A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens

Frederic D Sigoillot[1], Susan Lyman[1,3], Jeremy F Huckins[1,3], Britt Adamson[2], Eunah Chung[1], Brian Quattrochi[1,3] & Randall W King[1]

Because off-target effects hamper interpretation and validation of RNAi screen data, we developed a bioinformatics method, genome-wide enrichment of seed sequence matches (GESS), to identify candidate off-targeted transcripts in primary screening data. GESS analysis revealed a prominent off-targeted transcript in several screens, including *MAD2* (*MAD2L1*) in a screen for genes required for the spindle assembly checkpoint. GESS analysis results can enhance the validation rate in RNAi screens.

RNAi is a powerful discovery tool, but frequent false positives complicate analysis of genome-wide RNAi screens[1–3]. The problem arises because siRNAs can induce microRNA-like effects, downregulating expression of hundreds of genes nonspecifically[4,5]. Such effects can occur with as few as 6–7 nucleotides of sequence complementarity, although effects may become more pronounced with greater complementarity[6]. Some transcripts may be particularly susceptible to off-target silencing[7–9], but the identification of such 'off-targeted' transcripts typically occurs only after much effort has been expended to validate genes of interest. Therefore, new methods are necessary to identify these off-targeted transcripts earlier in the validation process.

We conducted an image-based high-throughput siRNA screen (**Supplementary Results 1** and **Supplementary Fig. 1**) to identify new genes required for the spindle assembly checkpoint (SAC)[10] in human cells. We determined that off-target effects were pervasive, as we could not validate any new genes from the primary screen despite having identified known components of the pathway. To understand the basis of the off-target effect, we tested 34 siRNAs with the strongest phenotype for their ability to downregulate known components involved in the SAC, and found that all 34 siRNAs strongly decreased *MAD2* (also known as *MAD2L1*) mRNA and protein levels in addition to those of their intended target (**Supplementary Results 2** and **Supplementary Fig. 2**). Half of these siRNAs contained a 7-mer seed sequence complementary to the *MAD2* 3′ untranslated region (UTR), indicating the potential for microRNA-like off-target regulation. We tested seven of these seed-match–containing siRNAs, and found that all downregulated a *MAD2* 3′ UTR reporter construct (**Supplementary Fig. 3**). Over half of all 324 siRNAs that inactivated the SAC (active siRNAs) in the screen contained a 7-mer seed match in the *MAD2* 3′ UTR, whereas only 8% of the siRNAs that did not alter the phenotype (inactive siRNAs) contained a seed match. These findings indicate that the majority of active siRNAs in our SAC component screen are likely to alter the phenotype by nonspecifically targeting the *MAD2* transcript.

To identify such potentially devastating off-target effects before the validation process, we developed an approach that uses primary screening data to identify transcripts that are sensitive to off-target effects (**Fig. 1**). Using phenotypic screen data, we separated the siRNAs into two groups: active ('with phenotype') and inactive ('without phenotype'). The program then calculates the seed-match frequency for active (SMF^a) and inactive (SMF^i) siRNAs for each transcript encoded in the genome (**Fig. 2**). In principle, transcripts that are sensitive to off-target regulation will bias the ratio of SMF^a:SMF^i, to which we refer as seed-match enrichment (SME), such that it exceeds 1, and we determined the statistical significance of this bias relative to other genes in the dataset. We refer to this approach as GESS analysis. It can be performed using genome-wide databases of full-length mRNAs or mRNA subregions (3′ UTRs, 5′ UTRs or coding sequence), although we only identified off-targeted transcripts using the 3′ UTR database, consistent with known rules of microRNA-based targeting.

We first evaluated the ability of the GESS analysis to identify *MAD2* as an off-targeted transcript in our SAC screen. Using GESS analysis we compared the seed-match frequency of the siRNAs that most strongly inactivated the SAC ($n = 49$) to the siRNAs that did not ($n = 9{,}856$). We analyzed each of 27,534 3′ UTR sequences in the human genome (**Fig. 2a**). When using a 7-mer seed match from either the antisense or sense strand seed sequences of an siRNA as a search criterion, we found that the 3′ UTR of the *MAD2* transcript had eightfold SME ($SMF^a = 65.3\%$; $SMF^i = 8.2\%$; $P = 4.2 \times 10^{-23}$). The only other significantly enriched transcript ($P = 1.3 \times 10^{-19}$) represented another *MAD2* sequence in the database. A GESS analysis in which all siRNA seed sequences were randomly scrambled showed no statistically significant outliers (corrected $P > 0.05$; **Supplementary Fig. 4**).

We determined how the GESS analysis of our SAC screen was affected by the following parameters: (i) strength of phenotype, (ii) the seed sequence length, (iii) the seed-match multiplicity, (iv) the source of inactive control siRNAs and (v) seed

[1]Department of Cell Biology, Harvard Medical School, Boston, Massachusetts, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]Present addresses: Gilead Sciences, Inc., Foster City, California, USA (S.L.), Department of Psychology and Brain Sciences, Dartmouth College, Hanover, New Hampshire, USA (J.F.H.) and Center for Academy Achievements, University of Massachusetts Medical School, Worcester, Massachusetts, USA (B.Q.). Correspondence should be addressed to R.W.K. (randy_king@hms.harvard.edu).

