A probability-based approach for the analysis of large-scale RNAi screens

Renate König[1,3], Chih-yuan Chiang[1], Buu P Tu[1], S Frank Yan[1], Paul D DeJesus[1,3], Angelica Romero[1], Tobias Bergauer[2], Anthony Orth[1], Ute Krueger[2], Yingyao Zhou[1] & Sumit K Chanda[1,3]

We describe a statistical analysis methodology designed to minimize the impact of off-target activities upon large-scale RNA interference (RNAi) screens in mammalian cells. Application of this approach enhances reconfirmation rates and facilitates the experimental validation of new gene activities through the probability-based identification of multiple distinct and active small interfering RNAs (siRNAs) targeting the same gene. We further extend this approach to establish that the optimal redundancy for efficacious RNAi collections is between 4–6 siRNAs per gene.

The discovery of RNAi as a pathway that allows modulation of gene expression has facilitated large-scale genetic screens in model-organism as well as in mammalian cells[1–7]. But off-target effects, where unintended mRNA targets with sequence homology to the RNAi oligonucleotide are degraded, have impeded the analysis of RNAi screens[8,9]. A strategy to establish RNAi target specificity is through the identification of multiple independently active siRNAs that target the same gene[10].

Presently available analysis methodologies for large-scale RNAi data sets typically rely on ranking screening data and are based on single siRNA activity or significance value[2,4,11]. These analyses focus on the identification of highly active siRNAs (wells), which typically fall within the top 1% of the assayed activities, and ignore much of the remainder of the data set. Furthermore, these strategies do not exploit redundancies in genome-scale libraries, which typically contain 2–4 siRNAs per gene. Thus, it is difficult to systematically identify genes for which multiple siRNAs are active across a screen, which do not fall within an upper threshold (that is, moderately active siRNAs).

We therefore developed a statistical score that models the probability of a gene 'hit' based on the collective activities of multiple siRNAs per gene. In this redundant siRNA activity (RSA) analysis, all wells in an assay are initially ranked according to their signals. Then, the rank distribution of all siRNAs (wells) targeting the same gene is examined and a *P*-value is assigned. Thus, *P*-value indicates the statistical significance of all wells targeting a single gene being unusually distributed toward the top ranking slots, calculated based on an iterative hypergeometric distribution formula (**Supplementary Methods** online; http://carrier.gnf.org/publications/RSA)[12]. Subsequently, all wells are ranked first based on this score, then by their individual activities. Therefore, wells clustered toward the top ranks are labeled as active, and the remaining ones are considered negative. Since the *P*-value is associated with a gene, all wells for the same gene are assigned identical *P*-values. In this probability-based approach, a gene with multiple moderately active wells is weighed more heavily than a gene with fewer active wells, although the latter class may still be favorably ranked given sufficiently high activities.

To assess the robustness of this approach, we applied this analysis to two genome-wide siRNA screens (screen A and screen B; **Supplementary Tables 1** and **2** online), and compared it to a conventional activity-based 'ranking' method (**Supplementary Fig. 1a** online). Each assay used identical siRNA libraries targeting approximately 19,628 genes, containing on average 3 wells/gene with 2 siRNAs/well (~6 siRNAs/gene, 53,860 wells total; **Supplementary Fig. 1b**). Experimental analysis of the silencing efficiency of this library indicated that at least 75% of siRNAs reduced their intended target mRNA levels by 75% or more. Thus we considered this to be a relatively efficacious reagent set (**Supplementary Fig. 1c**). All wells were assayed in either duplicate or triplicate, normalized to plate medians, and analyzed. Notably, both assays (screen A and screen B) were designed to detect siRNA antagonists, and also possessed similar dynamic ranges (**Fig. 1a**; **Supplementary Fig. 1d,e**). Screen A, however, had a greater number of (~45% more wells with activities >4 s.d. from the median) and more 'active' outliers than screen B (**Supplementary Fig. 1f**). From each screen we picked siRNAs contained in the ~55 top-ranking wells and 10 nonactive wells (2 siRNAs/well), as determined by each analysis methodology (ranking and RSA), and individually rearrayed the corresponding siRNAs in duplicate. Additionally, 42 siRNAs designed to contain limited homology to known human genes, as well as 3 commercially available oligonucleotides, served as negative controls (**Supplementary Fig. 1d,e** and **Supplementary Table 3** online). We ran both assays in duplicate, and normalized the data to the median signal of the negative control siRNAs. We next analyzed the reconfirmation rates for each analysis method. Each well in the original screens contained 2 siRNAs/well (**Supplementary Fig. 1a**), however, for our confirmation studies, each of

