Guilt by Association in Human RNAi Screens

Orland Gonzalez[1], Georg Malterer[2], Jürgen Haas[2], Ralf Zimmer[1]

August 16, 2011

[1]Institut für Informatik, Ludwig-Maximilians-Universität München, 80333 Munich, Germany. [2]Max von Pettenkofer Institut, Ludwig-Maximilians-Universität München, 80336 Munich, Germany.

Abstract

The development of RNAi libraries aimed at targeting the complete genomes of a number of organisms has enabled the function of genes to be interrogated at the genome-scale. Indeed, RNAi screens have proven to be effective at identifying genes associated with various biological processes, including cellular differentiation, cancer, signaling pathways, host-pathogen interactions, and many others. Here, we show how to exploit the concept of "guilt by association" in the analysis of these screens. In particular, we demonstrate that it is possible, within limits, to predict genes that lead to strong phenotypes upon perturbation, simply by looking at the phenotypes induced by the genes associated with them (in the context of functional association networks). On top of providing valuable context to the analysis and aiding hit prioritization, this also allows us to extend the coverage of a screen by identifying promising candidates from those that were originally not included. We demonstrate this by identifying and experimentally validating novel host factors of the human pathogen, Varicella Zoster virus (VZV). The cellular factors that we found include several proteasome subunits and genes associated with splicing and nuclear export. In addition, we also identified several host entities with antiviral activities. For example, we provide evidence that DHX9 constitutes part of the effective innate immune response that is mounted against VZV, likely by recognizing *pathogen-associated molecular pattern* (PAMP) elements within the viral DNA, and then inducing the expression of pro-inflammatory cytokines in response to this.

Introduction

Functional studies in mammalian cultured cells were hampered in the past by the lack of a powerful method for perturbing gene activities (Echeverri and Perrimon, 2006). This changed with the discovery of RNA interference (RNAi) and the subsequent development of siRNA libraries aimed at targeting complete genomes for a number of organisms (Birmingham *et al.*, 2009). Indeed, RNAi screens have proven to be effective at identifying genes associated with various biological processes, including cellular differentiation (Zhao and Ding, 2007; Hu *et al.*, 2009; Chia *et al.*, 2010), cancer (Zender *et al.*, 2008; Bauer *et al.*, 2010; Wurdak *et al.*, 2010), signaling (Berns *et al.*, 2004; DasGupta *et al.*, 2005), melanogenesis (Ganesan *et al.*, 2008), and host-pathogen interactions (Brass *et al.*, 2008; Zhou *et al.*, 2008; Li *et al.*, 2009; Tai *et al.*, 2009; Brass *et al.*, 2009). In this study, we demonstrate how the concept of "guilt by association" can be exploited for the analysis of these screens. In particular, we show that it is possible, within limits, to predict that a gene would lead to a strong phenotype upon perturbation, simply by looking at the phenotypes induced by perturbation of the genes associated with it. This observation carries implications that can be helpful for: (1) hit prioritization, i.e., identifying primary hits that are more likely to be confirmed; (2) providing valuable context to the primary hits, which could then be used to formulate hypotheses regarding possible mechanisms; and (3) extending the coverage of a screen by identifying candidates – from those that were originally not included – that are likely to induce strong phenotypes.

Guilt by association is a concept that is widely utilized in the field of systems biology (and biology in general). One of the more successful and prominent examples of this are methods that infer genes potentially involved in particular human diseases (see Oti and Brunner, 2007 for a review). These methods generally exploit the fact that the same or phenotypically similar diseases are often caused by functionally related genes (Brunner and van Driel, 2004; Lage *et al.*, 2007; Wood *et al.*, 2007; Lim *et al.*, 2006) – e.g., genes that belong to the same pathway, protein complex or PPI subnetwork – such that one could discover novel gene-disease associations simply by comparing candidate genes to those already known to be involved. Similarity metrics that have been proposed for this purpose include those based on sequence features (Adie *et al.*, 2005; Lopez-Bigas and Ouzounis, 2004), expression patterns (Bortoluzzi *et al.*, 2003), functional annotations (Perez-Iratxeta *et al.*, 2002; Turner *et al.*, 2003), literature citations (Hristovski *et al.*, 2005), physical interactions (Oti *et al.*, 2006), and combinations thereof (Aerts *et al.*, 2006; Franke *et al.*, 2006).

In our analysis of RNAi screens, we adopted a strategy that is often used by disease gene prediction methods: that of identifying candidates that are "close" to the known disease genes in the context of a functional association network (built from similarity metrics mentioned above). We designed and implemented several methods for this purpose, which we demonstrate by application to several published RNAi datasets from human cell lines. We show that "guilt by association" is indeed a useful concept that can be exploited for the analysis of RNAi datasets. Finally, we extended the coverage of a screen for host factors of the human pathogen, Varicella Zoster virus (VZV), by identifying promising candidates from among those that were originally not included (the screen covered only about 7,000 genes). We experimentally validated some of these predictions.

Results and discussion

Application to a HCV host factor screen

One of the most popular application of RNAi screens in human cell lines is the identification of host factors (HFs) of medically relevant viruses (Cherry, 2009; Mohr *et al.*, 2010). In this type of experiments, the typical phenotype read-out that is associated to each gene is a fold change that quantifies how much infection is inhibited or enhanced upon perturbation of that gene. To date, host factor screens have been performed for HIV (Brass *et al.*, 2008; König *et al.*, 2008; Zhou *et al.*, 2008), HCV (Tai *et al.*, 2009; Li *et al.*, 2009), Influenza (Brass *et al.*, 2009; Karlas *et al.*, 2010; Konig *et al.*, 2010), and West Nile virus (Krishnan *et al.*, 2008). We use the HCV dataset to illustrate our methods.

The main question that we wish to address is whether "guilt by association" is evident in human RNAi screens. Specifically, we ask whether it is possible to predict a strong perturbation effect for a gene simply by looking at the activities of those associated with it. To this end, we adopt the strategy of testing for the unexpectedness of the neighborhood of a gene in the context of a functional association network. Here, we use the STRING database (Jensen *et al.*, 2009) – which scores protein pairs for functional association based on various criteria, including coexpression, phyisical interactions, co-citation, genomic context, functional annotations – to provide the necessary context.

One of the simplest way to test for the enrichment of a gene's neighborhood is to compare how many screen hits it contains in relation to the number expected by chance. This is, in fact, the prototypical strategy for gene set analysis (GSA) (Dinu *et al.*, 2009; Liu *et al.*, 2007). A score (p-value) could very easily be derived from a hypergeometric distribution. Despite the simplicity of the approach however, it suffers from the fact that one needs to define the the set of "hits" beforehand. Moreover, it completely ignores the continuous nature RNAi datasets. Accordingly, we designed and implemented several methods that use the data values directly. For example, in one test that we refer to as *SOS-DN*, we normalize the data by performing a Z-score transformation, such that strongly inhibiting and enhancing perturbations get strongly negative and positive Z-scores, respectively, and then score each neighborhood by estimating the probability that the sum of squares of the data values of the genes that it contains are as high as they are. The different methods are described in detail in the *Methods* section

Hepatitis C virus (HCV) is a positive-sense single-stranded RNA virus that in humans causes its namesake disease, hepatitis C (Senecal and Morelli, 2007). About 3% of the world's population (270-300 million) is chronically infected with HCV. Of this number, about 30% will develop cirrhosis (liver scarring) within 20 years of initial infection, a condition that could then progress to life threatening complications, including liver failure and hepatocellular carcinoma. We use the recent screen for HFs of this medically relevant virus from Tai *et al.* (2009) to demonstrate our methods. Out of the approximately 21,000 knockdowns included in that dataset, we were able to map 17,821 to ENTREZ records. Of these, 13,104 participate in at least one interaction in the STRING-derived network. All subsequent analyses were limited to this subset.

Given that none of the methods that we used considered the phenotype of a gene itself in calculating its score (rather, the methods relied only on the the data values of neighboring genes), then one way of validating their effectiveness is to check that the genes that were ranked highly in each method also have strong data values themselves in the RNAi screen. We summarized the results of this analysis in row one of Figure 1. Each of the three graphs in the row shows the fraction of the top screen hits – defined as the genes with the top 1%, 2% and 5% strongest induced phenotypes, respectively – that is recovered as a function of the fraction of the total number of genes that is considered when the genes are prioritized according to each method. For example, we see that for certain methods (e.g., *SOS-DN-90*), we are already able to recover 25-30% of the top 1% genes with the strongest phenotype even if we consider only less than 2% of the genes. This indicates that at least for some of the hits, the strong phenotypes that they induce are reflected in the genes that they associate with.