RECEIVED 30 AUGUST 2011; ACCEPTED 27 JANUARY 2012; PUBLISHED ONLINE 19 FEBRUARY 2012; DOI:10.1038/NMETH.1898

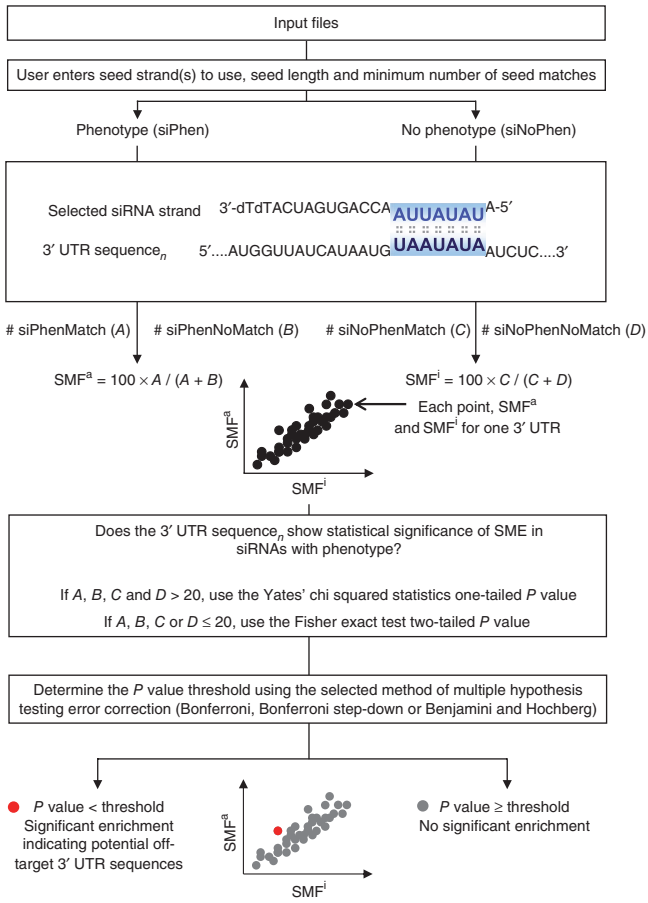

Figure 1 | Summary of the GESS method. Input files contain individually tested siRNA sequences, corresponding phenotype data and 3′ UTR sequence database. The GESS algorithm first splits the siRNAs into those with phenotype and those without phenotype. The user enters criteria for defining a matching transcript, including the siRNA strand(s) to use (either strand, antisense only, sense only or both strands), seed length (6, 7 or 8 for 6-mer, 7-mer or 8-mer, respectively) and seed-match multiplicity that must be found in a 3′ UTR to call the siRNA 'matching'. GESS calculates the percentage of siRNAs in each set that shows seed matching with each 3′ UTR in the genome-wide database. Statistical significance of SME among active siRNAs is compared to that of inactive siRNAs.

sequence strand choice (**Supplementary Results 3**). Relaxing the phenotype strength led to identification of additional outliers, yet *MAD2* remained the most significantly enriched transcript (**Supplementary Fig. 5**). Increasing the stringency of the method by lengthening the seed from seven to eight nucleotides also permitted specific identification of *MAD2* (**Supplementary Fig. 6**). Increasing the seed-match multiplicity, which increases stringency by requiring two seed matches per transcript, did not identify *MAD2* in some cases (**Supplementary Fig. 7**). Because most published RNAi screens do not provide the nucleotide sequences of all tested siRNAs, we developed an alternative method for generating a set of inactive seed sequences, in which we changed nucleotide 1 of the seed sequences of active siRNAs to its complement (P1c seeds), and found that this method also identified *MAD2* as an off-targeted transcript (**Supplementary Fig. 8**). Finally, considering seed matches from only the siRNA antisense strand showed better sensitivity but somewhat lower specificity than including each strand in the analysis (**Supplementary Fig. 9**).

We next tested whether a GESS analysis could identify off-targeted transcripts in other published screens. A recent screen had identified siRNAs that could overcome mitotic arrest induced by a small-molecule inhibitor of the mitotic kinesin Eg5 (ref. 11). Because mitotic arrest induced by this mechanism is SAC-dependent[12], we anticipated that *MAD2* could be an off-targeted transcript in this screen. In this case, the experimentally determined inactive siRNAs are not published, so we used P1c seeds to generate the inactive siRNAs. GESS analysis of this dataset, using 7-mer seeds and a seed-match multiplicity of 1, indeed identified the *MAD2* 3′ UTR as the strongest statistically

