Part I

RNAi HTS and Data Analysis

1

Introduction to Genome-Scale RNAi Research

1.1 RNAi: An Effective Tool for Elucidating Gene Functions and a New Class of Drugs

RNAi is a mechanism in living cells that helps determine which genes are active and how active they are. It is a naturally occurring pathway for the regulation of gene expression in which small RNA molecules lead to the destruction of *messenger RNA* (mRNA) with complementary nucleotide sequences [48;128]. RNAi has an important role in defending cells against parasitic genes – viruses and transposons – but also in directing development and gene expression in general.

Two types of small RNA molecules are central in the RNAi pathway (Figure 1.1). One is *small interfering RNA* (siRNA), sometimes known as short-interfering RNA or silencing RNA, a class of 20 to 25 nucleotide-long *double-stranded RNA* (dsRNA) molecules [48], and the other is *microRNA* (miRNA), a class of endogenous dsRNA molecules of about 21 to 23 nucleotides in length [89;91;92;128]. Both siRNA and miRNA can bind to other specific RNAs and either increase or decrease their activity, usually by preventing an mRNA from producing a protein.

siRNA. The RNAi pathway is controlled by endoribonuclease-containing complexes known as *RNA-induced silencing complexes* (RISCs) and initiated by an enzyme called *Dicer* in the cell's cytoplasm (Figure 1.1). In the initiation step, the Dicer cleaves long dsRNA molecules into siRNAs. An siRNA assembles into a RISC and unwinds into two single strands. One of the two strands, known as the *guide strand*, is then incorporated into the RISC. Later, the guide strand specifically pairs with a complementary mRNA molecule. This recognition event may produce one of the following two major outcomes: (i) post-transcriptional gene silencing [63;74] (i.e., the gene is not expressed) or (ii) epigenetic changes to a gene affecting the degree to which the gene is transcribed. Post-transcriptional gene silencing occurs when the pairing of the guide strand with mRNA induces cleavage by argonaute, the catalytic component of the RISC complex, in the region homologous to the siRNA.