A complementary view to analyzing the recovery of the top screen hits is to look directly where the top-ranked genes of each method falls along the main screen with respect to their data values. This is summarized in the middle row of Figure 1. Specifically, each of the graphs show how the top 1%, 2% and 5% top-ranked genes for each method are distributed in the main screen (sorted from left to right by decreasing phenotype). For example, we see in the leftmost graph that for the method *WCX-DN-90*, over 50% of the top 1%-ranked genes are among the 15% of genes that induce the strongest phenotypes. The statistical significance of these distributions, obtained via comparison to all genes in the network using Wilcoxon's Rank-Sum test, are shown in the corresponding plot in row 3 of the figure.

Some are more guilty than others

It is clear from Figure 1 that the correlation between a gene that induces a strong phenotype and those associated with it is stronger in some than in others. For example, we see in the top-left graph that only about 25-35% of the top 1% genes with the strongest phenotypes are highly predictable (i.e., recoverable). The question then is: what happened to the rest? To be sure, the sensitivity and accuracy of the methods that we use contribute to this discrepancy. However, there are also a number of other possible contibuting factors: First, the strong data values of some genes could simply be due to random noise. Indeed, most primary screens consist of only a few replicates (two in the case of HCV). Second, there could be not enough information on the function of some genes; i.e., there are not enough edges in the functional association network. Third, the critical function performed by a gene could be relatively isolated from others. Fourth, the function of the genes associated with a hit could be compensated for by others, which makes them essentially invisible. For example, consider two alternative pathways that converge at a hit gene. And last but not least, it is also possible that some of the strong phenotypes resulted from off-target effects; i.e., RNAi reagents that perturbed the activities of unintended genes due to partial sequence complementarity. Indeed, considering that wide-spread off-target effects have been reported in some studies (Ma *et al.*, 2006; Schultz *et al.*, 2011), even to the point that they dominated the primary hits (Ma *et al.*, 2006), it is very likely that this phenomenon contributes a great deal to the discrepancy.

In addition to the differences between genes, we also found that some screens tended to be more predictable than others (at least given the methods). For example, whereas substantial numbers of screen hits were found to be highly predictable in the HCV host factor (Figure 1) and the stem cell identity factor (Figure 2) screens, there were hardly any in the cancer chemosensitizer locus (Figure S1) and melanogenesis (Figure S2) screens. Indeed, the predictions for these latter two datasets were hardly better than random. Again, possible explanations include random and systematic noise, such as off target-effects, and poor data (technical) quality. In addition, it is also possible that the nature of the phenotypes themselves came into play. For example, there are arguably fewer human modules that are directly involved in melanogenesis than in HCV pathogenesis. That is, the universe of true potential hits is much smaller. A similar situation is probably also true for cancer chemosensitizer loci. The predictability of the different screens using some of the methods is summarized in Figure 3.

The experimental settings (and quality) used in a screen also seems to be a big determinant of predictability. For example, one of the phenotype readouts that were provided in the cellular division factor screen (Screen H) is the total amount of DNA. We can interpret this data as a measure of cellular viability. In this dataset, we found a substantial number of genes with strong phenotypes to be highly predictable (see Figure S3). However, when we analyzed some of the other cellular viability screens, such as screens J and L, we were not able to observe the same pattern (see figures S4 and S5, respectively). Indeed, very few of the genes in these two screens proved to be recoverable. Thus, we have a situation wherein different screens that measure the same biological activity (i.e., viability) exhibit very different levels of modularity (i.e., guilt by association) in their hits. As mentioned above, we believe this to be likely due to the different experimental settings that were used. For example, whereas screen H used DNA content to measure viability, screens J and L used ATP content and expression of a Renilla reporter construct, respectively. Moreover, and probably more importantly, screen J was optimized to detect perturbations that affect cellular division, which is closely-related to cellular viability. In contrast, the two other screens were optimized to detect far more removed phenotypes – i.e., host factors of HCV and members of the WNT pathway – since they were only meant to correct for toxic perturbations in the main screens that they accompanied.

Novel host factors of the Varicella Zoster virus

Using data from an ongoing screen for host factors of the human pathogen VZV, we identified novel candidates from those that were not included in the original screen. Specifically, for each gene g, we identified its neighbors in the functional association network that were included, and then calculated its enrichment score based on these (note that the methods use only the data values of related genes and not that of the reference gene itself). We experimentally tested 57 of the candidate host factors by performing the corresponding knockdowns using siRNA smartpools. The results are summarized in Figure 4. Of those tested, 24 (48%) inhibited viral growth by more than 50% (normalized to negative controls), compared to less than 4% in the main screen.

Several of the VZV-inhibiting perturbations that we found corresponded to subunits of the proteasome complex. In fact, 13 of the 15 components that we tested inhibited infection by more than 50%. Although one may expect such a situation, where perturbation of any subunit results in a strong phenotype, if the complex is truly linked to the virus, we still found this to be highly surprising given that such outcomes are very rarely observed in actual screens. For example, the eIF2 and eIF3 complexes are known to be utilized by HCV during duplication of its RNA genome (reviewed in Fraser and Doudna, 2007). However, in the recent screen for host factors of the virus (Tai *et al.*, 2009), only a very few subunits registered statistically significant effects (which were even moderate at best). In our own screen for host factors of Human Simplex virus 1 (HSV-1) (manuscript in preparation), subunits – of complexes known to be involved in the virus' pathogenesis – that induced strong phenotypes were again almost never the majority (although they could still be overrepresented in the hit list). Accordingly, the fact that perturbation of most proteasome subunits led to strong VZV-inhibiting phenotypes suggests that the virus is strongly dependent on the complex and/or the relevant function is very easy to disturb (i.e., the subunits are not able to effectively compensate for each other). Our results complement other studies that link the proteasome to VZV infection (Stallings *et al.*, 2006; Walters *et al.*, 2008).

The strongest phenotype we observed in our "secondary screen" belonged to perturbation of NXF1. Consistent with this, a scan through the literature revealed that the IE4 protein of VZV has been shown to interact with SR proteins in order to export viral mRNAs through the TAP/NXF1 pathway (Ote *et al.*, 2009). Nuclear export is likely also the mechanism behind the substantial effect of NUP93 (Nucleoporin 93kDa), one of the other candidates that we identified, on the virus.

In contrast to proviral host factors, we identified several genes that, upon perturbation, led to significant enhancement in viral growth. These include DHX9, TXNL4 and EIF3K. Their induced phenotypes indicate that in vivo, they likely perform antiviral activities. In the case of DHX9 (DEAH box protein 9), the gene has recently been reported to recognize CpG-containing microbial DNA in plasmacytoid dendritic cells (pDC) and induce activation of NF-κB through MYD88. Importantly, it was associated with the expression of pro-inflammatory cytokines after HSV-1 infection, and knockdown of the gene was demonstrated to inhibit pDC responses to the virus (Kim *et al.*, 2010). Our results thus strongly suggest that DHX9 is also able to recognize PAMP elements within the VZV DNA and that it constitutes part of the effective innate immune response that is mounted against the virus.

Methods

Data sources

To serve as context for the analysis of the RNAi data, we assembled a network by collecting interactions defined in the STRING database (Jensen *et al.*, 2009). Only those rated with at least a medium level of confidence (combined score$\geq$0.4) were included. All identifiers were mapped to ENTREZ records, using mapping information (Ensembl protein id to Entrez gene id) retrieved from ENSEMBL BioMart. In situations where more than one STRING record could be mapped to a gene pair, the strongest confidence value was assigned. This resulted in a network composed of 13,104 vertices (genes) and 330,523 edges (interactions). In certain parts of our analysis, we also considered higher confidence subnetworks by retaining only interactions with a combined score of at least 0.7 or 0.9.

RNAi datasets from human cell lines were collected directly from the literature. These include screens for genes involved in the WNT pathway (Tang *et al.*, 2008; Major *et al.*, 2008), cellular division (Kittler *et al.*, 2007), HCV (Tai *et al.*, 2009) and VZV pathogenesis, melanogenesis (Ganesan *et al.*, 2008), and stem cell differentiation (Chia *et al.*, 2010). Cell viability screens that accompanied some of these primary screens were included whenever available. As with the interaction network, all RNAi datasets (summarized in Table 1) were mapped to ENTREZ gene records.

Tests for neighborhood enrichment

We designed and implemented several statistical tests for the enrichment of the neighborhood of a gene g. The first, which we designate as *STF-DN*, is based on Stouffer's Z-score test. Specifically, let p_g be the p-value associated with perturbation of a gene g, we converted this to a Z-score z_g according to

$$z_g = \Phi^{-1}(p_g) \tag{1}$$

where $\Phi^{-1}(p)$ is the standard normal inverse cumulative distribution function evaluated at p. From this, we calculated the enrichment score of the neighborhood of a gene

g as

$$p_g^{STF-DN} = \Phi\left(\frac{1}{\sqrt{|N_g|}}\sum_{i\in N_g} z_i\right) \tag{2}$$

where N_g is the first order neighborhood of g; i.e., the set of genes that are directly adjacent to g in the functional association network. The p-value (for the unexpectedness of the neighborhood of g) that is calculated by Equation 2 is based on the fact that the sum of n independent standard normal random variables when normalized by $\sqrt{n}$ is also standard normally distributed.