[1]The Genomics Institute of the Novartis Research Foundation, 10675 John J Hopkins Drive, San Diego, California 92121, USA. [2]Qiagen GmbH, Qiagen Str. 1, 40724 Hilden, Germany. [3]Present address: Infectious & Inflammatory Disease Center, Burnham Institute for Medical Research, 10901 North Torrey Pines Road, La Jolla, California 92037, USA. Correspondence should be addressed to Y.Z. (yzhou@gnf.org) or S.K.C. (schanda@burnham.org).

RECEIVED 28 JUNE; ACCEPTED 16 AUGUST; PUBLISHED ONLINE 9 SEPTEMBER 2007; DOI:10.1038/NMETH1089

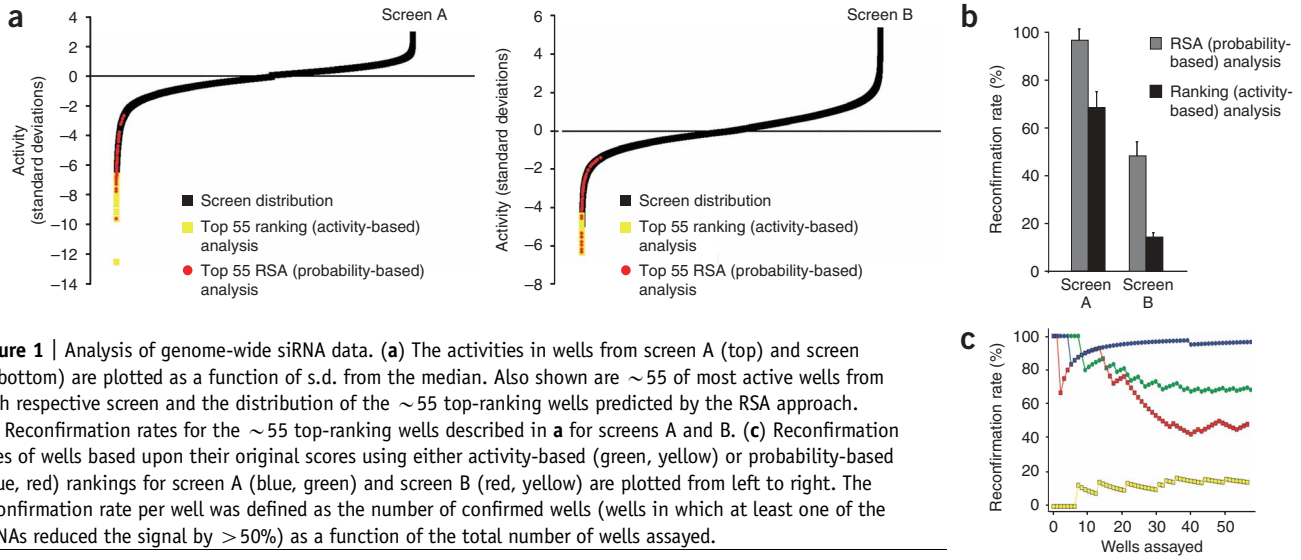

Figure 1 | Analysis of genome-wide siRNA data. (**a**) The activities in wells from screen A (top) and screen B (bottom) are plotted as a function of s.d. from the median. Also shown are ~55 of most active wells from each respective screen and the distribution of the ~55 top-ranking wells predicted by the RSA approach. (**b**) Reconfirmation rates for the ~55 top-ranking wells described in **a** for screens A and B. (**c**) Reconfirmation rates of wells based upon their original scores using either activity-based (green, yellow) or probability-based (blue, red) rankings for screen A (blue, green) and screen B (red, yellow) are plotted from left to right. The reconfirmation rate per well was defined as the number of confirmed wells (wells in which at least one of the siRNAs reduced the signal by >50%) as a function of the total number of wells assayed.