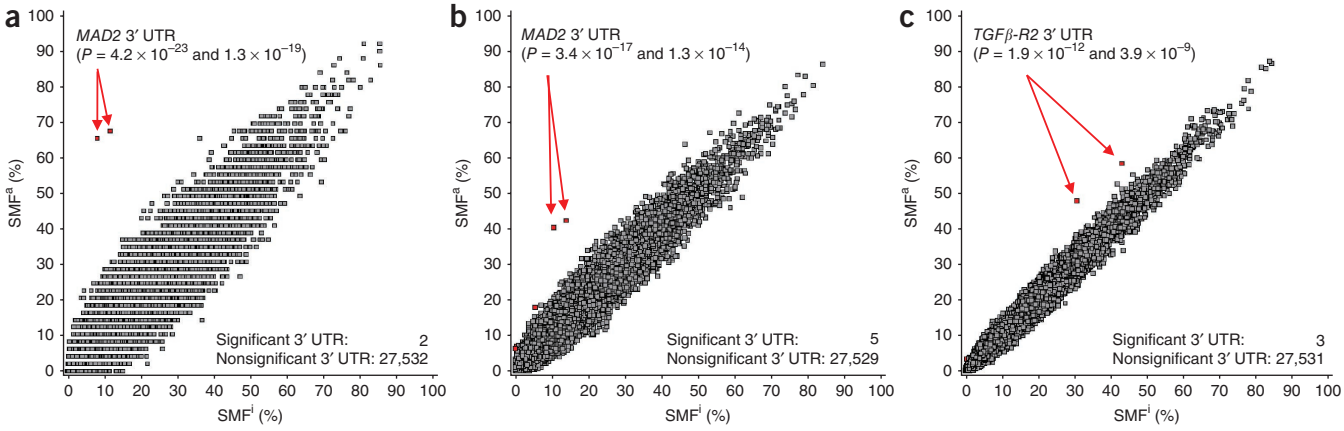

Figure 2 | Identification of major off-targeted transcripts in RNAi-screen datasets using the GESS method. (**a**) GESS analysis of 27,534 human mRNA 3′ UTRs from primary data of a screen that identified siRNAs inducing loss of SAC function. Each point represents one 3′ UTR and represents SMF^a value (percentage of 49 total active siRNAs) plotted against the SMF^i value (percentage of 9,856 total inactive siRNAs). (**b**) GESS analysis as in **a** on data published from an siRNA screen for components required for mitotic arrest upon inhibition of the mitotic kinesin Eg5 in HeLa cells[11]. P1c seeds were used as the source of inactive siRNAs (n = 308, for both active and inactive siRNAs). (**c**) GESS analysis as in **a**, on data published from an siRNA screen for genes involved in the TGFβ pathway[9]. Experimentally identified siRNAs that showed no phenotype (a cutoff of 2 s.d. of activity was used to separate active from inactive siRNAs) were used for the set of inactive siRNAs (391 active siRNAs, 18,869 inactive siRNAs). Significance threshold was determined independently for each data point, using the Benjamini and Hochberg (Simes') method to correct the baseline value of α which is 0.05. Significant outliers are highlighted in red.

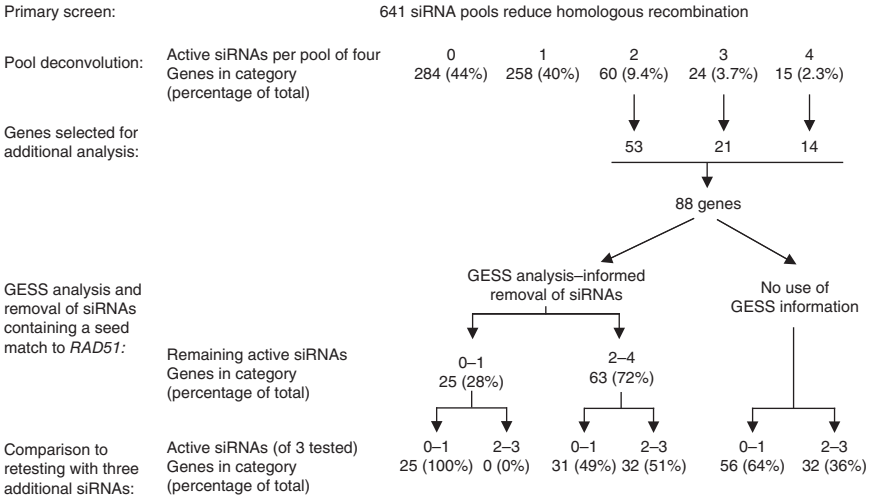

Figure 3 | GESS-informed selection of siRNA pools enriches for genes that reproduce the primary phenotype upon targeting with additional siRNAs. Schematic of RNAi screening and validation steps, reported in ref. 13, shows siRNA pools targeting 641 transcripts selected for defective homologous recombination in a primary siRNA screen. Upon deconvolution, pools targeting 99 genes showed a strong phenotype on at least two out of four siRNAs. Of these genes, 88 were evaluated further. GESS analysis showed that the *RAD51* 3′ UTR was sensitive to off-target effects[13]. The schematic shows the rate at which the strong phenotype was reproduced with and without removal of siRNAs that contain an antisense 7-mer seed match to *RAD51* 3′ UTR.

significant outlier, with SME of 3.9 ($P = 3.3 \times 10^{-18}$; **Fig. 2b**). A control analysis in which we randomly scrambled all active and inactive siRNA seed sequences showed no significant outliers (**Supplementary Fig. 10**).