The second test that we used is a weighted modification of Equation 2. We designate this as *SWT-NS*. For a given gene g, we calculate the enrichment of its neighborhood as

$$p_g^{SWT-NS} = \Phi\left(\frac{1}{\sqrt{\sum_{i\neq g} S_{gi}^2}}\sum_{i\neq g} S_{gi}\cdot z_i\right) \tag{3}$$

where

$$S_{ij} = \frac{|N_i \cap N_j|}{\max(|N_i|,|N_j|)} \tag{4}$$

is a measure of the similarity of the first order neighborhoods of two genes i and j. The motivation behind using Equation 3 is to let genes that have more similar neighborhoods to gene g contribute more to its score.

The third test that we implemented, which we designate as *SOS-DN*, is based on the fact that the sum of the squares of n independent, standard normal random variables is distributed according to a chi-square distribution with n degrees of freedom. For this test, we calculated gene z-scores differently than in the first two. Specifically, if x_g is the phenotype (e.g., fold change) value associated with the perturbation of g, then

$$z_g = \frac{x_g - \mu}{\sigma} \tag{5}$$

where μ and σ are the mean and standard deviation of the population, respectively. The main difference is that in this case, the "significant" genes are in both tails of the distribution (assuming that both positive and negative changes in phenotype are interesting). A Q statistic for the enrichment of the neighborhood of each gene g is calculated as

$$Q_g = \sum_{i\in N(g)} z_i^2 \tag{6}$$

from which a p-value is derived according to

$$\begin{aligned} p_g^{SOS-DN} &= \Pr(Q \leq Q_g) \\ &= 1 - F_{\chi^2_{|N_g|}}(Q_q) \end{aligned} \tag{7}$$

where $F_{\chi^2_{|N_g|}}$ is the chi-square cumulative distribution function with $|N_g|$ degrees of freedom.

In addition to the three tests described above, we also implemented one based on Wilcoxon's rank-sum test (*WCX-DN*). Here, we compared the ranks of the genes in the first order neighborhood of g (i.e., N_g) against the ranks of all the genes that were included in the screen and are in the functional association network (i.e., participate in at least one interaction/edge). Finally, we also implemented one of the best performing variants of the method described in Wang *et al.*, 2009, which was used in a related analysis of *Drosophila* RNAi screens. Specifically, for a given gene g, a score was calculated according to

$$N_g^{NPH-NS} = \frac{\sum_{i\neq g} S_{gi} H_i}{\sum_{i\neq g} S_{gi}} \tag{8}$$

where S_{gi} is calculated as in Equation 4, and H_i is a binary variable that is equal to 1 if gene i is a "hit" in the screen and 0 otherwise. Note that in contrast to the other tests, *NPH-NS* does not output a p-value. Moreover, it requires that some genes be designated as "hits" beforehand. How these hits were defined for each dataset is indicated in Table 1.

Experimental protocols

The siRNA knockdown experiments were performed in quadruplicates using black 96 well clear bottom assay plates (Costar). Briefly, 10,000 MeWo cells in 100 μl RPMI1640 supplemented with 10% FCS, Pen/Strep and Glutamine have been transfected with 5 pmol of siRNA smartPools (Thermo Scientific) using Lipofectamine 2000 (Invitrogen) in a final concentration of 0,4% according to manufacturer's manual. 48 hours after transfection the cells were infected with 100 pfu of rVZV-GFP (Zerboni and Arvin, 2000). 72 hours later, the GFP-signal was measured on a fluorescence reader (Fluostar Optima, BMG labtech). For deconvolution experiments, 5 pmol of single siRNAs were used instead of the smartPools.

References

Adie E. A., Adams R. R., Evans K. L., Porteous D. J., and Pickard B. S. (2005). Speeding disease gene discovery by sequence based candidate prioritization. *BMC bioinformatics*, **6**, 55.

Aerts S., Lambrechts D., Maity S., Van Loo P., Coessens B., De Smet F., Tranchevent L. C., De Moor B., Marynen P., Hassan B., Carmeliet P., and Moreau Y. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnol*, **24**, 537–544.

Bauer J. A., Ye F., Marshall C. B., Lehmann B. D., Pendleton C. S., Shyr Y., Arteaga C. L., and Pietenpol J. A. (2010). RNA interference (RNAi) screening approach identifies agents that enhance paclitaxel activity in breast cancer cells. *Breast Cancer Res.*, **12**, R41.

Berns K., Hijmans E. M., Mullenders J., Brummelkamp T. R., Velds A., Heimerikx M., Kerkhoven R. M., Madiredjo M., Nijkamp W., Weigelt B., Agami R., Ge W., Cavet G., Linsley P. S., Beijersbergen R. L., and Bernards R. (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature*, **428**, 431–437.

Birmingham A., Selfors L. M., Forster T., Wrobel D., Kennedy C. J., Shanks E., Santoyo-Lopez J., Dunican D. J., Long A., Kelleher D., Smith Q., Beijersbergen R. L., Ghazal P., and Shamu C. E. (2009).

Statistical methods for analysis of high-throughput RNA interference screens. *Nature Methods*, **6**, 569–575.

Bortoluzzi S., Romualdi C., Bisognin A., and Danieli G. A. (2003). Disease genes and intracellular protein networks. *Physiological Genomics*, **15**, 223–227.

Brass A., Dykxhoorn D., Benita Y., Yan N., Engelman A., Xavier R., Lieberman J., and Elledge S. (2008). Identification of Host Proteins Required for HIV Infection Through a Functional Genomic Screen. *Science*, **15**, 921–926.

Brass A. L., Huang I. C., Benita Y., John S. P., Krishnan M. N., Feeley E. M., Ryan B. J., Weyer J. L., van der Weyden L., Fikrig E., Adams D. J., Xavier R. J., Farzan M., and Elledge S. J. (2009). The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. *Cell*, **139**, 1243–1254.

Brunner H. G. and van Driel M. A. (2004). From syndrome families to functional genomics. *Nature Reviews*, **5**, 545–551.

Cherry S. (2009). What have RNAi screens taught us about viral-host interactions? *Curr Opin Microbiol*, **12**, 446–452.

Chia N. Y., Chan Y. S., Feng B., Lu X., Orlov Y. L., Moreau D., Kumar P., Yang L., Jiang J., Lau M. S., Huss M., Soh B. S., Kraus P., Li P., Lufkin T., Lim B., Clarke N. D., Bard F., and Ng H. H. (2010). A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature*, **468**, 316–320.

DasGupta R., Kaykas A., Moon R. T., and Perrimon N. (2005). Functional genomic analysis of the Wnt-wingless signaling pathway. *Science*, **308**, 826–833.

Dinu I., Potter J., Mueller T., Liu Q., Adewale A., Jhangri G., Einecke G., Famulski K., Halloran P., and Yasui Y. (2009). Gene-set analysis and reduction. *Brief. Bioinform.*, **10**, 24–34.

Echeverri C. J. and Perrimon N. (2006). High-throughput RNAi screening in cultured cells: a user's guide. *Nat. Rev. Genet.*, **7**, 373–384.

Franke L., van Bakel H., Fokkens L., de Jong E. D., Egmont-Petersen M., and Wijmenga C. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, **78**, 1011–1025.

Fraser C. S. and Doudna J. A. (2007). Structural and mechanistic insights into hepatitis C viral translation initiation. *Nat. Rev. Microbiol.*, **5**, 29–38.

Ganesan A. K., H Ho H., Bodemann B., Petersen S., Aruri J., Koshy S., Richardson Z., Le L. Q., Krasieva T., Roth M. G., Farmer P., and White M. A. (2008). Genome-wide siRNA-based functional genomics of pigmentation identifies novel genes and pathways that impact melanogenesis in human cells. *PLoS Genetics*, **4**, e1000298.

Hristovski D., Peterlin B., Mitchell J. A., and Humphrey S. (2005). Using literature-based discovery to identify disease candidate genes. *Int J Med Inform*, **74**, 289–98.

Hu G., Xu J. K. Q., Leng Y., Orkin S. H., and Elledge S. J. (2009). A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes Dev.*, **23**, 837–848.

Jensen L. J., Kuhn M., Stark M., Chaffron S., Creevey C., Muller J., Doerks T., Julien P., Roth A., Simonovic M., Bork P., and Merring C. (2009). STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–6.