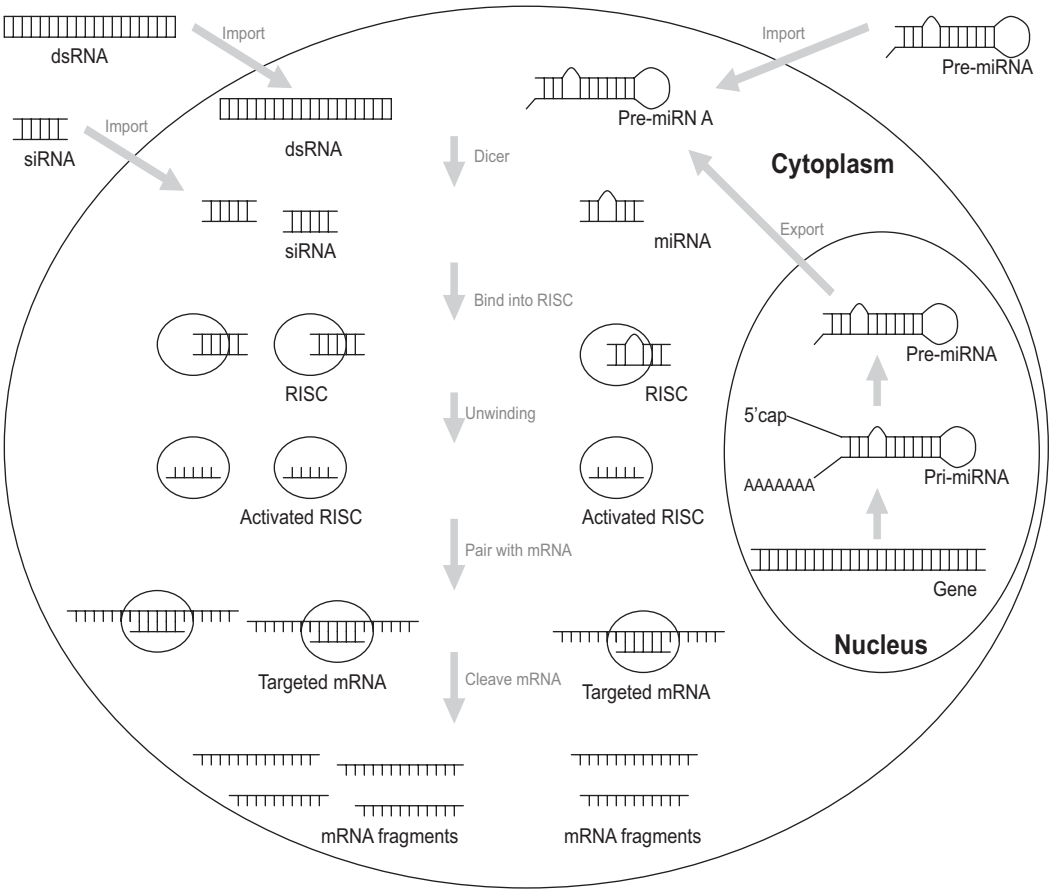

Figure 1.1 The RNA interference pathway. External double-stranded RNAs (dsRNA), siRNA, or pre-miRNA in the RNAi pathway may come from laboratory manipulation, which is the basis for the use of RNAi as an effective tool to knock down targeted genes.

miRNA. The miRNA molecule has a similar function to that of siRNA in the RNAi pathway (Figure 1.1). miRNA is a noncoding RNA. That is, miRNAs are encoded by genes from DNA from which they are transcribed, but miRNAs are not translated into protein; instead, each primary transcript (called a pri-miRNA) is processed into a short, 70-nucleotide stem-loop structure called a pre-miRNA in a cell's nucleus. The stem-loop structure is also called hairpin structure and, subsequently, a pre-miRNA is also called hairpin miRNA precursor [63] or short hairpin RNA (shRNA). A pre-miRNA is exported into cytoplasm and is then identified and cleaved by Dicers into a functional miRNA. Mature miRNAs are structurally similar to siRNAs. However, mature miRNA molecules are partially complementary to one or more mRNA molecules. The main function of miRNA is thought to be down-regulation of the translation of mRNA to protein.

In the RNAi pathway, the dsRNA may come from infection by a virus with an RNA genome or from laboratory manipulations. Therefore, this pathway can be co-opted by experimentally introducing synthetic dsRNAs designed to target specific mRNAs, thus knocking down the expression of the protein of interest [44;63;64]. The development of algorithms for siRNA design that produce potent and selective knockdown of targeted genes has led to a great deal of interest in using siRNA to elucidate gene function and identify novel targets for drug discovery. In medical research, in addition to drug target identification and validation, RNAi can also be harnessed to develop a whole new class of potential therapeutic agents [98]. In fact, RNAi is seen as the third class of drugs, after small molecules and proteins [24]. The importance of RNAi was further recognized when the Nobel Prize in Medicine and Physiology was awarded to A. Fire and C.C. Mello in 2006 for their work [48] on RNA interference in *Caenorhabditis elegans*, which they published in 1998. Galun [51] even parallels the article by Fire et al. [48] on RNAi to that of Watson and Crick [156] on the double helix of DNA.

1.2 High-Throughput Screening: A Vital Technology in Drug Discovery

High-throughput technologies such as microarrays, whole-genome single-nucleotide polymorphism chips, and *high-throughput screening* (HTS) play a central role in current molecular biological research and drug discovery. The ability of high-throughput technologies to simultaneously interrogate thousands of genes/compounds has led to important advances in solving a wide range of biological problems, including the identification of previously unknown genes involved in a biological pathway and the subsequent unveiling of new insights into developmental processes and pharmacogenomic responses, the evolution of gene regulation, and the discovery of new drug targets [18;54;100]. Likewise, RNAi can be utilized on a genome-wide scale via HTS technology, which allows thousands of siRNAs to be tested simultaneously to identify previously unknown genes involved in a biological pathway [8;16;64;96;138;158;176;180].