We tested the GESS method on a previously published RNAi screen of 6,000 human genes for new components of the TGFβ pathway, which did not identify any new components of the pathway and was plagued by off-target effects[9]. In that study, the vast majority of active siRNAs tested (89%; 172 of 193 tested) had been experimentally confirmed to reduce the amounts of transcripts encoding either the TGFβ receptor 1 or 2, with the latter being more sensitive. We performed a GESS analysis on primary data of the screen, using the 391 active siRNAs and 18,869 inactive siRNAs. Using at least one 7-mer seed match as a search criterion, the GESS analysis revealed *TGFβR2* (represented by two sequences in the database) as the major outlier in the analysis with SME values of 1.6 ($P = 1.9 \times 10^{-12}$) and 1.4 ($P = 3.9 \times 10^{-9}$), whereas *TGFβR1* (two sequences in the database) showed no significant enrichment (SME = 0.97, $P = 0.664$ and SME = 0.99, $P = 0.832$) (**Fig. 2c**). We identified a third weak outlier, but there is no evidence that it is involved in the TGFβ pathway. A control GESS analysis with randomly scrambled seed sequences for all siRNAs showed no significant outliers (corrected $P \geq 0.05$; **Supplementary Fig. 11**). We also investigated the effect of varying GESS parameters on the outcome of the analysis (**Supplementary Results 4** and **Supplementary Fig. 12**).

Finally, in a separate study[13] GESS analysis revealed *RAD51* as a potential off-targeted transcript in a screen for genes required for homologous recombination[13], and off-target regulation of *RAD51* had been confirmed experimentally[13]. To examine whether a GESS analysis can help prioritize hits from siRNA screens, we investigated the consequences of removing siRNAs that contain a seed match against the *RAD51* 3′ UTR (**Fig. 3**). The primary screen, followed by pool deconvolution, identified 88 candidate genes using a criterion of at least 2–4 siRNAs producing the phenotype. After removing siRNAs that contain a 7-mer antisense seed match to *RAD51* 3′ UTR, 63 candidate genes retained at least two active siRNAs. We compared the performance of the original 88 candidates to the 63 'GESS analysis–selected' candidate genes using data from a validation analysis in which three additional independent siRNAs to each gene had been evaluated[13]. In this analysis, 32 of 88 genes scored with at least two of three additional siRNAs, a confirmation rate of 36%. When we restricted the analysis to the GESS analysis–selected candidates, 32 of 63 candidate genes were confirmed (51%). None of the 25 candidates eliminated by GESS showed a phenotype with more than one of three new siRNAs. When we repeated this process using ten randomly selected genes containing a 3′ UTR of similar length, we observed no positive effect on validation rate (**Supplementary Table 1**). This analysis indicates the value of taking into account potential off-targeted transcripts identified in a GESS analysis in prioritizing genes for validation in siRNA screens.

There is no tool other than the GESS method, to our knowledge, that can be used to systematically examine screen data to directly identify potential off-targeted transcripts. A previously described approach to identify off-target effects in screens searches for siRNA seed sequences that are overrepresented in the set of active siRNAs as compared to inactive siRNAs[7,8,14] but does not identify which transcripts might be targeted. We compared the GESS method to seed sequence enrichment analysis. For screens in which GESS analysis had identified a biologically confirmed, significant outlier, we attempted to identify 7-mer seeds that were overrepresented in active siRNAs compared to inactive siRNAs (**Supplementary Table 2a**). In our SAC screen, we identified eight such seed sequences (**Supplementary Table 2b**), indicative of a potential off-target effect. However, seed sequence enrichment analysis alone did not highlight the extent of off-target silencing in the screen, as only 35 of 324 active siRNAs (11%) contained a seed sequence that was significantly enriched (corrected $P < 0.05$). Furthermore, this analysis could not directly identify the *MAD2* 3′ UTR as the relevant off-targeted transcript in this dataset. Analysis of the dataset from the TGFβ pathway RNAi screen[9] identified one 7-mer seed sequence that was significantly enriched ($P = 9.7 \times 10^{-7}$) among active siRNAs (**Supplementary Table 2c**), present in only five of 391 (1.28%) active siRNAs as compared to five of 18,869 inactive siRNAs (0.03%). Analysis of the Eg5 inhibitor override screen[11], as well as the homologous recombination screen[13], did not identify statistically overrepresented seed sequences. In summary, the GESS method appears to be more sensitive in identifying potential off-target effects compared to simple seed-sequence analysis and can directly identify the sensitive transcript(s). Because GESS

analysis does not require that active siRNAs contain a common seed sequence, it can detect off-target effects even if no particular seed sequence is enriched among active siRNAs. The GESS method uses the sequence of an mRNA transcript to 'integrate' the information contained among different active siRNAs.

In total, we analyzed data from 13 screens (**Supplementary Table 3**), and identified four screens, described here, in which we identified significant outliers and established microRNA-based off-target effects to be problematic. Nine published RNAi screen datasets showed either no significant outliers or a few weakly significant outliers whose biological importance has not been investigated. The sequences of inactive siRNAs are not published for five of these nine screens, and thus we used P1c-seed sequences as a source of inactive siRNAs. However, this approach is not as information-rich as using the experimentally determined inactive siRNAs because the statistical significance of enrichment in the GESS analysis depends not only on an increase in the frequency of seed matches to a transcript among active siRNAs, but also a corresponding decrease in frequency of seed matches among inactive siRNAs. Furthermore, GESS analysis of genome-wide screens is most informative if siRNAs are screened individually rather than as pools. Because screens in *Drosophila melanogaster* and *Caenorhabditis elegans* use multiple siRNAs generated from long dsRNAs (~500 base pairs), GESS is unlikely to be informative in these systems.