Karlas A., Machuy N., Shin Y., Pleissner K., Artarini A., Heuer D., Becker D., Khalil H., Ogilvie L., Hess S., Mäurer A., Müller E., Wolff T., Rudel T., and Meyer T. (2010). Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature*, **463**, 818–822.

Kim T., Pazhoor S., Bao M., Zhang Z., Hanabuchi S., Facchinetti V., Bover L., Plumas J., Chaperot L., Qin J., and Liu Y. J. (2010). Aspartate-glutamate-alanine-histidine box motif (DEAH)/RNA helicase A helicases sense microbial DNA in human plasmacytoid dendritic cells. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 15181–15186.

Kittler R., Pelletier L., Heninger A. K., Slabicki M., Theis M., Miroslaw L., Poser I., Lawo S., Grabner H., Kozak K., Wagner J., Richter V. S. C., Bowen W., Jackson A. L., Habermann B., Hyman A. A., and Buchholz F. (2007). Genome-scale RNAi profiling of cell division in human tissue culture cells. *Nature Cell Biology*, **9**, 1401–1412.

König R., Zhou Y., Elleder D., Diamond T., Bonamy G., Irelan J., Tu C. C. B., De Jesus P., Lilley C., Seidel S., Opaluch A., Caldwell J., Weitzman M., Kuhen K., Bandyopadhyay S., Ideker T., Miraglia L., Bushman F., young J., and Chanda S. (2008). Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, **135**, 49–60.

Konig R., Stertz S., Zhou Y., Inoue A., Hoffmann H. H., Bhattacharyya S., Alamares J. G., Tscherne D. M., Ortigoza M. B., Liang Y., Gao Q., Andrews S. E., Bandyopadhyay S., De Jesus P., Tu B. P., Pache L., Shih C., Orth A., Bonamy G., Miraglia L., Ideker T., Garcia-Sastre A., Young J. A., Palese P., Shaw M. L., and Chanda S. K. (2010). Human host factors required for influenza virus replication. *Nature*, **463**(7282), 813–817.

Krishnan M. N., Ng A., Sukumaran B., Gilfoy F. D., Uchil P. D., Sultana H., Brass A. L., Adametz R., Tsui M., Qian F., Montgomery R. R., Lev S., Mason P. W., Koski R. A., Elledge S. J., Xavier R. J., Agaisse H., and Fikrig E. (2008). RNA interference screen for human genes associated with West Nile virus infection. *Nature*, **455**(7210), 242–245.

Lage K., Karlberg E. O., Storling Z. M., Olason P. I., Pedersen A. G., Rigina O., Hinsby A. M., Tumer Z., Pociot F., Tommerup N., Moreau Y., and Brunak S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnol.*, **25**, 309–316.

Li Q., Brass A. L., Ng A., Hu Z., Xavier R. J., Liang T. J., and Elledge S. J. (2009). A genome-wide genetic screen for host factors required for hepatitis C virus propagation. *Proc. Natl. Acad. Sci. U.S.A.*, **106**(38), 16410–5.

Lim J., Hao T., Shaw C., Patel A. J., Szabo G., Rual J. F., Fisk C. J., Li N., Smolyar A., Hill D. E., Barabasi A. L., Vidal M., and Zoghbi H. Y. (2006). A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, **125**, 801–814.

Liu Q., Dinu I., Adewale A. J., Potter J. D., and Yasui Y. (2007). Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, **8**, 431.

Lopez-Bigas N. and Ouzounis C. A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.*, **32**, 3108–3114.

Ma Y., Creanga A., Lum L., and Beachy P. A. (2006). Prevalence of off-target effects in Drosophila RNA interference screens. *Nature*, **443**, 359–363.

Major M. B., Roberts B. S., Berndt J. D., Marine S., Anastas J., Chung N., Ferrer M., Yi X., Stoick-Cooper C. L., von Haller P. D., Kategaya L., Chien A., Angers S., MacCoss M., Cleary M. A., Arthur W. T., and Moon R. T. (2008). New regulators of Wnt/beta-catenin signaling revealed by integrative molecular screening. *Science Signaling*, **1**, ra12.

Mohr S., Bakal C., and Perrimon N. (2010). Genomic screening with RNAi: results and challenges. *Annu Rev Biochem*, **79**, 37–64.

Ote I., Lebrun M., Vandevenne P., Bontems S., Medina-Palazon C., Piette E. M. J., and Sadzot-Delvaux C. (2009). Varicella-zoster virus IE4 protein interacts with SR proteins and exports mRNAs through the TAP/NXF1 pathway. *PLoS One*, **4**, e7882.

Oti M. and Brunner H. G. (2007). The modular nature of genetic diseases. *Clinical Genetics*, **71**, 1–11.

Oti M., Snel B., Huynen M. A., and Brunner H. G. (2006). Predicting disease genes using protein-protein interactions. *J Med Genet*, **43**, 691–698.

Perez-Iratxeta C., Bork P., and Andrade M. A. (2002). Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, **31**, 316–319.

Schultz N., Marenstein D. R., De Angelis D. A., Wang W. Q., Nelander S., Marks A. J. D. S., Massague J., and Sander C. (2011). Off-target effects dominate a large-scale rnai screen for modulators of the tgf-beta pathway and reveal microrna regulation of tgfbr2. *Silence*, **2**, 3.

Senecal D. L. and Morelli J. (2007). Hepatitis C virus infection: a current review. *JAAPA*, **20**, 21–25.

Stallings C. L., Duigou G. J., Gershon A. A., Gershon M. D., and S J Silverstein S. (2006). The cellular localization pattern of Varicella-Zoster virus ORF29p is influenced by proteasome-mediated degradation. *J. Virol.*, **80**, 1497–1512.

Tai A., Benita Y., Peng L., Kim S., Sakamoto N., Xavier R., and Chung

R. (2009). A Functional Genomic Screen Identifies Cellular Cofactors of Hepatitis C Virus Replication. *Cell Host Microbe*, **5**, 298–307.

Tang W., Dodge M., Gundapaneni D., Michnoff C., and Lum M. R. L. (2008). A genome-wide RNAi screen for Wnt/beta-catenin pathway components identifies unexpected roles for TCF transcription factors in cancer. *Proc Natl Acad Sci USA*, **105**, 9697–9702.

Turner F. S., Clutterbuck D. R., and Semple C. A. (2003). POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biology*, **4**, R75.

Walters M. S., Kyratsous C. A., Wan S., and Silverstein S. (2008). Nuclear import of the varicella-zoster virus latency-associated protein ORF63 in primary neurons requires expression of the lytic protein ORF61 and occurs in a proteasome-dependent manner. *J. Virol.*, **82**, 8673–8686.

Wang L., Tu Z., and Sun F. (2009). A network-based integrative approach to prioritize reliable hits from multiple genome-wide RNAi screens in Drosophila. *BMC Genomics*, **10**, 220.

Whitehurst A. W., Bodemann B. O., Cardenas J., Ferguson D., Girard L., Peyton M., Minna J. D., Michnoff C., Hao W., Roth M. G., Xie X. J., and White M. A. (2007). Synthetic lethal screen identification of chemosensitizer loci in cancer cells. *Nature*, **446**, 815–819.

Wood L. D., Parsons D. W., Jones S., Lin J., Sjoblom T., Leary R. J., Shen D., Boca S. M., Barber T., Ptak J., Silliman N., Szabo S., Dezso Z., Ustyanksky V., Nikolskaya T., Nikolsky Y., Karchin R., Wilson P. A., Kaminker J. S., Zhang Z., Croshaw R., Willis J., Dawson D., Shipitsin M., Willson J. K., Sukumar S., Polyak K., Park B. H., Pethiyagoda C. L., Pant P. V., Ballinger D. G., Sparks A. B., Hartigan J., Smith D. R., Suh E., Papadopoulos N., Buckhaults P., Markowitz S. D., Parmigiani G., Kinzler K. W., Velculescu V. E., and Vogelstein B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.

Wurdak H., Zhu S., Romero A., Lorger M., Watson J., Chiang C. Y., Zhang J., Natu V. S., Lairson L. L., Walker J. R., Trussell C. M., Harsh G. R., Vogel H., Felding-Habermann B., Orth A. P., Miraglia L. J., Rines D. R., Skirboll S. L., and Schultz P. G. (2010). An RNAi screen identifies TRRAP as a regulator of brain tumor-initiating cell differentiation. *Cell Stem Cell*, **6**, 37–47.

Zender L., Xue W., Zuber J., Semighini C. P., Krasnitz A., Ma B., Zender P., Kubicka S., Luk J. M., Schirmacher P., McCombie W. R., Wigler M., Hicks J., Hannon G. J., Powers S., and Lowe S. (2008). An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer. *Cell*, **135**, 852–864.

