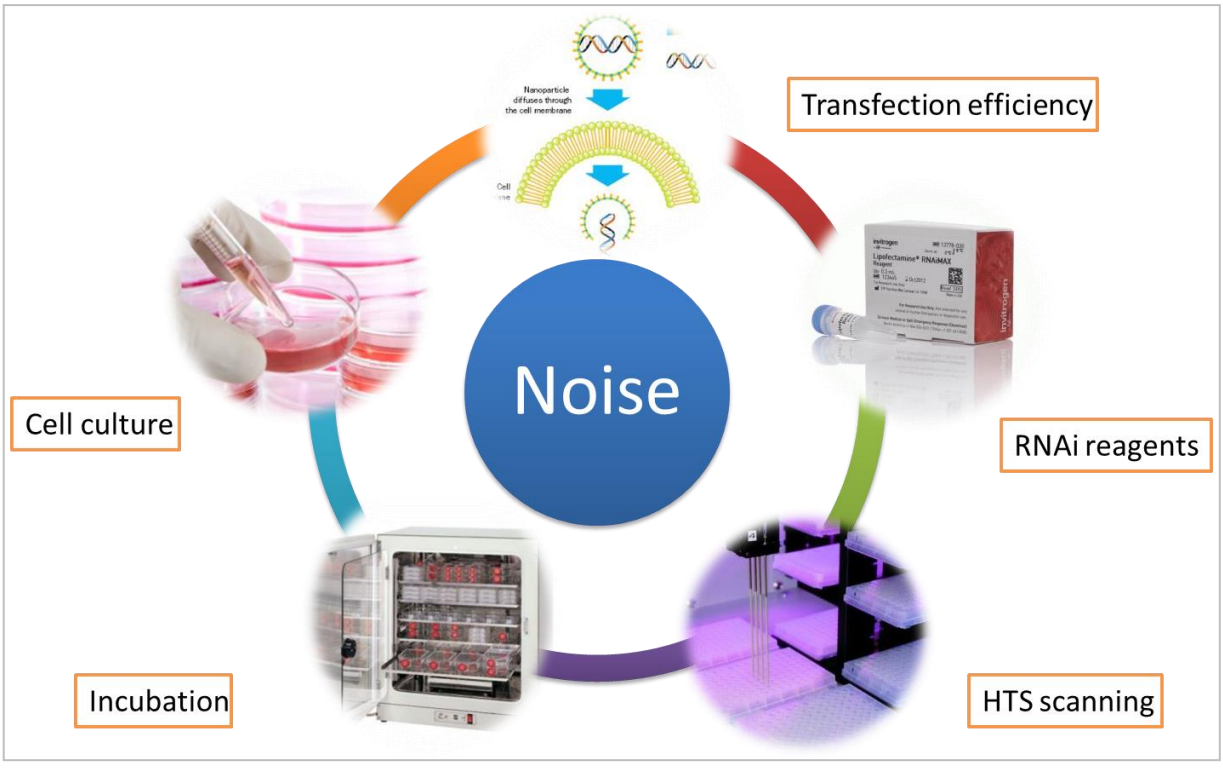


Figure 11. Sources of noise in high-throughput RNAi screening. Factors such as transfection efficiency, cell culture conditions, incubation environments, HTS screening or the RNAi reagent itself can bring in spatial noise in high-throughput RNAi screening projects.

Ignoring position effects leads to low signal-to-noise ratios and hampers sensitivity. Recently, spatial noise elements such as the edge effect have been taken into account in data normalization of high-throughput RNAi screening data (Carralot, et al., 2011); however, only row and column effects have been adjusted using analysis of variance (ANOVA) in the only existing approach to address the global spatial background noise across a plate using B score statistics (Malo, et al., 2006). This simplified model can be an effective approach for simple row and column effects. In our preliminary analysis of the high-throughput RNAi screening data, complicated spatial patterns were shown for many plates from high-throughput RNAi screening projects. Thus, over-simplified modeling of background noise approaches may often result in over-correction for some wells and under-correction for others. To help tackle this problem, we attempted to adopt advanced statistical models to accurately quantify and correct complex spatial background noise in high-throughput RNAi screening experiments.

As a well-established statistical model to fit observed data with spatial distribution patter, Kriging interpolation (Banerjee, et al., 2003) is widely used in geostatistics. In this study, we employed a Kriging model to quantitatively identify and correct spatially-correlated background noise in high-throughput RNAi screening data, and we implemented a user-friendly software package available on a Galaxy platform, SbacHTS, for open-source implementation of the Kriging correction. Meanwhile, intuitive data visualization and quality assessment tools are also available in our package. We discovered that SbacHTS software can effectively identify and correct spatial background noise, enhance the signal-to-noise ratio and increase statistical detection power for RNAi screening experiments.

2.2 Methods and Materials

2.2.1 Geostatistical modeling

SbacHTS software adopted Kriging interpolation to fit spatial noise patterns to identify and correct spatial background noise in high-throughput screening data. For each individual plate, at well s, observed intensity (e.g. cell viability readout from the well) Y_s is modeled as below:

$$Y_s = X_s + \varepsilon_s \ . \tag{9}$$

Here X_s is the signal from the well s and ε_s is defined as spatially-correlated background noise. Our assumption is that X_s is from a normal distribution:

$$X_s \sim N\left(\mu_s, \sigma_s^2\right) \tag{10}$$

And here μ_s is the mean of siRNAs in well s.

For ε_s, we model it from a multivariate Gaussian distribution,

$$\varepsilon_s \sim \mathbb{N}\left(\vec{0}, \Sigma\right) \tag{11}$$

Where

$$\Sigma = \sigma^2 \rho\left(\phi, d_{i,j}\right) + \tau^2 I \tag{12}$$

And $\rho\left(\phi, d_{i,j}\right)$ is a function of distance between plate well i and j with $d_{i,j}$ defined as below

$$d_{i,j} = \left|s_i - s_j\right| \tag{13}$$

with parameters ϕ and $\tau^2 I$ that model independent Gaussian white noise. From this model, we can estimate the signal and spatial background noise. For details about statistical inference and parameters estimation, reader can refer to (Banerjee, et al., 2003).

2.2.2 Synthetic lethal screening

Chemotherapeutic drug resistance has become a treatment hurdle for cancer therapy. Genes that are required for drug resistance might be interesting drug targets in that inhibition of such genes may sensitize resistant cancer cell lines to otherwise sub-lethal concentration of traditional chemotherapeutic drugs (Figure 12). We refer to these genes as "chemo-sensitizers" (Whitehurst, et al., 2007).

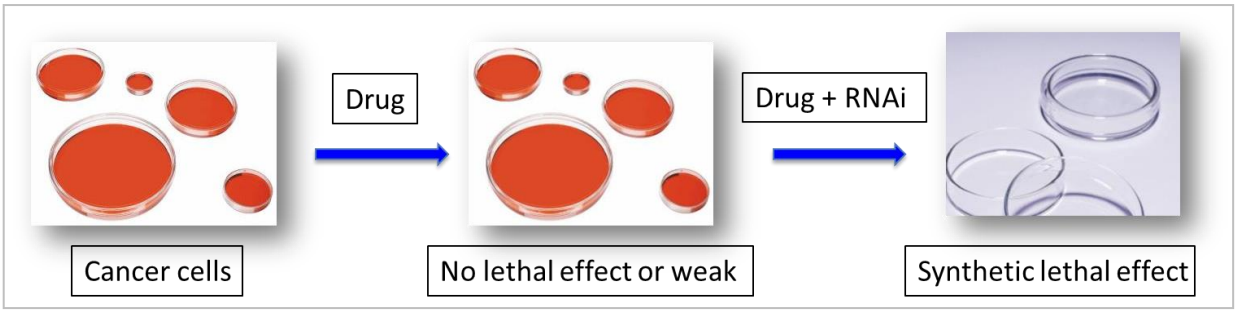


Figure 12. Screening scheme for synthetic screening. HST was used to identify chemo-sensitizers that might help cancer cell lines overcome drug resistance.

2.2.3 Screening paradigm

Chemotherapeutic drug resistance has become a treatment hurdle for cancer therapy. Our collaborator Michael White's group (Whitehurst, et al., 2007) conducted a high-throughput RNAi screening project for identification of chemotherapeutic sensitizers. A non-small-cell lung cancer cell line was established and used in this screen. At the beginning of experiments, cells

were transfected with a siRNA library in a genome-wide pattern. At the end of three days, the drug was added into each well of the plates at a concentration under which cell lines have a sub-lethal effect. By the end of five days, end-point assays were performed and cell viability was measured. We had a media-only control group in this screening for comparison. Experiments were carried out in triplicate. A t-test was used to compare between the treatment and control groups to test for significant difference in cell viability. If significant difference appeared, the gene might be a potential chemotherapeutic sensitizer.

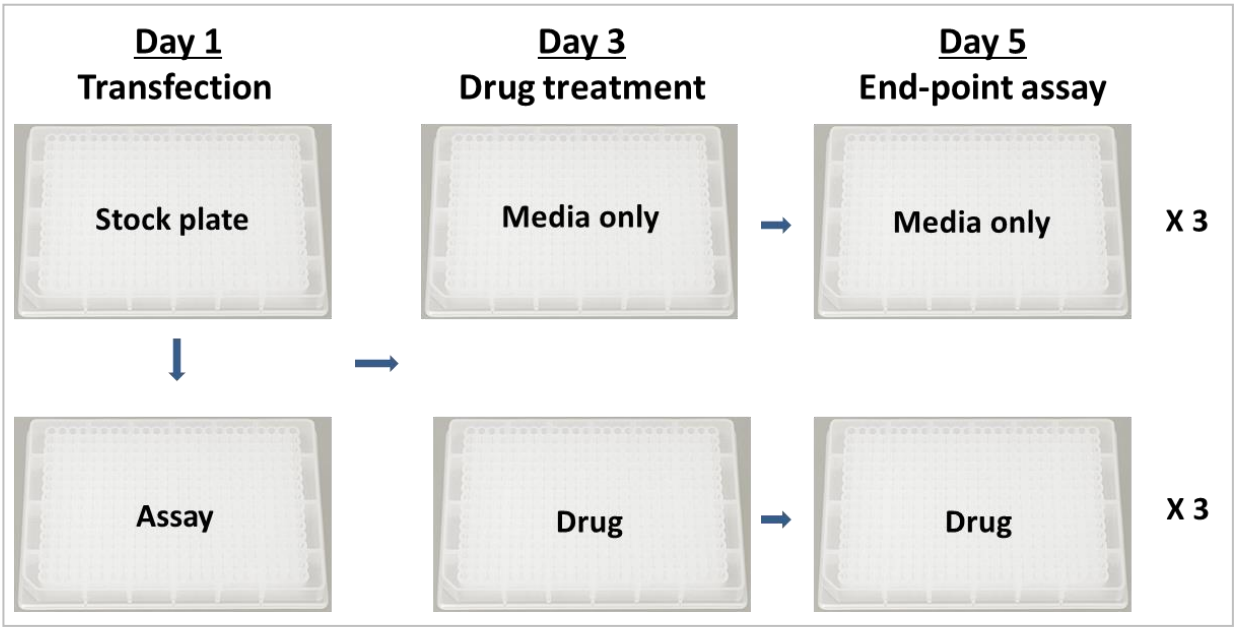


Figure 13. Screening paradigm for synthetic screen. Cells were transfected with siRNA library and at the end of three days, the drug was added into each well at a sub-lethal concentration. By the end of five days, cell viability was used as an end-point assay.

2.3 Results

Datasets from Michael White's experiments are performed with whole-genome arrayed siRNA library on 267 96-well micro plates (Whitehurst, et al., 2007). Paclitaxel-treated non-

small-cell lung cancer cells (experimental group) were compared with vehicle-treated cells (control group) and each was carried out in triplicate (Figure 13).

2.3.1 Spatial noise pattern visualization

For high-throughput RNAi screening, noise is inevitable because many factors can contribute to the noise pattern across a plate, such as transfection efficiency, reagent activity, and incubation and cell culture conditions. Theoretically, random noise is expected such that there should be no specific noise pattern; however, based on our exploratory observations the HTS noise often presents a spatially-correlated pattern (Figure 14). This is reasonable since the oxygen concentration may be higher on the edge during incubation, while siRNAs concentration might be lower in the middle of a plate, and such effects could accumulate and superimpose over time.

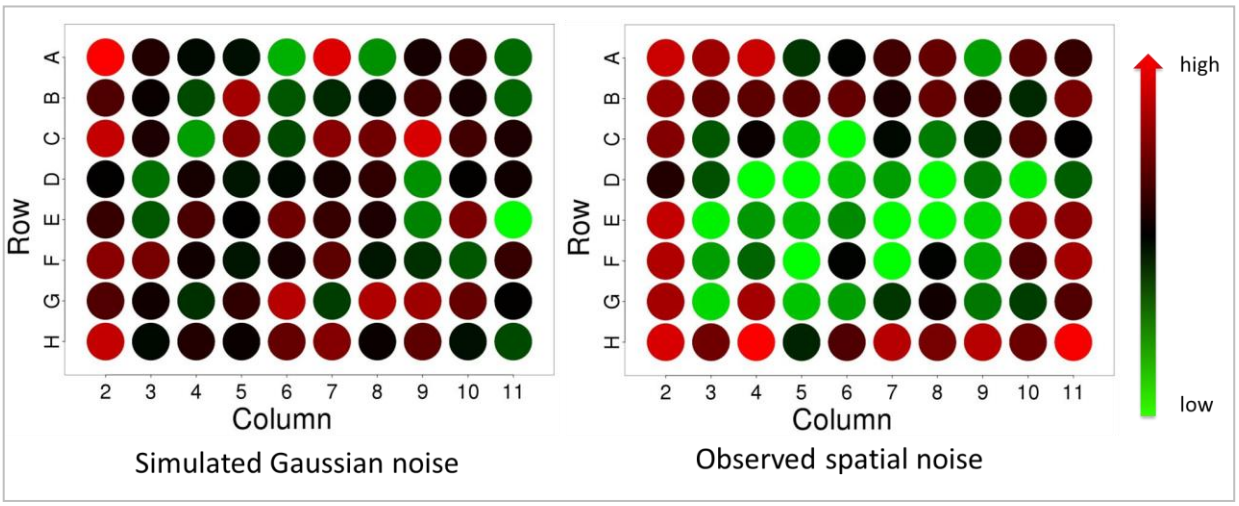


Figure 14. Simulated Gaussian noise VS observed spatial noise. Left, Gaussian noise is simulated to compare with the observed spatial noise that is frequently present in high-throughput RNAi screening.

SbacHTS is able to display the observed spatial background noise pattern across each plot. After identification of such spatially-correlated background noise, we fit data into the SbacHTS model. Fitted spatial background noise is also available for visualization and analysis (Figure 15), which gives an opportunity for fast and intuitive recognition of the spatial pattern of background noise and distribution of fitted values customized for each plate across a whole HTS project.

After identification of such spatially correlated noise, we can remove them from the original intensity and produce a purer read-out that is free of such spatial background noise. We will later show the benefits after removing such noise.

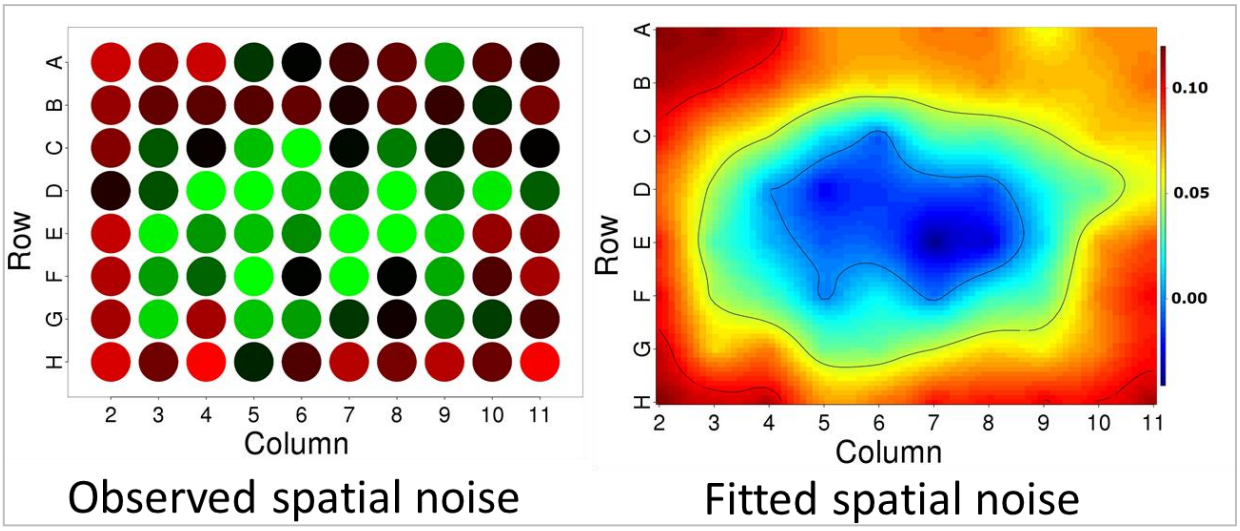


Figure 15. Observed spatial noise and fitted spatial noise pattern. From right, we can tell the fitted spatial noise accurately captures the noise pattern from observed data, and after fitting such a noise pattern can be removed from raw data to help enhance downstream analysis.

2.3.2 Improvement of coefficients of variation and statistical detection power

As a measurement of the signal-to-noise ratio, the coefficient of variation (CV) is commonly used and defined by the ratio between the standard deviation and mean as below:

$$C_v = \frac{\sigma}{\mu} \tag{14}$$

It is estimated using the sample mean and sample standard deviation as below:

$$\hat{C}_v = \frac{s}{\bar{x}} \tag{15}$$

From our dataset, experiments were performed in triplicate, which therefore gave us enough to estimate the coefficient of variation for each gene. The coefficients of variation were therefore estimated for each siRNA pool.

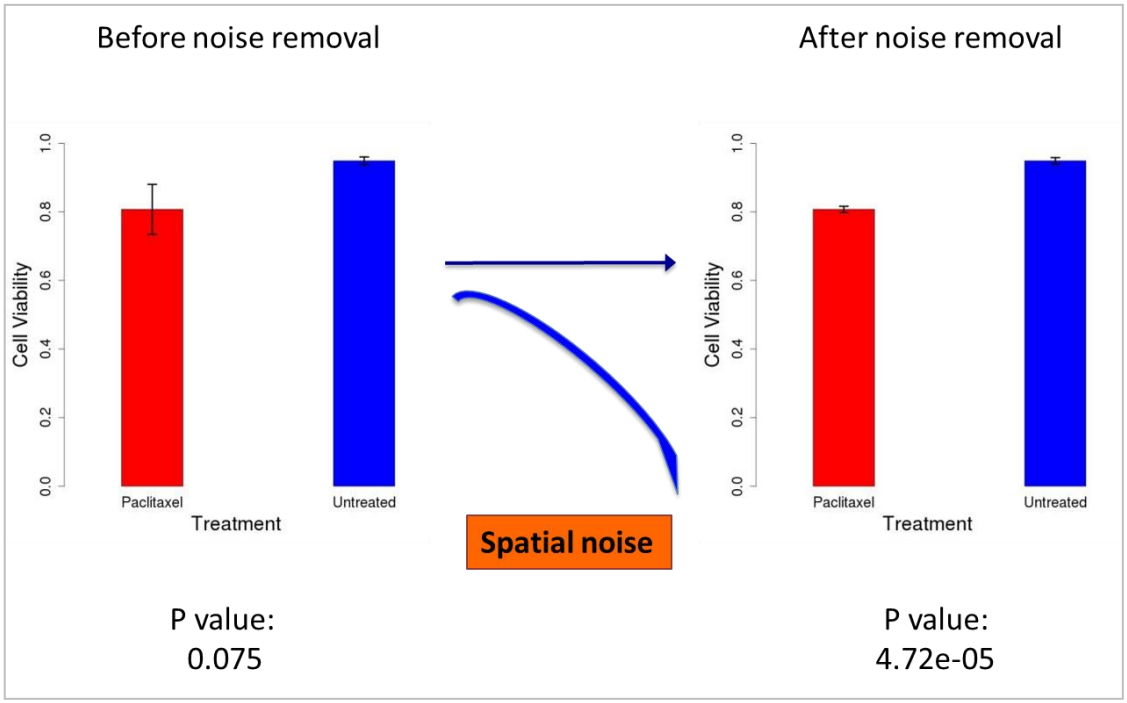


Figure 16. Increased detection power of one single siRNA from HTS. The p value is enhanced after removing spatially-correlated background noise.

For a single gene, removal of spatial noise could lead to a reduction of the variation and subsequently increase statistical power (Figure 16). For example, the P value could be improved from 0.075 to 4.72×10^{-5} under the same biological phenotypic strength for both experimental and control groups.

Globally, reduction of spatial noise resulted in an overall decrease in the coefficients of variations (Figure 17 left). For example, the 90th percentile of CV was reduced from 0.044 (original data) to 0.039 (corrected data), and consequently the statistical detection power of 10% change between the experimental and control group was increased from 0.88 (original data) to 0.94 (corrected data) for 90% of the genes from whole HTS project. Meanwhile, the type I error rate is controlled at the same level of 5%.