Why some transcripts are particularly sensitive to miRNA-like off-target effects remains unclear. *MAD2* is average among known spindle checkpoint genes in terms of 3′ UTR length or (A+U)-richness, ruling out trivial explanations. The *MAD2* 3′ UTR may contain specific secondary structures or bind to specific proteins that render it particularly sensitive to off-target effects. Alternatively, the function of the SAC may be particularly sensitive to small changes in MAD2 protein levels. Consistent with this idea, MAD2 is a haplo-insufficient tumor suppressor *in vivo*, and cells lacking one copy of *MAD2* show decreased ability to arrest in mitosis in the presence of microtubule inhibitors[15]. Similarly, the process of homologous recombination may be particularly sensitive to *RAD51* gene dosage, explaining why *RAD51* was identified by GESS as a prominent off-targeted transcript in an siRNA screen for genes involved in homologous recombination[13]. Finally, similar observations have been reported for the TGFβ pathway RNAi screen[9] in which minor reductions of the TGFβ receptor transcripts appear to have a major effect on the screen assay. Together these findings suggest it may be useful to assemble a database of genes whose transcripts are highly sensitive to off-target effects and incorporate this information into the design algorithms used to generate siRNAs. Incorporation of GESS as a routine component of the analysis of high-throughput screens should enable investigators to counter-screen for down-regulation of sensitive transcripts and reduce the false positive rate during the validation process. Identification of transcripts sensitive to off-target effects will also enable a better understanding of the rules that govern miRNA-like targeting and help improve the design of siRNA reagents for RNAi screens.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

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

ACKNOWLEDGMENTS

We thank members of the Institute of Chemistry and Cell Biology and C. Shamu for providing siRNA sequences for hits from the Eg5-inhibitor RNAi screen as well as the use of facility equipment for our screening experiments, S. Natesan and P. August for helpful discussions in early stages of this work, S. Elledge for helpful discussions and for critical reading of the manuscript and J. Ware at the Harvard Catalyst Biostatistics consulting group for help in devising the statistical analysis workflow in the present manuscript. Funding for statistical analysis was supported in part by grant 1 UL1 RR025758-01, Harvard Clinical and Translational Science Center, from the US National Center for Research Resources; the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources or the US National Institutes of Health. This research was funded by a Sanofi-Aventis grant and US National Institutes of Health grant GM66492 to R.W.K.

AUTHOR CONTRIBUTIONS

F.D.S., S.L. and R.W.K. conceived the study. F.D.S., S.L., E.C., B.Q. and B.A. performed the experiments. F.D.S. and J.F.H. wrote the GESS program code. F.D.S. and R.W.K. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Jackson, A.L. & Linsley, P.S. *Nat. Rev. Drug Discov.* **9**, 57–67 (2010).
2. Mohr, S., Bakal, C. & Perrimon, N. *Annu. Rev. Biochem.* **79**, 37–64 (2010).
3. Sigoillot, F.D. & King, R.W. *ACS Chem. Biol.* **6**, 47–60 (2011).
4. Jackson, A.L. *et al. Nat. Biotechnol.* **21**, 635–637 (2003).
5. Tschuch, C. *et al. BMC Mol. Biol.* **9**, 60 (2008).
6. Bartel, D.P. *Cell* **136**, 215–233 (2009).
7. Lin, X. *et al. Oncogene* **26**, 3972–3979 (2007).
8. Lin, X. *et al. Nucleic Acids Res.* **33**, 4527–4535 (2005).
9. Schultz, N. *et al. Silence* **2**, 3 (2011).
10. Musacchio, A. & Salmon, E.D. *Nat. Rev. Mol. Cell Biol.* **8**, 379–393 (2007).
11. Tsui, M. *et al. PLoS ONE* **4**, e7339 (2009).
12. Kapoor, T.M., Mayer, T.U., Coughlin, M.L. & Mitchison, T.J. *J. Cell Biol.* **150**, 975–988 (2000).
13. Adamson, B., Smogorzewska, A., Sigoillot, F.D., King, R.W. & Elledge, S.J. *Nat. Cell. Biol.* advance online publication, doi:10.1038/ncb2426 (19 February 2012).
14. Sudbery, I., Enright, A.J., Fraser, A.G. & Dunham, I. *BMC Genomics* **11**, 175 (2010).
15. Michel, L.S. *et al. Nature* **409**, 355–359 (2001).

ONLINE METHODS

Software and data. The GESS standalone package used in this work is available as a compressed archive (**Supplementary Software 1–5**). Software packages for updated versions will be available at http://king.med.harvard.edu/. All siRNA sequences and associated phenotype data used to perform GESS analyses described in this manuscript are available in **Supplementary Data 1**. Excel result files for the main GESS analyses in this manuscript are available in **Supplementary Data 2**. Transcript database files for the human and mouse genomes are available in **Supplementary Data 3**.