Zerboni, L. S. M. W. C. and Arvin A. (2000). Varicella-zoster virus infection of a human CD4-positive T-cell line. *Virol.*, **270**, 278–285.

Zhao Y. and Ding S. (2007). A high-throughput siRNA library screen identifies osteogenic suppressors in human mesenchymal stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 9673–9678.

Zhou H., Xu M., Huang Q., Zhang X. D., Castle J. C., Stec E., Ferrer M., Strulovici B., Hazuda D., and Espeseth A. (2008). Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe*, **4**, 495–504.

Table 1: RNAi Datasets.

Code	Phenotype	Scale	Hit list used	Comments	Reference
A	Chemosensitization	genome-wide	author supplied high confidence list	NCI-H1155 cells	Whitehurst *et al.*, 2007
B	HCV host factors	genome-wide	author supplied primary hit list	Huh7 cells	Tai *et al.*, 2009
C	Melanogenesis	genome-wide	$\|Z\| \geq 2$	MNT-1 cells	Ganesan *et al.*, 2008
D	Stem cell identity	genome-wide	author supplied deconvolution list	hESC cell lines	Chia *et al.*, 2010
E	VZV host factors	druggable	$\|Z\| \geq 2$	MeWo cells	unpublished
F	WNT signaling	genome-wide	author supplied primary hit list	DLD1 cells	Major *et al.*, 2008
G	WNT signaling	genome-wide	$\|Z\| \geq 2$	HeLa cells	Tang *et al.*, 2008
H	Cell viability	genome-wide	$\|Z\| \geq 2$	HeLa cells	Kittler *et al.*, 2007
I	Cell viability	genome-wide	$\|Z\| \geq 2$	Companion[a] to B	Whitehurst *et al.*, 2007
J	Cell viability	genome-wide	$\|Z\| \geq 2$	Companion[a] to C	Tai *et al.*, 2009
K	Cell viability	druggable	$\|Z\| \geq 2$	Companion[a] to F	unpublished
L	Cell viability	genome-wide	$\|Z\| \geq 2$	Companion[a] to H	Tang *et al.*, 2008

[a] A secondary screen used to filter nonviable perturbations in the primary screen.

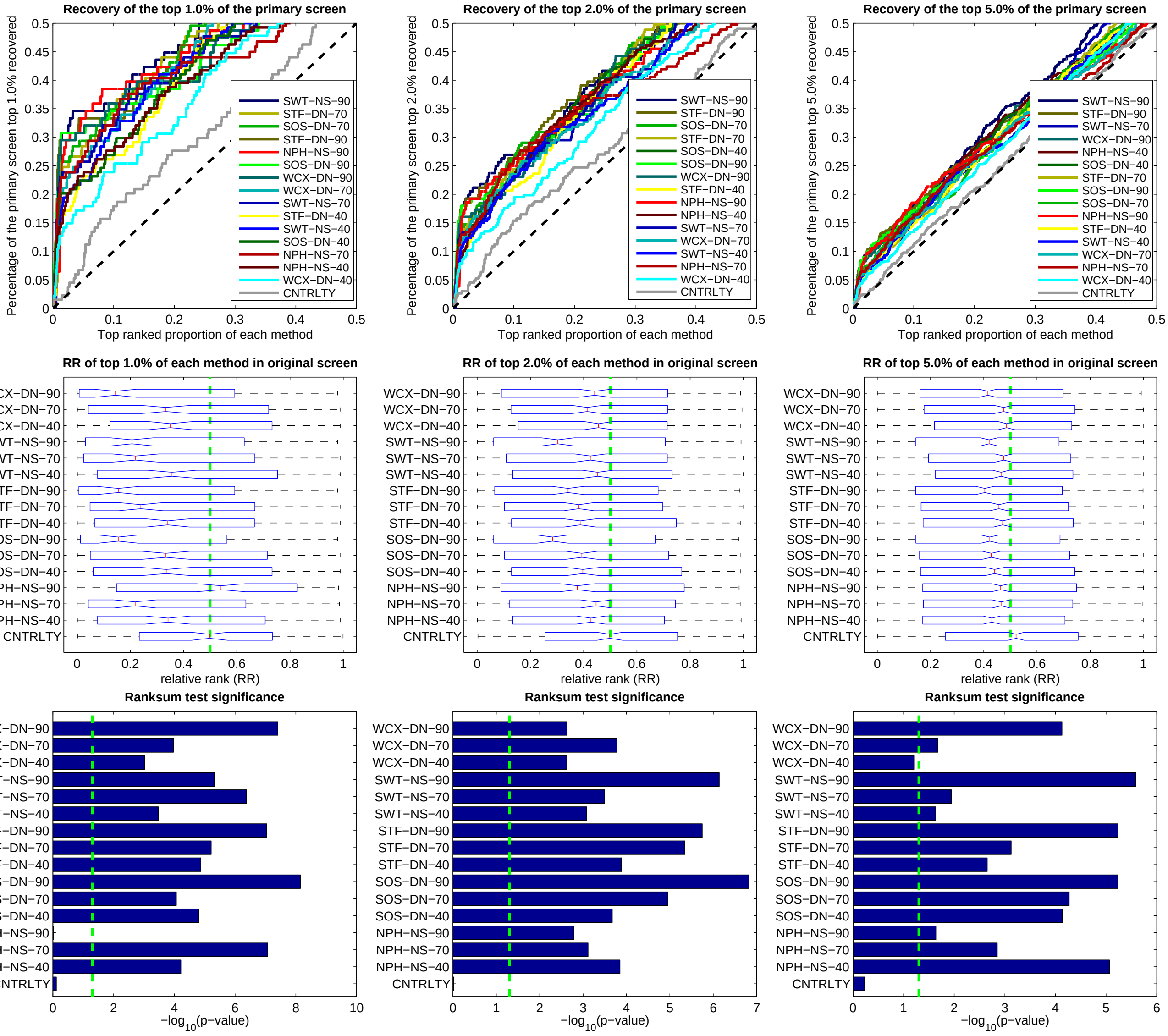
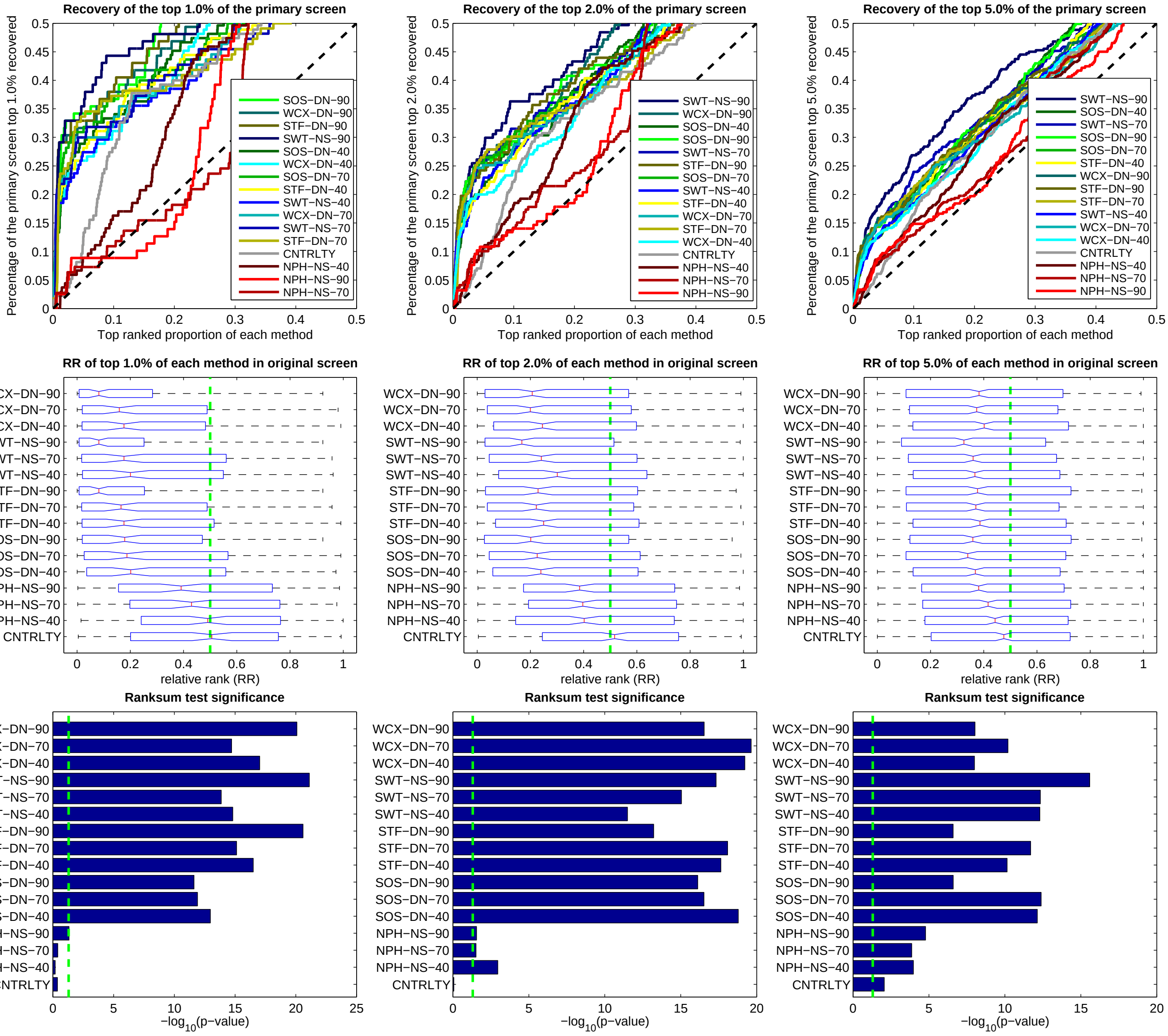