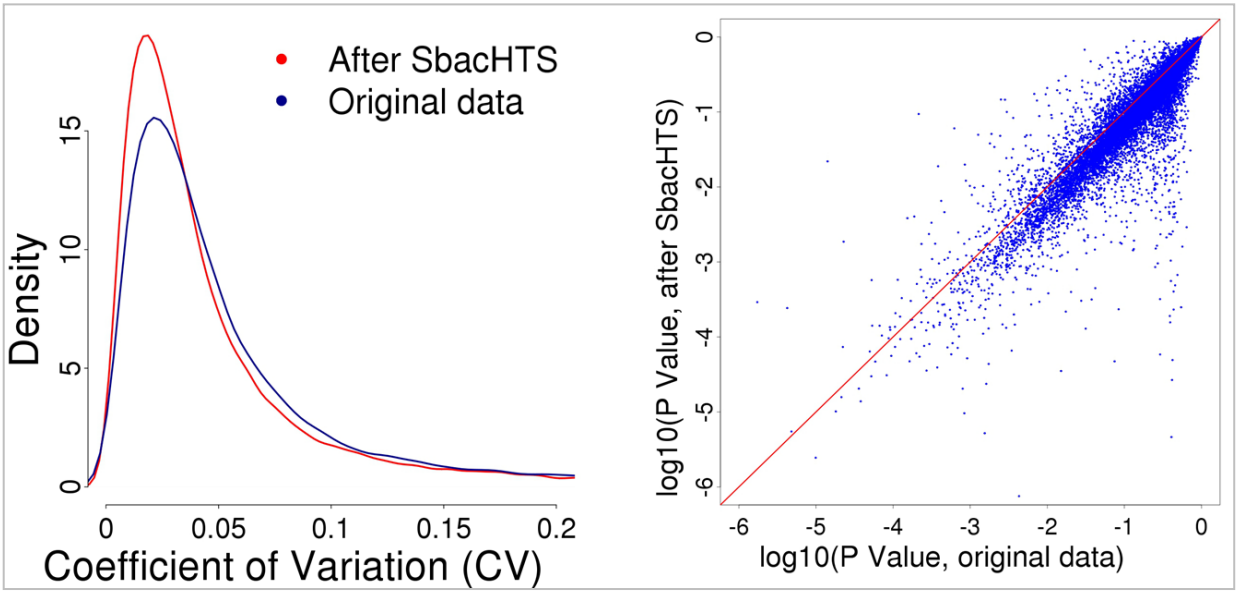


Figure 17. Increased signal-to-noise ratio and statistical detection power. Left, overall distribution of the coefficient of variation after SbacHTS modeling is reduced compared with the original data. Right, scatterplot of each individual siRNA with P values before and after SbacHTS modeling. Data was log-transformed at base 10.

Locally, we also scatterplotted P values for each individual siRNA before and after SbacHTS modeling (Figure 17 right). Overall, the P values before and after correction were consistently linear with each, indicating there is no overcorrection using SbacHTS. However, we do see statistical power was enhanced from the down-size shift pattern with respect to the diagonal line (red line). For most changesP values were decreased, while for some they were increased, which suggests that spatial background might result in false positives from high-throughput RNAi screening.

The ultimate goal of the HTS project is to identify hits from primary screening, and in our case we tried to identify siRNAs that significantly decrease cell viability under a sub-lethal concentration of chemotherapeutic drug on non-small-cell lung cancer. We used t-tests to compare cell viability between experimental and control groups. A P value was given for each gene. A Beta Uniform model (Pounds and Morris, 2002) was used to evaluate the false discovery rate (FDR). We could only identify 101 hits from the original data; however, after SbacHTS modeling, we could identify 867 hits with the same FDR level less than 0.05 (Figure 18). Therefore, we successfully helped enhance statistical detection power while controlling the same false discovery rate.

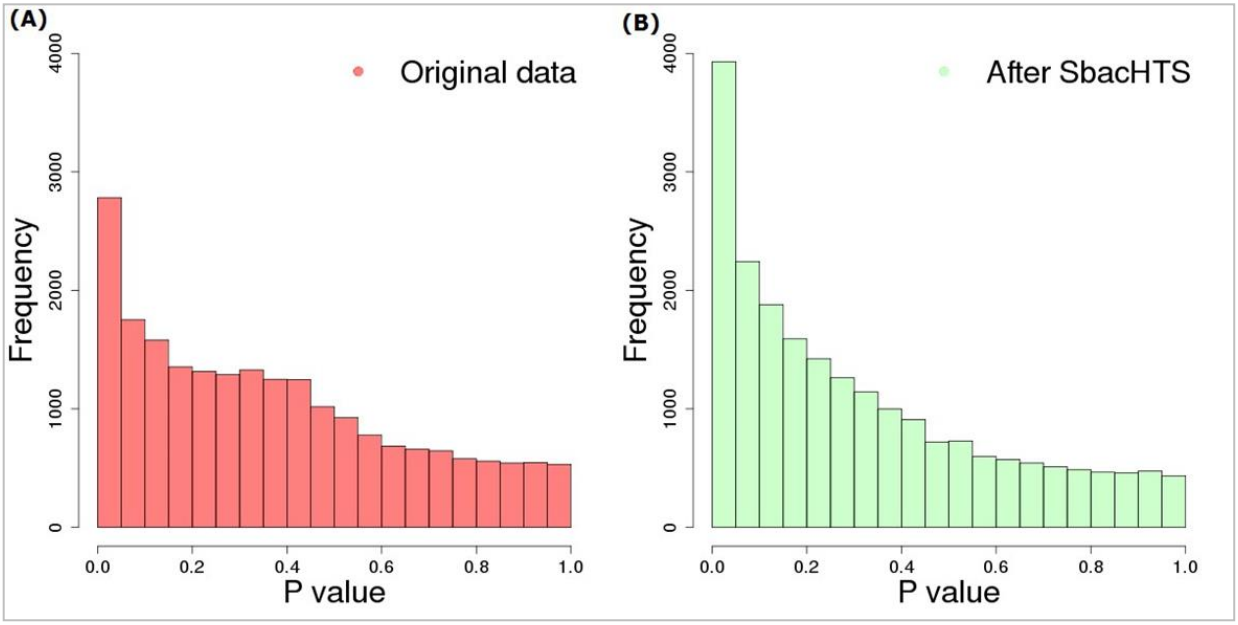


Figure 18. Histogram of P values before and after SbacHTS. A) Before SbacHTS, we can only identify 101 hits with a controlled false discovery rate of less than 5%. B) After SbacHTS modeling, we could identify 867 hits under the same criterion. The result indicates that the spatial background correction decreases the noise and improves the statistical power.

2.3.3 Visualization of batch effects

Batch effect, a systematic experimental error, is pervasive in high-throughput RNAi screening. The visualization of batch effects is the one of the most powerful approaches for detecting it. Therefore, we also included this type of functionality in SbacHTS software to provide an approach to summarize the measurements (such as observed readouts or normalized robust z scores) from each plate (Figure 19). Measurement scores are grouped across the plate and allow users to detect systematic bias or batch effects originating within the experimental procedures.

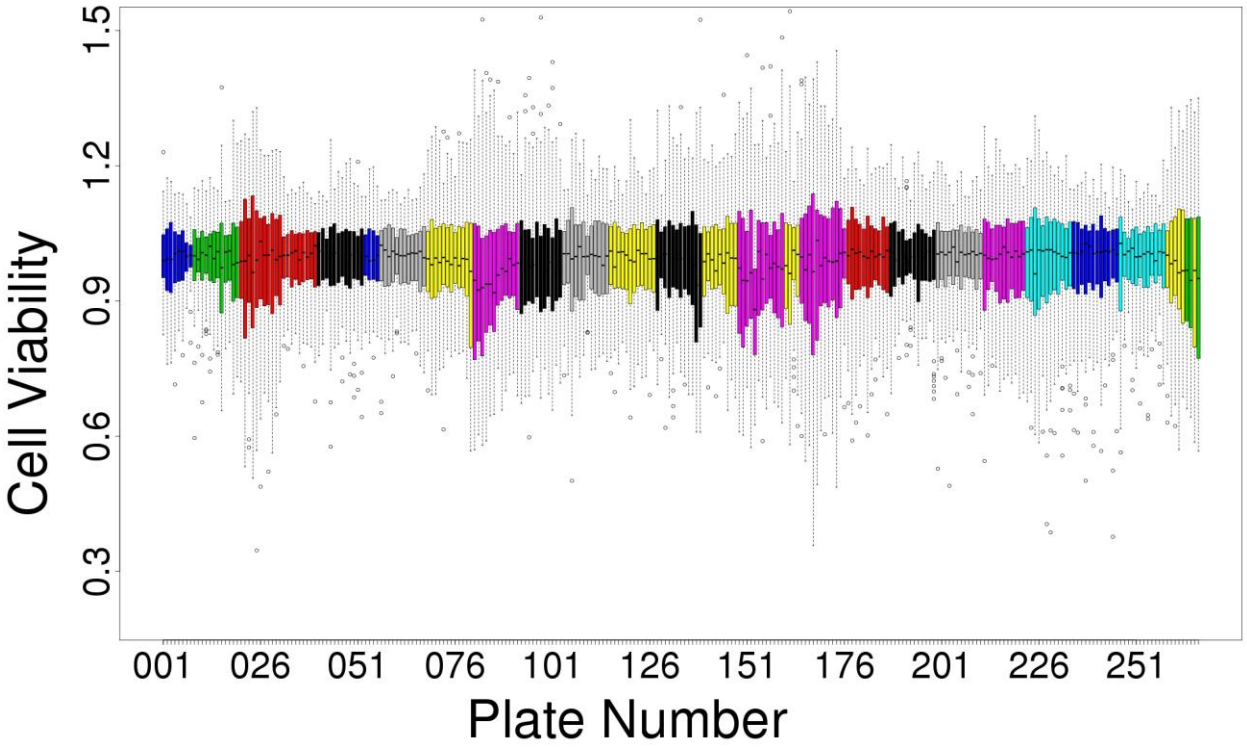


Figure 19. Batch effects from a HTS. Data from a whole genome screening was visualized using a box plot for identification of batch effects across different plates.

2.3.4 Implementation

We developed SbacHTS in R and implemented it as a web-based user-friendly Galaxy tool (Giardine B, Riemer C et al 2010) (Figure 20), available at http://www.galaxy.qbrc.org/. The user's manual is also available online.

Analysis results show that SbacHTS can identify and correct spatial background noise, enhance the signal-to-noise ratio and help with hit identification from high-throughput screening experiments. In addition, SbacHTS is computationally efficient, and only needs less than 5 minutes to process the 267-plate data from a whole genome project.

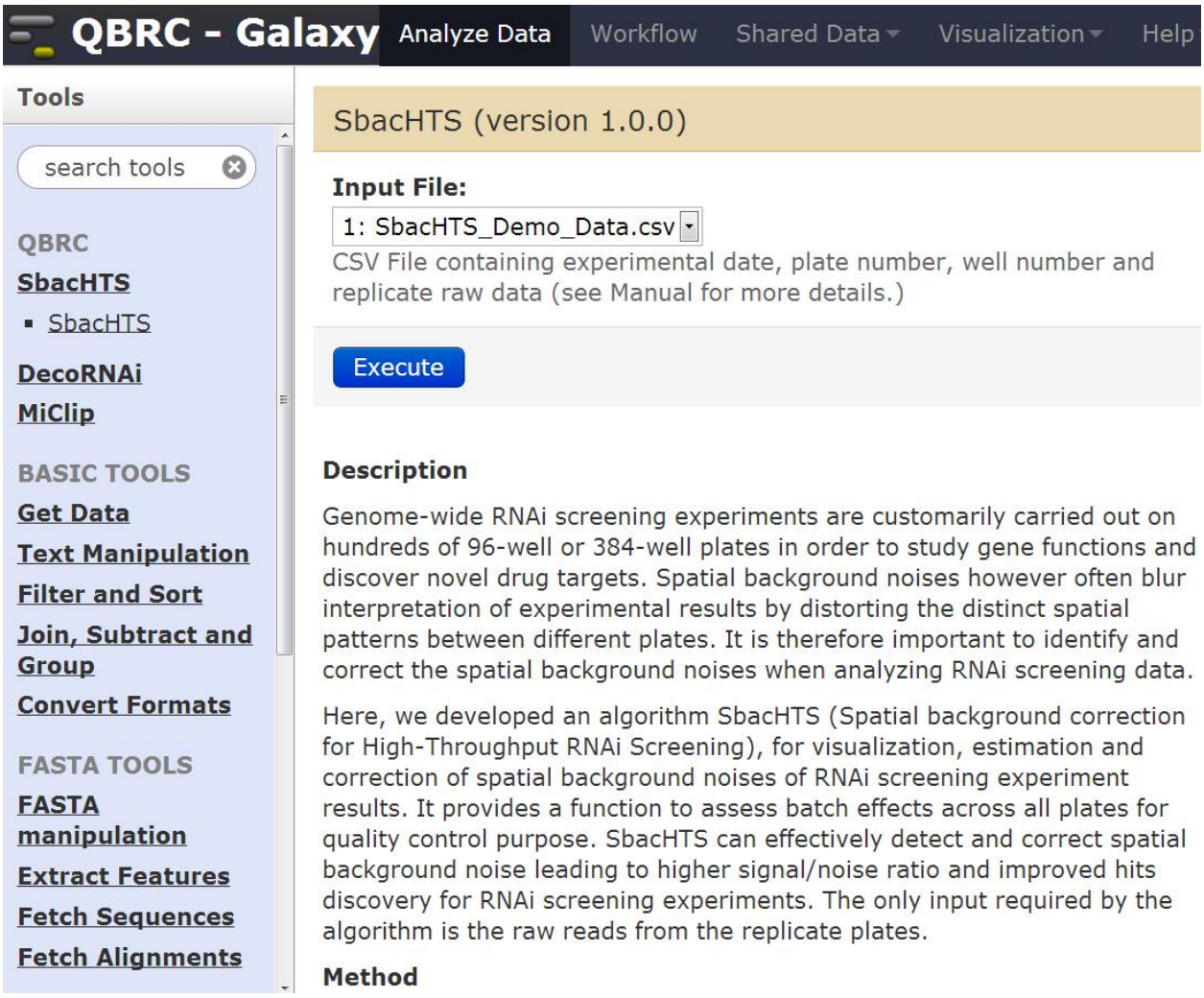


Figure 20. Snapshot of web-based software SbacHTS. We developed SbacHTS and implemented it as a Galaxy tool available within the scientific community for wider application of our algorithm in the analysis pipeline of high-throughput RNAi screening.

2.4 Discussion

In order to help with high-throughput cell-based RNAi screening data visualization and analysis, we developed a novel statistical modeling method, SbacHTS (spatial background noise correction in high-throughput RNAi screening), to help with identifying and correcting noise in HTS. HTS has been widely used for discovering new drug targets and annotating gene functions, but measurements are blurred by spatial background noise whose patterns can differ across each individual plate.

Identification and correction of such position effects becomes a computational challenge in analysis pipeline of HTS projects, and therefore we want to substantially enhance measurement accuracy and screening success by modeling HTS data. We built SbacHTS software for the visualization, estimation and correction of spatial background noise in HTS. SbacHTS is available as a web-based user-friendly bioinformatics tool on the Galaxy open source framework with open access web interface on our public Galaxy webpage. We found that SbacHTS software could effectively detect and correct spatial background noise, reduce the signal-to-noise ratio and enhance statistical detection power in high-throughput RNAi screening experiments.

Although SbacHTS was developed and demonstrated with high-throughput RNAi datasets, our approach could be readily generalized to other formats of high-throughput screening, such as small-molecule screening in which noise is also pervasive. As long as experiments are performed in triplicate or more, SbacHTS should be able to identify and detect spatially-correlated noise from experimental procedures. Therefore it can be anticipated that SbacHTS will have more applications with advancing screening technologies.

2.5 Bibliography

Banerjee, S., Gelfand, A.E. and Carlin, B.P. (2003) *Hierarchical Modeling and Analysis for Spatial Data* Chapman & Hall/CRC Monographs on Statistics & Applied Probability Series. Taylor & Francis.

Birmingham, A., *et al.* (2009) Statistical methods for analysis of high-throughput RNA interference screens, *Nat Methods*, **6**, 569-575.

Boutros, M. and Ahringer, J. (2008) The art and design of genetic screens: RNA interference, *Nat. Rev. Genet.*, **9**, 554-566.

Boutros, M., Bras, L.P. and Huber, W. (2006) Analysis of cell-based RNAi screens, *Genome Biol.*, **7**, R66.

Breitling, R., *et al.* (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, *FEBS Lett.*, **573**, 83-92.

Brideau, C., *et al.* (2003) Improved statistical methods for hit selection in high-throughput screening, *J. Biomol. Screen.*, **8**, 634-647.

Carralot, J.P., *et al.* (2011) A Novel Specific Edge Effect Correction Method for RNA Interference Screenings, *Bioinformatics*.

Chung, N. (2008) Median absolute deviation to improve hit selection for genome-scale RNAi screens, *J. Biomol. Screen.*, **13**, 149-158.

DasGupta, R., *et al.* (2005) Functional genomic analysis of the Wnt-wingless signaling pathway, *Science*, **308**, 826-833.

Dragiev, P., Nadon, R. and Makarenkov, V. (2012) Two effective methods for correcting experimental high-throughput screening data, *Bioinformatics*, **28**, 1775-1782.

Fawcett, T. (2006) An introduction to ROC analysis, *Pattern Recognit. Lett.*, **27**, 861-874.

Forster, T., Roy, D. and Ghazal, P. (2003) Experiments using microarray technology: limitations and standard operating procedures, *J. Endocrinol.*, **178**, 195-204.

Konig, R. (2007) A probability-based approach for the analysis of large-scale RNAi screens, *Nat. Methods*, **4**, 847-849.

Malo, N., *et al.* (2006) Statistical practice in high-throughput screening data analysis, *Nat Biotechnol*, **24**, 167-175.

Manly, K.F., Nettleton, D. and Hwang, J.T. (2004) Genomics, prior probability, and statistical tests of multiple hypotheses, *Genome Res.*, **14**, 997-1001.

Muller, P., *et al.* (2005) Identification of JAK/STAT signalling components by genome-wide RNA interference, *Nature*, **436**, 871-875.

Ogier, A. and Dorval, T. (2012) HCS-Analyzer: open source software for high-content screening data correction and analysis, *Bioinformatics*, **28**, 1945-1946.

Orvedahl, A., *et al.* (2011) Image-based genome-wide siRNA screen identifies selective autophagy factors, *Nature*, **480**, 113-117.

Possemato, R., *et al.* (2011) Functional genomics reveal that the serine synthesis pathway is essential in breast cancer, *Nature*, **476**, 346-350.

Wagner, E.K. (2002) Practical approaches to long oligonucleotide-based DNA microarray: lessons from herpesviruses, *Prog. Nucleic Acid Res. Mol. Biol.*, **71**, 445-491.

Whitehurst, A.W., *et al.* (2007) Synthetic lethal screen identification of chemosensitizer loci in cancer cells, *Nature*, **446**, 815-819.

Wiles, A.M., *et al.* (2008) An analysis of normalization methods for Drosophila RNAi genomic screens and development of a robust validation scheme, *J. Biomol. Screen.*, **13**, 777-784.

Zhang, J.H., Chung, T.D. and Oldenburg, K.R. (1999) A simple statistical parameter for use in evaluation and validation of high throughput screening assays, *J. Biomol. Screen. B*, **4**, 67-73.

Zhang, X.D. and Zhang, Z. (2013) displayHTS: a R package for displaying data and results from high-throughput screening experiments, *Bioinformatics*.

Zhang, X.H.D. (2006) Robust statistical methods for hit selection in RNA interference high-throughput screening experiments, *Pharmacogenomics*, **7**, 299-309.

Zhang, X.H.D. (2007) A new method with flexible and balanced control of false negatives and false positives for hit selection in RNA interference high-throughput screening assays, *J. Biomol. Screen.*, **12**, 645-655.

Zhang, X.H.D. (2007) A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays, *Genomics*, **89**, 552-561.

Zhang, X.H.D. (2007) The use of strictly standardized mean difference for hit selection in primary RNA interference high-throughput screening experiments, *J. Biomol. Screen.*, **12**, 497-509.

Zhang, X.H.D. (2008) Genome-wide screens for effective siRNAs through assessing the size of siRNA effects, *BMC Res. Notes*, **1**, 33.

Zhang, X.H.D. (2008) Hit selection with false discovery rate control in genome-scale RNAi screens, *Nucleic Acids Res.*, **36**, 4667-4679.

Zhang, X.H.D., Marine, S.D. and Ferrer, M. (2009) Error rates and powers in genome-scale RNAi screens, *J. Biomol. Screen.*, **14**, 230-238.

CHAPTER THREE

COMPUTATIONAL DETECTION AND SUPPRESSION OF SEQUENCE-SPECIFIC OFF-TARGET PHENOTYPES FROM WHOLE GENOME SCREENS

Even though high-throughput RNAi screening has been widely accepted and used in biomedical and biological research, computational challenges remain in data mining. One of those challenges is the biological pleiotropy that comes from multiple modes of action of siRNAs and transfection reagents. A major blurring feature of these reagents is the microRNA-like translational inhibition resulting from a hexamer of as short as 6 nucleotides with complementarity to many different mRNAs. We developed a computational approach, Deconvolution Analysis of RNAi Screening data (DecoRNAi), for identification and correction of siRNA-mimic-miRNA off-target effects (OTE) in primary RNAi screening data sets. Substantial reduction of false positive rates was experimentally validated in five distinct datasets from difference biological contexts across different genome-wide siRNA libraries. We also implemented a public-access graphical-user-interface that was constructed to facilitate application of our algorithm within the scientific community.