these siRNAs were split into two wells. Therefore, we considered the original siRNA activity confirmed if one of these two siRNAs reduced the assay signal more than 50% when compared to the median value of the negative controls. As screen A had a larger number of outliers than screen B (**Supplementary Fig. 1f, Supplementary Tables 1** and **2** online), 'hitpicks' based on activity-based ranking of screen A were significantly outside the normal distribution of the data set (–7 to –13 s.d.; **Fig. 1a** and **Supplementary Fig. 1f**), whereas those from screen B were closer to the normal distribution curve (–4 to –6 s.d.; **Fig. 1a** and **Supplementary Fig. 1f**). Compared to the activity-based method, the RSA approach identified siRNAs distributed deeper into the data set, but siRNAs identified in screen A, on average, had higher activities than those identified in screen B (**Fig. 1a** and **Supplementary Fig. 1f**). Consistent with these observed characteristics, screen A displayed a two- to fourfold higher reconfirmation rate than screen B (**Fig. 1b,** and **Supplementary Tables 4** and **5** online). Surprisingly, siRNAs identified using the RSA methodology had higher rates of reconfirmation for both screens (**Fig. 1b**). The rate of reconfirmation as a function of ranking in the initial screen analysis by either the RSA or the ranking method for screen B shows the major discrepancy between the two approaches can be attributed to low reconfirmation levels of the most active siRNAs predicted by the ranking approach (**Fig. 1c**). Specifically, the confirmation rate for the most active siRNAs in screen B determined by the ranking analysis was initially around 0%, and then gradually increased. This phenomenon is contrary to the popular presumption that more active wells are necessarily more confirmable. This behavior, however, has been documented for small-molecule high-throughput screens, and is known as an 'abnormal' confirmation curve[13,14]. These false positive activities are usually observed in the wells with high activity and can be attributed to experimental or logistical artifacts. Although this trend is not as dramatic in screen A, we observed similar behavior in an additional screen (**Supplementary Fig. 1g**). Notably, the reconfirmation rates for the RSA approach plateaued at higher levels for both screens, suggesting the RSA method can also be used to enrich for true

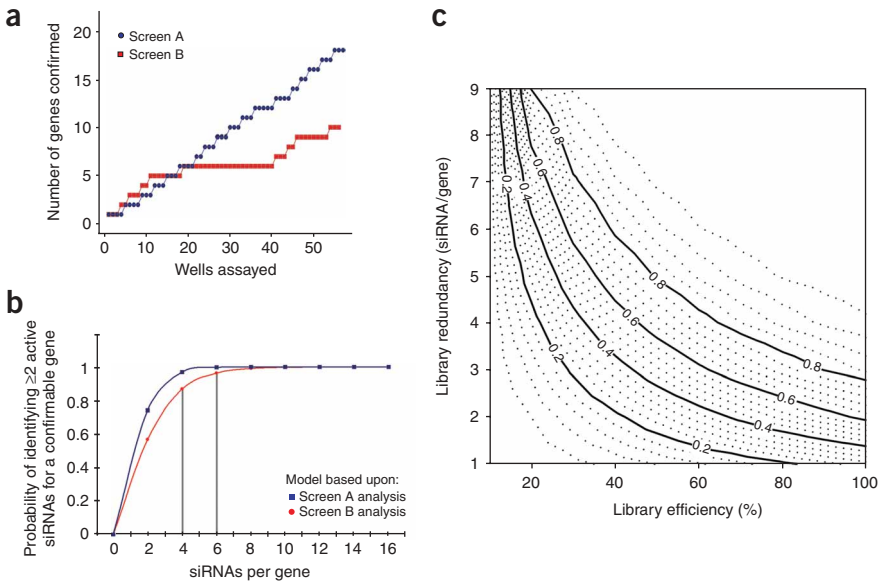

Figure 2 | Gene-centered analysis of large-scale RNAi data. (**a**) Number of confirmed gene activities, as defined by 2 or more active siRNAs targeting the same gene, is plotted as a function of original ranking by RSA analysis for screen A and screen B. (**b**) Confirmation data were used to model the relationship between library redundancy (siRNAs/gene) and the probability of identifying 2 or more siRNAs for a confirmable (true positive) gene in a large-scale screen. (**c**) This model was extended to predict siRNA redundancies required for libraries with varying average efficiencies of target mRNA knockdown (percentage of siRNAs that reduce target mRNA by 75% or more). Probabilities are shown on contour lines (also see **Supplementary Methods**).

positive activities through a deeper sampling of large-scale siRNA data sets (**Fig. 1c**).