Figure 1: Summary of results for the HCV host factor screen (Screen B). Top row: Summarizes the recovery of the top hits in the screen (i.e., genes with the strongest induced phenotype) by each method. For example, the leftmost graph shows the fraction of the top 1% genes with the strongest induced phenotypes that is recovered as a function of the fraction of the total number of genes that is considered when the genes are prioritized according to the corresponding method scores. Note that none of these methods used the phenotype of a gene itself in calculating its score; only the data values of the gene's neighbors were considered. The different methods in the legend are ordered according to the corresponding AUC (area under the curve). **Middle row:** Shows the distribution of the top ranked genes for each method in the primary screen (sorted from the strongest to the weakest phenotype). RR stands for relative rank. **Bottom row:** Shows the statistical significance of the corresponding distribution in the middle row. Specifically, we plotted the logarithms of the p-values obtained via comparison of the phenotypes of the top-ranked genes for each method to the phenotypes of all the genes in the network using Wilcoxon's Rank-Sum test. Please see the methods section for the method codes used. The number that follows each code refers to the confidence threshold that was used in assembling the functional association network from the STRING database.


Recovery of the top 1.0% of the primary screen

Recovery of the top 2.0% of the primary screen

Recovery of the top 5.0% of the primary screen

RR of top 1.0% of each method in original screen

RR of top 2.0% of each method in original screen

RR of top 5.0% of each method in original screen

Ranksum test significance

Ranksum test significance

Ranksum test significance

Figure 2: Summary of results for the stem cell identity factors screen (Screen D). Top row: Summarizes the recovery of the top hits in the screen (i.e., genes with the strongest induced phenotype) by each method. The different methods in the legend are ordered according to the corresponding AUC (area under the curve). **Middle row:** Shows the distribution of the top ranked genes for each method in the primary screen (sorted from the strongest to the weakest phenotype). RR stands for relative rank. **Bottom row:** Shows the statistical significance of the corresponding distribution in the middle row.

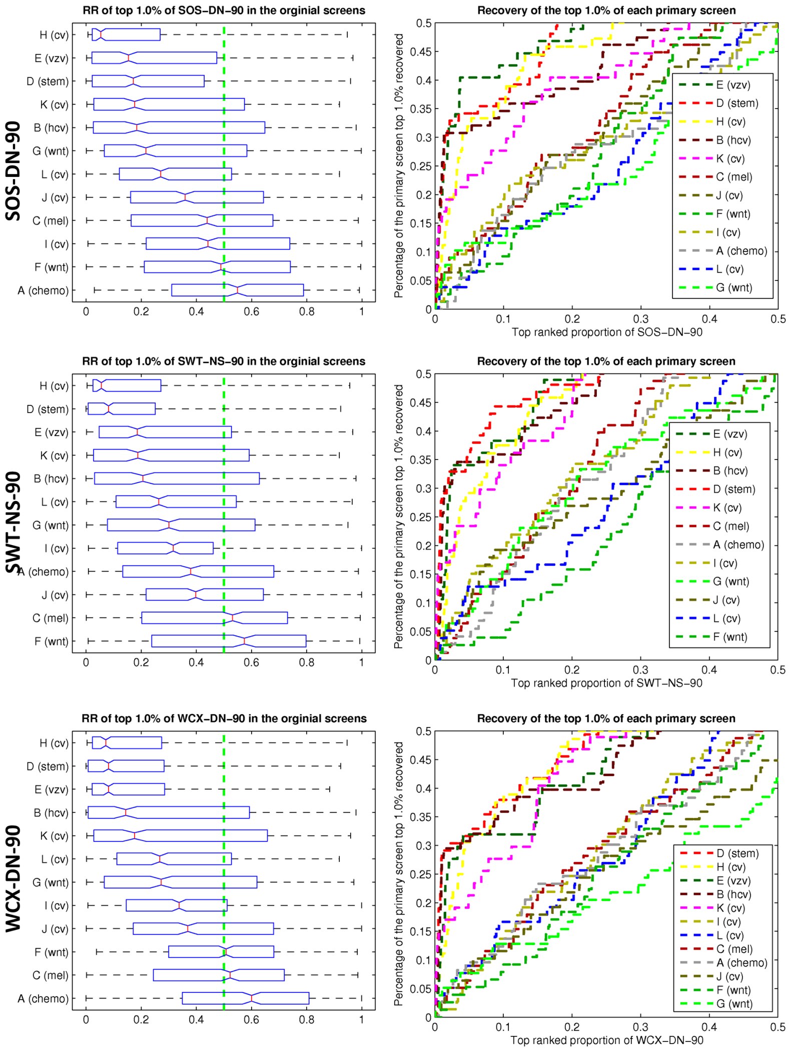


Figure 3: Predictability of the human RNAi screens. Summarizes the predictability of the different screens using SOS-DN-90 (top), SWT-NS-90 (middle) and WCX-DN-90 (bottom). The graphs on the left show the distribution of the top 1% ranked genes by each method in the primary screens (sorted from left to right by decreasing phenotype strength). The graphs on the right show the recovery of the top 1% hits of each screen by the three methods. Clearly, the correlations between the genes that induce strong phenotypes and those associated with them are more evident in some screens than in others.

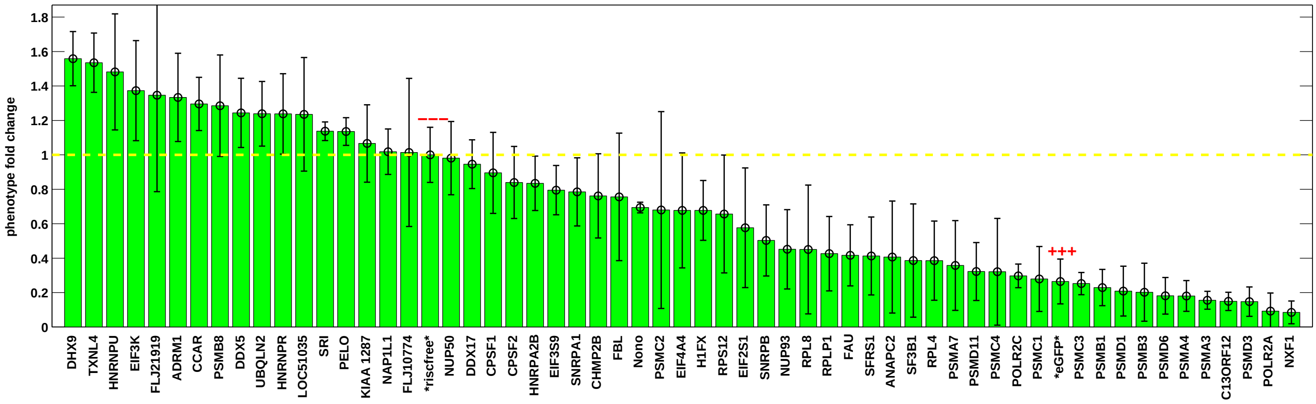


Figure 4: Validation of VZV predictions. We experimentally tested 57 of the candidate host factors for VZV by performing the corresponding knockdowns using siRNA smartpools. 24 (48%) inhibited viral growth by more than 50% (normalized to negative controls), compared to less than 4% in the main screen. Several were also found to have significantly enhanced infection, including DHX9, TXNL4, and EIF3K. Positive and negative controls are marked accordingly.

Supporting Information

Guilt by Association in Human RNAi Screens

Gonzalez et. al.