Tissue culture. HeLa H2B-GFP cells were grown from low passage in Dulbecco's modified Eagle medium (DMEM; Cell-Gro) supplemented with 10% fetal bovine serum (FBS; Atlanta Biologicals), in a humidified incubator at 37 °C and 5% CO_2.

siRNA library. The Qiagen 'Druggable' genome siRNA library V1.0, consisting of two individual siRNAs for 5,090 human genes, was used in our primary siRNA screen. Nontargeting siRNA control 3 (D-001210-01-20) and *MAD2* (GGAACAACUGAAAGAUUGGdTdT, custom synthesis) were from Dharmacon. The sequences of all siRNAs used in this study are reported in **Supplementary Data 1** along with corresponding phenotypic data.

siRNA transfections and image-based screening. HeLa H2B-GFP cells were plated at 1,000 cells per well in 30 µl OptiMEM containing 1% FBS, supplemented with penicillin, streptomycin and 2 mM glutamine (pen-strep-glu) in 384-well plates (Corning), 16–18 h before transfection. The cells were washed twice with OptiMEM containing pen-strep-glu and then incubated in 40 µl OptiMEM containing pen-strep-glu per well for 1–4 h before transfection. For each well to be transfected with siRNA, 8 µl OptiMEM, 0.5 µl GeneSilencer diluent and 0.25 µl GeneSilencer (Genlantis) was first premixed and then added to 2 µM siRNA. The siRNA-reagent mix was incubated at room temperature for 15 min and then added to cells, yielding a final siRNA concentration of 150 nM. Four-to-six hours after transfection, 20 µl DMEM containing 30% FBS were added. Taxol (150 nM final concentration) was added to the cells 32–36 h after transfection. Twenty-four hours later, cells were fixed, and nuclei were stained by adding one volume of Dulbecco's phosphate-buffered saline (DPBS) fix/stain solution (final concentrations: 3.7% formaldehyde, 250 ng/ml Hoechst 33342 (Molecular Probes) and 0.1% Triton X-100). After 20 min incubation at room temperature, the cells were washed 2–3 times with DPBS. Fluorescence images of nuclei were obtained using a CellWoRx high-content screening microscope (Applied Biosystems). Nuclear morphology was analyzed by manual inspection of images. Untransfected cells and those transfected with control siRNA remain arrested in mitosis under these conditions (**Supplementary Fig. 1**). In contrast, cells treated with a positive control siRNA targeting the essential SAC component MAD2 exited mitosis, as indicated by the presence of interphase cells with multilobed nuclei (SAC bypass; **Supplementary Fig. 1**). Each siRNA was transfected in duplicate wells and each well was imaged in one location. Each image was given a penetrance phenotype (P) reflecting the number of cells affected by the siRNA. The penetrance categories were: 3 (80–100% cells affected), 2.5 (60–80% cells), 2 (40–60% cells), 1.5 (15–40% cells) and 1 (>0% to 15% cells). A sub-rating (SR), reflecting the proportion of affected cells showing SAC bypass, was also assigned using similar categories from 3 to 1. The penetrance and subrating category values were multiplied to reflect the overall rate of bypass in each image, and the higher rate of the two replicates per siRNA was retained. Three phenotype thresholds were considered in the present analyses: a high threshold ($P \times SR = 9$), yielding 49 active siRNAs; a relaxed threshold ($P \times SR \geq 7.5$), yielding 137 siRNAs; and a low threshold ($P \times SR \geq 2$), yielding 324 active siRNAs.

Plasmid constructs. Total RNA was isolated and purified from HeLa *H2B-GFP* cells using the RNeasy kit (Qiagen). A cDNA library was generated by reverse-transcribing the total RNAs using a reverse transcription system (Promega) following manufacturer's protocol. *MAD2* mRNA sequences were PCR-amplified from the cDNA library. The PCR primers contained XbaI sites at both ends. XbaI-digested PCR fragments were cloned into the pGL3-control vector (Clontech) digested with XbaI. This results in expression of an mRNA coding for the Firefly luciferase with *MAD2* sequences downstream of the stop codon(*). The BglII-BamHI cassette from pRL-TK vector (Clontech), containing the *Renilla* luciferase gene under the control of HSV thymidine kinase promoter, was nondirectionally cloned into the BamHI site of the pGL3-control and pGL3-control-*-*MAD2* sequences vectors. Resulting plasmids sequences were verified by DNA sequencing.

Luciferase reporter assays. HeLa *H2B-GFP* cells were plated in 24-well plates (BD Falcon 353047) at 30,000 cells per well in 500 µl OptiMEM containing 1% FBS and pen-strep-glu. Sixteen hours after plating, cells were transfected with 50 nM siRNAs with GeneSilencer as follows. The cells were washed with OptiMEM and then incubated 1–4 h in 150 µl OptiMEM containing pen-strep-glu in the absence of FBS. GTS diluent (2.5 µl) and GeneSilencer reagent (1.25 µl) were premixed in 40 µl OptiMEM and added to 5 µl of siRNA stocks (2 µM) for each well. The siRNA transfection mix was incubated 15 min at room temperature and added to the cells. DMEM containing 20% FBS (200 µl) was added to each well 4–6 h after siRNA transfection. Twenty-four hours later, the siRNA transfection medium was replaced with 500 µl growth medium (without pen-strep-glu) and plasmid transfection was initiated. Plasmids (500 ng per well) were transfected with Fugene 6 (Roche) using a reagent ratio of 5 µl Fugene 6: 2 µg plasmid. OptiMEM (100 µl) was mixed with 0.75 µl Fugene 6 and preincubated at room temperature for 5 min. The premix was added to 500 ng plasmid. The plasmid-reagent mix was then added to the cells after 15 min incubation at room temperature. Dual luciferase assays were performed 24–48 h after initiating plasmid transfections, following the manufacturer's protocol (Dual-Glo system, Promega). Luminescence measurements were performed on an Envision plate reader (PerkinElmer).