We next determined the frequency of genes for which two or more independent siRNAs reconfirmed in its original assay, suggesting true 'on-target' activities. For screen A, we used the RSA approach to identify 57 wells corresponding to 19 genes. We were able to confirm the activities of at least 2 siRNAs for 18/19 genes (95%; **Fig. 2a**). In most cases, 3 or more siRNAs per gene recapitulated the original phenotype in screen A or B (**Supplementary Fig. 2a** online). Additionally, quantitative RT-PCR revealed a strong correlation between the phenotypic activity of siRNAs and their ability to silence their cognate transcript (**Supplementary Fig. 2b**). Moreover, we observed a 96% validation rate of these siRNAs in a relevant secondary assay for screen A (**Supplementary Fig. 2c**,**d**).

We next analyzed screen B, which had a significantly lower confirmation rate (48%), to determine if this methodology could be successfully applied to less robust data sets. RSA analysis identified 56 wells targeting 28 genes. We were able to identify 10 genes (36% of total or 75% of confirmed hitpicks) for which at least 2 siRNAs possessed potent activities (**Fig. 2a**), suggesting that artifacts inherent to high-throughput screening can account for a considerable fraction of false positives predicted by the RSA analysis.

To determine the degree of complexity required in a library to optimize the detection of multiple siRNA activities against one gene (a 'true hit'), we developed a mathematical model to understand the relationship between confirmation rate of a true hit and library redundancy. We applied this model to the RSA analysis to calculate probability of identifying multiple siRNAs targeting the same gene in libraries of varying redundancies. This analysis indicates that siRNA libraries should contain 4–6 siRNAs targeting each gene analyzed, and libraries of greater redundancy provide diminishing levels of additional validation (**Fig. 2b**). Additionally, we calculated redundancies for siRNA libraries with varying average efficiencies of target mRNA knockdown based on these data (**Fig. 2c**).

Taken together, our results demonstrate the utility of performing RSA analysis on large-scale RNAi data sets. Similar approaches have been successfully used for analysis of microarray data, but there currently exists no parallel methodology for the analysis of large-scale RNAi screens[15]. As the RSA methodology analyzes the collective behavior of all wells targeting a gene, it is a powerful approach to circumvent potential off-target effects through the identification of multiple active siRNAs. Although this approach is limited by false positive activities, which are likely attributed to screening artifacts, our results demonstrate that RSA outperforms activity-based analyses in the identification of confirmable activities. Therefore, application of this methodology should considerably enhance the interpretation of large-scale RNAi data through the reduction of false-positive activities derived from both experimental artifacts and off-target activities.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank L. Miraglia for helpful discussions and oversight of screens, J. Zhang for excellent technical assistance, S. Batalov (Genomics Institute of the Novartis Research Foundation) and P. Aza-Blanc (Burnham Institute) for the identification of negative control siRNA sequences, E. Lader (Qiagen) for facilitating collaboration, D. Elleder (Salk Institute) for providing the MLV supernatant, N.R. Landau (New York University, School of Medicine) for providing pNL43-luc-r$^+$e$^-$, and N. Somia (University of Minnesota) for the gift of pCMVgp. R-language implementation of the RSA algorithm was provided by B. Zhou (Genomics Institute of the Novartis Research Foundation). This work was supported by the Novartis Research Foundation and a grant from the US National Institutes of Health (1 R01 AI072645-01).

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Berns, K. *et al. Nature* **428**, 431–437 (2004).
2. Boutros, M. *et al. Science* **303**, 832–835 (2004).
3. Brummelkamp, T.R. *et al. Nat. Chem. Biol.* **2**, 202–206 (2006).
4. Moffat, J. *et al. Cell* **124**, 1283–1298 (2006).
5. Paddison, P.J. *et al. Nature* **428**, 427–431 (2004).
6. Sonnichsen, B. *et al. Nature* **434**, 462–469 (2005).
7. Westbrook, T.F. *et al. Cell* **121**, 837–848 (2005).
8. Birmingham, A. *et al. Nat. Methods* **3**, 199–204 (2006).
9. Jackson, A.L. *et al. Nat. Biotechnol.* **21**, 635–637 (2003).
10. Echeverri, C.J. *et al. Nat. Methods* **3**, 777–779 (2006).
11. Kittler, R. *et al. Nature* **432**, 1036–1040 (2004).
12. Yan, S.F., Asatryan, H., Li, J. & Zhou, Y. *J. Chem. Inf. Model.* **45**, 1784–1790 (2005).
13. Fay, N. & Ullmann, D. *Drug Discov. Today* **7**, S181–S186 (2002).
14. Bajorath, J. *Nat. Rev. Drug Discov.* **1**, 882–894 (2002).
15. Shedden, K. *et al. BMC Bioinformatics* **6**, 26 (2005).