HTS technology uses automation (including robotics, data processing, and control software, liquid handling devices, and sensitive detectors) to run an assay screen against a library of candidate compounds or siRNAs. An assay is a test for specific biochemical activity such as the inhibition or stimulation of a biochemical or biological mechanism. The biochemical activity can be represented by measured responses such as the reflectivity of polarized light shined on cells or the intensity of emission from labeled particles. A typical compound HTS-screening library contains more than 100,000 small molecules. A genome-scale siRNA library may contain 60,000 or more siRNAs that are pooled to target about 25,000 genes. Usually, three siRNAs targeting the same gene are pooled together. Hence, using HTS, one can rapidly identify active compounds, antibodies, genes, or effective siRNAs that modulate a

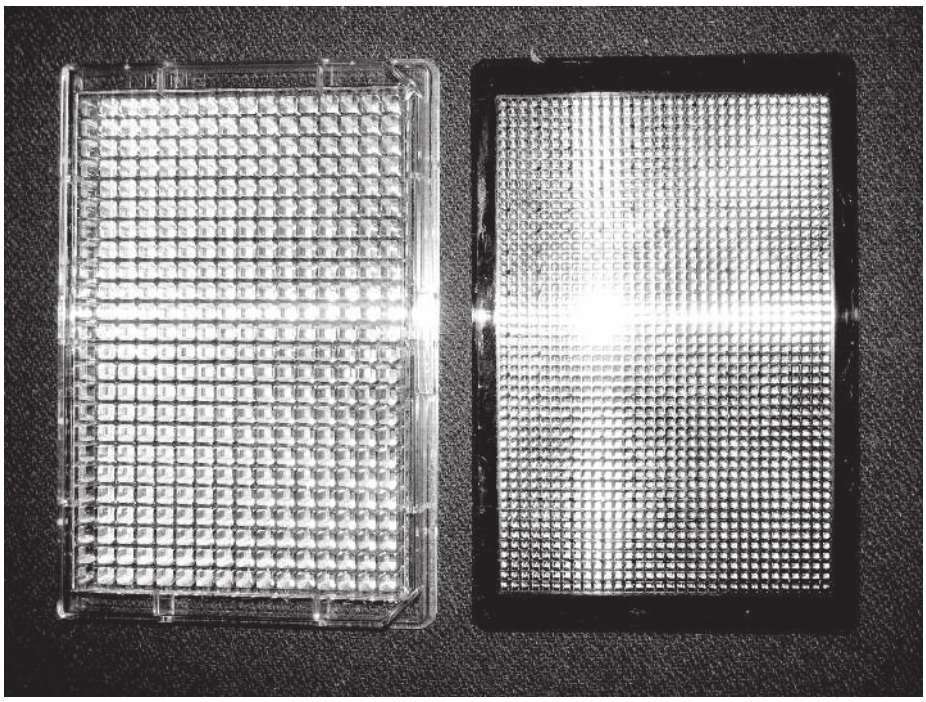

Figure 1.2 Two types of microtiter plates commonly used in high-throughput screens. Left: a 384-well plate. Right: a 1,536-well plate.

particular biomolecular pathway and thus can discover the interaction or role of a particular biochemical process in biology.

The key testing vessel of HTS is a small container with a grid of small, open divots, called wells. This container is called a microtiter plate or microplate of about 5 inches long and $3\frac{1}{8}$ inches wide and is usually disposable and made of plastic (Figure 1.2). The microplates for HTS generally have 384, 1,536, or 3,456 wells, although they may have 96 wells in some experiments. Most of the wells contain test compounds (i.e., small molecules) or siRNAs, with one compound or siRNA per well, although some of the wells contain controls to indicate the quality of the assays. The test compounds or siRNAs are also called sample compounds or sample siRNAs. A screening facility usually has a library of source microtiter plates holding the compounds or siRNAs that have been carefully chosen, arranged, and cataloged. The source plates are also called stock plates, and they are either created by the lab or obtained from a commercial source. The source plates themselves are not directly used in experiments. During experiments, copies of selected source plates are created by pipetting a small amount of liquid (often measured in nanoliters) from the wells of a source plate to the corresponding wells of a completely empty plate. The copied plates that are actually used in the experiments are called *assay plates*.

To conduct an HTS experiment, the researchers fill each well of an assay plate with some biological matter, such as protein, cells, or an animal embryo, and incubate

them for a certain time so that the biological matter can absorb, bind to, or have other reactions with the compounds/siRNAs in the wells. Then the response representing the biochemical reaction (e.g., the intensity of light emitted by labeled particles) is measured across the wells, usually by an automated machine. The machine outputs the result as a grid of numeric values, with each number mapping to the value obtained from a single well of an assay plate in an experiment. A high-capacity machine can measure dozens of assay plates in a few minutes, generating very quickly thousands of experimental data points. One of the most important features in HTS technology is automation, which relies on robotics and high-speed computers. Typically, an integrated HTS system consisting of one or more robots has the ability to transport assay plates from station to station for sample and reagent addition, mixing, incubation, and readout. Therefore, an HTS system can usually simultaneously prepare, incubate, and output many plates, subsequently testing a large number of compounds/siRNAs (e.g., up to 100,000 compounds per day) and generating a huge amount of data. The term *ultra high-throughput screening* (uHTS) has been created to refer to an HTS facility that can screen in excess of 100,000 compounds a day on a routine basis.

