Minimizing the risk of reporting false positives in large-scale RNAi screens

Christophe J Echeverri[1], Philip A Beachy[2], Buzz Baum[3], Michael Boutros[4], Frank Buchholz[5], Sumit K Chanda[6], Julian Downward[7], Jan Ellenberg[8], Andrew G Fraser[9], Nir Hacohen[10,11], William C Hahn[10,12], Aimee L Jackson[13], Amy Kiger[14], Peter S Linsley[13], Lawrence Lum[15], Yong Ma[2], Bernard Mathey-Prévôt[16], David E Root[8], David M Sabatini[8,17], Jussi Taipale[18], Norbert Perrimon[16,19] & René Bernards[20]

Large-scale RNA interference (RNAi)-based analyses, very much as other 'omic' approaches, have inherent rates of false positives and negatives. The variability in the standards of care applied to validate results from these studies, if left unchecked, could eventually begin to undermine the credibility of RNAi as a powerful functional approach. This Commentary is an invitation to an open discussion started among various users of RNAi to set forth accepted standards that would insure the quality and accuracy of information in the large datasets coming out of genome-scale screens.

In recent years the large-scale application of RNAi to functional genomic screens in *Caenorhabditis elegans*, *Drosophila melanogaster* and, more recently, mammalian cells has led to the publication of a fast-growing number of 'hit lists', that is, genes that elicit a positive response (or score) in various functional assays. As with other 'omic' approaches, these studies have inherent rates of false positives and negatives that can be caused by the quality of reagents, sensitivity of the assay, data sampling issues and so forth. The explosive speed at which RNAi technology has been adopted in many research fields, combined with the rapid evolution in our understanding of the underlying silencing pathways, has made it particularly difficult for users to keep up with what represents 'best scientific practice' in this field. Though general guidelines have been discussed recently to insure sound experimental approaches to maximize RNAi specificity[1,2], many of us involved in genome-scale screens have felt a need for a public discussion on adopting ever-improving quality standards in such screens to minimize the contamination of false positives in future RNAi datasets. We propose, by way of this Commentary, to trigger

an open and hopefully constructive dialog throughout the community centered around some key issues that we, as contributors to this field, have come to appreciate as important caveats in RNAi experiments to date.

Specificity of RNAi phenotypes: experimental controls lead the way

The first issue to complicate the interpretation of RNAi datasets in recent years has been the targeting specificity inherent to the various types of dsRNA reagents used to direct the RNAi pathway's silencing activity[2]. Several studies (reviewed in ref. 1) have documented various so-called off-target effects, particularly in mammalian systems, linking them to virtually every type of RNAi reagent used to date, including synthetic small interfering RNAs (siRNAs) and vector-based short hairpin RNAs (shRNAs). Collectively, off-target effects comprise all detectable phenotypic consequences arising from unintended interactions, whether dependent on nucleotide sequence or not, between the silencing molecules and various cellular components, including proteins (for example, interferon pathway machinery) or nontargeted mRNAs.

A primary source of sequence-dependent off-target effects, at least in mammalian systems, appears to be the relatively high tolerance for mismatches between

the siRNA guide strand, the ultimate targeting molecule, and the complementary target mRNA sequence, outside of the short 'core' targeting region, that is, bases 2–8, known as the guide strand's 'seed region'[3,4]. Indeed, this type of partial complementarity has been shown to underlie the mechanism of target silencing exhibited by microRNAs (miRNAs; for example, see ref. 5). Attempts to predict such sequence-dependent off-target effects through *in silico* sequence analyses have so far failed. Alternative RNA backbone chemistries, structural variants and advanced design algorithms have resulted in a reduction of the occurrence of these effects, yet none of these approaches can reproducibly eliminate them altogether. It is therefore currently impossible to rule out off-target risks for RNAi experiments through reagent design alone.

