nature | **methods**

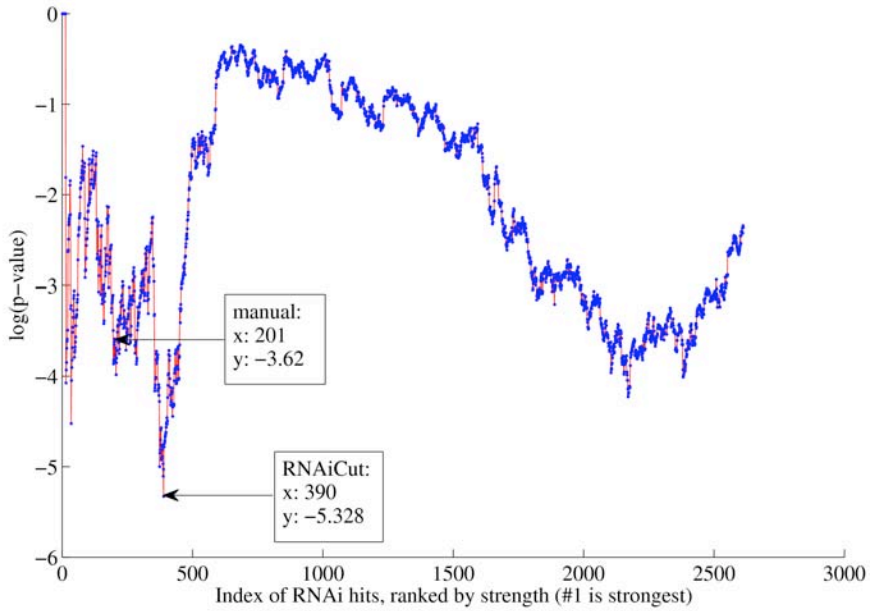
RNAiCut: automated detection of significant genes from functional genomic screens

Irene M Kaplow, Rohit Singh, Adam Friedman, Christopher Bakal, Norbert Perrimon & Bonnie Berger

Supplementary figures and text:

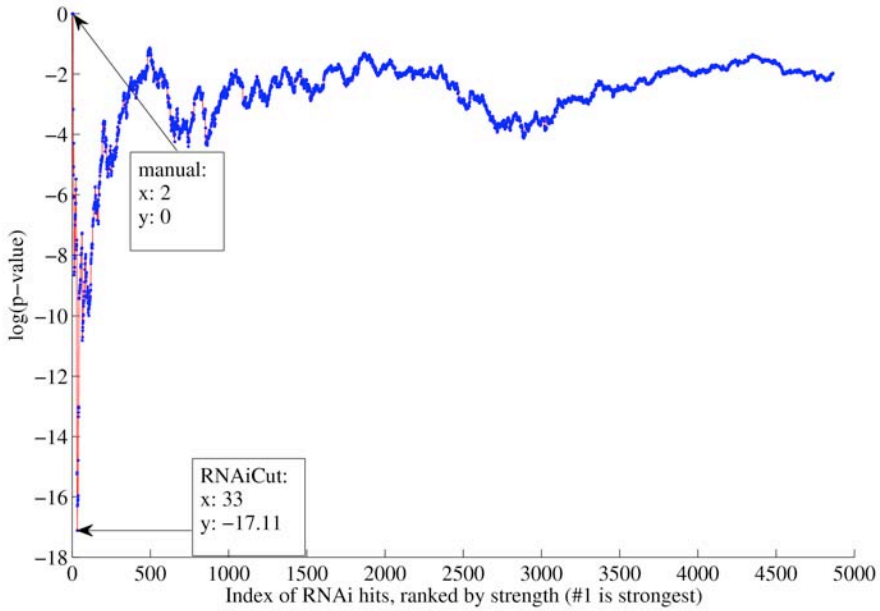
Supplementary Figure 1	Wingless signaling for genes with positive scores
Supplementary Figure 2	Wingless signaling for genes with negative scores
Supplementary Figure 3	Hedgehog signaling for genes with positive scores
Supplementary Figure 4	Hedgehog signaling for genes with negative scores
Supplementary Figure 5	Protein secretion for genes with positive scores
Supplementary Figure 6	Protein secretion for genes with negative scores
Supplementary Figure 7	Cell titer for genes with positive scores
Supplementary Figure 8	Cell titer for genes with negative scores
Supplementary Figure 9	Calcium entry for genes with positive scores
Supplementary Figure 10	Calcium entry for genes with negative scores
Supplementary Figure 11	Local scrambling comparison
Supplementary Figure 12	Hedgehog signaling on multi-species PPI network
Supplementary Table 1	Comparison of manual screener and RNAiCut cutoffs
Supplementary Table 2	GO enrichment for manual screener and RNAiCut cutoffs
Supplementary Table 3	Locations of canonical genes in RNAi screen lists
Supplementary Table 4	Local scrambling results
Supplementary Table 5	Numbers of genes with each gene identifier
Supplementary Results	
Supplementary Methods	