3.1 Introduction

3.1.1 False positives in primary screening in HTS

Genome-wide high-throughput RNAi screening has been widely used in biomedical and biological research for discovery of novel drug targets, identification of pathway components or

investigation into unknown molecular machinery, and has proven to be an effective and powerful means for functional annotation of protein-coding genes in a variety of biological contexts from both normal and disease samples and cell lines (Birmingham, et al., 2006; Kim, et al., 2013; Orvedahl, et al., 2011; Tang, et al., 2008; Ward, et al., 2012; Whitehurst, et al., 2007).

However, regardless of the wide acceptance and usage of HTS in biological and biomedical research, the false positive rate has continually blurred the interpretation of primary high-throughput RNAi screening. The off-target effect has been observed and recognized in research resulting from both siRNAs themselves and delivery vehicles (Jackson and Linsley, 2010). A tradeoff always appears between minimizing the false-positive rate and increasing the false-negative rate when a simple cutoff is used as a select ion criterion of hits in primary screening (Mohr, et al., 2010). Studies have shown that the false positive rate comes from both sequence-independent and sequence-dependent off-target effects (Sigoillot and King, 2011).

When cell lines were transfected with siRNAs designing to target the same genes despite their difference in sequence design (Figure 21), microarray analysis of gene expression showed that four individual siRNAs generated a dramatic gene expression pattern that was a confirmation of off-target effects from primary screening, and evidence was observed that the off-target effect is highly associated with short sequence region within the 5’ end of siRNAs. It presented itself as a new computational challenge for high-throughput RNAi screening data analysis and hit selection (Birmingham, et al., 2006; Jackson, et al., 2006). The ultimate goal of a primary screening from HTS projects is to pick as many true positives as possible without excessively comprising the false-negative. Simply using a stringent cut-off from primary

screening doesn't satisfy this need, which must be overcome before high-throughput RNAi screening can have a wider application, such as clinical trials of siRNA as novel therapeutics.

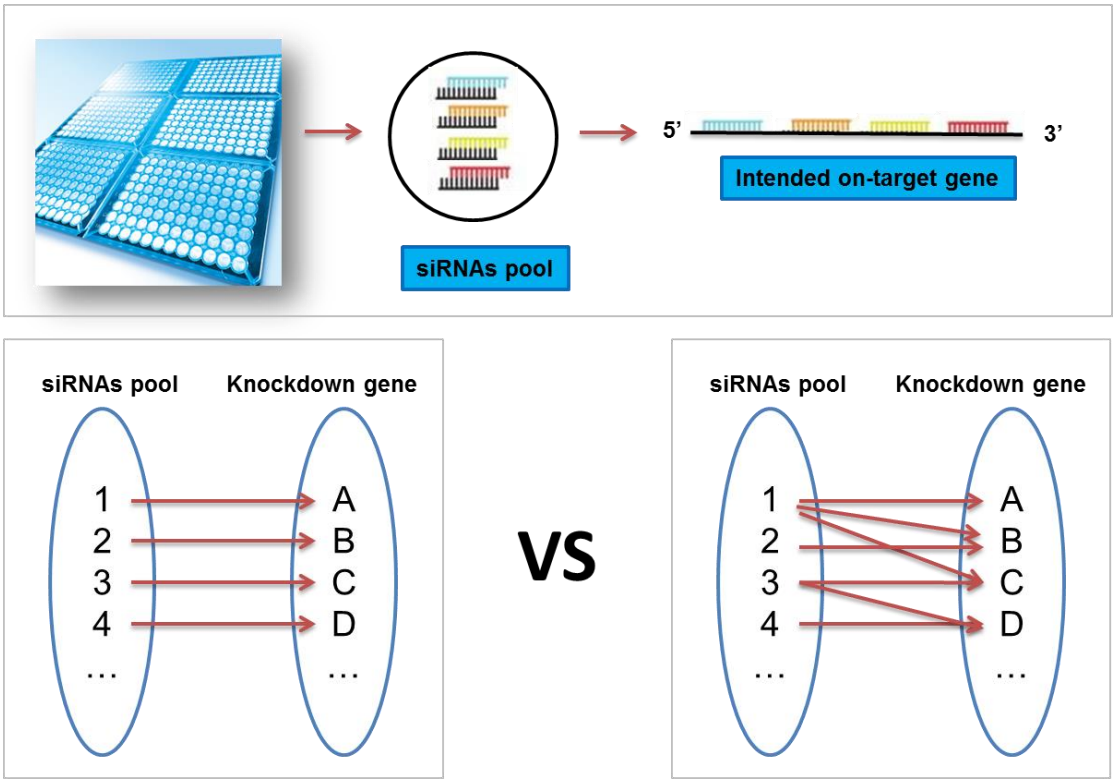


Figure 21. Off-target effect in high-throughput RNAi screening. Despite the fact that all pooled siRNAs are designed to target the same gene regardless of their sequence difference, evidence is observed that one siRNA pool might simultaneously inhibit multiple genes, causing an off-target effect in HTS projects and blurring the interpretation of most primary screening.

3.1.2 siRNA-mimic-miRNA off-target effect in HTS

miRNA is another small RNA molecule that in its mature form is about 20~26 nucleotides long. One miRNA can target multiple mRNAs, and seed match is a major determinant of miRNA/mRNA complementarity. "Seed" refers to the 2`7 hexamer sequence on the 5' end of miRNAs. Given that both mature miRNA and siRNA share structure similarity, it is

quite possible that siRNA might mimic miRNA to raise off-target effects in primary high-throughput RNAi screening. This becomes a pressing challenge for studies where the goal is to maximize the return of accurate gene-specific information. However, one individual siRNA often interferes with the hundreds of gene expression through partial sequence complementarity between hexamers on the 5' end of siRNAs and mRNAs (Jackson, et al., 2006; Sigoillot, et al., 2012). Therefore the phenotypic results from siRNA screens usually consist of the intentional "on-target" effects of target gene depletion together with unintentional "off-target" effects that are hexamer-sequence-dependent, but target-gene-independent (Figure 22).

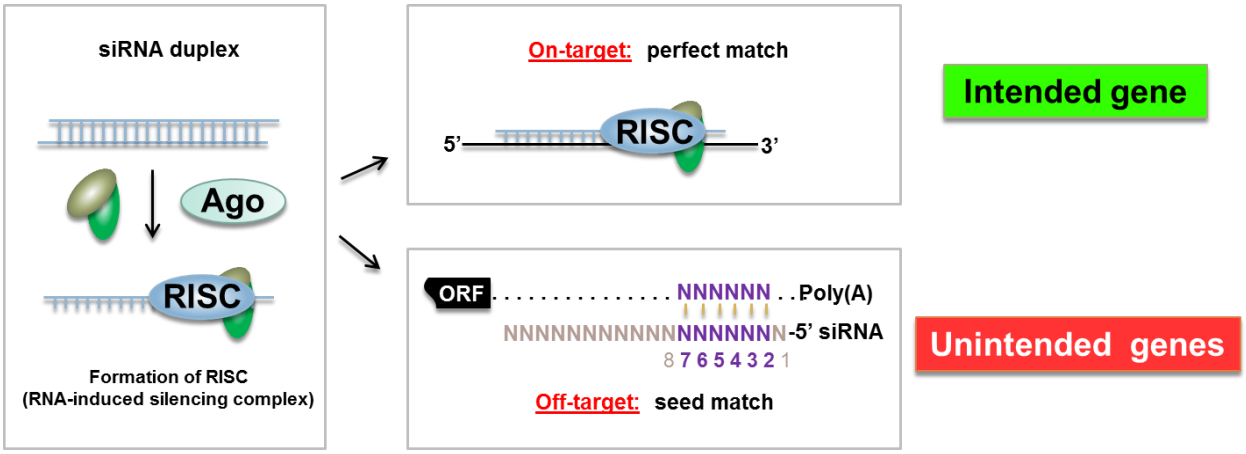


Figure 22. On-target vs. off-target effects. In a siRNA knock down event, siRNAs might cause both on-target and off-target effects. On one hand, they can inhibit intended gene through perfect match while on the other, they can also off-target unintended genes via seed match.

siRNA-mimic-miRNA off-target effect can lead to many false positives that consequently obscure interpretation of the overarching screen results. Time- and resource-intensive experimental approaches for target validation therefore often define the limits of the reliable gene-level information from any given screen. Computational approaches have been designed which can help identify off-targeted transcripts within a given screening effort, and

therefore lead to discovery of new genes or pathways associated with the phenotype under investigation (Bartel, 2009; Buehler, et al., 2012). However, directly addressing high false positive rates and deconvolution of off-target phenomena is still a major bottleneck restraining the pace of discovery for functional genomics efforts. Here we developed a computational approach to identify and correct siRNA-mimic-miRNA off-target effects from high-throughput RNAi screening.

3.2 Methods and Materials

3.2.1 Data processing

All data processing and z-score derivations were consistent with the original publications (Kim, et al., 2013; Orvedahl, et al., 2011; Tang, et al., 2008; Ward, et al., 2012; Whitehurst, et al., 2007)

(1). For the H1155 toxicity screens (Whitehurst, et al., 2007), host modulators of H1N1-cytopathogenicity (Ward, et al., 2012) and the HCC4017 toxicity screens (Kim, et al., 2013), raw cell viability data were transformed to a robust Z score (formula shown below) and adjusted for batch effects. That is, raw data were grouped by experimental batch and within each group, the sample median and median absolute deviation were used to calculate a robust Z score. Annotation of all siRNA/miRNAs pools and their associated z-scores can be found in the corresponding publications.

$$z\ score = \frac{cell\ viability - sample\ median}{median\ absolute\ deviation\ (MAD)} \tag{16}$$

$$MAD = median_i\left(\left|X_i - sample\ median\right|\right) \quad (17)$$

(2) For the WNT pathway siRNA screen (Tang, et al., 2008), Z scores were calculated as a standard score centered on the population mean of each screening run as described by the average of each triplicate experiment minus the standard deviation (SD). Annotation of all siRNA pools and their associated z-scores can be found in publications.

(3) For the selective autophagy siRNA screen (Orvedahl, et al., 2011), the mitochondrial mass for each cell was approximated by the following formula: Mitochondrial Mass ~ β0 + β1 Parkin + β2 siRNA + β3 Parkin ×siRNA. Two-way ANOVA models were used to identify siRNAs that decreased Parkin-mediated mitophagy:

$$y_{ijk} = \mu + \tau_i + \theta_j + (\tau\theta)_{ij} + \varepsilon_{ijk} \quad (18)$$

and Z scores were calculated as the statistical significance. Annotation of all siRNA pools and their associated z-scores can be found in publications.

3.2.2 DecoRNAi analysis

The LASSO (least absolute shrinkage and selection operator) regression approach was adapted to quantify the strength of seed-link effects. For this analysis, each Z score was modeled as a linear combination of on-target and seed sequence-based off-target effects. The LASSO regression model was defined as below:

$$\vec{Z} = X\vec{\beta} + \vec{Y},\ \text{subject to}\ \left|\vec{\beta}\right| < s \quad (19)$$

where Zi is the i[th] original Z score, β_j is the estimated off-target effect of the j[th] seed family, Y_i is the corrected Z score (on-target effect) and λ is the penalty parameter, X is denoted as below:

$$X = \left[x_{ij} \right], xi_j = \begin{cases} 1, \text{if } i \in j \text{ family} \\ 0, \text{otherwise} \end{cases} \tag{20}$$

And the solution is given as:

$$\beta = \arg\min_{\beta} \left[\left\| \vec{Z} - X\vec{\beta} \right\|^2 + \lambda \left| \vec{\beta} \right| \right] \tag{21}$$

For each seed family, we can thus estimate the coefficient that indicates the strength and direction of predicted off-target effects. A negative coefficient means the seed family tends to lower Z scores and vice versa. Based on empirical experience, λ is set to 0.001 as the default. We annotate those coefficients with an absolute value > 1 as indicating candidate off-target effects for all four datasets shown in this manuscript. However, all the parameters and cutoff values are tunable by users.

For LASSO-selected off-target seed families, we further examine the statistical significance using the Kolmogorov-Smirnov test (KS-test). Taking $\vec{Z}$ as a vector of original Z scores from the primary screening, the empirical distribution function F_{n_s} for Z scores from seed family S is defined as:

$$F_{n_s}(z) = \frac{1}{N_s} \sum_{i=1}^{N_s} I_{Z_i \le z} \tag{22}$$

Where I_{Z_i} is the indicator function, equal to 1 if $Z_i \le z$ and equal to 0 otherwise, and N_s is the total number of Z scores from seed family S. The Kolmogorov–Smirnov statistic for a given cumulative distribution function F(z) is as follows:

$$D_{n_s} = \sup_z \left| F_{n_s}(z) - F(z) \right| \tag{23}$$

The statistical significance (p value) was then determined by the Kolmogorov-Smirnov statistic.

3.2.3 Web-based application (Galaxy)

The DecoRNAi application is available at http://galaxy.qbrc.org/root?tool_id=sirna_offtarget, which is an open web-based interface. Analysis parameters can be specified by users as below:

- InputFile: CSV File containing response variable and siRNA sequence data.
- Strand: Specify the strand orientation for analysis.
- Lambda: Penalty parameter used in the model.
- Seed Range: 1-14. Specify the seed region to be used.
- Library: Specify siRNA library. Default is custom which requires user input sequences.
- Strength: Specify the cutoff for strength of seed-linked effect. Must be a positive value.
- Significance: Specify the cutoff for significance (P-value).

3.2.4 Tissue culture, oligo transfection and cell viability assays

H1155 cells were grown in RPMI 1640 (Gibco®) supplemented with 5% fetal bovine serum (FBS; Atlanta Biologicals) and 1% penicillin/streptomycin (Gibco®). All siRNAs were purchased from Dharmacon. The library contains 24 sets of 4 siRNAs each. The oligos targeting transmembrane protein 114 (TMEM114) from Dharmacon were used for siRNA-negative control. The miR 4633-5p and the synthetic miRNA were from Ambion. Nontargeting miRNA control (IN-001005-01-05) was from Dharmacon. For reverse transfection, 1ul siRNA (10uM) in 30ul serum free media (SFM) was mixed with 0.4ul RNAi Max (Invitrogen) in 10ul SFM. 40ul siRNA-reagent mix per well and 5000 cells per well, from a single cell suspension, were delivered in 100ul media in 96-well microtiter plates. Cell viability was measured 96 hours post-transfection with CellTiter-Glo (Promega) according to the manufacturer's specifications.

3.3 Results

3.3.1 siRNA-mimic-miRNA off-target effects

To the best of our knowledge of miRNA/mRNA complementarity, a major determinant of translational inhibition of mRNA by a given miRNA is seed match (Figure 3) between mRNA and miRNA. "Seed" refers to a 6-nucleotide hexamer on the 5' end of the miRNA (Bartel, 2009). Therefore, if a sRNA mimics miRNA mechanism to off-target unintended genes, the "seed sequence" should also apply to it. To test this hypothesis, we examined a dataset from a whole-genome HTS toxicity screen on non-small-cell lung cancer cell (Whitehurst, et al., 2007), H1155, and used it as a benchmark to examine the strength of this "seed sequence" mimic

phenomena. In this screen, we employed an arrayed one-gene/one-well commercial siRNA library with pooled siRNA oligonucleotide duplexes (4 siRNAs per well). The end-point assay was to measure cell viability and identify the genes required for non-small-cell lung cancer survival and growth. In our library, a siRNA is 19 nucleotides long and if we define any continuous hexamer as a potential "seed sequence", we have 14 different locations to define a "seed" (Figure 23 left). A Kolmogorov–Smirnov test (KS-test) was used to examine the seed sequence/phenotype association for all hexamer windows and only 1-6, 2-7, and 3-8 had a statistically significant association at a controlled false discovery rate 5% (Figure 23 right). In order to keep with the current best understanding of dominant determinants of the miNRA targeting mechanism (Bartel, 2009), we used a 2-7 hexamer window as the defined "seed sequence" in the following studies.

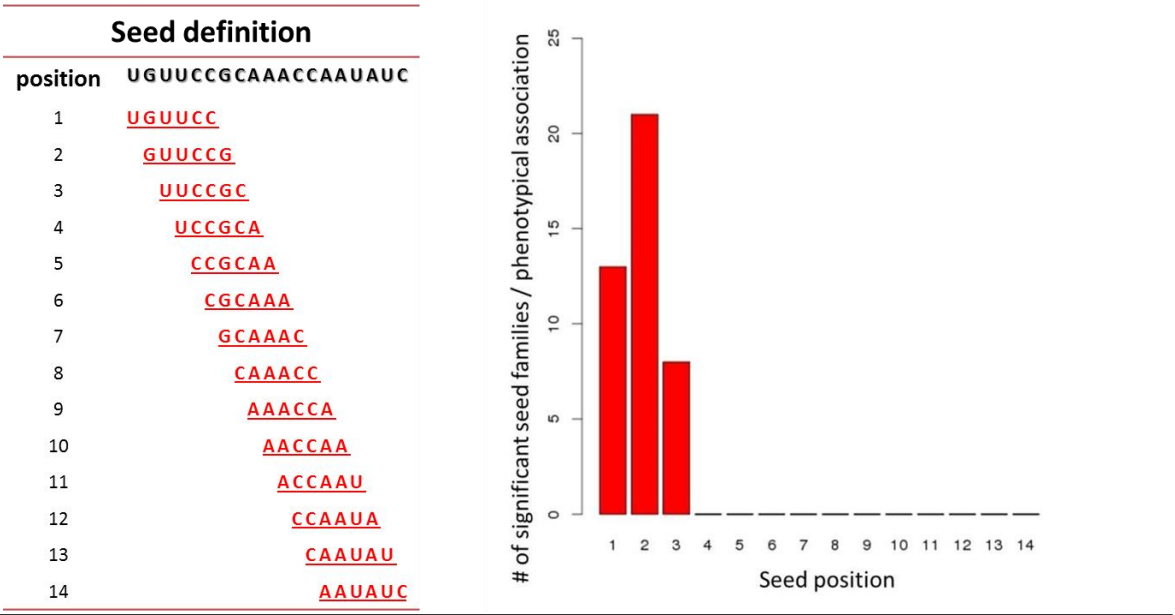


Figure 23. Define seed sequences on siRNAs. Left, an illustration of a hexamer sliding window for different definition of seed sequence along a 19 nt-long siRNA; and right, a seed sequence/phenotype association for different seed sequence at controlled false discovery rate, by which we can tell the 2~7 has the strongest association.

Seed sequence membership for each of the 168,992 oligonucleotides in the library was separately defined for each of the siRNA sequences. siRNAs sharing the same seed sequence are grouped as a seed family (Figure 24 left). In a genome-wide siRNA library, a sum total of 4^6 (4,096) possible non-redundant "seeds" are present and on average for each seed family there are almost 40 siRNA oligos (Figure 24 right). The presence of a given seed within such family size gives us an opportunity to estimate and identify siRNA-mimic-miRNA off-target effects.

Seed Family: **GUUCCG**

Gene Symbol	siRNA sequence	Z score
C21ORF49	U**GUUCCG**CAAACCAAUAUC	-9.11
HRK	A**GUUCCG**CCUGCCGAGCAG	-5.94
NSMCE2	U**GUUCCG**UCAGCUUCUCUG	-5.78
MBNL3	U**GUUCCG**CCCAUUAAUCUC	-5.69
PSG6	A**GUUCCG**ACUCUUAGGUUC	-5.51
CCDC12	A**GUUCCG**CAGCCUAAGUUC	-5.27
XKR6	U**GUUCCG**CCAUGGAUGAUC	-4.85
C14orf115	U**GUUCCG**UGGAGCAUCUUC	-4.71
LOC286129	U**GUUCCG**AGAGAUAAGACA	-3.06
NCAPH	U**GUUCCG**UAAUCUGCGUGU	-1.98
LOC400684	A**GUUCCG**UAAAUCAGGCCC	-1.20
LOC388129	U**GUUCCG**ACGGCUAAGACC	-0.68
SLC25A1	U**GUUCCG**GUAUUUGUGCGC	-0.40
HLA-G	U**GUUCCG**UGUCUCCUCUUC	0.62
LOC400512	G**GUUCCG**CCUCGUGGGUUA	0.80

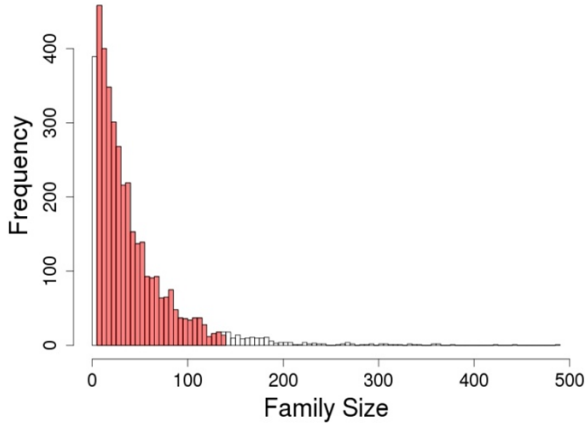


Figure 24. Seed family. Left, a demonstration of a seed family. All siRNAs originally targeting different genes with different sequences share the common seed sequence "GUUCCG"; right, a frequency distribution of seed family size from whole-genome siRNA library.