Branched DNA (bDNA) assay for mRNA level quantification. mRNA levels were measured, in duplicate, 48 h after siRNA transfection of 20,000 HeLa *H2B-GFP* cells per well in 24-well plates (as described for the luciferase assays). The bDNA assay (QuantiGene/Panomics) was conducted following the manufacturer's protocol and using probe sets specific to *MAD2* (PA-11305-02; NM_002358), *BUBR1* (PA-11159-01;

NM_001211), *BUB1* (PA-11577-01; NM_004336), or the housekeeping genes *GAPDH* (PA-10382-02; NM_002046) and *PPIB*, which encodes a subunit of Cyclophilin (PA-10384-02; NM_000942). Duplicate measurements were averaged, normalized per average housekeeping *PPIB* mRNA measurement. The normalized ratio for control siRNA–transfected cells was used as 100% reference for determination of relative changes in the ratio for other siRNAs. The results were displayed as a heat map with indicated scale using Spotfire DecisionSite.

Quantitative western blotting. MAD2 and GAPDH protein levels were determined by SDS-PAGE separation of proteins followed by western blotting. The proteins were detected using a rabbit antibody to MAD2 (Bethyl, A300-301A) and mouse antibody to GAPDH (AbCam, ab8245). Secondary antibodies coupled to fluorophores (anti-mouse Alexa-Fluor750 and anti-rabbit Alexa-Fluor680, Invitrogen) were used to detect both signal on the same membrane using an Odyssey (Li-Cor Biosciences) scanner. Quantifications were performed using the Odyssey program and are reported as MAD2/GAPDH signal ratio, normalized to control treatment.

Sequence databases. Genome-wide sequences for human 5′ UTRs, coding sequences or 3′ UTRs were retrieved from the Ensembl database using the online tool Martview (http://www.biomart.org/; Ensembl Genes 61, Human genome built GRCh37 or earlier) or Refseq using the UCSC Genome Bioinformatics table browser (http://genome.ucsc.edu/cgi-bin/hgTables; Refseq 44). Sequences of 19 nucleotides or less and duplicate sequences were removed, and the remaining sequences were formatted into a text file with one sequence per line, and corresponding identity information was formatted into an Excel file with the same name (**Supplementary Data 3**).

Genome-wide enrichment of seed sequences (GESS) bioinformatics tool. Matlab (MathWorks) 2007 and more recent versions were used to program and run the GESS seed-match search tool. The program is available as Matlab m-code files and as standalone versions packaged with the appropriate Matlab Compiler Runtime (MCR) for Windows 32-bit and 64-bit systems, Linux 32-bit and 64-bit systems, and Mac OS (**Supplementary Software 1**–**5**). The packages were compiled using Matlab Compiler version 4.14 or later. Input data files examples available in **Supplementary Data 1** consist of two text files that can be generated following the GESS manual provided along with the program (in **Supplementary Software 1**–**5**). One input data file in **Supplementary Data 1** contains a list of either all 19-nucleotide siRNA sequences (sense strand sequence or target sequence) in upper case with ATGC code (no U) one sequence per line, or only the sequence of active siRNAs if inactive siRNA sequences were unavailable. The second file in **Supplementary Data 1** contains phenotypic data (1 for siRNA with phenotype, 0 for siRNA with no phenotype) in the same order as the siRNA sequences, one number per line when providing both active and inactive siRNA sequences. If only active siRNA sequences are provided, GESS generates control siRNA seed sequences by changing nucleotide 1 of each seed to its complement and no phenotypic data file is required. To run a GESS methodology negative control run, siRNA seed sequences of both active and inactive siRNAs can optionally be randomly scrambled by the program. The program requests the user to define the length of seed sequences to analyze (typically 6–8 nucleotides with the default being 7) and the minimal number of seed matches (multiplicity) an siRNA must show toward a target sequence to consider it matching. The program allows selection of the strand(s) of the siRNA that should be used in the analysis. Either strand is considered by default, meaning that a seed sequence derived from either the sense or antisense strand must contain a seed match to the target sequence for the siRNA to be considered matching. Alternatively, the antisense strand only, the sense strand only or both strands (each strand must satisfy the seed-matching parameters) can be analyzed. The program also asks the user to indicate which genome-wide transcript sequence database text file should be used (3′ UTR, 5′ UTR and coding sequence sequence databases for human and mouse are provided as **Supplementary Data 3**). The transcript sequences must be in upper case, one sequence per line with ATGC code (no U). Three different multiple-hypotheses testing correction methods can be selected in the analysis, as below. If a significant outlier (corrected $P < 0.05$) is detected in the primary GESS run, the analysis can be repeated after removing the major outlier, as this approach might enable other, less prominent off-target effects to be detected. The user simply chooses the option to exclude siRNAs matching a sequence and provides a text file containing the outlier sequence.