Supplementary Figure 1: Wingless signaling for genes with positive scores



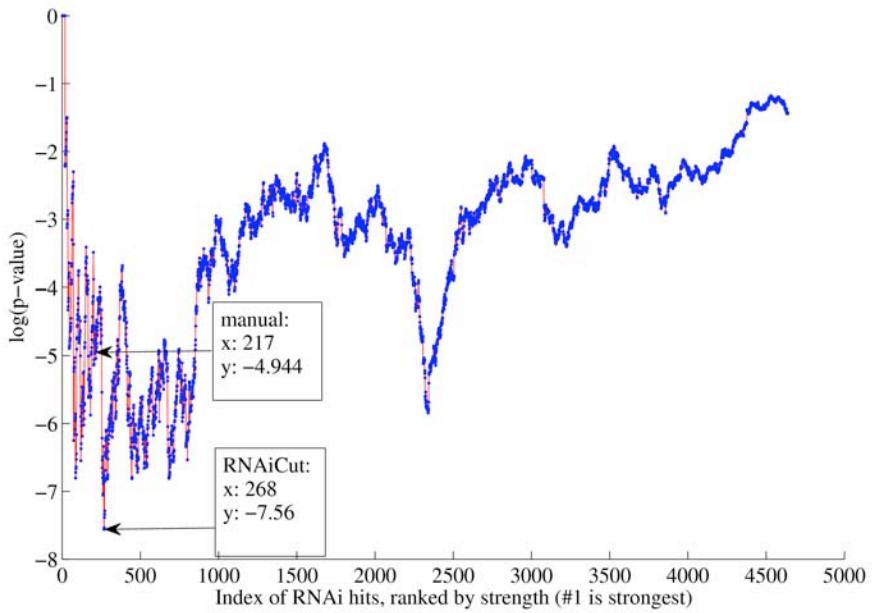
RNAiCut Results for genes with positive Z-scores for wingless signaling screen in *D. melanogaster*.[1] Genes are ordered on the x-axis from left to right based on the decreasing magnitude of Z-scores from the RNAi screen. The y-axis denotes the p-value, as a function of k, of finding a random PPI subnetwork as well-connected as the one containing the k highest-scoring genes from the RNAi screen. These results are based on the *D. melanogaster* PPI network.

Supplementary Figure 2: Wingless signaling for genes with negative scores



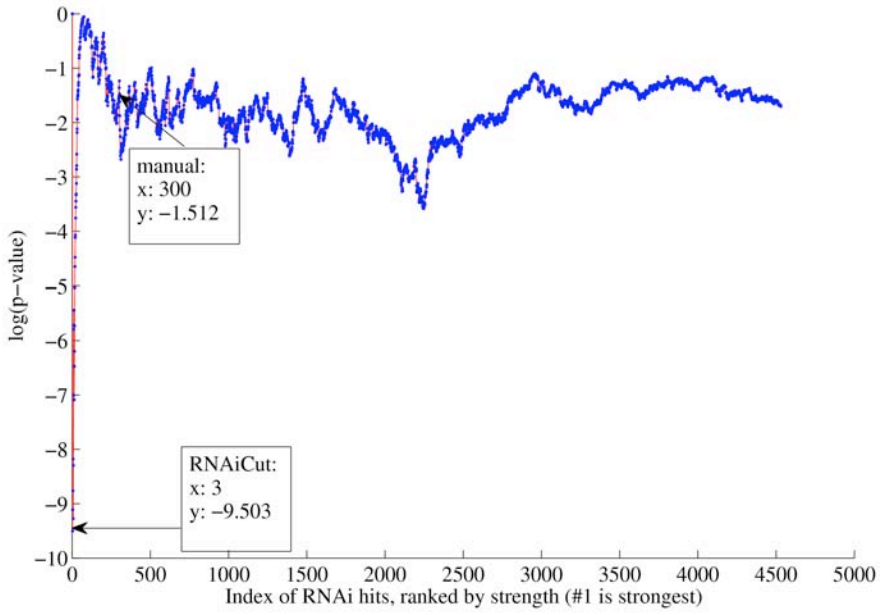
RNAiCut Results for genes with negative Z-scores for wingless signaling screen in *D. melanogaster*.[1] Genes are ordered on the x-axis from left to right based on the decreasing magnitude of Z-scores from the RNAi screen. The y-axis denotes the p-value, as a function of k, of finding a random PPI subnetwork as well-connected as the one containing the k highest-scoring genes from the RNAi screen. These results are based on the *D. melanogaster* PPI network.

Supplementary Figure 3: Hedgehog signaling for genes with positive scores



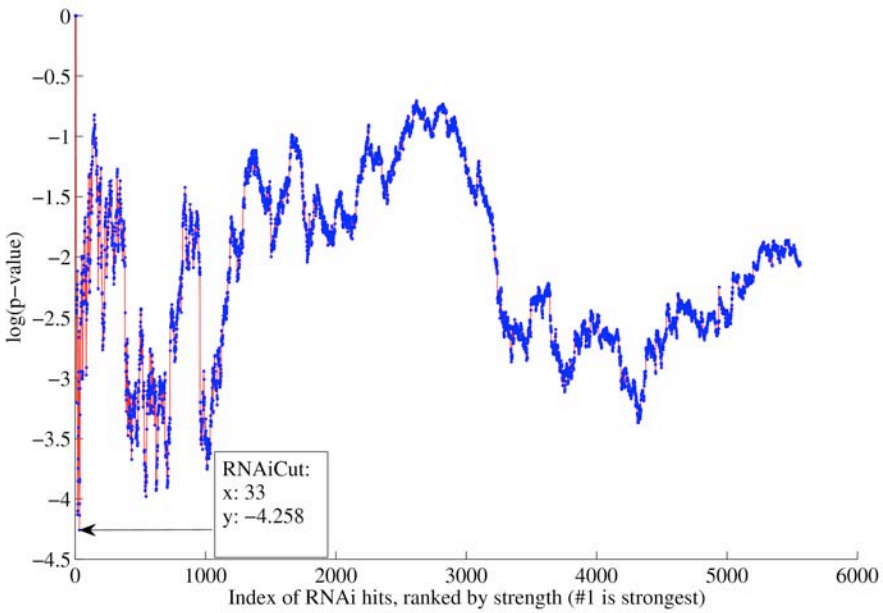
RNAiCut Results for genes with positive Z-scores for hedgehog signaling screen in *D. melanogaster*.[2] Genes are ordered on the x-axis from left to right based on the decreasing magnitude of Z-scores from the RNAi screen. The y-axis denotes the p-value, as a function of k, of finding a random PPI subnetwork as well-connected as the one containing the k highest-scoring genes from the RNAi screen. These results are based on the *D. melanogaster* PPI network.