1.3 Genome-Scale RNAi Screens

Genome-scale RNAi screens can be conducted in different organisms. Three that have been intensively studied are *C. elegans* (a small roundworm), *Drosophila* (fruit flies), and human cells.

Human genome-scale RNAi screens are currently conducted in human cells, including stem cells, and a variety of immortal cell lines [41]. Long dsRNAs activate interferon responses, which leads to apoptosis (cell death) in somatic cells. Short dsRNAs do not activate the interferon response; thus RNAi in human cells must use short dsRNA of less than 30 nucleotides such as synthetic siRNA, vector-expressed *short-hairpin RNA* (shRNA, and *endoribonuclease-derived siRNA* (esiRNA). Even shorter dsRNAs may be needed: a recent study [85] found that dsRNAs of just 21 nucleotides long triggered an immune response through toll-like receptor three (TLR-3), which ultimately inhibited angiogenesis, and that simply shortening the siRNA to fewer than 18 nucleotides seemed to eliminate TLR-3 recognition.

RNAi screens in human cells can use a diverse set of phenotypic measurements such as homogenous cell viability, alternations in reporter-gene expression, high-content readouts using automated microscopy, and immunofluorescence signal from a highly specific antibody. In situations where the focus is on a single measured response, the design and analysis should be similar regardless of differing types of phenotypic measurements. In other cases, we may consider multiple measurements simultaneously in an experiment, especially for high-content readout. More details about different analyses will be provided in Chapter 5.

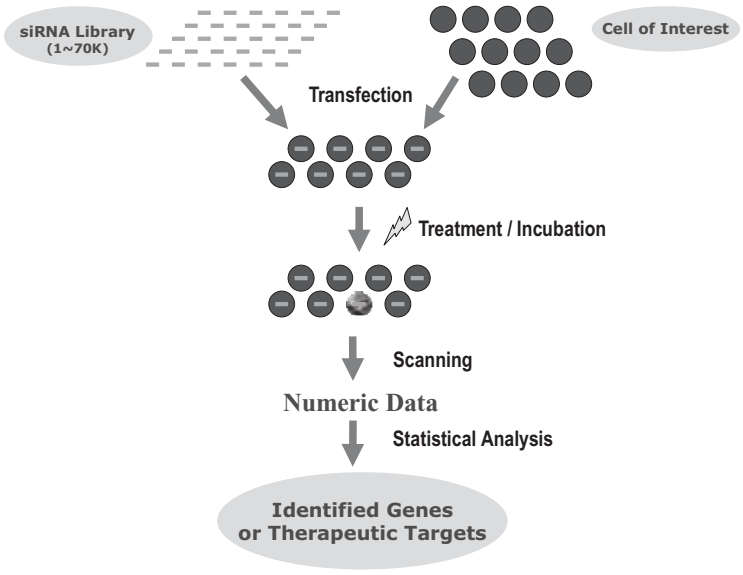

Figure 1.3 Procedure of genome-scale RNAi screens.

The HTS technology applied to RNAi has made it feasible to use cell-based assays to query every gene in the genome for its potential function in a given biological process of a cell. The general procedure for cell-based RNAi screens is demonstrated in Figure 1.3. The first step is to choose an RNAi library and a robust and stable type of cell, which can be a stem cell, primary cell, established cell line, or engineered cell line. Primary cells have limited passaging capacity (thus siRNAs are hard to get into the cells), whereas established cell lines such as Hela and HEK are capable of indefinite passaging under proper cell culture conditions. Engineered cell lines are modified to over-express or underexpress native, tagged, or engineered proteins.