Statistical analysis of GESS results. Because some of the sequences analyzed contained low seed match event numbers, we calculated the chi squared statistic with correction for continuity (Yates' chi squared statistic), which compensates for low event numbers.

$$\chi^2_{\text{Yates}} = \frac{N(|N_{\text{siPhenMatch}} \times N_{\text{siNoPhen}} - N_{\text{siNoPhenMatch}} \times N_{\text{siPhen}}| - N/2)^2}{N_{\text{siPhen}} \times N_{\text{siNoPhen}} \times N_{\text{siMatch}} \times N_{\text{siNoMatch}}}$$

N is the total number of siRNAs tested in the GESS analysis. N_{siPhen} is the number of siRNAs with phenotype. N_{siNoPhen} is the number of siRNAs with no phenotype. $N_{\text{siPhenMatch}}$ is the number of siRNAs with phenotype with seed matching to tested sequence. $N_{\text{siNoPhenMatch}}$ is the number of siRNAs with no phenotype with seed matching to tested sequence. N_{siMatch} is the number of siRNAs with seed matching to tested sequence. $N_{\text{siNoMatch}}$ is the number of siRNAs with no seed matching to tested sequence. $N_{\text{siPhenNoMatch}}$ is the number of siRNAs with phenotype with no seed matching to tested sequence. $N_{\text{siNoPhenNoMatch}}$ is the number of siRNAs with no phenotype with no seed matching to tested sequence.

The one-tailed probability (*P* value) of the Yates' chi squared statistics was calculated (with a degree of freedom of 1). The chi square was set to zero if the chi squared calculation denominator was null (the Yates' chi square cannot be calculated). The corresponding *P* value is then equal to 1. If any of $N_{\text{siPhenMatch}}$, $N_{\text{siPhenNoMatch}}$, $N_{\text{siNoPhenMatch}}$ or $N_{\text{siNoPhenNoMatch}}$ was less than or equal to 20, the Fisher's exact test two-sided *P* value was determined instead of the Yates's chi squared *P* value. The genomic sequences were ranked from the one with lowest *P* value (rank = 1) to the one with highest *P* value (rank = A, the number of genomic sequences analyzed).

Three multiple-hypotheses testing correction methods were implemented in GESS. The Benjamini and Hochberg false discovery rate correction[16] was used as default as it is considered a good balance between limiting report of false positive and false

doi:10.1038/nmeth.1898

negative off-targeted transcripts. The null hypothesis (there is no difference between the frequency of siPhen and siNoPhen containing a seed match to a given sequence) was rejected if the *P* value calculated above was less than the corrected *P* value threshold ($\alpha \times$ rank of sequence/A), in which α is set as 0.05 by default (more stringent α values can be input by the user) and A is the number of genomic sequences analyzed. The number of sequences passing or failing the test is indicated on each graph. Two additional methods are available for analysis by the user, namely, the Bonferroni[17] and the Bonferroni step-down[18] methods. The corrected *P* value thresholds are (α/A) and (α/(A + 1 − rank of sequence)), respectively. These methods are more stringent than the Benjamini and Hochberg method. Although these methods can be used to limit the rate of false positive off-targeted transcripts identified, in the analysis weaker genuine off-targeted transcripts may be missed as false negatives. Corrected *P*-value thresholds and associated statistical significance status for the three methods are reported in the result file generated by GESS (an example is provided in **Supplementary Data 2**).

Data visualization. The program plots the percentage of siRNAs containing a seed match to a transcript of interest, comparing the siRNAs with phenotype (*y* axis) to those without phenotype (*x* axis). Each genomic sequence is represented by one point on the graph and statistical enrichment of significance is indicated in red. Alternatively, Spotfire DecisionSite was used to generate the graphs. Sequences with significant SME (corrected *P* value < 0.05) were depicted in red, and nonsignificant sequences were depicted in gray. The numbers of significant and nonsignificant outliers are provided in the result graphs generated by GESS (see **Fig. 2**).

siRNA seed sequence enrichment analysis (SSEA). The GESS algorithm was adapted to be applied to siRNA seed sequences as follows. A list of 16,384 7-mer seed sequences was generated and stored as a text file and Excel file in the same format as the transcript sequences database files. These text files were used instead of the genome-wide transcript sequence databases to search for seed presence in the active and inactive siRNAs. All calculations and statistical decisions were performed similarly as for the GESS method. Provided fewer events are expected to be counted as compared to a GESS analysis, the multiple hypothesis testing error correction was restricted to the Benjamini and Hochberg method.

16. Benjamini, Y. & Hochberg, Y. *J. Royal Stat. Soc. B* **57**, 289–300 (1995).
17. Bonferroni, C. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8**, 3–62 (1936).
18. Holm, S. *Scandinavian Journal of Statistics* **6**, 65–70 (1979).