Supplementary Figure 4: Hedgehog signaling for genes with negative scores



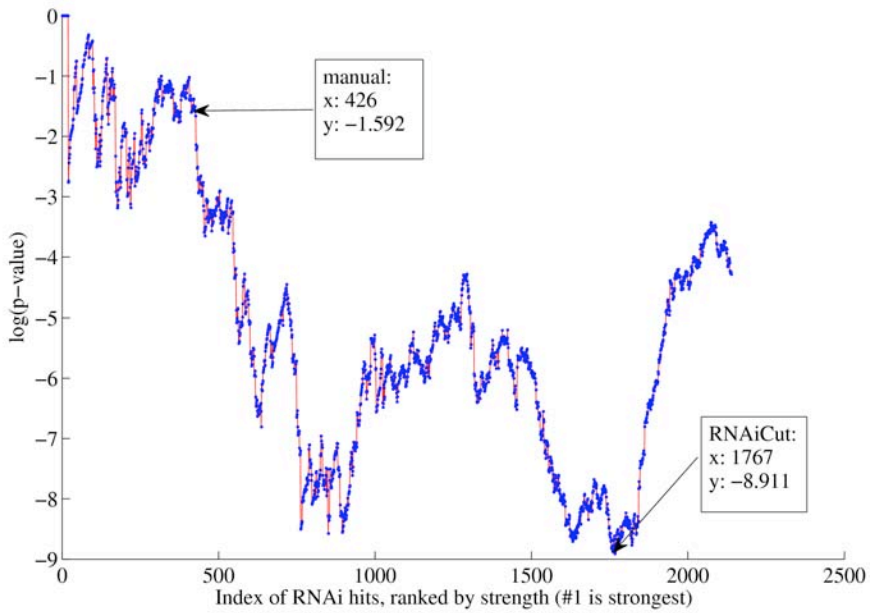
RNAiCut Results for genes with negative Z-scores for hedgehog signaling screen in *D. melanogaster*.[2] Genes are ordered on the x-axis from left to right based on the decreasing magnitude of Z-scores from the RNAi screen. The y-axis denotes the p-value, as a function of k, of finding a random PPI subnetwork as well-connected as the one containing the k highest-scoring genes from the RNAi screen. These results are based on the *D. melanogaster* PPI network.

Supplementary Figure 5: Protein secretion for genes with positive scores



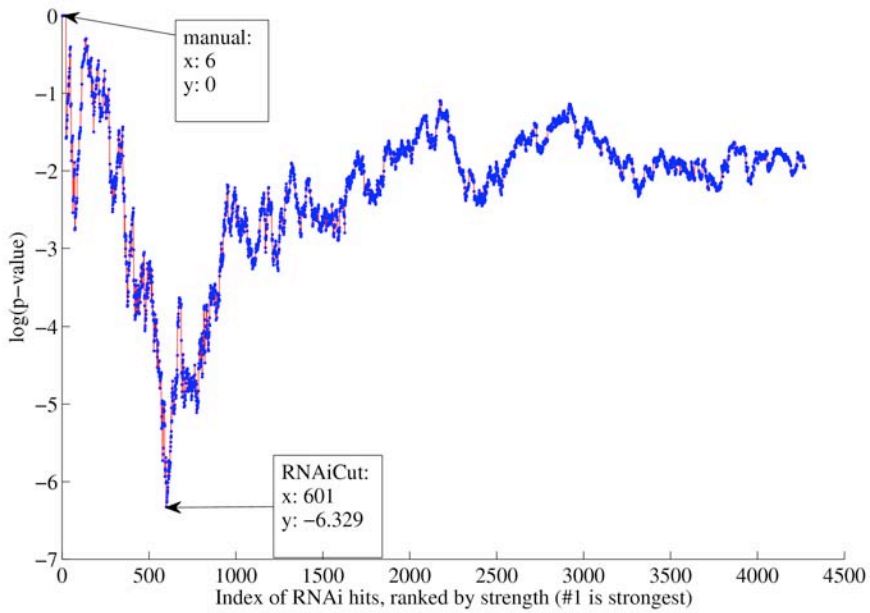
RNAiCut Results for genes with positive Z-scores for protein secretion screen in *D. melanogaster*.[3] Genes are ordered on the x-axis from left to right based on the decreasing magnitude of Z-scores from the RNAi screen. The y-axis denotes the p-value, as a function of k, of finding a random PPI subnetwork as well-connected as the one containing the k highest-scoring genes from the RNAi screen. These results are based on the *D. melanogaster* PPI network.

Supplementary Figure 6: Protein secretion for genes with negative scores



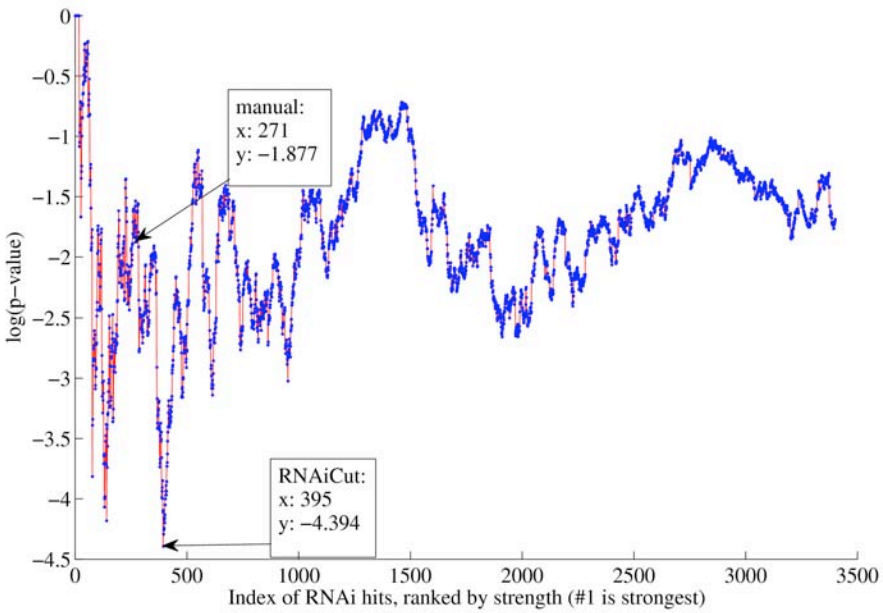
RNAiCut Results for genes with negative Z-scores for protein secretion screen in *D. melanogaster*.[3] Genes are ordered on the x-axis from left to right based on the decreasing magnitude of Z-scores from the RNAi screen. The y-axis denotes the p-value, as a function of k, of finding a random PPI subnetwork as well-connected as the one containing the k highest-scoring genes from the RNAi screen. These results are based on the *D. melanogaster* PPI network.