Though we started with the KS-test, a liability of the KS-test is in that it is quite sensitivity to family size, which is a common problem with most statistical tests (Figure 25), which leads to false positive discovery since the P value is small for large family size though biological strength (off-target effect) is not strong. Therefore, we developed a novel algorithm, DecoRNAi (de-convolution analysis of RNAi screening data), to estimate the strength and direction of seed-associated off-target effects using LASSO (least absolute shrinkage and

selection operator) as a penalized linear regression model. DecoRNAi is robust to large family size and based on the biological phenotypic strength of seed association. We used the H1155 dataset as a benchmark to develop a scoring system and as development of a methodology. We also applied it to a wide spectrum of datasets from different biological contexts across different siRNA libraries.

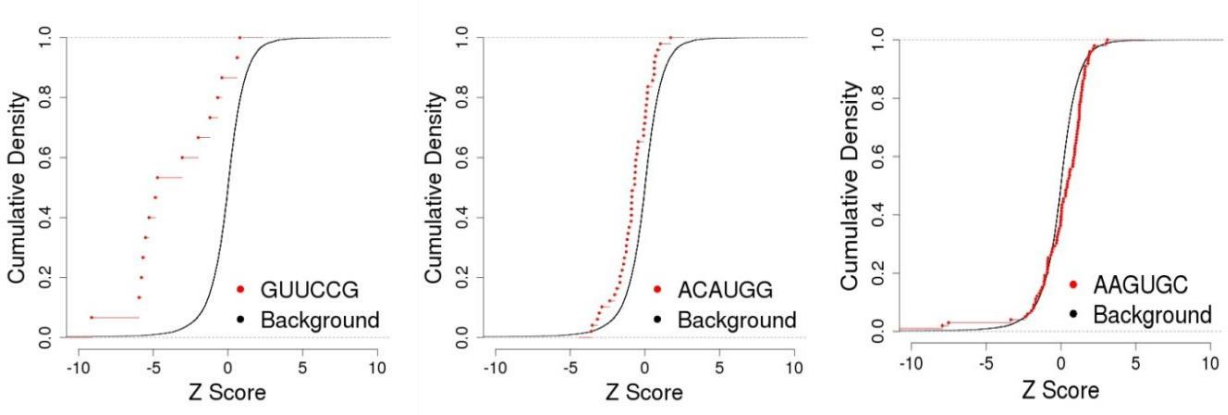
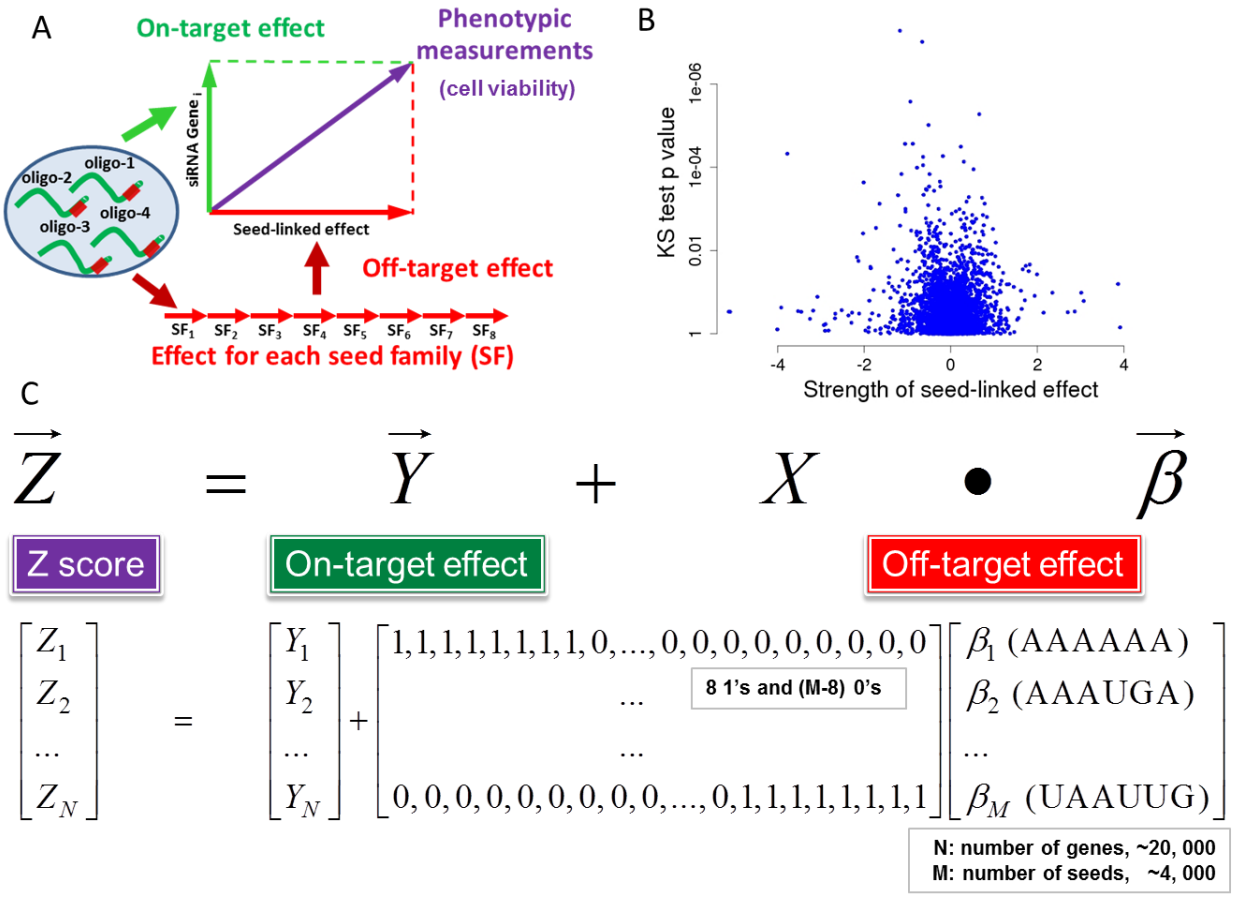


Figure 25. The liability of the KS-test depends on sample size. Examples of the empirical distribution plot for three different seed families of increasing family size (15, 49, and 100, respectively) show similar KS-test P values (~10^{-5}). The KS-test is sensitive to family size and motivates us to develop novel algorithm to identify siRNA-mimic-miRNA off-target effects.

In summary, we tried to project the phenotypic measurements onto two dimensions. One is the on-target effect coming from the knock-down of intended gene; the other is from seed-driven off-target effects (Figure 26 A). Because of the presence of 8 single-stranded siRNAs in one pool, the off-target effect is the combinatory sum of 8 potential off-target seed families. Mathematically, we were partitioning the phenotypic readout, a robust Z score, into two parts, on-target effect and off-target effect (Figure 26 C). The off-target effect is cumulative from 8 seed families. Usually in one siRNA library we have ~4,000 seed families and they are estimated. Only 8 are present in one Z score, which determines the design matrix (Figure 26 C).

A volcano plot shows the resulting seed family scores after DecoRNAi modeling (Figure 26 B). The seed-linked effect was plotted against statistical significance, and globally we can tell the distribution of the off-target effects for each seed family. From which, based on empirical experience, we identify 13 off-target seed families (Table 1) and follow-up experimental validations were carried out to test the validity of our approach.



$$\vec{Z} = \vec{Y} + X \bullet \vec{\beta}$$

$$\begin{bmatrix} Z_1 \\ Z_2 \\ \dots \\ Z_N \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_N \end{bmatrix} + \begin{bmatrix} 1,1,1,1,1,1,1,1,0,\dots,0,0,0,0,0,0,0,0,0 \\ \dots \\ 0,0,0,0,0,0,0,0,0,\dots,0,1,1,1,1,1,1,1,1 \end{bmatrix} \begin{bmatrix} \beta_1\ (\mathrm{AAAAAA}) \\ \beta_2\ (\mathrm{AAAUGA}) \\ \dots \\ \beta_M\ (\mathrm{UAAUUG}) \end{bmatrix}$$

Figure 26. Mathematical demonstration of DecoRNAi. A) original phenotypic readout Z scores are projected onto both on-target effect (green) and off-target effect (red) to perform a deconvolution analysis. B) global visualization of seed-linked effect and statistical significance from a H1155 study is plotted via volcano plot. C) mathematical demonstration of deconvolution analysis of high-throughput RNAi screening results is shown in illustration.

Out of 661 pools with significant Z scores (≤ -3), 10.29% (68/661) contained identified off-target effects from 13 identified off-target seed families. They correspond to 365 siRNA pools from a whole-genome screen. This indicates these off-target effects are quite pervasive in high-throughput RNAi screening hit selection. In order to evaluate the experimental performance of seed-driven effects, we chose four "off-target" seed families (GUUCCG, UCCAGG, UUGCAG, UAUGCC) (Figure 27) for a secondary individual oligo screen.

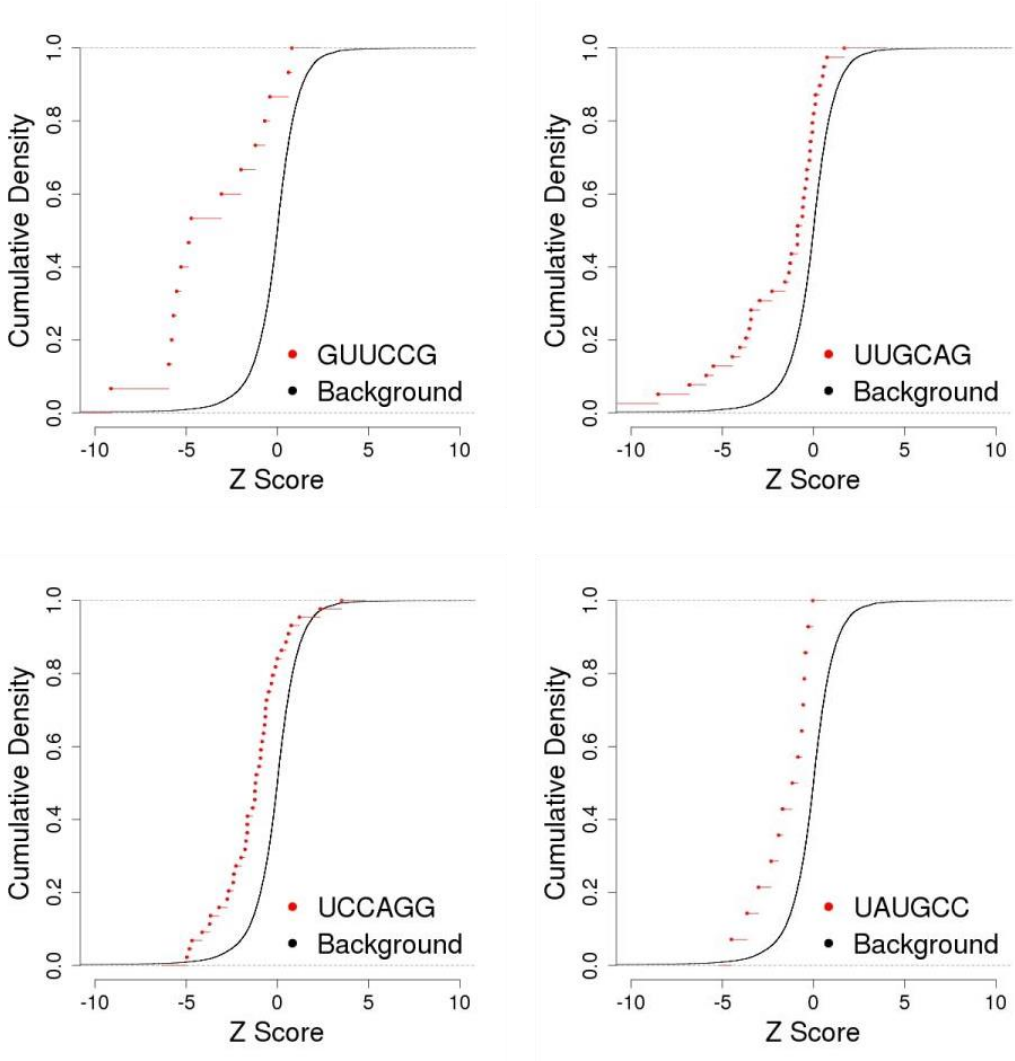


Figure 27. Seed families to be re-tested. Four selected off-target seed families for secondary screening are to be evaluated for identified off-target effect from HTS projects.

Seed family	Strength of seed-linked effect	Family size	Significance (P value)
GUUCCG	-3.77	15	4.72E-05
GUGUAC	-2.02	7	3.88E-03
UACUCC	-2.01	36	2.31E-04
ACAUGC	-1.72	21	2.87E-03
UUGCAG	-1.64	39	7.51E-04
CCCGCA	-1.32	11	9.46E-03
UAUGCC	-1.21	14	4.70E-04
UCCAGG	-1.17	44	5.24E-08
UCAGUU	-1.17	8	2.53E-03
UUCACC	-1.14	29	1.45E-04
UAUAGG	-1.05	69	2.74E-05
UAGGAG	-1.05	43	9.99E-04
ACUAGU	-1.04	31	1.16E-03

Table 1. Summary of identified off-target seed families from H1155.

From the four selected off-target seed families, twenty-four genes were selected for further analysis. For each gene, four individual siRNAs were separately evaluated for their efficacy on H1155 cell viability upon successful transfection (Figure 28). Therefore for each pool, siRNAs are classified into off-target siRNAs (red) with a predicted killing effect and same-pool siRNAs (green) without killing effect.

Based on our prediction, off-target siRNAs should have lower cell viability than those same-pool siRNAs. Secondary screen results confirmed our prediction in that for each individual gene, off-target oligos almost always have the lowest cell viability (Figure 28, left). Identified off-target seeds were strongly associated with consequences on cell viability. Consistent with individual-gene evaluation, the cumulative density function also suggests a dramatic difference between off-target siRNA Z score distribution and same-pool siRNA Z score distribution (Figure 28, right, P value < 0.01).

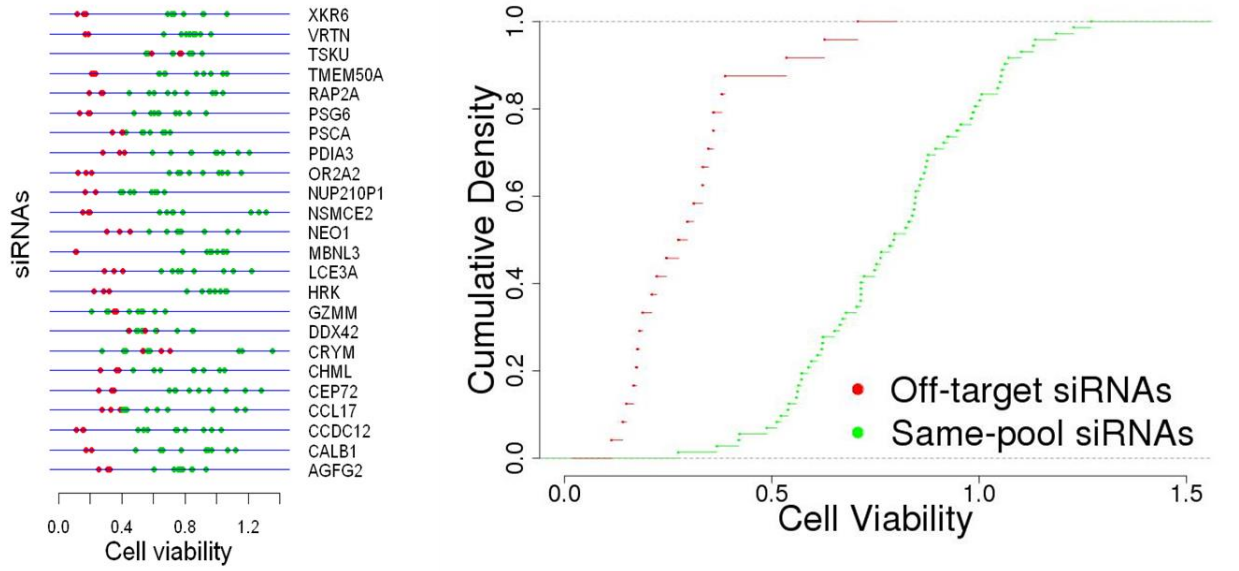


Figure 28. Experimental validation of identified off-target effects. Red dots represent cell viability of off-target siRNAs on H1155 and green dots represent cell viability of same-pool siRNAs. Left, off-target siRNAs are stronger associated with killing effect than same-pool siRNAs within each gene. Right, cumulative density also indicates dramatic difference between off-target siRNAs and same-pool siRNAs with P value < 0.01. Data provided by JiMi Kim from Dr. Michael White's lab.

At present we have identified off-target seed families based on siRNA-mimic-miRNA hypothesis. One interesting question is what would occur if a known miRNA shares the same seed with an identified off-target seed family. Of the 13 detected seed families, only one (UAUGCC) is found to be present within the seed sequence of an annotated human miRNA (hsa-miR-4633-5p). Consistent with the prediction, introduction of the miRNA mimic of miR-4633-5p resulted in a similar viability defect (Figure 29, top). In order to further evaluate the sufficiency of seed sequence-dependent induction of off-target phenotypes, we had a synthetic miRNA corresponding to the validated seed family GUUCCG (Figure 29, bottom). It also effectively inhibited cell viability on H1155 cells. The evidence is strong that siRNA does mimic the miRNA mechanism to cause off-target effects in primary screen.

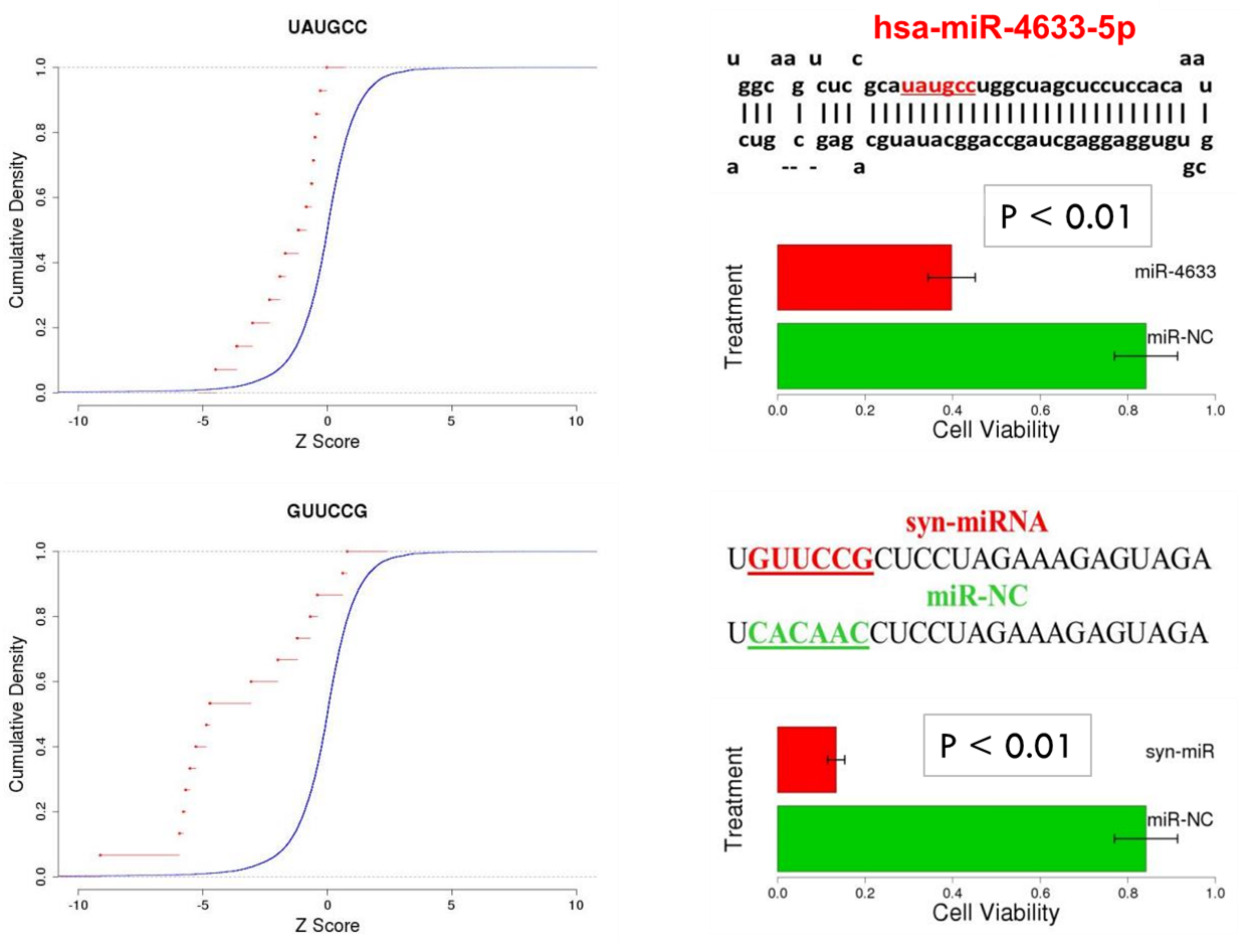


Figure 29. Seed sequence-dependent induction of off-target effect. Top, an miRNA mimic of an has-miR-4633 shared seed with identified off-target seed family UAUGCC that significantly inhibited H1155 cell viability (P value < 0.01). Bottom, a synthetic miRNA mimic containing the predicted off-target seed GUUCCG also significantly inhibited H1155 cell viability, while negative control of an identical sequence with the exception of the seed region did not (P value < 0.01). Data provided by JiMi Kim from Dr. Michael White's lab.