Transfection. The delivery of siRNAs in a library into cells is called *transfection* of siRNAs into the cells in an RNAi screen. For transfection, we need to identify relevant siRNA controls and HTS-compatible transfection methods that give optimal gene knockdown and cell viability. The transfection can be conducted using either suspension-mode electroporation or lipid reagents. Suspension-mode electroporation is good for cells that are difficult to transfect but currently is limited to the 96-well format. Lipid reagents are easily scalable and mostly highly efficient with varying degrees of cytotoxicity and stability. In addition to choosing the best transfection method, we need to determine cell density (i.e., number of cells per well), time of lipid/microRNA complex formation, assay incubation time post-transfection, transfection efficiency (i.e., the performance of assay with respect to positive and negative control siRNAs), detection reagent stability at the working concentration, effect of cell passage number, and so on.

There are various types of RNAi-detection assay technologies. Two commonly used types are well-based assays using photomultiplier tube– or charge-coupled

device–based cameras and cell imaging based on fluorescent labeling of macromolecule of interest. To minimize the need to continually change platform instrumentation for new assays, RNAi assays can be broken into portions, with one robotic platform handling transfection, a second handling detection, and, possibly, a third carrying out a high-content read, which allows each robot to specialize in a particular aspect of the RNAi HTS process.

1.4 An Example of Genome-Scale RNAi Research

In genome-scale RNAi screens, a primary goal is to select siRNAs with a desired effect size. The siRNA effect is represented by the magnitude of difference between the intensity of an siRNA and that of a negative reference in RNAi HTS experiments [183]. For screens using the common platform of 384-well plates, limitations of experimental time and cost usually do not allow a single experiment to have more than two hundred 384-well plates, whereas two hundred 384-well plates is usually the minimal requirement for conducting a genome-wide screen with replicates (i.e., each siRNA is measured multiple times). Therefore, currently, a typical RNAi HTS project starts with a first screen (called *primary screen*) of single or pooled siRNAs targeting about 20,000 genes, most of which have no replicate. The single or pooled siRNAs identified (called hits) in the primary screen are further investigated using one or more secondary screens (called *confirmatory screens*) in which each siRNA or pool has replicates. A typical primary screen has fifty to one hundred fifty 384-well plates, and a typical confirmatory screen has three to twenty 384-well plates.

For example, a genome-scale RNAi project for hepatitis C virus (HCV) started with a primary screen, in which a total of about 22,000 siRNA pools were tested across 97 plates [180]. The experiment was designed to identify host factors associated with HCV replication using the HCV replicon assay system described in Zuck et al. [185]. The negative control used in the experiment was a nonsilencing siRNA. Two positive control siRNAs were used: (i) a very strong one that targeted the HCV replicon [121] and (ii) a weaker one that targeted hVAP33 [61;97].

Following the primary RNAi HTS experiment and using the methods described in Zhang et al. [161;162;174;180], a total of 640 siRNAs were identified as hits. These siRNAs were transfected into 384-well plates with controls set up as in the primary experiment. HuH-7 cells containing an HCV genotype 1b replicon were transfected with these 640 siRNAs as described for the primary screen, but the transfections were carried out in triplicate of every source plate for improved statistical robustness in a confirmatory screen. In addition, a second confirmation screen was carried out using HuH-7 cells expressing the genotype BK-2b HCV replicon. The replicon-containing cells were transfected in triplicate and assayed in the same manner as the HuH-7 genotype 1b HCV replicon containing cells in the primary screen and in the first confirmatory screen [173].

1.5 Challenges in Genome-Scale RNAi Research

Genome-scale RNAi screens have two major advantages over classical genetic screens: (i) the sequences of all identified genes are immediately known, and (ii) lethal mutations are easier to identify because it is unnecessary to recover mutants [16]. Classical genetic screens for elucidating gene function largely rely on the recovery of lethal mutation, which is usually time-consuming and may be difficult. Now RNAi screens directly measure the knock-down impact of siRNAs on their targeted genes, thus making it unnecessary to recover mutants as in classical genetic screens. Thus genome-scale RNAi screens have great promise for elucidating gene function and for discovering new drug targets in our post-genome era. Meanwhile, as a technology that is still under development, genome-scale RNAi screens face many challenges. Four key challenges are (i) controlling optimal experimental time, (ii) identifying moderate or weak hits, (iii) reducing off-target effects, and (iv) gleaning biologically relevant information from a large body of data.