Supplementary Figure 7: Cell titer for genes with positive scores



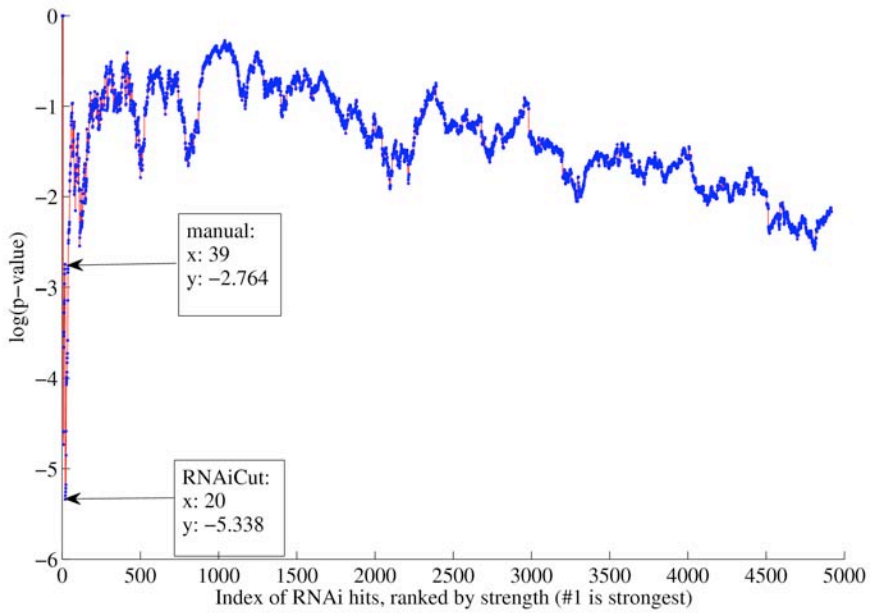
RNAiCut Results for genes with positive Z-scores for cell titer screen in *D. melanogaster*.[4] Genes are ordered on the x-axis from left to right based on the decreasing magnitude of Z-scores from the RNAi screen. The y-axis denotes the p-value, as a function of k, of finding a random PPI subnetwork as well-connected as the one containing the k highest-scoring genes from the RNAi screen. These results are based on the *D. melanogaster* PPI network.

Supplementary Figure 8: Cell titer for genes with negative scores



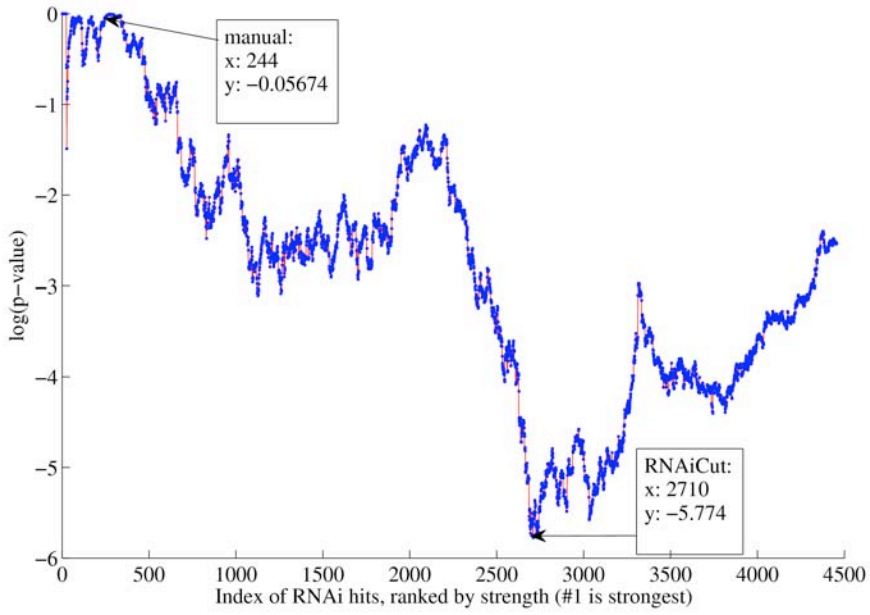
RNAiCut Results for genes with negative Z-scores for cell titer screen in *D. melanogaster*.[4] Genes are ordered on the x-axis from left to right based on the decreasing magnitude of Z-scores from the RNAi screen. The y-axis denotes the p-value, as a function of k, of finding a random PPI subnetwork as well-connected as the one containing the k highest-scoring genes from the RNAi screen. These results are based on the *D. melanogaster* PPI network.

Supplementary Figure 9: Calcium entry for genes with positive scores



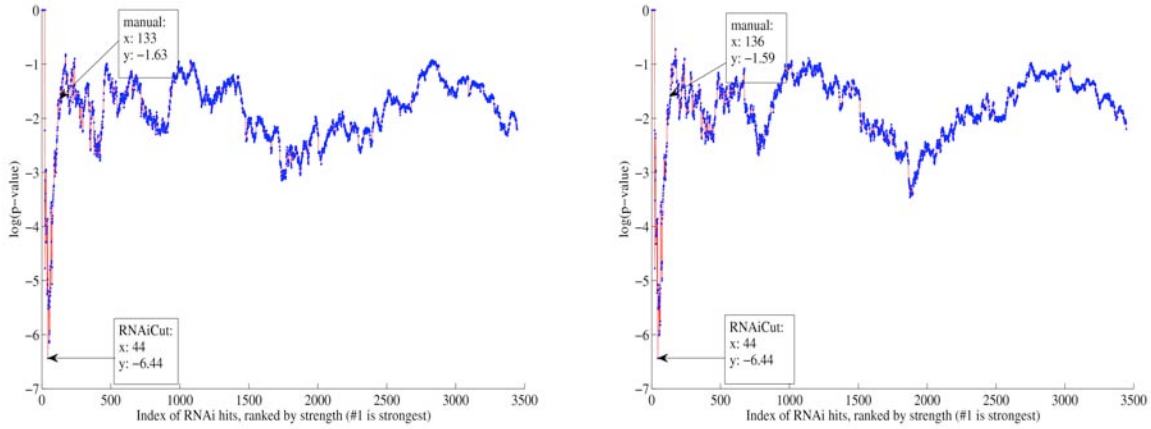
RNAiCut Results for genes with positive Z-scores for calcium entry screen in *D. melanogaster*.[5] Genes are ordered on the x-axis from left to right based on the decreasing magnitude of Z-scores from the RNAi screen. The y-axis denotes the p-value, as a function of k, of finding a random PPI subnetwork as well-connected as the one containing the k highest-scoring genes from the RNAi screen. These results are based on the *D. melanogaster* PPI network.