As a methodology, we sought to test our algorithm on different siRNA library platforms. To this end, we evaluated the performance of DecoRNAi using a distinct genome-wide siRNA library, and we examined an additional toxicity screen designed to identify genes required for lung cancer cell viability but using another non-small-cell lung cancer cell line

HCC4017 on siRNA library platform from the Ambion (Kim, et al., 2013). DecoRNAi identified 10 off-target seeds from the library enriched in these screens (Table 2).

Seed family	Strength of seed-linked effect	Family size	Significance (P value)
ACAUGU	-1.29	43	4.42E-06
ACUACG	-1.37	7	7.39E-03
AGGUCG	-1.28	10	6.47E-04
AUGUCC	-1.02	16	6.48E-03
GUAGUU	-1.04	45	1.61E-06
UAGGUC	-1.46	63	1.28E-13
UAGUUG	-1.06	81	7.45E-08
UCGUAC	-1.34	10	3.64E-03
UCUGAC	-1.68	33	1.47E-04
UGCUCU	-1.20	10	1.68E-03

Table 2. Summary of identified off-target seed families from HCC4017 screen.

The global score was visualized (Figure 30, A). We also tested individual siRNAs to evaluated identified off-target effects. 60 individual siRNA oligonucleotides were retested for their killing effect upon cell line HCC4017 and consistently showed that off-target siRNAs have dramatic consequences on cell viability as compared to other same-pool siRNAs targeting the same genes (Figure 30, B).

Sometimes scientists and researchers are trapped in situations such that knock-down is working well and phenotype is present. However, follow-up experiments don’t support the hypothesis. Everything works well except that the results have nothing to do with target gene. Here we even measured the mRNA expression level for selected siRNAs, and results show that even though all individual siRNAs successfully silenced gene expression (Figure 30, C), only off-target siRNAs had the phenotype of interest. This is strong evidence that the observed phenotype was uncoupled with on-target gene. It was the off-targeted genes or pathways that

were producing observations. Therefore identification of such off-target effects is important to remove false positive hits.

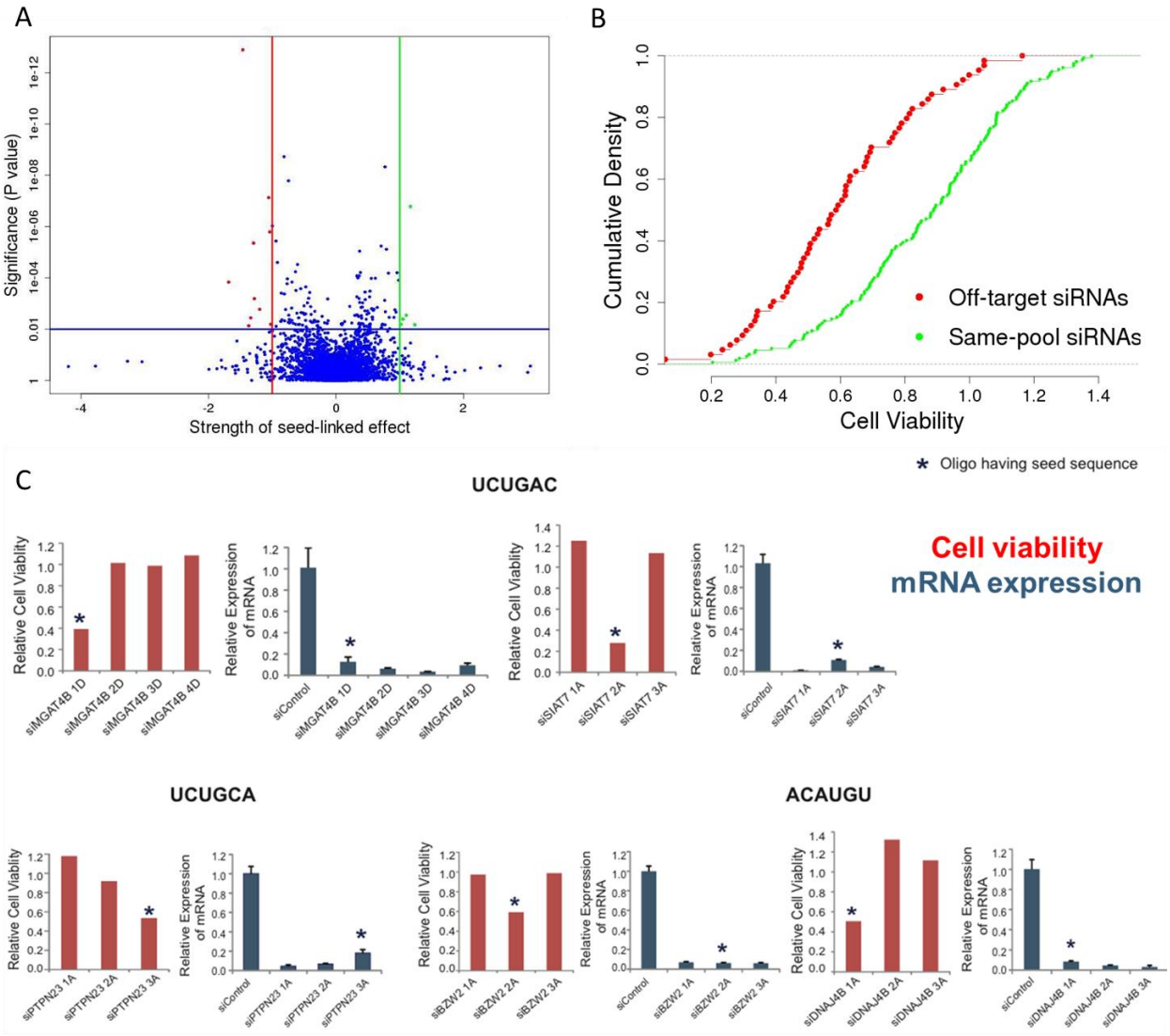


Figure 30. Off-target effect validations. A) global visualization of scores from HCC4017 HTS project. X axis is strength of seed-linked effect and y axis is statistical significance. B) experimental validation of identified off-target effect using individual siRNAs. C) mRNA expression showed observed phenotype was uncoupled with on-target genes. Data provided by JiMi Kim from Dr. Michael White's lab.

As in the previous example, we tested annotated miRNA and synthetic miRNA as well and phenotypes were consistent with a dominant seed-sequence dependent mode of action (Figure 31), further supporting the validity of our approach. miRNA mimic hsa-miR-4256 shared identified off-target seed UCUGAC, and synthetic miRNA contained seed ACAUGU. Both had a killing effect on cell line HCC4017.

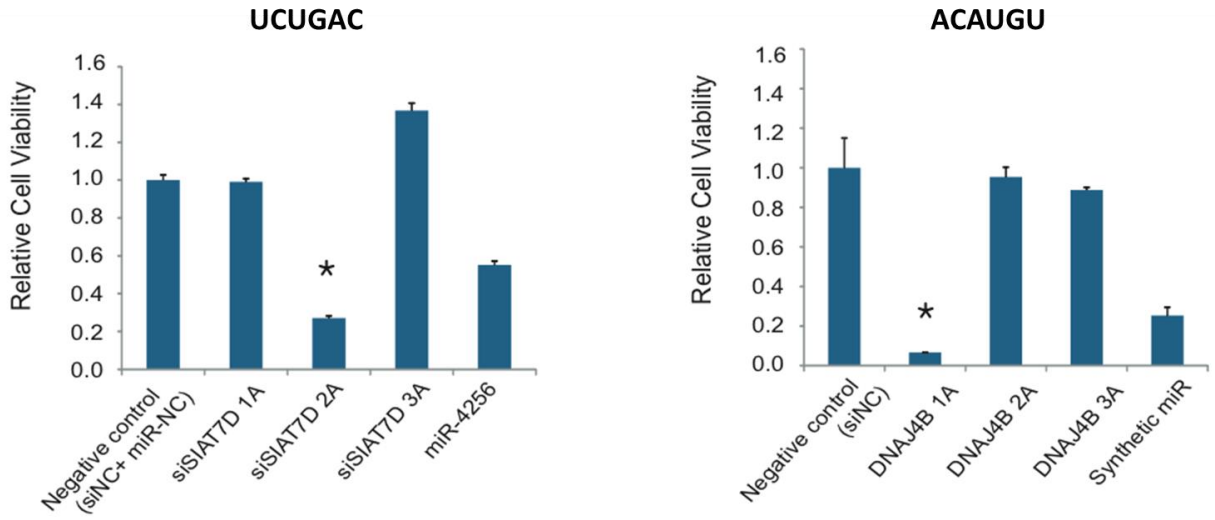


Figure 31. siRNA-mimic-miRNA validation. Validation of seed-sequence-dependent induction of off-target effect by miRNA mimic and synthetic miRNA. Data provided by JiMi Kim from Dr. Michael White's lab.

So far we have been talking about lung cancer cell survival and growth, and therefore the potential off-targeted genes and pathways are fairly numerous. We can identify off-target siRNA from such screening projects. One question is what would occur if we narrowed down the potential off-targeted biological context. In theory, off-target effects should be much less.

Here we applied DecoRNAi to a WNT screen project that attempted to return genes modulating WNT pathway activation, and therefore used a very specific endpoint assay based on a WNT-specific and a WNT-independent reporter gene combination (Tang, et al., 2008).

Consistent with our hypothesis, only one seed-sequence association was identified among reagents that selectively reduced WNT relevant activity. This may be suggestive of the narrower biological space that can be off-targeted and interfere with the endpoint assay employed in this screening effort, and therefore less off-target effect was observed.

Seed family	Strength of seed-linked effect	Family size	Significance (P value)
GCAUGG	-1.99	10	7.13E-03
CGUCAG	-1.43	6	9.83E-03
UAGGCA	-1.15	43	1.22E-04
AGGCAU	-1.15	24	7.82E-03
CCGAAU	-1.06	8	3.38E-03
UGUUGG	-1.04	35	9.97E-04

Table 3. Summary of identified off-target seed families from autophagy screen.

For primary high-throughput RNAi screening, the ultimate goal is to identify as many true positives as possible. In other words, we want to reduce the false positive rate. How can we help that? Based on our model, after we identify the off-target effects, we can remove them from the primary Z score and have the corrected Z score. Here the autophagy screen is an image-based high-throughput RNAi screen to identify gene products required for virus-induced autophagy by measuring colocalization of the Sinbis virus capsid protein with autophagolysosomes at the single cell level (Orvedahl, et al., 2011). We could identify six significant seed-sequence associations with inhibition of selective autophagy corresponding to 125 siRNA pools from primary screening (Table 3). In secondary individual oligo screening, off-target siRNAs trended towards lower Z scores than those same-pool siRNAs (Figure 32, B). The merit of this screen is that they have a relatively large-scale secondary individual screen, and therefore we can help evaluate the performance of the corrected Z score (Figure 32, C). The corrected Z-score is a better measure to prioritize the targets for further validation than the original Z-scores and reduce

the false positive rates. With that said, no matter how many hits are selected, the corrected Z score will always have a lower false positive rate. For example, the false positive rate for the top 20 “hits” rank ordered by the primary Z-score is 24%, and it was reduced to 17% when using the corrected Z score. In summary, we helped to reduce the false positive rate from primary screening.

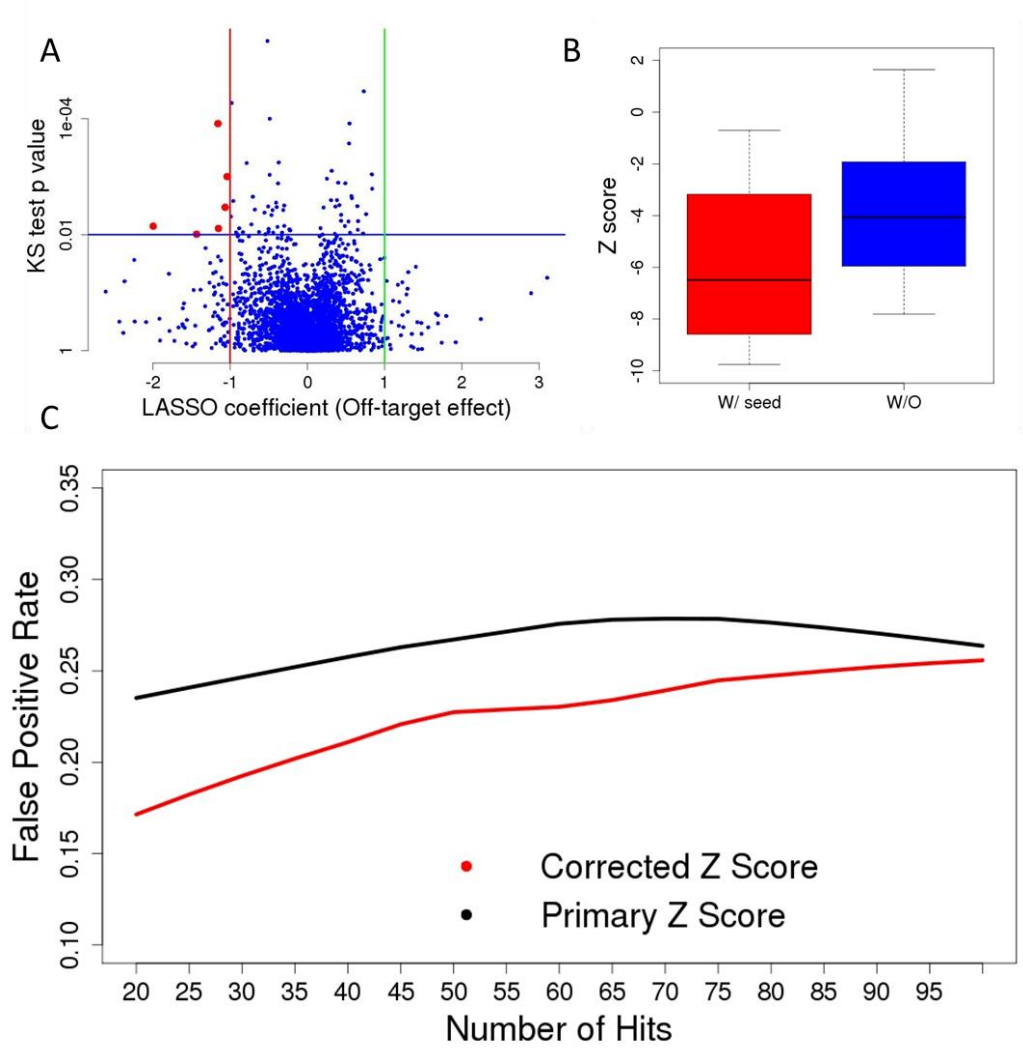


Figure 32. Off-target effect in autophagy screen. A) global visualization of DecoRNAi scores from an autophagy screen. B) in a secondary screening, off-target siRNAs have lower Z score than same-pool siRNAs. C) after identification of off-target effects, the false positive rate was reduced.

Seed family	Strength of seed-linked effect	Family size	Significance (P value)
UGCGGU	-1.58	13	1.61E-03
UGGUAG	-1.39	22	7.06E-03
ACGUGG	-1.32	47	1.45E-03
UUCUGC	-1.13	34	5.67E-03
ACUGGG	-1.04	35	8.39E-04
AUCUGG	-1.02	139	6.32E-11
UGCUGU	-1.02	33	4.42E-03
UCAUGG	-1.00	31	3.47E-04
UUGGGU	1.01	32	4.03E-04
UAAUGC	1.01	25	5.27E-04
UACCCG	1.14	23	1.05E-03
UUGGUC	1.14	10	3.56E-04
UCCGUA	1.80	6	4.20E-03

Table 4. Summary of identified off-target seed families from a virus infection screen.

In our last example, the H1N1-cytopathogenicity screen project attempted to identify genes that modulate influenza virus replication in human bronchial epithelial cell line HBEC30 (Ward, et al., 2012). Besides primary high-throughput RNAi screening, we also have parallel miRNA mimic screen and H1N1-induced cytopathogenicity that were measured using cell viability as the endpoint assay in our experiments. From the primary siRNA screen, DecoRNAi identified 13 significant seed families corresponding to 8 synthetic lethal off-target seed family (353 siRNA pools) and 5 synthetic viable off-target seed family (96 siRNA pools) (Table 4). Only 1 of 8 synthetic lethal off-target seeds shared a common seed with a human miRNA; hsa-miR-491. Of note, the hsa-miR-491 mimic has the lowest cell viability out of all reagents from the miRNA mimic screen (Figure 33, right). This is strong evidence that the underlying hypothesis of this siRNA-mimic-miRNA mechanism is true, and additionally our approach is validated again in that DecoRNAi is effective in identification of such off-target effects from primary whole-genome RNAi screening.

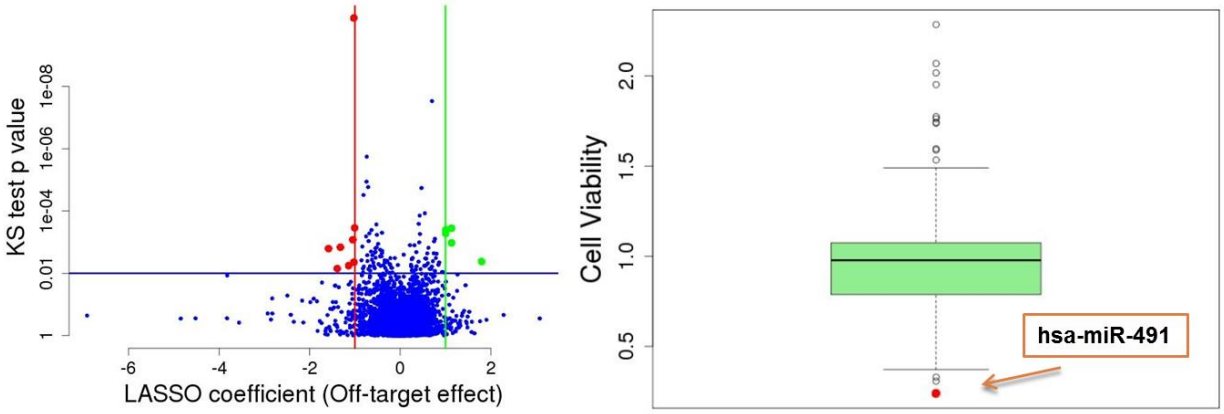


Figure 33. Off-target effects in virus infection HTS. Left, a global visualization of DecoRNAi scores from a virus infection screen project. Right, the miRNA mimic screen validated that hsa-miR-491, a miRNA mimic that shares an identified off-target seed, has the lowest cell viability.

3.3.2 Web interface implementation

To help users implement the DecoRNAi algorithm, we have created a public access web-based graphical user interface at http://galaxy.qbrc.org/root?tool_id=sirna_offtarget for custom analysis (Figure 34). Users only need to input siRNA pool identifiers and phenotypic measurements (for example, Z score), and we will conduct the analysis. We have also included pre-computed seed sequence families for 3 commonly employed commercial siRNA libraries (Ambion and Dharmacon). Custom collection analysis is also available, and the tool will compute seed sequence from a user-supplied reagent sequence table. The default parameters were provided for the DecoRNAi online tools based on the empirical performance, but all parameters are able to be re-defined by users. The output files include global score visualization, identified seed families, the siRNA pools containing off-target effect, corrected z-scores and the annotated miRNAs with phenotypes of interest.