Experimental times. Compared with small-molecule HTS experiments, RNAi HTS experiments have more challenges in controlling the optimal experimental time so that all the potent siRNAs are measured in their effective peaks. RNAi targets mRNA and depletes it from the cell. However, the levels of mRNA vary across a wide dynamic range; consequently, the depletion of different mRNAs may take different lengths of time. Furthermore, after the depletion of mRNAs in the cell, the residual protein can remain for an extended time, which is protein dependent. As a result, the knockdown of genes by RNAi reagents has a wide temporal range (e.g., within 12–120 hours after transfection) [62]. When a large number of siRNAs are used in a single assay as in a genome-scale RNAi screen, the onset of action of potent siRNAs can occur at different times, and there is no time point in the assay that is optimal for all the potent siRNAs. If a potent siRNA is not measured at or nearly at its effective peak, it may act like an impotent siRNA during measurement and may not be identified as a hit in the assay, which then leads to a higher false-negative rate in RNAi screens than in small-molecule screens.

Identifying hits. A unique feature of RNAi is that its effect on genes is to knock down, not knock out completely. Compared with the effect of knockout in classical genetic screens, the size of effect on a measured phenotype is moderate or weak for many potent siRNAs. Furthermore, in some cases, the siRNAs with moderate effects are more biologically relevant [173]. Although many of the siRNAs might affect the outcome of the assay, in most cases, a small percentage of the effective siRNAs truly target genes involved in the phenotype under investigation. That is, the true-positive hits are buried in a large body of data containing substantial noise. siRNAs with moderate or weak effects are more likely to be buried among siRNAs with extremely weak or no effects. In other words, a nonhit tends to behave more like a hit in RNAi HTS experiments. In contrast, the effects of potent small molecules

are usually strong. Therefore, RNAi HTS experiments tend to have a higher false-positive rate than both classical genetic screening experiments and small-molecule HTS experiments.

Off-target effects. When using RNAi as a gene-silencing tool, we want an RNAi reagent to specifically knock down a target gene but not to interfere with other genes. However, an siRNA can silence not only the target gene, but also other genes with similar sequences. The silencing effect of an siRNA on nontarget genes is called an *off-target effect*. Off-target effects can produce false positives, leading to misleading results and erroneous conclusions about the genes that are involved in a biological pathway, when RNAi experiments are used to elucidate gene functions [78–80].

Detecting biologically relevant data. One of the major advantages of HTS technologies is their ability to simultaneously interrogate thousands of genes/compounds. With the ability to generate large amounts of data per experiment, HTS technologies have led to an explosion in the rate of data compiled in recent years [9;70]. Consequently, one of the most fundamental challenges of HTS biotechnologies is to glean biological significance from large volumes of data [16;25;38;43;70;82;136;161;176;180]. There are many analytic questions to be solved. For example, systematic errors (e.g., row and/or column effects) and outliers are not uncommon in HTS experiments. How should we address them? For quality control (QC) in genome-scale RNAi experiments, signal-to-background ratio, signal-to-noise ratio, signal window, assay variability ratio, and Z-factor have been adopted to evaluate data quality [17;39;77;99;116;123;148;150;159;180;185]. How well do these QC metrics work? For hit selection, z-score and t-statistic are commonly used. Do these methods work well? If not, can we develop better analytic methods? Should we perform analyses on a plate-by-plate basis (called plate-wise) or on all the plates in an experiment (called experiment-wise)? How should we address multiplicity issues of hit selection in genome-scale RNAi screens?

To face all these challenges, we must adopt appropriate experimental designs and suitable analytic methods so that we can obtain optimal results in genome-scale RNAi research. For example, the z-score and t-statistic are both based on testing the null hypothesis of exactly no effects on average. However, owing to the network of gene interactions, many genes may have some degree of impact on a measured biochemical response [167;178]. Better analytic metrics are required for assessing the size of siRNA effects rather than testing the null hypothesis of no effect on average. I have developed a statistical parameter, *strictly standardized mean difference* (SSMD), for effectively measuring the size of siRNA effects [161;162;165]. On the basis of SSMD, I have also proposed an error-control method for maintaining a balanced control of both false-positive and false-negative rates [161;174;175;178]. In addition, we need to adopt effective sequence designs, such as the design of multiple individual siRNAs per gene and siRNA pooling to address off-target effects [16]. We also need better plate designs to account for positional effects in RNAi screens [166;173]. In the