Supplementary Figure 10: Calcium entry for genes with negative scores



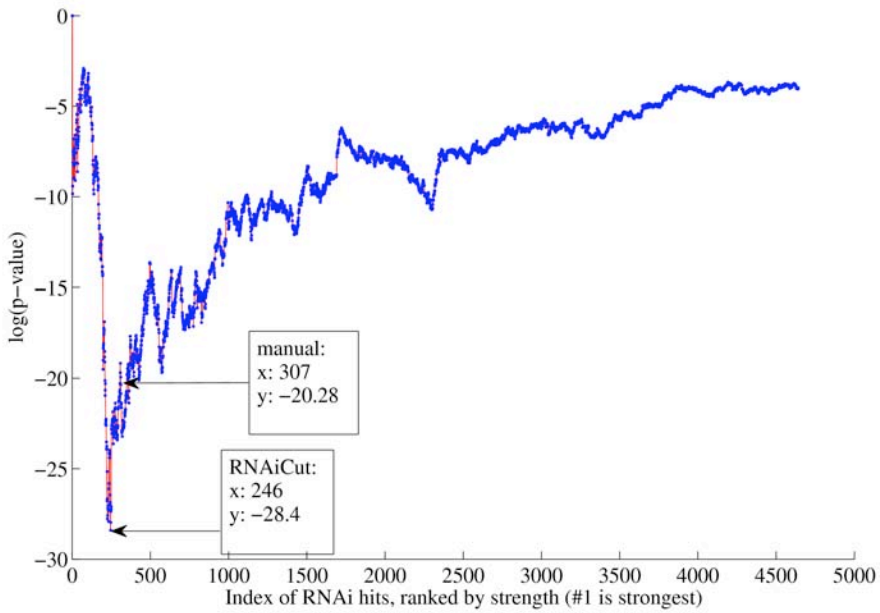
RNAiCut Results for genes with negative Z-scores for calcium entry screen in *D. melanogaster*.[5] Genes are ordered on the x-axis from left to right based on the decreasing magnitude of Z-scores from the RNAi screen. The y-axis denotes the p-value, as a function of k, of finding a random PPI subnetwork as well-connected as the one containing the k highest-scoring genes from the RNAi screen. These results are based on the *D. melanogaster* PPI network.

Supplementary Figure 11: Local scrambling comparison



RNAiCut Results for genes with positive Z-scores (left) and for locally scrambled genes with positive Z-scores (right) for insulin-triggered MAPK pathway screen in *D. melanogaster*.[6] Genes are ordered on the x-axis from left to right based on the decreasing magnitude of z-scores from the RNAi screen. The y-axis denotes the p-value, as a function of k, of finding a random PPI subnetwork as well-connected as the one containing the k highest-scoring genes from the RNAi screen. These results are based on the *D. melanogaster* PPI network.

Supplementary Figure 12: Hedgehog signaling on multi-species PPI network



RNAiCut Results for genes with negative Z-scores for hedgehog signaling screen in *D. melanogaster*.[1] Genes are ordered on the x-axis from left to right based on the decreasing magnitude of z-scores from the RNAi screen. The y-axis denotes the p-value, as a function of k, of finding a random PPI subnetwork as well-connected as the one containing the k highest-scoring genes from the RNAi screen. These results are based on the multi-species PPI network.

Supplementary Tables

Supplementary Table 1: Locations of canonical genes in RNAi screen lists

RNAi scores	Canonical Gene Data	Insulin	Wg signaling	Hh signaling
Negative Scores	Number of canonical genes in screen	21	19	8
	Percentage of canonical genes in top 1000 genes	71%	21%	50%
	Percentage of canonical genes not in top 1000 genes	29%	79%	50%
Positive Scores	Number of canonical genes in screen	11	8	9
	Percentage of canonical genes in top 1000 genes	30%	25%	44%
	Percentage of canonical genes not in top 1000 genes	70%	75%	56%

This table compares the percentages of canonical genes in screens ranked in the top 1,000 genes to the percentages of canonical genes ranked after the 1,000th gene for signaling screens and for genes with negative and positive scores. This table also shows the number of canonical genes in each signaling screen.

Supplementary Table 2: Comparison of manual screener and RNAiCut cutoffs

RNAi scores	Method	Before/ After	Insulin	Wg signaling	Hh signaling	Protein secretion	Cell titer	Calcium entry
Negative Scores	Manual Screener	Cutoff	-1.5	-2.0	-2.0	-1.5	-2.2	-3.0
		Number of genes before	329	5	441	592	341	306
		Number of genes after	4,530	9,397	7,800	3,268	5,684	8,010
	RNAiCut	Cutoff	-2.38	-1.39	-5.09	-0.09	-1.72	-0.47
		Number of genes before	139	52	3	3,115	510	4,961
		Number of genes after	4,720	9,350	8,238	745	5,515	3,355
Positive Scores	Manual Screener	Cutoff	1.5	2.0	3.0	NA	3.0	3.0
		Number of genes before	225	300	400	NA	16	68
		Number of genes after	6,285	4,293	8,292	NA	8,120	9,674
	RNAiCut	Cutoff	2.35	1.26	2.76	1.42	1.25	3.34
		Number of genes before	67	631	484	66	1,141	33
		Number of genes after	6,443	3,962	8,208	10,672	6,995	9,709

This table shows the manual screener cutoffs and the RNAiCut cutoffs for each screen and for genes with negative and positive scores. It also shows numbers of genes before and after the manual and RNAiCut thresholds.