As a result, the standard of care to demonstrate the specificity of an observed RNAi-induced phenotype (namely the suppression of the targeted gene alone, rather than a reagent-specific off-target effect) warrants closer scrutiny. At greater risk are conclusions relying on RNAi experiments in the absence of any other supporting data from non-RNAi approaches. Thankfully, the application of well-designed controls can, in virtually all cases, insure minimal risks of mis-
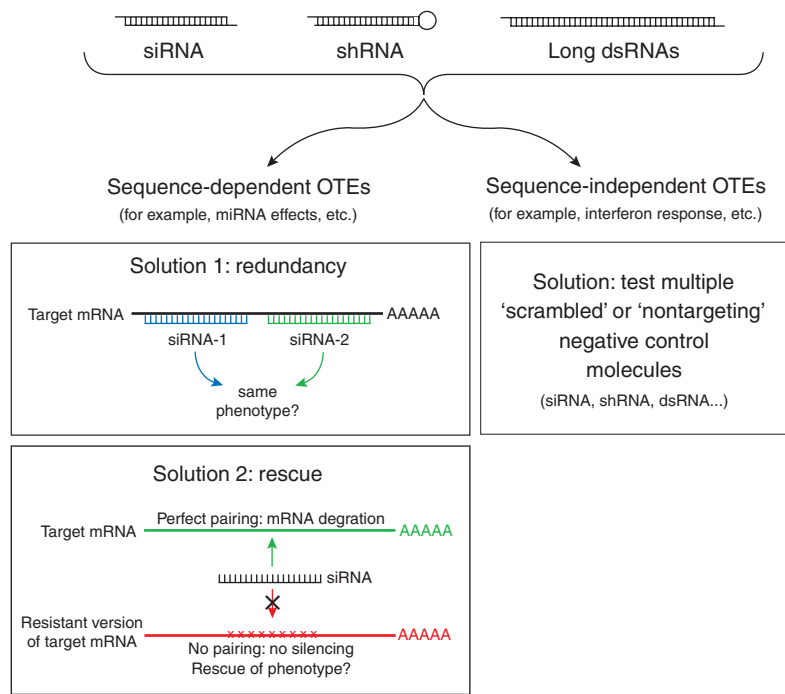
interpretation[2]. In as much as it is practical, internal validation, as with any other type of experiment, is the preferred standard of care for all RNAi datasets, from single-gene studies to genome-scale screens. Whereas these steps can easily be implemented in small-scale studies, genome-scale RNAi studies present challenges of their own, as they generate large datasets in need of validation.

Here we propose that the only ways of adequately addressing sequence-dependent off-target effects within RNAi experiments themselves are 'the two Rs' (**Fig. 1**): rescue or redundancy. Rescue experiments, whereby the RNAi-induced phenotype is countered by expression of a functional version of the target gene that is resistant to the silencing reagent, offer the most convincing control. However, they remain technically challenging and cannot be carried out on a large scale. Reagent redundancy experiments, in which multiple distinct silencing reagents targeting the same gene cause the same phenotype, therefore offer a more universally applicable control. As the specific pattern of cross-silenced transcripts for any given siRNA or shRNA derives directly from its own nucleotide sequence, the probability will be very low that several siRNAs or shRNAs with completely distinct sequences will share the same sequence-dependent off-target effects. Therefore, confirmation of an observed phenotype with redundant silencing reagents, offers the most straightforward and compelling way of demonstrating RNAi target specificity in large-scale screens.

Obviously, this remedy raises the question of how much redundancy is enough to confirm gene specificity: are 2 siRNAs or shRNAs sufficient, or do we need 3 or 4 of them to yield the same phenotype? In theory, this will depend on several factors including, first and foremost, the design and type of silencing reagent, the organism, the pathways being probed and the depth of analysis used to document the resulting phenotypes. Thus, without comprehensive, large-scale analyses to shed further light on this complex issue, this remains an open question. Until then, we strongly advocate that published reports of RNAi-induced phenotypes should be controlled through at least 2 distinct siRNA or shRNA sequences, and/or confirmed through rescue experiments.

These precautions are technically feasible in virtually all situations, and represent a prudent and acceptable way of minimizing the risks of false positives for future RNAi datasets. The publication of any unvalidated 'single-hit' RNAi phenotypes should explicitly acknowledge the higher risk that these may represent false positive results. When several reagents are tested before two can be found to yield matching results, the proportion of positives with respect to the total number tested represents another important measure of the result's reliability, and should also be disclosed. In those cases, a discussion of the relative gene knockdown and extent of phenotype observed with the various reagents can serve as further criterion for data validation.