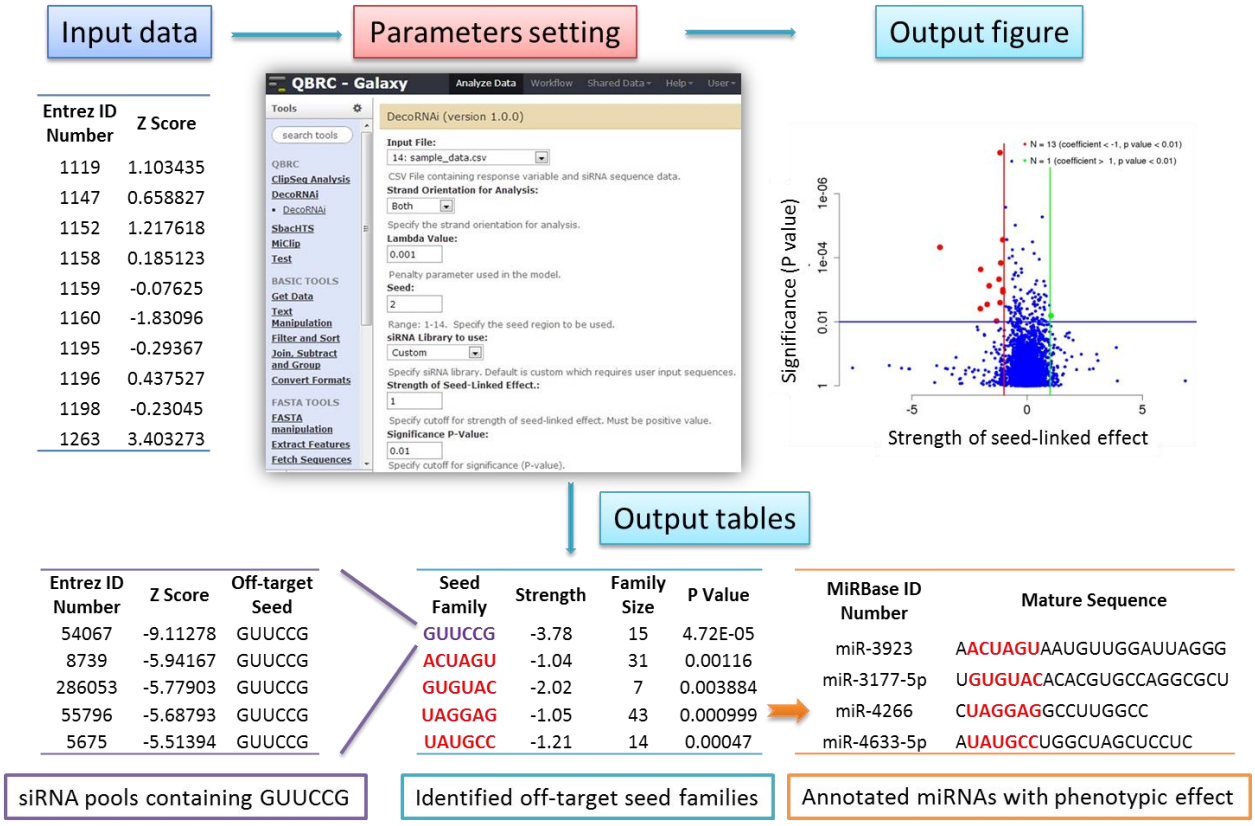


Figure 34. Illustrations of the web-based graphical user interface DecoRNAi. Seed families are pre-computed for the Dharmacon Library circa 2005, Dharmacon Library circa 2009, and Ambion Silencer Select. For screens employing these reagents, the only required input is the quantitative screen measurement for each reagent (for example, normalized z-score). Other libraries can be analyzed upon uploading the library-wide sequence information for each oligonucleotide or processed shRNA. Parameter settings are user-selected. The output files include a global visualization of seed family behavior, the predicted off-target seed families, the siRNA pools containing off-target seed families, the potential miRNAs sharing common seeds with identified off-target seed families, and the corrected z-scores.

3.3.3 Summary

Here we have designed a data-driven computational approach, Deconvolution Analysis of RNAi Screening data (DecoRNAi), to quantify the strength and direction of siRNA-mimic-miRNA off-target effects as well as statistical significance and correction of off-target effects

from whole-genome RNAi screens to reduce the false positive rate (Figure 35), all validated on multiple datasets from different biological contexts across different siRNA libraries.

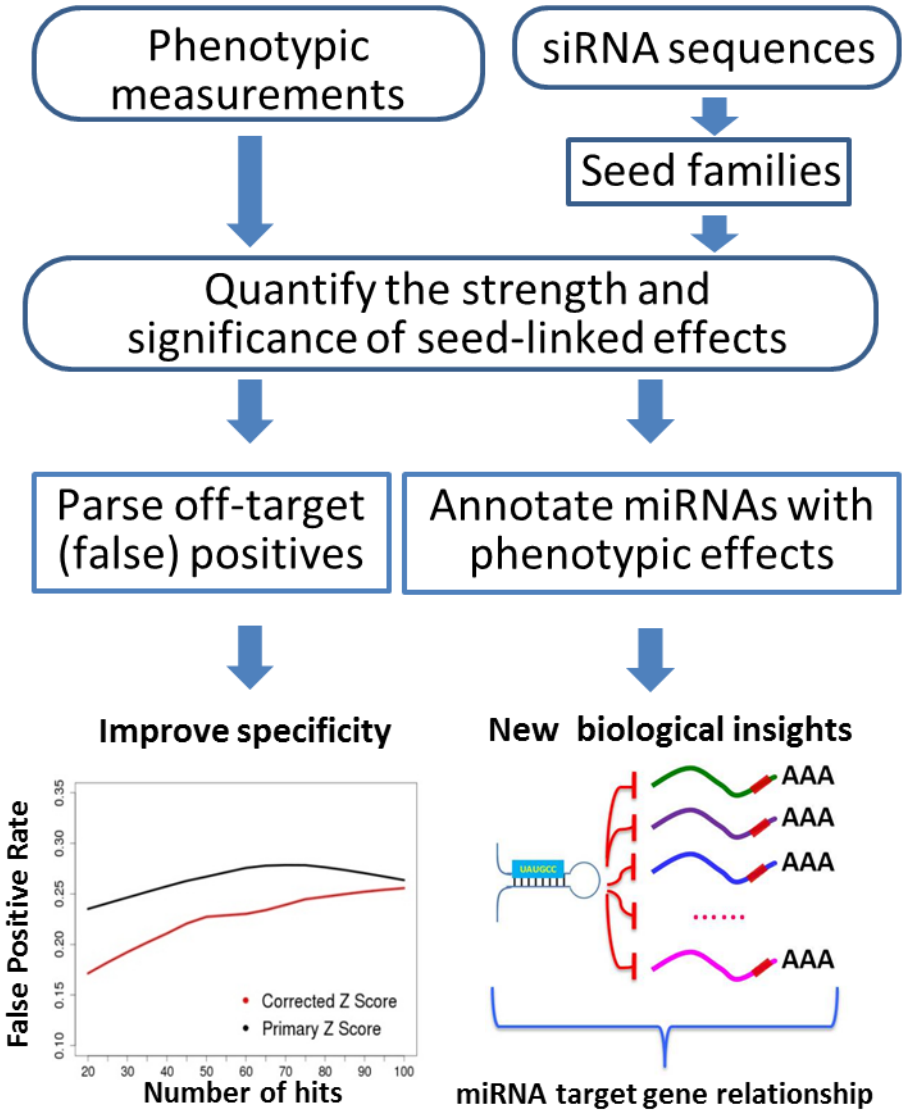


Figure 35. Workflow of DecoRNAi analysis. From the phenotypic measurements and siRNA sequences, we can quantify the strength and direction of seed-linked off-target effects as well as statistical significance.

In order to simultaneously estimate gene-specific on-target effects and siRNA-mimic-miRNA off-target effects, we developed a deconvolution algorithm to partition phenotypic measurements from primary high-throughput RNAi screening datasets. We tested our approach on 5 independent whole-genome siRNA screens, and found that microRNA (miRNA) mimicry by siRNA oligonucleotides is a pervasive source of "off-target" biological phenomenon in HTS projects. Application of DecoRNAi significantly enhanced the accurate return of single gene-specific observations at whole-genome scale, and is provided here an open-source tool to enhance lead discovery accuracy from high-throughput RNAi screening studies.

We applied DecoRNAi to five whole-genome siRNA screens employing distinct biological contexts and endpoint assays. These included a non-small-cell lung cancer screen for genes required for lung cancer cell growth and survival, a siRNA and miRNA mimic screen for host modulators of H1N1- cytopathogenicity, a siRNA screen for modulators of WNT reporter gene activation, a siRNA image-based screen for selective autophagy factors, and one additional screen for lung cancer drug target discovery using a distinct whole-genome siRNA library.

3.4 Discussion

We constructed DecoRNAi to quantitatively identify seed-dependent off-target effects by modeling the enrichment of oligonucleotide sequence-specific effects from genome-wide RNAi primary screen data. We don't require arbitrary phenotypic threshold selection, and we attempt to combine the statistical significance of population separation with phenotypic effect size to return biologically meaningful correlations. We found that the algorithm performed well on multiple datasets across diverse phenotypic assays and within distinct reagent collections. As

expected, siRNA-mimic-miRNA behavior of siRNA oligonucleotides was a pervasive feature associated with primary screening phenotypes. This was detectable by DecoRNAi, experimentally verifiable, and could be imitated with appropriately designed synthetic miRNA-like molecules.

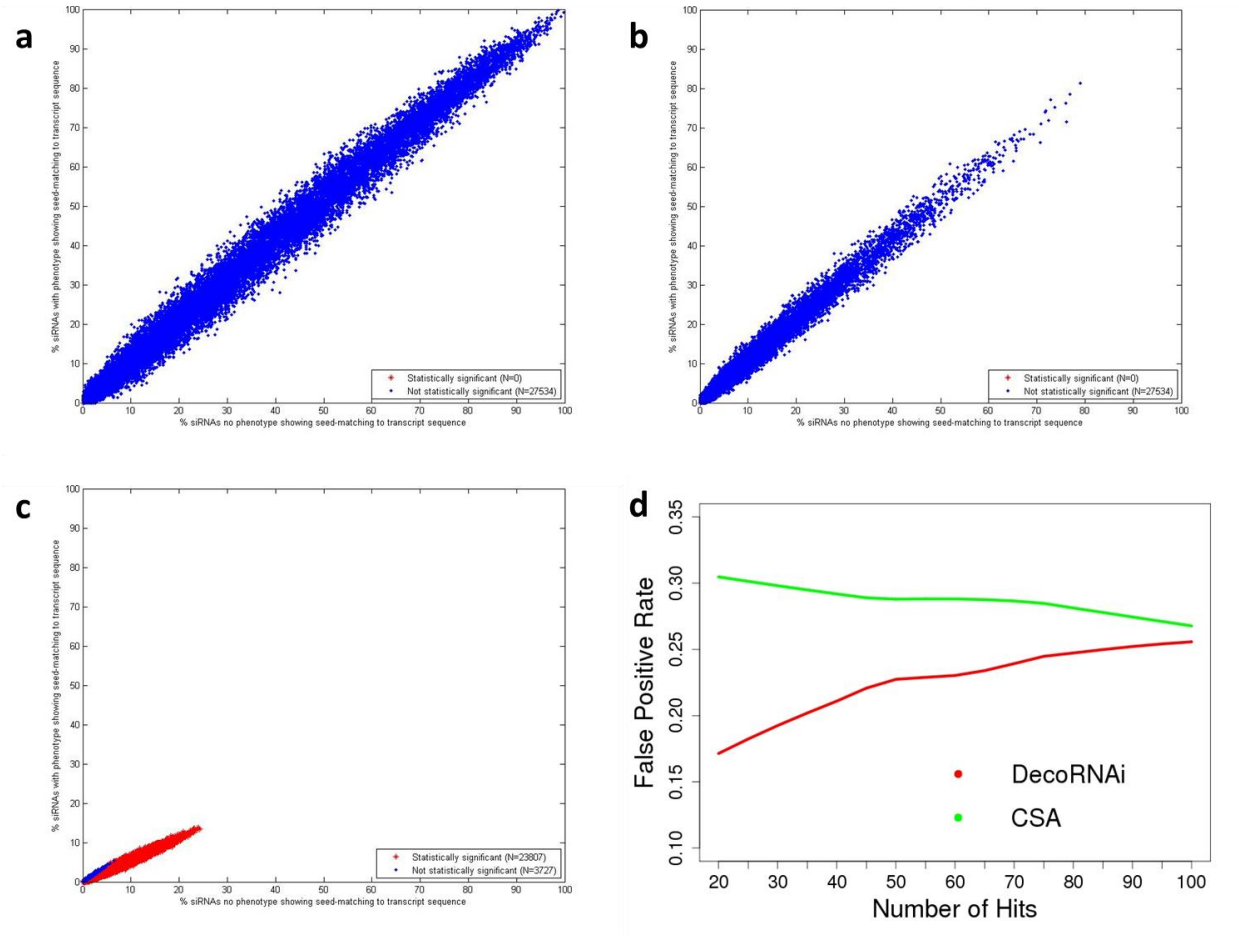


Figure 36. Comparison with GESS and CSA analysis. GESS analysis of human mRNA 3′ _UTRs from primary data of the H1155 toxicity screen (a), the selective autophagy screen (b), and the H1N1 cytopathogencity screen (c). Each point represents one 3' _UTR and represents the SMFa value plotted against the SMFi value. (d). DecoRNAi-mediated Z score corrections reduce false positive rates compared to the CSA approach from the selective autophagy screen. Here, gene targets scoring positive with 2 or more confirmed siRNAs out of a total of 4 are considered to be true positives. The X-axis indicates arbitrary "hit" thresholds based on rank-ordered Z-scores after applying the DecoRNAi approach or the CSA approach, and the Y-axis indicates the corresponding false positive rate.

GESS (Genome-wide enrichment of seed sequence matches) is a recently reported computational tool designed to identify off-targeted transcripts rather than to isolate and correct off-target phenotypes (Sigoillot, et al., 2012). We applied the GESS algorithm in an attempt to employ it later. However, this method identified no off-target siRNA pools from either the H1155 toxicity screen or the selective autophagy screen. In stark contrast, this approach identified 23,807 off-target siRNA pools from the H1N1 cytopathogencity screen (Figure 36 a-c). However, we anticipate that GESS's intended utility will compensate DecoRNAi, providing a mechanism to help identify gene cohorts that are responding to siRNAs responsible for seed-sequence driven phenotypes.

Two additional computational efforts designed to deflect spurious gene-level annotations from large-scale RNAi screens are ATARiS (Analytic Technique for Assessment of RNAi by Similarity) (Shao, et al., 2013) and CSA (Common Seed Analysis) (Marine, et al., 2012). ATARiS was developed to detect coherent behavior from multiple shRNAs targeting the same gene. While effective, the method is less generalizable outside of pooled shRNA screens and requires multi-sample RNAi screens (at least 10 samples in their publication). CSA, like DecoRNAi, detects correlated biological behavior of siRNAs that share the same seed sequence. However, CSA does not account for family-size bias with its statistical significance metric. Integration of statistical significance with the strength and direction of biological phenotypes is likely an important consideration for optimized detection of false positives (Supplementary Figure 4c-e). Furthermore, DecoRNAi quantifies seed-driven off-target effects by modeling the on-target effects and off-targets from all individual siRNA oligo duplexes in the same gene pool, which is more efficient than looking at individual siRNA seed families separately. In support of

these considerations, we found that the DecoRNAi-corrected Z scores had a significantly better true positive rate than CSA corrections (Figure 36 d).

A limitation of DecoRNAi is appropriate representation of seed families within a given screening collection to reach sufficient statistical power for detection of phenotypic associations. However, from the cumulative analysis of 5 different whole genome siRNA screens, we estimate that the DecoRNAi approach will cover ~85% of the seed sequence families present in a typical commercial arrayed siRNA library (Figure 37, A and B). To facilitate automated application of DecoRNAi to siRNA and shRNA library screening efforts, we have embedded pre-computed seed family annotations for three commonly used commercial RNAi libraries (Dharmacon Library circa 2005, Dharmacon Library circa 2009, and Ambion Silencer). In addition, we provided a tool for automated generation of seed family annotation of user-specific siRNA or shRNA oligonucleotide collections (http://galaxy.qbrc.org/root?tool_id=sirna_offtarget). The tool is based on Galaxy open source framework and accepts the phenotypic measures (such as z-scores) from the primary screen as input, and users can easily apply different parameters for analyzing the data. All of the user-specified parameters are well documented, and the intermediate outputs are provided, in order to make it convenient for users to trace back the analysis steps.

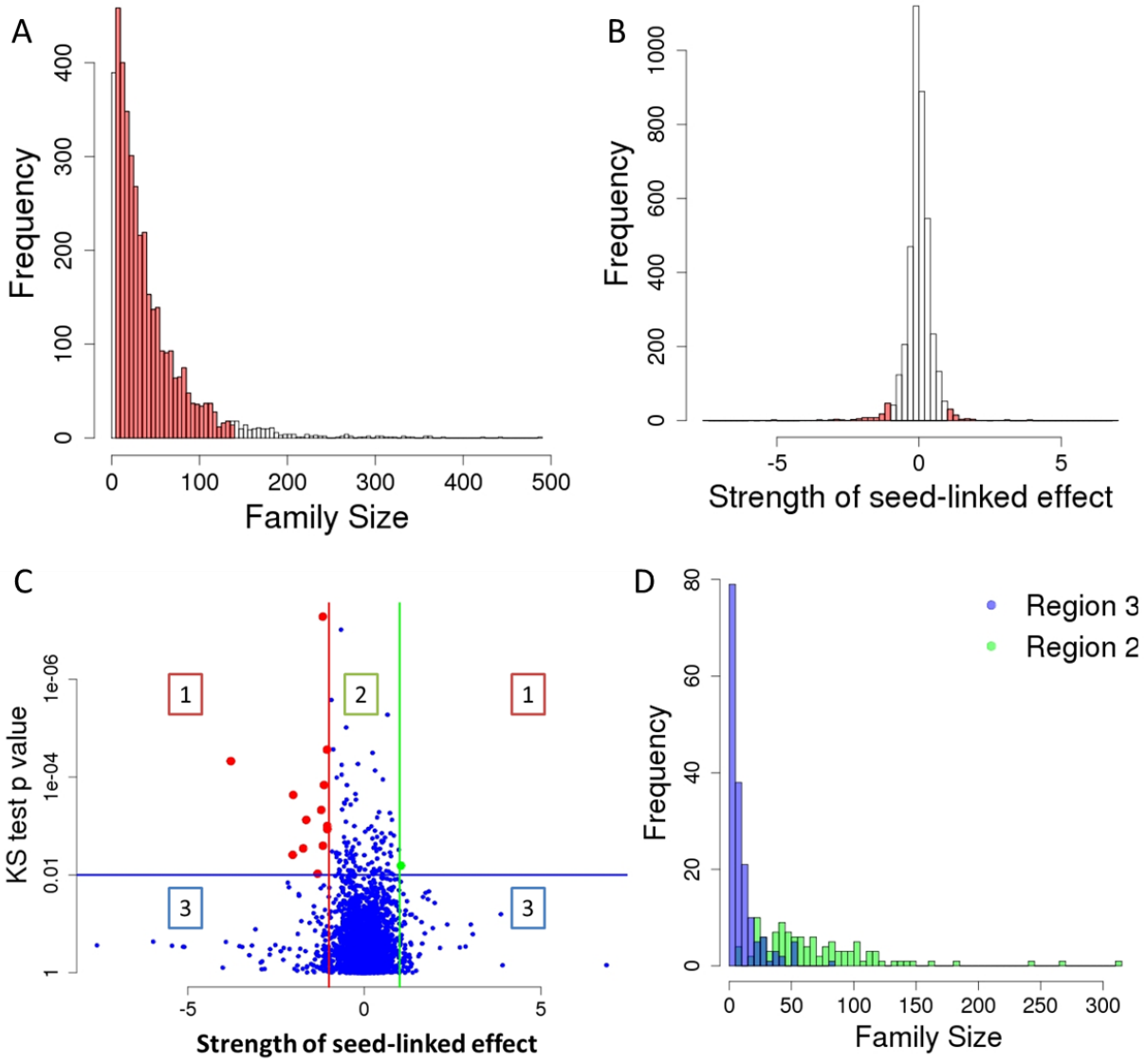


Figure 37. Optimization of DecoRNAi. A) of ~4000 seed families, seed family sizes ranging from 6 to 139 (red bars, 86% of total) were detectable within typical siRNA screens employing the Dharmacon library. B) for these studies, the threshold selection for significant LASSO coefficients included 3% of each tail. This threshold is investigator tunable. C) classification of seed families based on LASSO coefficients and KS p values. Region 1: considered to be both biologically and statistically significant; Region 2: considered to be statistically significant but biologically insignificant; Region 3: considered to be biologically significant but statistically insignificant. D) from the family size distribution, most seed families from region 3 have a family size of less than 6, and those from region 2 are enriched for the largest family sizes.

In summary, DecoRNAi is a computational tool that fills an important unmet need for the functional genomics research community as it enhances the return of rigorous biologically meaningful observations downstream of screening efforts that are otherwise consuming enormous quantities of time and reagents by following bad leads or by weeding them out using strictly empirical approaches. Substantial reduction of off-target rates was experimentally validated in 5 distinct biological screens across different genome-wide siRNA libraries. A public-access graphical user interface has been constructed to facilitate application of this algorithm within the functional genomics community.

3.5 Bibliography

Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions, *Cell*, **136**, 215-233.

Birmingham, A., *et al.* (2006) 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets, *Nat Methods*, **3**, 199-204.

Buehler, E., *et al.* (2012) siRNA off-target effects in genome-wide screens identify signaling pathway members, *Sci Rep*, **2**, 428.

Jackson, A., *et al.* (2006) Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity, *RNA*, **12**, 1179 - 1187.

Jackson, A.L. and Linsley, P.S. (2010) Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application, *Nature reviews. Drug discovery*, **9**, 57-67.

Kim, H.S., *et al.* (2013) Systematic identification of molecular subtype-selective vulnerabilities in non-small-cell lung cancer, *Cell*, **155**, 552-566.

Marine, S., *et al.* (2012) Common seed analysis to identify off-target effects in siRNA screens, *J Biomol Screen*, **17**, 370-378.

Mohr, S., Bakal, C. and Perrimon, N. (2010) Genomic screening with RNAi: results and challenges, *Annual review of biochemistry*, **79**, 37-64.

Orvedahl, A., *et al.* (2011) Image-based genome-wide siRNA screen identifies selective autophagy factors, *Nature*, **480**, 113-117.