Supplementary Table 3: Numbers of genes with each gene identifier

RNAi scores	Numbers of Genes	Insulin	Wg signaling	Hh signaling	Protein secretion	Cell titer	Calcium entry
Negative Scores	Number of genes with DRSC identifiers	N/A	14,963	13,398	6,794	9,753	14,403
	Number of genes with CG numbers	5,204*	9,895	8,681	4,129	6,381	8,786
	Number of genes with internal identifiers	4,859	9,402	8,241	3,860	6,025	8,316
Positive Scores	Number of genes with DRSC identifiers	N/A	6,995	14,256	18,059	13,661	16,628
	Number of genes with CG numbers	6,963*	4,884	9,166	11,270	8,584	10,258
	Number of genes with internal identifiers	6,510	4,593	8,692	10,738	8,136	9,742

*Insulin genes originally had official gene symbol identifiers, and those identifiers were converted directly to our internal identifiers.

This table compares the numbers of genes before and after each gene identifier conversion for each screen and for genes with negative and positive scores. The difference between the number of genes with each identifier is the number of gene names lost in the conversion.

Supplementary Table 4: GO enrichment for manual screener and RNAiCut cutoffs

RNAi scores	Method	Before/After	Insulin	Wg signaling	Hh signaling
Negative Scores	Manual Screener	Cutoff	-1.5	-2.0	-2.0
		Screen/GO enrichment before	111	0	5
		Screen/GO enrichment after	0	3	3
	RNAiCut	Cutoff	-2.38	-1.39	-5.09
		Screen/GO enrichment before	113	69	11
		Screen/GO enrichment after	1	3	5
Positive Scores	Manual Screener	Cutoff	1.5	2.0	3.0
		Screen/GO enrichment before	17	11	0
		Screen/GO enrichment after	3	10	6
	RNAiCut	Cutoff	2.35	1.26	2.76
		Screen/GO enrichment before	18	9	0
		Screen/GO enrichment after	3	5	8

This table shows the manual screener and RNAiCut z-score cutoffs for RNAi screens in *D. melanogaster* for different pathways (in columns) for genes with negative (top of table) and positive (bottom of table) scores. It also gives the number of enriched Gene Ontology (GO) functions relevant to each screen, before and after the screener and RNAiCut thresholds (in rows).

Supplementary Table 5: Local scrambling results

RNAi scores	Local Scrambling Trial	Insulin	Wg signaling	Hh signaling
Negative Scores	Original	108	33	3
	1	108	34	3
	2	110	39	3
	3	108	32	6
	4	108	41	3
	5	106	42	6
	6	106	40	3
	7	106	35	6
	8	106	33	6
	9	100	39	3
	10	109	34	3
Positive Scores	Original	44	390	268
	1	44	407	267
	2	53	384	267
	3	55	404	268
	4	52	397	269
	5	52	387	263
	6	51	398	266
	7	55	389	269
	8	44	386	262
	9	53	398	270
	10	44	415	269

This table shows the locations of the global minimums on RNAiCut plots for signaling screens and genes with negative and positive scores on the *D. melanogaster* PPI network after each local scrambling trial. “Original” is the location of the original (no scrambling) global minimum on the plot.

Supplementary Results:

Interpreting the Plots

Low p-values correspond to higher statistical significance, so a low y-coordinate at a specific x-coordinate on our plot would indicate that the subgraph of genes through the rank of that x-coordinate is highly connected. Our plots are roughly V-shaped, and this matches our intuition of how RNAi hit-strength relates to PPI connectivity. For low values of k, the p-value is high because, since there are very few nodes in the subgraph, the subgraph has few edges. As k increases, the p-value decreases, indicating higher connectivity in the PPI subgraph. We think that this occurs because additional nodes added to the subgraph correspond to genes likely to be involved in the pathway or process. Since PPI connectivity correlates with function, this is reflected in lower p-values. For sufficiently high k, the genes beyond index k do not play a substantial role in the pathway or process, so the p-value increases. Thus, our plots have a major dip, and the global minimum is the cutoff, which is our estimate of the index of the last gene potentially involved in our pathway or process.

Sometimes, the plots have multiple dips. This is likely the result of having two or more sets of highly inter-connected genes within the screen. When this occurs, we choose the global minimum, meaning the minimum of the deepest dip, as our cutoff. Always choosing the lowest point on the plot as our cutoff enables our method to be fully automated. However, because our result is graphed, these multiple dips are visible and can be manually analyzed by the researcher to understand the reason for these multiple sets of highly connected genes, which might have biological significance.

Some genes from the RNAi screens are not in the fly PPI network. When we count the number of genes before the cutoff, we count the number of genes ranked higher than the gene at the global minimum on the plot, and we include genes that are not in the PPI network. Therefore, the number of genes before the cutoff (Supplementary Table 2) is greater than the x-coordinate at the global minimum (Supplementary Figures 1-12).

Comparison of RNAiCut versus Manual Cutoffs

We found the cutoffs from the manual screener and from RNAiCut. For each screen and for genes with positive and negative scores, after we found the cutoff on the plot, we identified the gene at the cutoff and found that gene in our original ranked list. The number of genes before the cutoff is the number of genes before (and including) the gene at the cutoff in the original list (Supplementary Table 2).

GO enrichment

We found the Gene Ontology (GO)[7] enrichment for each signaling screen (Supplementary Table 4). The numbers in the table are the numbers of enriched pathway-relevant functions; pathway-relevant functions were determined by using DAVID[8, 9] to find GO terms enriched for the canonical genes in the pathway (see Supplementary Methods). Thus more pathway-relevant GO terms associated with the hit list (before the cutoff) suggests identification of more true positives. A later cutoff does not necessarily lead to higher enrichment before the cutoff. If moving the cutoff later will add genes that are not associated with GO terms relevant to the pathway, then moving the cutoff later will decrease the enrichment because the added genes will decrease the statistical significance of the pathway-relevant GO terms. This means that, when our cutoff is later than the manual cutoff and our enrichment is as good or better than the manual enrichment, we have identified genes involved in the pathway or process that the manual screener left out. For example, we can justify our later cutoff for the wingless negative screen because the GO enrichment before our cutoff is substantially better than the GO enrichment before the manual cutoff, meaning that the manual cutoff likely resulted in more false negatives for real pathway modulators. In fact, our cutoff includes two canonical genes for wingless signaling, *arm* and *wg*, that the manual cutoff leaves out. The enrichment charts from DAVID[7] are available at the "Link to Enrichment Results Tables" on the RNAiCut website, http://rnaicut.csail.mit.edu.