Additionally, sequence-independent off-target effects must also be controlled. So-called 'scrambled' or 'nontargeting negative control' siRNA or shRNAs are the best available tools for this purpose as they are designed to avoid targeting any transcripts expressed in the chosen sample. One important caveat however is that state-of-the-art design for siRNAs still cannot rule out the risk that such negative control siRNAs or shRNAs may themselves elicit unintended sequence-dependent off-target effects. Therefore, several 'candidate' molecules should be tested for each new assay or cell combination to empirically validate that the chosen control reagent accurately reflects the cellular, undis-turbed baseline. Consequently, published RNAi datasets should include assay results observed in cells left untreated or treated with validated negative control siRNAs (in some cases, mock transfected controls will add confidence), insuring that any difference between them does not invalidate the observed gene-specific phenotypes.

Controlling RNAi specificity in nonmammalian models

Although off-target risks associated with long dsRNAs, typically used in *C. elegans* and *D. melanogaster,* have been less-well studied than those of siRNAs and shRNAs, emerging *D. melanogaster* data now confirm that the issue is non-negligible in this model system, and present knowledge offers no reason to expect differently from *C. elegans*. Recent evidence indeed suggests that longer dsRNAs can also give rise to off-target effects through segments of complementarity as short as 16 nt, or through dsRNA segments containing simple trinucleotide repeats such as $(CAN)_n$, which are widespread in many fly genes[6]. Furthermore, in this issue Kulkarni *et al.*[7] show that the presence of perfect 19-nt matches to nontargeted transcripts, within long dsRNAs, statisti-

Figure 1 | Appropriate experimental controls to minimize risks of misinterpretation of RNAi data due to off-target effects (OTEs). siRNA-like molecules, vector-based shRNAs and long dsRNAs trigger detectable off-target effects in all major systems studied to date, from mammalian cells to *D. melanogaster* and *C. elegans*. Simple solutions are available to minimize the risk that an observed phenotype may arise from an off-target effect rather than the targeted gene's loss of function.

cally correlates with a higher incidence of phenotypically detectable cross-silencing, when compared to dsRNAs designed to avoid such matches.

The overriding message from these observations is that we should not delay applying these lessons and avoid, at least in *D. melanogaster*, the use of long dsRNAs containing (CAN)$_n$ motifs or exact matches of 19 nt or longer to nontargeted transcripts. Beyond this, the aforementioned reagent redundancy approach is still highly desirable, especially when large screening 'hit lists' are reported. We nonetheless recognize that designing multiple redundant dsRNAs of several hundred base pairs in length can be challenging, particularly with small genes. Thus, in our view, all reports of RNAi phenotypes based on single dsRNA datasets should make clear mention of it, explicitly indicating that specificity controls, which could be used to further rule out the risk of off-target effects, were not completed.

Using non-RNAi datasets to support RNAi results

The basic priority should always be that all feasible controls are carried out to internally minimize risks of misinterpretation of the stand-alone RNAi dataset. When this is not possible, relying on other types of data, especially if they come from other 'omics' analyses (which are themselves fraught with false positives) carries a notable risk of misinterpretation, owing to circular arguments stemming from preconceptions or inadvertent stretching of interpretations to fit a desired hypothesis. Nonetheless, some types of experiments such as genetic knockouts and other loss-of-function methodologies that are readily available in model organisms, offer a safe and compelling basis for either challenging or confirming RNAi data, keeping in mind that negative RNAi results are inherently very difficult to interpret. Beyond gene-specific loss-of-function methodologies or support-

ing datasets from well-diversified, independent experimental sources, we would argue for strong caution in relying on non-RNAi data to support RNAi results that otherwise carry only weak internal controls.

Specificity through uniqueness: the value of phenotypically rich assays

At the heart of these issues is one that tends to be overlooked: the specificity of the phenotypic characterization itself, that is, the 'distinctiveness' of the reported RNAi phenotype. Although the relevance of this factor may not be exclusive to RNAi experimentation, it does have a direct impact on the likelihood that two randomly chosen siRNAs might be judged as causing 'the same phenotype'. In other words, how many general molecular events affected by RNAi might lead to a drop in cell proliferation in HeLa cells, or a decrease in glucose-stimulated insulin secretion in an insulinoma cell line? Are certain assays more prone than others to be affected by off-target effects? Indeed, most will agree that assays designed to cast a 'wide net', that is, monitoring phenotypic parameters that can be modulated by a very broad range of cellular changes, bear a higher risk of detecting off-target effects than those assays that monitor very precise, narrowly controlled parameters.