Shao, D.D., *et al.* (2013) ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens, *Genome research*, **23**, 665-678.

Sigoillot, F.D. and King, R.W. (2011) Vigilance and validation: Keys to success in RNAi screening, *ACS chemical biology*, **6**, 47-60.

Sigoillot, F.D., *et al.* (2012) A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens, *Nat Methods*, **9**, 363-366.

Singh, N.K., *et al.* (2013) siMacro: a fast and easy data processing tool for cell-based genome-wide siRNA screens, *Genomics Bioinform*.

Tang, W., *et al.* (2008) 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.

Ward, S.E., *et al.* (2012) Host modulators of H1N1 cytopathogenicity, *PLoS One*, **7**, e39284.

Whitehurst, A.W., *et al.* (2007) Synthetic lethal screen identification of chemosensitizer loci in cancer cells, *Nature*, **446**, 815-819.

CHAPTER FOUR

STATISTICAL MODELING AND VISUALIZATION OF IMAGE-BASED HIGH-THROUGHPUT RNA INTERFERENCE SCREENING RESULTS

Image-based high-content screening has enabled us to describe complex multivariate cellular phenotypes on the single-cell level. Recently, the combination of image-based screen and high-throughput RNAi screen has led to genome-wide functional annotation in a wide spectrum of biological research and drug target discovery. However, statistical modeling and visualization tools are still lacking in that no specific tools are available for image-based high-throughput RNAi screening. We developed iScreen (image-Based High-Throughput RNAi Screen Analysis Tool) R package for statistical modeling and visualization of image-based high-throughput RNAi Screen. iScreen is available on CRAN for user download. Experimental data demonstrates the capability and efficiency of iScreen.

4.1 Introduction

RNAi is a loss-of-function technique with wide applications in biomedical research, mediated via either small interfering RNAs (siRNAs) or short hairpin RNAs (shRNAs). Advancing technology enables genome-wide high-throughput RNAi screening for functional genomics or drug target discovery. Recently, in combination with high-content microscopy, image-based high-throughput RNAi screening has had a breakthrough in that it can accurately describe complex multivariate cellular phenotypes and provide a high level of cellular

information, which has been widely used for finding novel pathway components or drug targets (Giuliano, et al., 2004; Orvedahl, et al., 2011).

4.1.1 Image-based screening

Traditionally, high-throughput screening is based on single read-out system, regardless of whether the system is molecule screening or RNAi screening. Commonly used phenotype readouts include cell viability, the chemical activity of metabolic substrates, or the strengths of pathway activity. Such an oversimplified representation of complex biological processes remains an issue for understanding a phenotype of interest, though this process does provide faster computation and relatively greater detection power in many HTS projects, often at a lower cost.

However, for complex cellular phenotypes it has become necessary to simultaneously measure multiple features that might be relevant for biological or therapeutic purposes (Young, et al., 2008). For example, in the study of autophagy activity, 400~500 cells are plated in each well of microplate and for each cell within one well, autophagosomes have to be recognized and their number counted for quantification and analysis (Orvedahl, et al., 2011). For small-molecule compound screening, it is also beneficial to profile cellular phenotypes after treatment since this provides comprehensive information about therapeutic effects (Young, et al., 2008).

Within drug discovery projects in many pharmaceutical and biotechnology companies, it has become more and more necessary to enhance the quality of drug screening procedures and results in the use of high content screening (HCS) based on automated image-based read-out system to measure multiple cellular features. HCS is especially popular on projects focused on multi-feature cellular phenomena such as signaling, cell change transformation and cellular

toxicology, making it a powerful tool for studying conditions like neurological disorders and autoimmune disease (Lang, et al., 2006; Mitchison, 2005; Nichols, 2007).

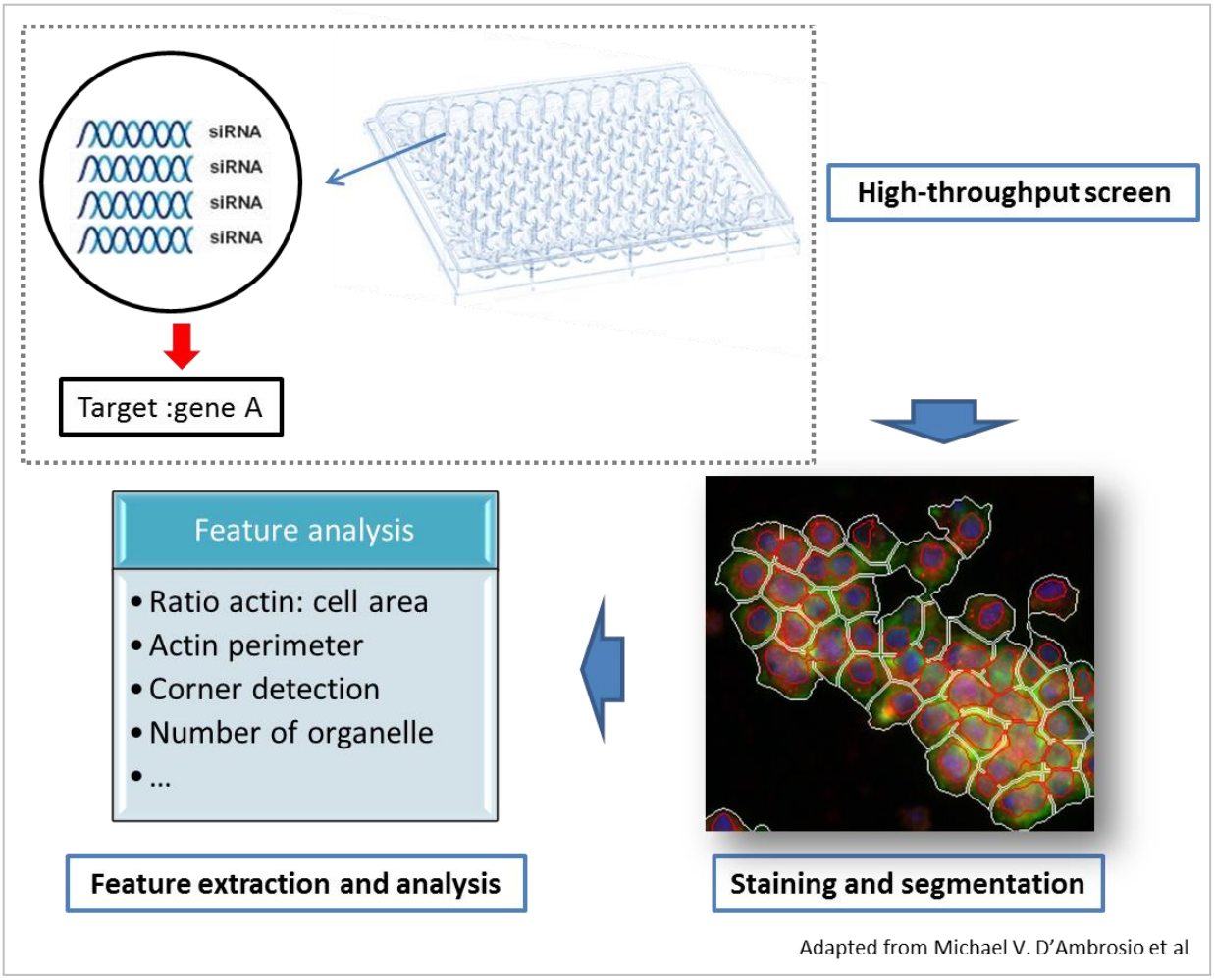


Figure 38. Work flow of cellular image-base screening. In each well of the screening plate, pooled siRNAs are used to knock down genes in a genome-wide pattern. Afterward, automatic staining and cellular segmentation are performed for downstream feature extraction and analysis.

Recently, instrument development and quantification algorithms have made several advancements (Giuliano, et al., 2003; Lee and Howell, 2006; Young, et al., 2008). For example, CellProfiler, which was developed by MIT, has helped addressed a number of image quantifications of biological phenotypes such as cell count, cell size, protein levels, cell shape or

DNA and protein staining pattern (Carpenter, 2006). Therefore the analysis results are featured with high-dimensional output, which requires quantification and special handling.

In high-content high-throughput RNAi screening, the first part of the workflow is very similar to that of ordinary screening. Pooled siRNAs are used to knock down genes in a genome-wide pattern. However, downstream operation and analysis are more complex than usual. Cells have to be stained to reveal components such as DNA, tubulin and actin, before automatic segmentation can be performed. Features such as the actin perimeter, number of autophagosomes, corner detection, and ratio will be measured. Statistical modeling and machine learning have to be trained to operate classification and further analysis (Figure 38).

4.1.2 Commercially available software

Commercially available imaging software has made it possible to perform high-throughput image-based screening. For example, Cellomics (Ghosh, et al., 2004; Giuliano and Taylor, 1998; Kapur, 2002) has developed ArrayScan HCS for the fluorescent-protein biosensors screening used to determine the molecular dynamics of macromolecules, ions and metabolites, thereby enhancing drug discovery (Ghosh, et al., 2004).

GE-Healthcare also developed automated confocal microscopy techniques capable of analyzing one million data points per day (Lang, et al., 2006). This has been used to provide screening of higher density plates with fewer reagents, allowing for rapid analysis of high density formats and ultra-high-throughput screening of a wider range of biological assays (Fowler, et al., 2000; Oakley, 2002).

Commercially available software also includes Evotec (Jager, 2003) and TTP (Grepin, 2003), the former for automated confocal microscopy screening and the latter for laser scanning fluorescence microplate cytometers (Lang, et al., 2006).

4.1.3 Comparison of methods for high-content screening

With the help of high-content image-based screening, we now have a means of broadly characterizing cellular response to compound treatment and RNAi knockdown. Given that multiple features from a single cell are measured out of the hundreds or even thousands of cells within a single well from a microplate, high-content screening presents itself with huge high-dimensional datasets, which makes analysis a bottlenecking step in research and projects. Quantitative scientists and researchers have been developing and applying a variety of new or available methods to modeling and analyzing such big data (Ljosa, et al., 2013).

In some screening projects, either negative control or positive control is available and presents itself as a distribution. Therefore, in order to detect the difference between the sample distribution and control distribution, Kolmogorov-Smirnov (KS) statistics can be used to quantify and test the difference between two cumulative density functions (CDFs) that are based non-parametric statistics (Figure 39), in which statistics D is defined as below:

$$D = \sup_z \left| F_1(z) - F_2(z) \right| \tag{24}$$

where $F_1()$ and $F_2()$ are two CDFs and D is test statistics.

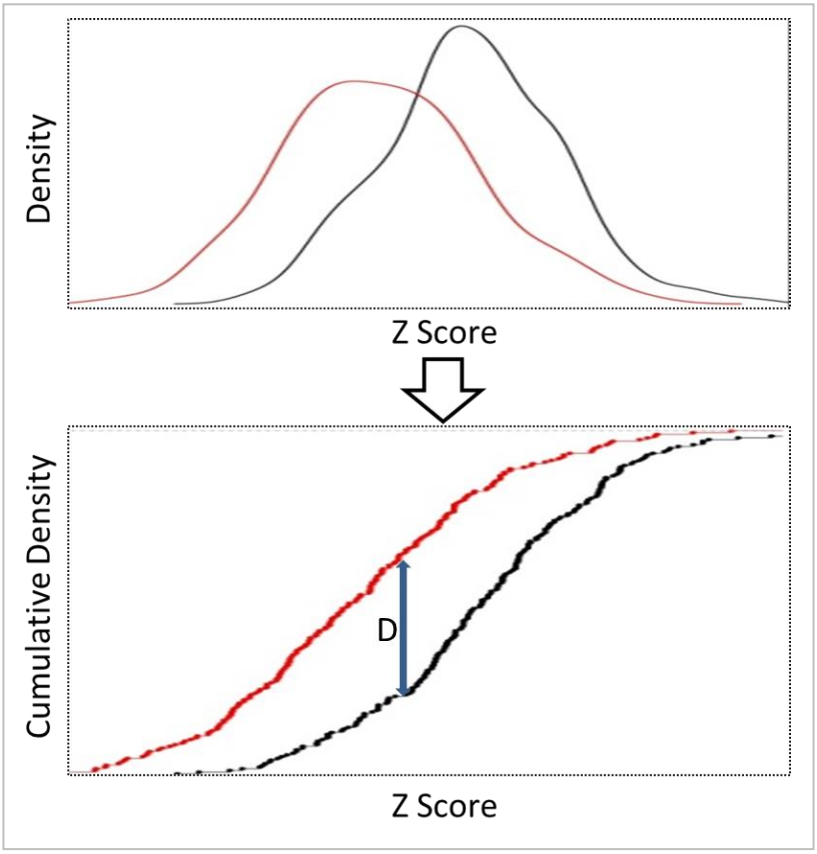


Figure 39. Demonstration of KS statistics. We can transform an ordinary density plot into a cumulative density plot, which provides better visualization and power to detect difference since it is based on non-parametric statistics.

In Perlman's study, KS statistics were used to profile the dose-dependent phenotypic effect of drug treatments in high-content screening and to identify suggested targets for blinded drugs. They provided a systematic and comprehensive pipeline toward profiling at the single-cell level, facilitating the discovery of a new drug mechanism (Perlman, 2004).

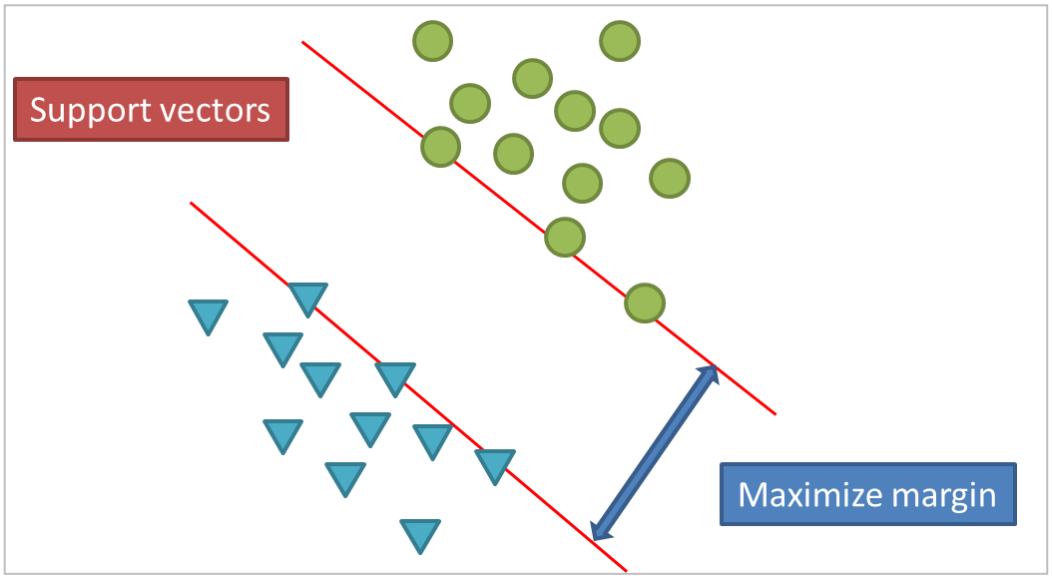


Figure 40. Demonstration of SVMs. Machine learning approach SVMs could be used for classifying cells in each sample from high-content screening projects.

As a supersized machine learning approach, support vector machine (SVM) provides a means of classification and regression analysis (Figure 40). Given a training data set with each object marked with a class identifier, the model can help predict classes for new incoming data sets. For example, in high-content screening SVM can be used to distinguish drug-treated cells and vehicle-treated cells. Dr. Loo's group employed SVM to characterize compound activities in measuring drug effect and identifying cellular responses to dose-dependent drug treatment. (Loo, et al., 2007).

A Gaussian mixture (GM) model is another methodology for classification and regression analysis. The basic idea is to distinguish two Gaussian populations when they are mixed and to select a cut-off for pattern recognition, such as in classification of drug-resistant and -sensitive cell lines. A Bayesian information criterion (BIC) can help identify the number of subgroups, and then the GM model can identify criterion for classification (Figure 41).

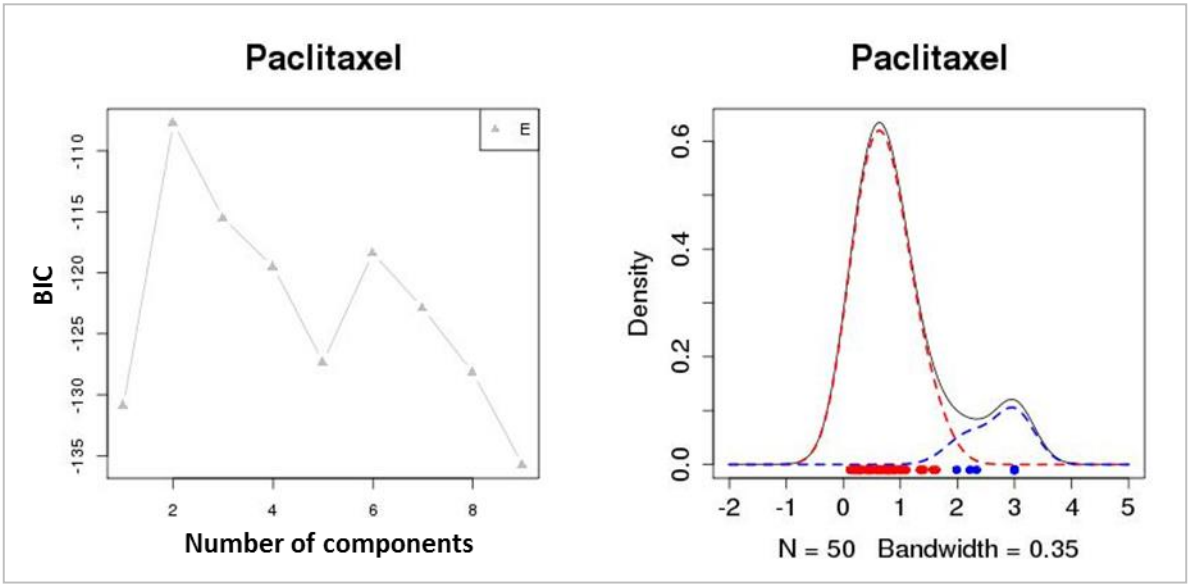


Figure 41. Demonstration of Gaussian Mixture model. 50 non-small-cell lung cancer cell lines are treated with Paclitaxel and drug responses are measured. BIC showed two significantly distinguishable groups, and cell lines are classified into resistant and sensitive subsets.

Factor analysis has a history of almost one hundred years (Spearman, 1904) and has been widely applied in fields such as high-dimensional image data (Carroll and Schweiker, 1951; Floyd and Widaman, 1995; Malinowski, 2002; Stewart, 1981; Tinsley and Tinsley, 1987).

In the factor analysis model, observed sample X's are modeled as a hidden transformation $A\,y+\mu$ of the factors and a case-specific noise term ν dictated as below:

$$x = Ay + \mu + \nu \tag{25}$$

Using the expectation maximization (EM) algorithm, we can estimate A, μ and the corresponding variance matrix and therefore calculate the maximum posterior of y as below:

$$E\left[y \mid x_n\right] = A^T \left(AA^T + \Sigma\right)^{-1} \left(x_n - \mu\right) \tag{26}$$

Those algorithms and approaches have been accepted as a standard analysis pipeline and used widely in a series of high-content screening (Azegrouz, et al., 2013; Ljosa, et al., 2013; Pau, et al., 2013; Zhong, et al., 2013). However, so far no R package or tools have been developed specifically for image-based high-throughput RNAi screening.

4.1.4 Challenge of high-content screening

For current algorithms and tools designed for high-content imaging, the focus has been on feature extraction and classification of heterogeneous cell types via machine learning. Similarly, although image-based high-throughput RNAi screen generates huge amounts of feature data, the study process is gene-centered in that researchers experiment to annotate pathway- or phenotype-related genes while controlling for other confounding factors. Furthermore, a data visualization tool is lacking to control experimental quality and plot analysis results.

Given that fact that image data comes from a variety of distributions, we implemented our package via generalized linear regression to cover both continuous and categorical data. In addition, we also made iScreen quite flexible. Users can provide customized functions to implement analyses.

4.2 Methods and Materials

4.2.1 Experimental procedure

Data was provided by collaborator Xiaonan Dong from Dr. Beth Levine's lab. A HeLa cell line was used to screen for genes required for virus-induced autophagy. An autophagosome was used as an indicator of autophagy activity.

4.2.2 Statistical modeling

For image-based high-throughput RNAi screening, data can assume a variety of distributions from the exponential family, such as normal, Poisson, gamma or binomial. Therefore, we implement via generalized linear models (GLMs) in our package to handle varying data. Generalized linear models are a large class of statistical models for relating responses to predictor variables in a linear pattern, including many commonly encountered types of dependent variables and error structures as special cases. In addition to regression models for continuous dependent variables, models for rates and proportions, binary, ordinal and multinomial variables and counts can be handled as GLMs. In GLMs,

$$E(Y) = g(\eta) \tag{27}$$

where

$$\eta = X\beta \tag{28}$$

Y is the response variable, η is the specified linear predictor and g(•) is the link function, determined by the probability distribution of Y. In order to facilitate custom analysis, we also allow the user to provide self-defined functions

$$Y = f\left(X, \theta\right) \tag{29}$$

to model and analyze data such that customized functions can be integrated into our analysis pipeline and make use of other analysis and visualization tools in our package.