Benefit of Using PPI Network

Screener-determined thresholds depend on the availability of prior information about the pathway being studied and the subjective view of the screener. Note that the z-score chosen by the manual screener varies substantially from screen to screen (Supplementary Table 2); because RNAi screens are noisy, the same z-score cannot be used for every screen. Thus, a naive, fixed cutoff strategy would not be successful.

Additionally, rankings in RNAi screens can be substantially inaccurate. The canonical genes in signaling screens are not generally found near the top of the ranked list; in fact, they are often scattered throughout the list (Supplementary Table 1). The fact that, for every signaling screen we tested other than insulin positive, at least half of the canonical genes are not found near the top of the ranked list demonstrates that ranking does not always accurately show the significance of a gene (assuming that the "canonical" genes are the most involved in a pathway). This means that additional information should be used to improve determination of the threshold. RNAiCut uses PPI data results in addition to ranking to find a threshold customized to the dataset, without any need of prior pathway knowledge or subjective decisions.

Robustness Determination

To evaluate the influence of noise in the RNAi observations on the RNAiCut results, we generated multiple randomized locally scrambled datasets by randomly re-ordering genes whose z-scores were within a +/- 0.05 range. We did this for all three signaling screens and for genes with negative and positive scores. The effect of this is to introduce localized scrambling in the overall ordering (by z-score) of RNAi hits. We analyzed this locally scrambled dataset with RNAiCut on the fly PPI network. An example is shown above, comparing the original plot with the one corresponding to the locally scrambled dataset. As can be seen, the broad characteristics of the two plots, including the approximate position of the global minimum, are similar (Supplementary Figure 13). In fact, the global minimum after local scrambling was always within 25 out of 2,500 - 5,000 genes of the original global minimum (Supplementary Table 5). This

consistency shows that the results of RNAiCut are resistant to some noise in the RNAi observations.

Significance of RNAiCut's Success

RNAiCut's success demonstrates that genes with related functions are highly connected in PPI networks. RNAiCut thus verifies the key biological observation driving our approach, that PPI data provides orthogonal biological information to functional genomic data, with both datasets identifying well-connected genes regulating a single biological process.

By using PPI data, the RNAiCut tool provides an unbiased, automated approach to identifying pathway-relevant hits in functional genomic screens, and it can be applied to any specific-function screen using RNAi or cDNA reagents in organisms with available connectivity data. In contrast, a manual approach to estimating a threshold requires some knowledge of the relevant genes beforehand; it is a subjective process that uses only z-score-based RNAi screen gene rank. RNAiCut is therefore especially valuable for choosing genes for further analysis from screens for which very little *a priori* knowledge exists for the relevant pathway or biological process. As more PPI data becomes available, RNAiCut will likely become more accurate.

Potential Limitations

Resolving gene synonyms

A common – and difficult – problem when integrating multiple datasets is resolving gene synonyms. Not every gene has a name under every type of gene identifier, so genes are often lost when translating from one set of gene identifiers to another. In our study, fly dsRNA names were first mapped to CG numbers or FlyBase identifiers, and these were then mapped to our internal identifiers. Both the steps (dsRNA -> CG numbers and CG numbers -> internal identifiers) lose some information. Typically, the information loss is not substantial, and the majority of the genes always remain.

However, in this particular case, the information loss was sometimes substantial, especially the information loss in the dsRNA identifiers -> CG numbers conversion (Supplementary Table 3). As both RNAi and PPI datasets become more mature, we expect that such information loss will diminish.

RNAi screens for general "housekeeping" processes

RNAiCut performs well when pathway-relevant hits from the RNAi screen all perform a specific function. While this may be true in many cases (e.g., a screen to identify genes in a specific pathway), in some cases the set of genes involved in a pathway or process may span many distinct sets of functions. In such cases, we expect there to be multiple dips in the graph, corresponding to various sets of functions.

The protein secretion negative and calcium entry negative screens shown previously had two dips each and especially late global minimums (Supplementary Figures 8 and 12). In each case, the second dip contains the global minimum. We think that these screens, unlike the signaling screens, are are not identifying genes involved in one specific pathway, but rather genes in a general cell biological process that influences and is influenced by multiple pathways at various times. Thus, there might be multiple groups of closely connected genes within these screens. The two dips that we see could be dips for different sets of closely connected genes. Therefore, our method may be less effective when determining a cutoff for a housekeeping function screen than it is when determining a cutoff for a screen for a specific pathway.

Sparseness of PPI network

The *D. melanogaster* PPI network is continuing to be updated, so some edges will likely be added in the near future. With the current fly PPI network, we obtained reasonable cutoffs for most of our screens. For hedgehog signaling negative, however, our cutoff for the fly PPI network included only 3 genes and corresponded to a z-score of -5.09, which is substantially greater in magnitude than we would expect (Supplementary Figure 6). When looking at the numbers of degrees and edges of these genes in the fly

PPI network, we found that the first three genes have twelve, twenty, and eleven degrees but share two out of three possible edges. The probability of this occurring is especially low, which is why the global minimum occurs at this point.