It is therefore generally advisable for users to maximize the reliability of RNAi results by monitoring specific and endogenous-based cell readouts to increase the selectivity of the phenotype, and thereby help to sort out false positives. One solution is to generate, whenever possible, multiparametric measurements either through single high-content readouts or multiple parallel analyses: monitoring multiple markers, examining kinetics via multiple time points, comparing phenotypes from panels of multiple cell lines or from different mutant backgrounds, and others. Of course, this level of detailed analysis is seldom feasible during the first

pass of major screens, which therefore will generate higher rates of false positives. This can and should be addressed by following up such first screening passes by applying multiple complementary phenotypic assays in secondary and tertiary screening passes.

Conclusions

We are concerned about the variability in the standards of care applied to genome-scale screens. The proliferation of hit lists containing significant numbers of false positives will, if left unchecked, erode the credibility of RNAi studies. This would be particularly grievous as it would effectively squander the unprecedented power of RNAi, including one of the biggest advantages it now offers—when well-controlled—over most other 'omics' technologies: the ability to achieve extremely low rates of false positives. Although the present discussion focuses on genome-scale screens in which the potential for misleading readers through inadequate controls is greatest, similar or more exacting standards should be expected in small-scale studies. We therefore hope that the present Commentary, if it succeeds in triggering more in-depth discussions within the research community on these and other related issues, will help insure that RNAi experimentation and especially RNAi screening continues to live up to its full potential as one of the most powerful functional genomics tools available to date.

1. Echeverri, C.J. & Perrimon, N. *Nat. Rev. Genet.* **7**, 373–384 (2006).
2. Cullen, B.R. *Nat. Methods* **3**, 677–681 (2006).
3. Jackson, A.L. *et al. Nat. Biotechnol.* **21**, 635–637 (2003).
4. Birmingham, A. *et al. Nat. Methods* **3**, 199–204 (2006).
5. Doench, J.G., Petersen, C.P. & Sharp, P.A. *Genes Dev.* **17**, 438–442 (2003).
6. Ma, Y., Creanga, A., Lum, L., & Beachy, P.A. *Nature* advance online publication 10 September 2006 (doi: 10.1038/nature05179).
7. Kulkarni, M.M. *et al. Nat. Methods* **3**, 833–838 (2006).

[1]Cenix BioScience GmbH, Tatzberg 47, Dresden, 10307, Germany. [2]Howard Hughes Medical Institute, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [3]Morphogenesis Group, Ludwig Institute for Cancer Research UCL Branch, Courtauld Building, 91 Riding House Street, London, W1W 7BS, UK. [4]Signaling and Functional Genomics, German Cancer Research Center (DKFZ/B110), Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany. [5]Max Plank Institute of Molecular Cell Biology and Genetics, 108 Pfotenhauerstrasse, 01307 Dresden, Germany. [6]Division of Cellular Genomics, The Genomics Institute of the Novartis Research Foundation, 10675 John J. Hopkins Drive, San Diego, California 92121, USA. [7]Cancer Research UK London Research Institute, 44 Lincoln's Inn Fields, London WC2A 3PX, UK. [8]Gene Expression Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany. [9]The Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, UK. [10]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [11]Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital Boston, 55 Fruit Street, Massachusetts 02114, USA. [12]Department of Medical Oncology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA. [13]Rosetta Inpharmatics, 401 Terry Ave. N, Seattle, Washington 98109, USA. [14]Department of Cell and Developmental Biology, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. [15]Department of Cell Biology, UT Southwestern Medical Center, 5323 Harry Hines Blvd. Dallas, Texas 75390, USA. [16]Drosophila RNAi Screening Center and Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [17]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [18]Molecular and Cancer Biology Program, Biomedicum Helsinki, PO Box 63 (Haartmaninkatu 8), FI-00014 University of Helsinki, Finland. [19]Howard Hughes Medical Institute, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [20]Division of Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.