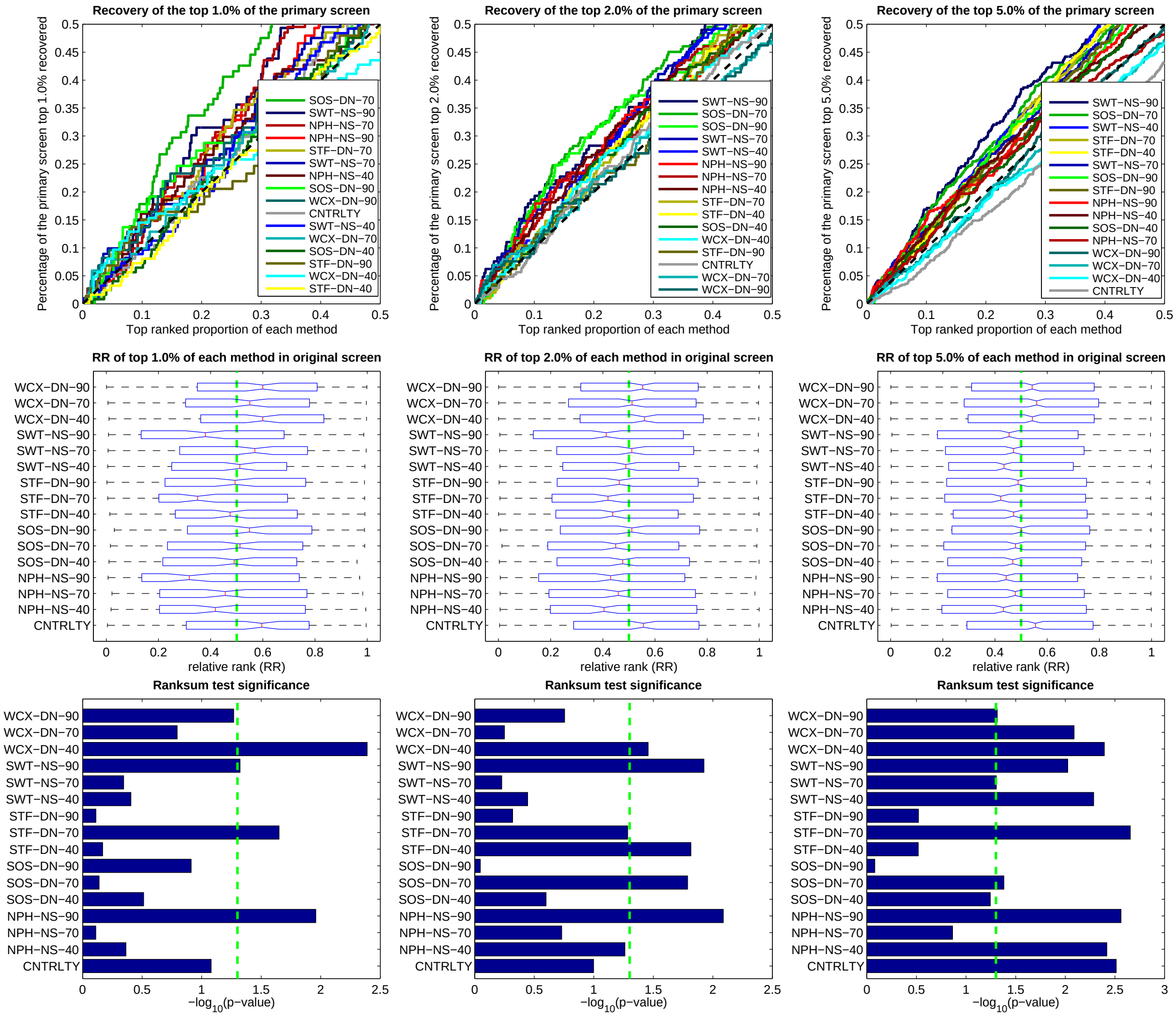


Figure S1: Results for the chemosensitizer loci screen (Screen A). Top row: Summarizes the recovery of the top hits in the screen (i.e., genes with the strongest induced phenotype) by each method. The different methods in the legend are ordered according to the corresponding AUC (area under the curve). **Middle row:** Shows the distribution of the top ranked genes for each method in the primary screen (sorted from the strongest to the weakest phenotype). RR stands for relative rank. **Bottom row:** Shows the statistical significance of the corresponding distribution in the middle row.

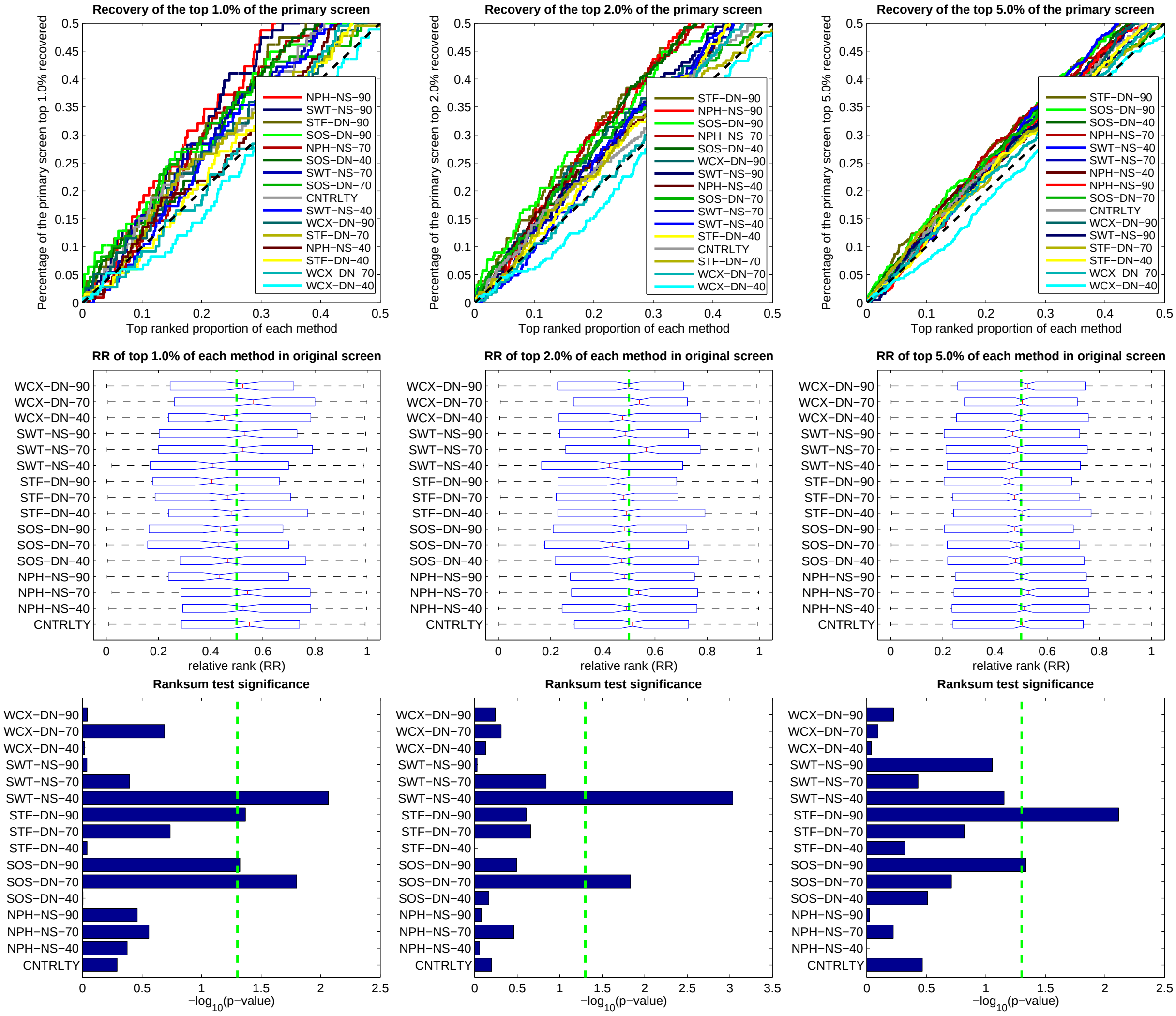


Figure S2: Results for the melanogenesis factors screen (Screen C). Top row: Summarizes the recovery of the top hits in the screen (i.e., genes with the strongest induced phenotype) by each method. The different methods in the legend are ordered according to the corresponding AUC (area under the curve). **Middle row:** Shows the distribution of the top ranked genes for each method in the primary screen (sorted from the strongest to the weakest phenotype). RR stands for relative rank. **Bottom row:** Shows the statistical significance of the corresponding distribution in the middle row.