4.3 Results

4.3.1 Visualization

For high-content image-based screening, visualization is often the preliminary step toward data analysis (Figure 42). Shown below is a snapshot from an autophagy study in which researchers were interested in the co-localization of the red-labelled Sindbis virus and green-labelled autophagosome. We have a built-in visualization tool for plotting such data. Users can also specify different types of plotting parameters such as diameter, shape and color (Figure 43, left).

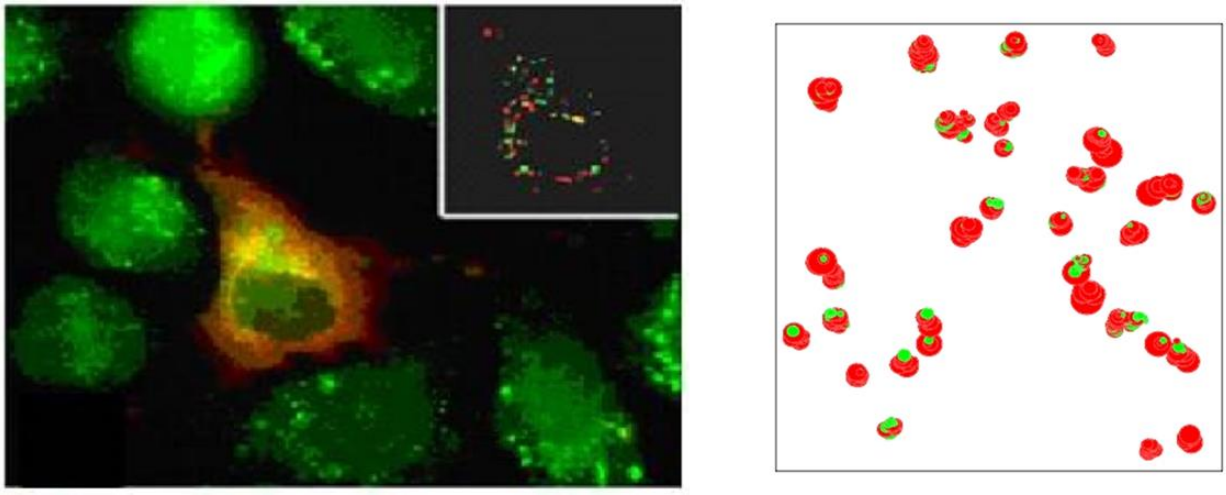


Figure 42. Data visualization from an image-based screening project. Red, Sindbis virus; green, autophagosomes.

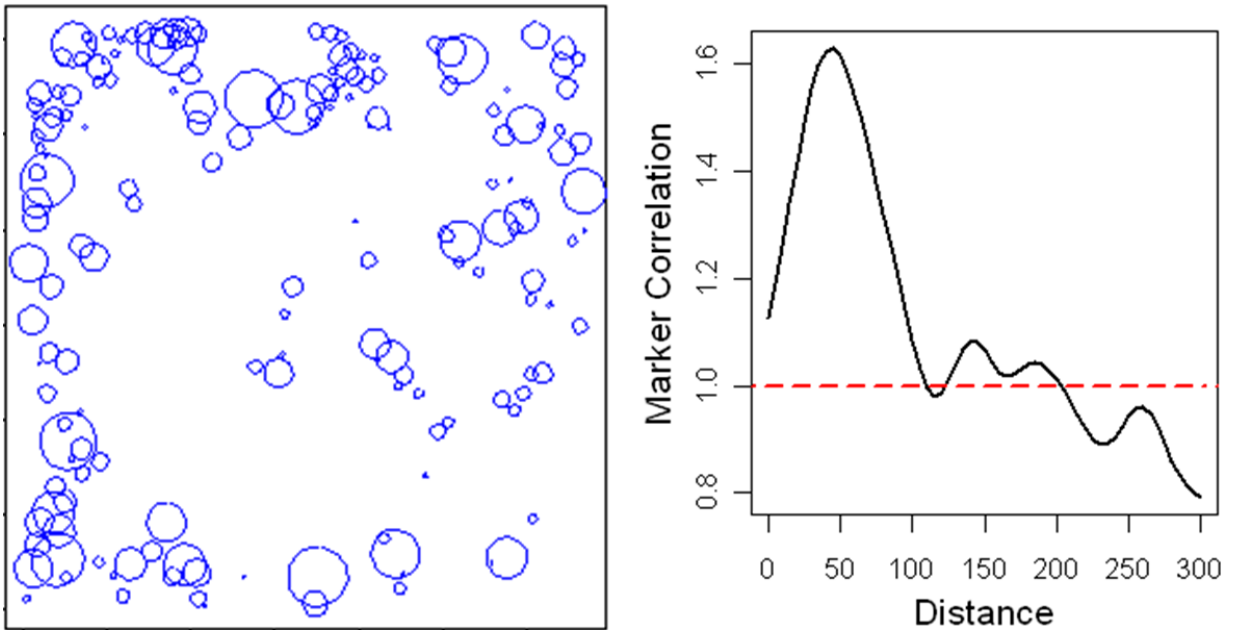


Figure 43. Customized analysis and visualization. Left, a customized data plot with tunable parameters such shape, diameter and background color. Right, a plot of marker correlation with respect to the distance between two markers.

4.3.2 Case Study

Here we applied our package to a second autophagy study. In this case, experiments were carried out on 96-well microplates, and in each well, 200~500 cells were planted. At the end of the experiments, the number of autophagosomes was counted for each cell. Accordingly, we chose a Poisson regression to model and analyze data since the response variable was in the form of count numbers. Therefore, a link function log was used. In addition, we used the negative controls on each plate as a reference in our model. For each well on the plate, we estimated a coefficient that measured the strength and direction of how the distribution of count numbers in each well deviated from the negative controls.

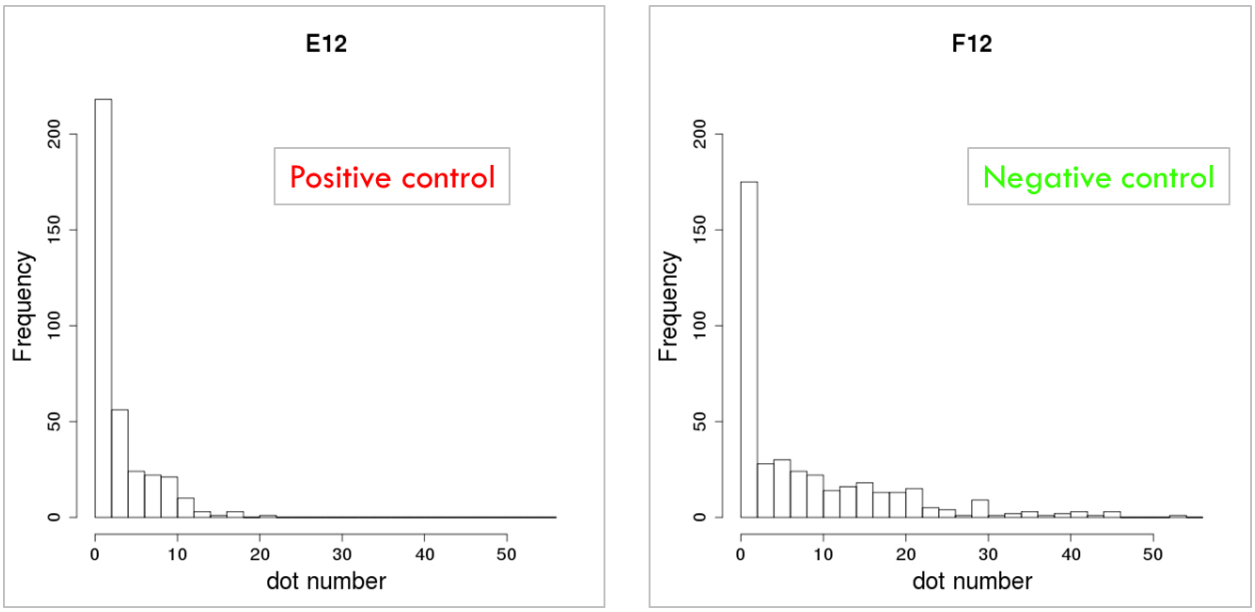


Figure 44. Poisson distribution modeling of experimental data. Left, positive control; right, negative control.

4.3.3 Quality control

In order to visualize analysis results and perform quality control, we developed methods to plot the original data and analysis results and to use them as a means of visualizing row or column effects for quality control (Figure 45). This functionality provides users with a means to check real-time data quality in order to explore any problems that might occur during screening.

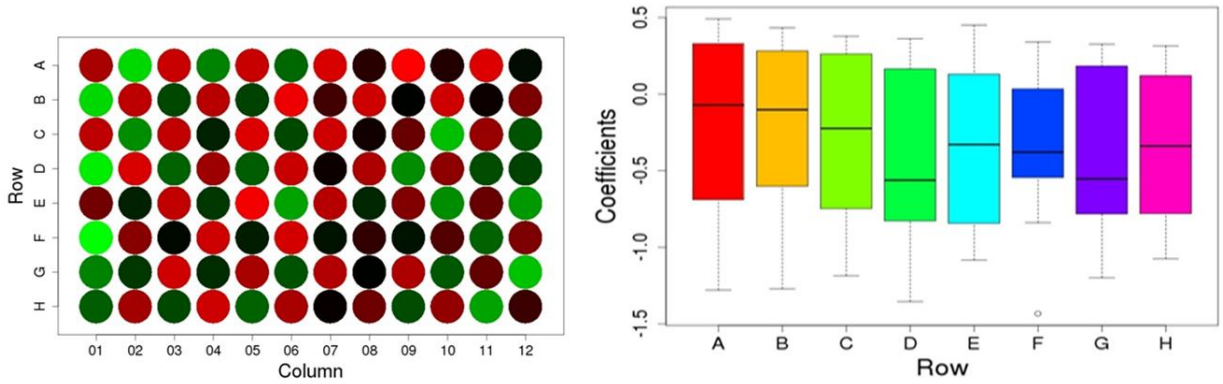


Figure 45. Visualization for quality control. Left, the estimated score visualization in a plate pattern. Right, the row effect visualization for quality control.

4.3.4 ROC curve

In our preliminary screen, experiments were designed such that positive control and negative control were planted alternately in wells across the plate. Therefore we were able to evaluate our approaches using the ROC (receiving operating characteristics) curve. As shown (Figure 46), the AUC (area under the curve) is almost one. We came to the conclusion that iScreen is powerful and efficient enough to identify true positives from such image-based high-throughput RNAi screening.

4.3.5 User-defined function

In our package, we also provided flexibility so that users can incorporate user-defined functions into our analysis pipeline. For example, in the co-localization study we can implement a marker correlation (Figure 43, right) to determine the significance level if dots of two colors are co-localized.

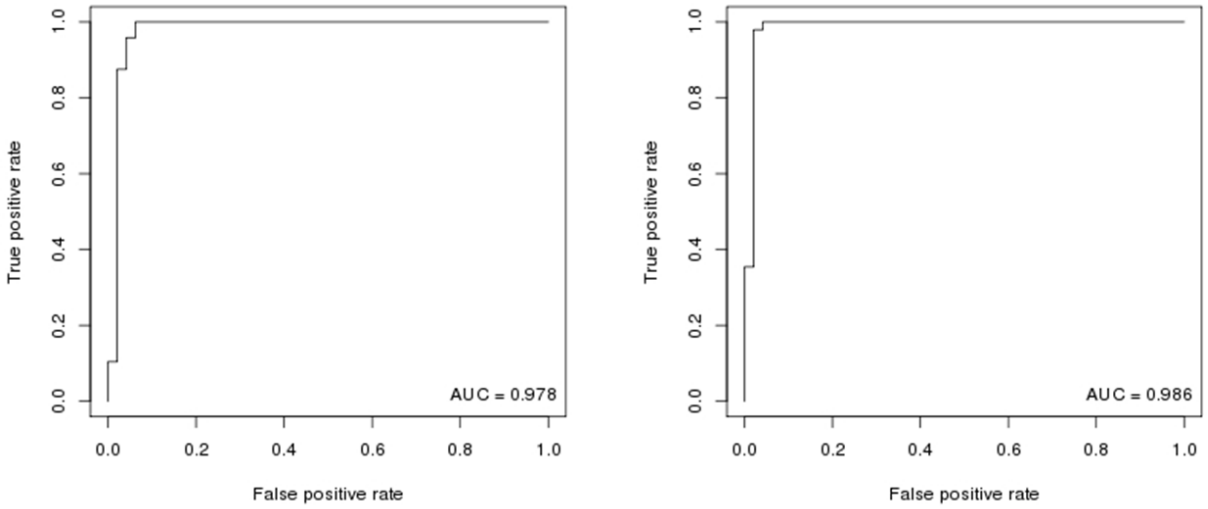


Figure 46. ROC plot. An ROC plot from two exploratory studies in which we knew the true situation for each well. Therefore we could evaluate the performance of our analysis.

4.4 Discussion

Recently, image-based high-throughput RNAi screening has emerged as a novel technique to identify pathway- or phenotype-relevant genes. Given the fact that specific tools were lacking for such analysis, we developed the iScreen R package to facilitate data analysis and visualization. In our package we implemented a generalized linear regression to handle a variety of distribution in the forms of both continuous and discrete data. Experimental validation showed the competency and capability of our package through an autophagy study. Besides, we also allow the user to provide a customized model and make use of our analysis pipeline and visualization tool via integrating a user-defined function into our package.

4.5 Bibliography

Azegrouz, H., *et al.* (2013) Cell-Based Fuzzy Metrics Enhance High-Content Screening (HCS) Assay Robustness, *Journal of biomolecular screening*, **18**, 1270-1283.

Carpenter, A. (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes, *Genome Biol.*, **7**, R100.

Carroll, J.B. and Schweiker, R.F. (1951) Factor analysis in educational research, *Rev. Educ. Res.*, **21**, 368-388.

Floyd, F.J. and Widaman, K.F. (1995) Factor analysis in the development and refinement of clinical assessment instruments, *Psychol. Assess.*, **7**, 286-299.

Fowler, A., Davies, I. and Norey, C. (2000) A multi-modality assay platform for ultra-high throughput screening, *Curr. Pharm. Biotechnol.*, **1**, 265-281.

Ghosh, R.N., Grove, L. and Lapets, O. (2004) A quantitative cell-based high-content screening assay for the epidermal growth factor receptor-specific activation of mitogen-activated protein kinase, *Assay. Drug Dev. Technol.*, **2**, 473-481.

Giuliano, K.A., Chen, Y.T. and Taylor, D.L. (2004) High-content screening with siRNA optimizes a cell biological approach to drug discovery: defining the role of P53 activation in the cellular response to anticancer drugs, *Journal of biomolecular screening*, **9**, 557-568.

Giuliano, K.A., Haskins, J.R. and Taylor, D.L. (2003) Advances in high content screening for drug discovery, *Assay Drug Dev. Technol.*, **1**, 565-577.

Giuliano, K.A. and Taylor, D.L. (1998) Fluorescent-protein biosensors: new tools for drug discovery, *Trends Biotechnol.*, **16**, 135-140.

Grepin, C. (2003) Increasing the quality of compounds isolated during primary screening: high content screening with Acumen Explorer, *Curr. Drug Discov.*, **3**, 37-42.

Jager, S. (2003) A modular, fully integrated ultra-high-throughput screening system based on confocal fluorescence analysis techniques, *J. Biomol. Screen.*, **8**, 648-659.

Kapur, R. (2002) Fluorescence imaging and engineered biosensors: functional and activity-based sensing using high content screening, *Ann. NY Acad. Sci.*, **961**, 196-197.

Lang, P., *et al.* (2006) Cellular imaging in drug discovery, *Nat. Rev. Drug Discov.*, **5**, 343-356.

Lee, S. and Howell, B.J. (2006) High-content screening: emerging hardware and software technologies, *Methods Enzymol.*, **414**, 468-483.

Ljosa, V., *et al.* (2013) Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment, *Journal of biomolecular screening*, **18**, 1321-1329.

Loo, L.H., Wu, L.F. and Altschuler, S.J. (2007) Image-based multivariate profiling of drug responses from single cells, *Nature methods*, **4**, 445-453.

Malinowski, E.R. (2002) Factor Analysis in Chemistry. Nature Publishing Group, pp. 1-432.

Mitchison, T.J. (2005) Small-molecule screening and profiling by using automated microscopy, *ChemBioChem*, **6**, 33-39.

Nichols, A. (2007) High content screening as a screening tool in drug discovery, *Methods Mol. Biol.*, **356**, 379-387.

Oakley, R.H. (2002) The cellular distribution of fluorescently labeled arrestins provides a robust, sensitive, and universal assay for screening G protein-coupled receptors, *Assay. Drug Dev. Technol.*, **1**, 21-30.

Orvedahl, A., *et al.* (2011) Image-based genome-wide siRNA screen identifies selective autophagy factors, *Nature*, **480**, 113-117.

Pau, G., *et al.* (2013) Analysis of high-throughput microscopy-based screens with imageHTS, *Bioconductor*, **R package version 1.8.0**.

Perlman, Z.E. (2004) Multidimensional drug profiling by automated microscopy, *Science*, **306**, 1194-1198.

Spearman, C. (1904) [ldquo]General intelligence[rdquo], objectively determined and measured, *Am. J. Psychol.*, **15**, 201-293.

Stewart, D.W. (1981) The application and misapplication of factor analysis in marketing research, *J. Mark. Res.*, **18**, 51-62.

Tinsley, H.E.A. and Tinsley, D.J. (1987) Uses of factor analysis in counseling psychology research, *J. Couns. Psychol.*, **34**, 414-424.

Young, D.W., *et al.* (2008) Integrating high-content screening and ligand-target prediction to identify mechanism of action, *Nature chemical biology*, **4**, 59-68.

Zhong, R., *et al.* (2013) SbacHTS: spatial background noise correction for high-throughput RNAi screening, *Bioinformatics*, **29**, 2218-2220.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary

The ultimate goal of my dissertation research was to develop an analysis pipeline for high-throughput RNAi screening. Since its discovery, RNAi has opened up a wide spectrum of biomedical and biological research discoveries, including enhanced functional annotations, identification of drug targets and identification of novel therapeutic approaches. At the University of Texas Southwestern Medical Center we have several high throughput screening cores, and I have closely collaborated with the high-throughput RNAi screening center. I have been developing statistical models of high-throughput RNAi screening data to tackle computational challenges in data analysis and visualization. As part of my dissertation research, I have completed three major parts of the projects.

First, I focused on modeling and correcting spatial correlated background noise from high-throughput RNAi screening. I employed a well-established geostatistical model, Kriging interpolation, to fit high-throughput RNAi screening data from duplicated experiments. We have shown that removal of such spatial background noise helps enhance statistical detection power by reducing variation within data. Experimental validation demonstrated that we can identify false negatives from high-throughput RNAi screening.

After data normalization, identifying hits from HTS is the ultimate goal that will allow us to pick up as many true positives as possible. However, the interpretation of high-throughput RNAi screening result has been hampered by off-target effects, which could be due to reagent concentrations, immune response to transfection, or sequence-dependent off-targeting. One of

the sequence-dependent off-targeting effects is from the observation that siRNA might mimic miRNA to raise off-target effects in high-throughput RNAi screening. In order to identify such off-target effects, I developed a deconvolution analysis approach to model data from HTS projects where pooled siRNAs are used to knock down genes. To develop this new methodology, I tested our novel algorithm on multiple datasets from different biological contexts across different siRNA libraries. All identified off-target candidates were experimentally validated. This approach is thus far the most comprehensive analysis of siRNA-mimic-miRNA off-target effects. We even implemented our algorithm as a web-based user-friendly Galaxy tool available online within UT Southwestern and outside to the general scientific community.

Due to the advanced development of imaging instruments and technology, high-content screening has been integrated into high-throughput RNAi screening, which enables multiple features from a single cell and comprehensive descriptive quantification of complex biological process due to loss-of-function interference. Consequently, new methodologies and analysis pipelines are needed. Therefore, we developed a new R package, "iScreen", that is now available to the general scientific community for data visualization and analysis. "iScreen" is curated on CRAN for downloading.

In summary, I created an analysis pipeline for high-throughput RNAi screening and made online tools for implementing our algorithm and approaches. Experimental validation confirmed the validity of our methodologies.

5.2 Future work

Advances in genome-wide high-throughput RNAi screening have sped up the discovery of novel drug targets; however, prioritization of phenotype-associating genes remains challenging, making it the rate-limiting step in analysis of such studies and even blurring the interpretations and experimental validations.

In future work, we will develop an integrated analysis of RNAi screening data with complementary genomic data, such as genome-wide functional gene networks, to prioritize RNAi screen hits and provide a system-level understanding of how gene perturbations affect phenotypes of interest from a network point of view. Specifically, we will integrate RNAi screening data with tissue/phenotype-specific functional gene network data to 1) prioritize candidate phenotype-related genes for experimental validation and 2) provide a comprehensive network view of gene perturbations for phenotypic outcomes.

Our results will show if integrating RNAi screening data with tissue/phenotype specific functional networks is more robust and accurate for finding phenotype-related genes and sub-networks enriched with cellular processes, which are important for phenotypes, compared to those that use only RNAi screen data or a protein-protein interaction network. Thus they can bring a novel phenotype-specific perspective for further investigation.