Since having only three pathway-relevant genes does not seem reasonable, we ran RNAiCut for hedgehog signaling negative on a multi-species network that included PPI connections from human and worm. To create this network, *Drosophila* PPI were supplemented with interologs from human and worm PPI data mapped to *Drosophila* functional orthologs. The mapping between the genes across various species was computed by a global alignment of the respective species-wide PPI networks. This results in an orthology mapping that incorporates both sequence similarity and similarity in the network structure.[10]

With the multi-species network, we found a new cutoff, -2.20, which is close to the manual cutoff, -2.0 (Supplementary Figure 14). We also found that, when using the multi-species PPI network, RNAiCut includes two canonical genes for hedgehog signaling, hh and cos, that it leaves out when using the fly PPI network. cos has a greater degree in the multi-species PPI network than it has in the fly PPI network (11 versus 4). We think that, because the fly PPI network may still be missing connections, as the fly PPI network is updated, the cutoff for hedgehog signaling on the fly PPI network for genes with negative scores will improve. We have provided this multi-species network on our website, http://rnaicut.csail.mit.edu, so that users can select this network if they obtain unreasonable results using the fly PPI network.

Supplementary Methods

Determining Cutoffs

RNAiCut determines pathway-relevant genes from functional genomic data by introducing the use of the connectivity of subgraphs of protein-protein interaction (PPI) networks. We are not changing the z-scores of individual hits; instead, we are using connectivity of a PPI network to find a z-score cutoff that indicates which genes' relationship to the pathway or process should be subject to further study. The PPI network was constructed with *D. melanogaster* PPI data from DIP and Biogrid.[11, 12] These networks have, as nodes, *D. melanogaster* genes and, as edges, the PPIs between them. In our analysis, we eliminated genes whose corresponding dsRNAs have >10 off-targets (defined by a 19 nt homology stretch), as this filtering substantially reduces off-target effects.[13, 14, 15]

RNAiCut first sorts the scores (either raw assay or z-scores) from a functional genomic experiment in descending order of strength (i.e., the strongest hit is ranked first) after separating genes with positive scores from genes with negative scores. It then eliminates repeats of genes, leaving in only the gene with greatest (or least for negative) score. For each of the first k nodes (k = 1, 2, 3,…) in a set, we asked the question: “compared to k randomly chosen nodes, how much more inter-connected is the subgraph formed by these k nodes in the PPI graph?” We use the number of edges in a subgraph as the measure of its inter-connectedness. This measure can be computed more efficiently than another popular measure of inter-connectedness, mean graph path length.

We then compute the number of edges in the PPI subgraph induced by the first k nodes and apply a theoretical model to quickly yet reliably approximate the p-value of obtaining this number by chance. Computing the p-value of obtaining the number of edges in a subgraph induced by the top k nodes is typically done through simulations that generate random samples by repeated, random re-wiring of the edges in the network and then compute the p-value of obtaining the number of edges in the subgraph induced by the k nodes. Unfortunately, these simulations can be time-

consuming and cumbersome, especially for genome-scale screens. We therefore applied an algorithm that approximates the probability distribution of edges in a subgraph with a Poisson distribution, as described by Pradines *et al*.[16]

These p-value computations (one for each value of k) are summarized as a plot where the X-axis consists of the RNAi hits ordered in decreasing order of significance (i.e., k) and the Y-axis, the natural log of the p-value of observing a subgraph of at least the size (i.e., edge count) of the one induced by the k highest scoring nodes (Supplementary Figures 1-12). (When the p-value is 0, we plot 0 on the Y-axis; this occurs for low values of k, when the subgraph does not yet have edges.) This plot is typically V-shaped, with a clear global minimum. We took the global minimum in the graph as the cutoff value, that is, all the genes to the left of and including the global minimum were classified as hits relevant to the pathway or process. We also classify genes not in the PPI network that are ranked higher than our global minimum according to z-score as pathway/process-relevant hits.

Quantifying Relevant Enrichment

We use the Database for Annotation, Visualization and Integrated Discovery (DAVID)'s functional annotation chart to determine which functions were enriched in each signaling screen (NIAID/NIH).[8, 9] We first use DAVID's functional annotation chart to find the enriched Gene Ontology (GO) functions for the canonical genes in each signaling screen. We choose enriched terms with p-value <0.05 as the pathway-relevant functions. Then, we find the enriched GO functions for the genes before the cutoff and for the genes after the cutoff. We count the number of significantly enriched pathway-relevant functions. We determine which functions are significant (p-value < 0.05) using the Benjamini-Hochberg FDR procedure for adjusting for multiple hypotheses, with n equal to the number of pathway-relevant functions for the screen. Finding substantially more pathway-relevant functions before the threshold than after the threshold suggests that the threshold is accurate.

References:

1. Dasgupta, R., Kaykas, A., Moon, R. T. and Perrimon, N. *Science* 308 826-832 (2005).
2. Nybakken, K., Vokes, S.A., Lin, T. Y., McMahon, A. P., Perrimon, N. *Nature Genetics* 37 12 1323-1332 (2005).
3. Bard, F. et al. *Nature* 439 (7076), (2006).
4. Boutros, M. et al. *Science* 303 (5659), (2004).
5. Vig, M. et al. *Science* 312 (5777), (2006).
6. Friedman, A. and Perrimon, N. *Nature* 444 (7116) (2006).
7. The Gene Ontology Consortium *Nature Genetics* 25(1) 25-29 (2000).
8. Dennis, G. Jr. et al. *Genome Biology* 4(5) 3 (2003).
9. Huang, D. W., Sherman, B. T., Lempicki, R. A. *Nature Protocols* 4(1) 44-57 (2009).
10. Singh, R., Xu, J., and Berger, B., Proc Natl Acad Sci U S A., *in press* (2008).
11. Salwinski L., et al. *Nucleic Acids Research* 32 D449-451 (2004).
12. Stark C, et al. *Nucleic Acids Research* 34 D535-539 (2006).
13. Kulkarni, M. M., et al. *Nature Methods* 3 833-838 (2006).
14. Perrimon, N. and Mathey-Prevot, B. *FLY* 1 1-5 (2007).
15. Ramadan, N., Flockhart, I., Booker, M., Perrimon, N. and Mathey-Prevot, B. *Nature Protocols* 2 2245-2264 (2007).
16. Pradines, J., Dancik, V., Ruttenberg, A., and Farutin, V. Lecture Notes in Bioinformatics, 4453, 296 (2007).