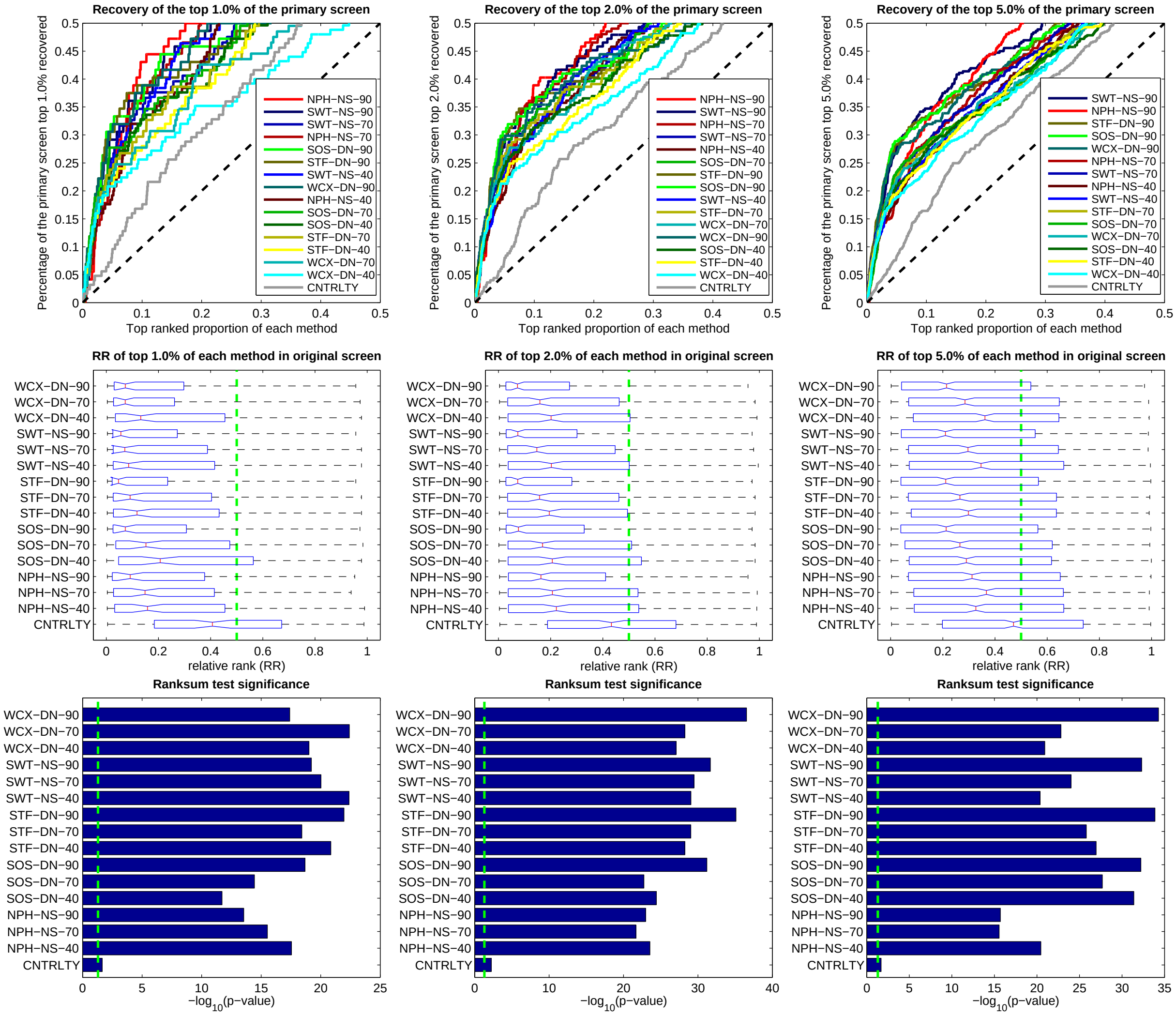
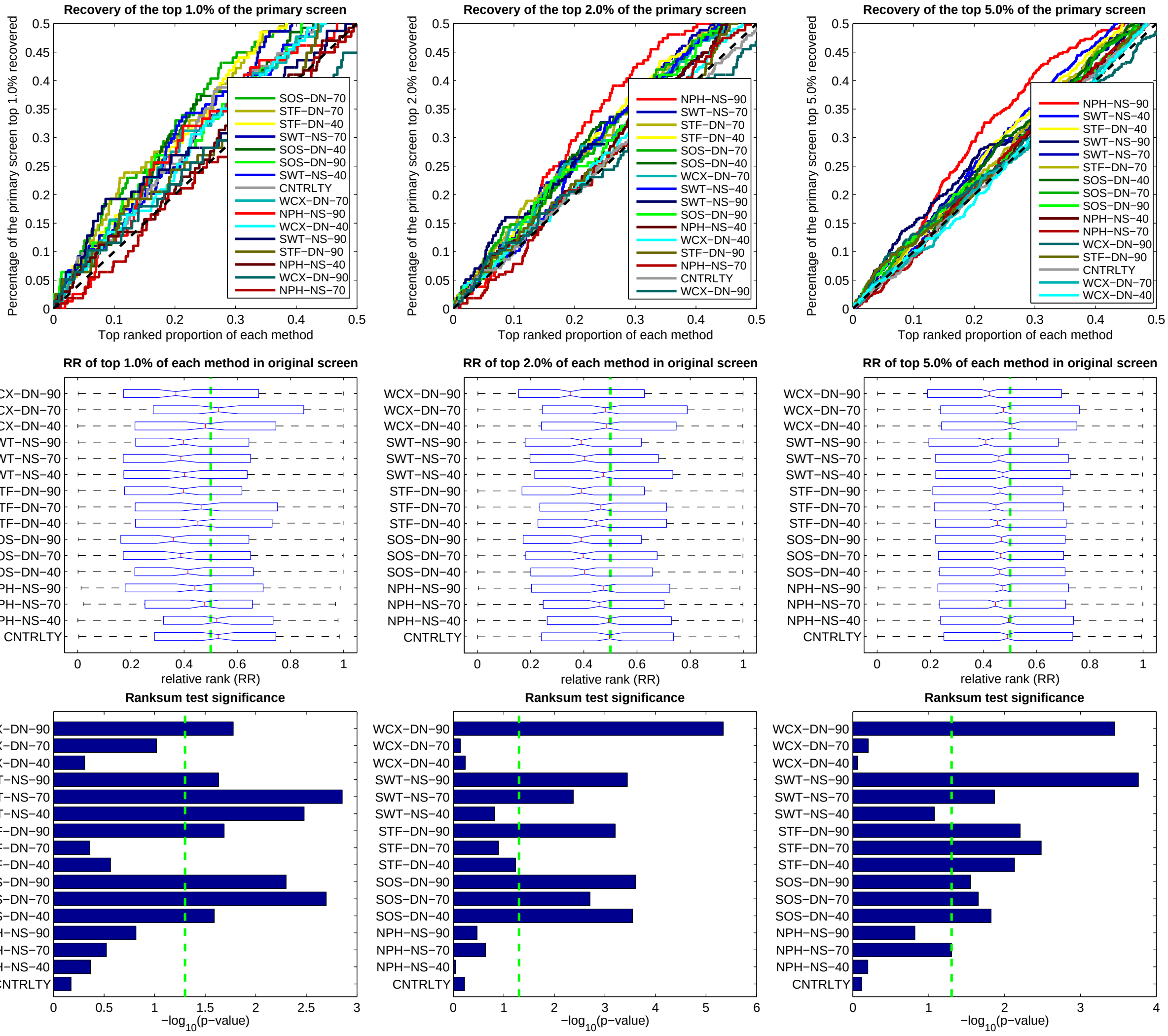


Figure S3: Results for the Kitler et al. (2007) cell viability screen (Screen H). Top row: Summarizes the recovery of the top hits in the screen (i.e., genes with the strongest induced phenotype) by each method. The different methods in the legend are ordered according to the corresponding AUC (area under the curve). **Middle row:** Shows the distribution of the top ranked genes for each method in the primary screen (sorted from the strongest to the weakest phenotype). RR stands for relative rank. **Bottom row:** Shows the statistical significance of the corresponding distribution in the middle row.



Recovery of the top 1.0% of the primary screen

Recovery of the top 2.0% of the primary screen

Recovery of the top 5.0% of the primary screen

RR of top 1.0% of each method in original screen

RR of top 2.0% of each method in original screen

RR of top 5.0% of each method in original screen

Ranksum test significance

Ranksum test significance

Ranksum test significance

Figure S4: Results for the Tai et al. (2009) cell viability screen (Screen J). Top row: Summarizes the recovery of the top hits in the screen (i.e., genes with the strongest induced phenotype) by each method. The different methods in the legend are ordered according to the corresponding AUC (area under the curve). **Middle row:** Shows the distribution of the top ranked genes for each method in the primary screen (sorted from the strongest to the weakest phenotype). RR stands for relative rank. **Bottom row:** Shows the statistical significance of the corresponding distribution in the middle row.

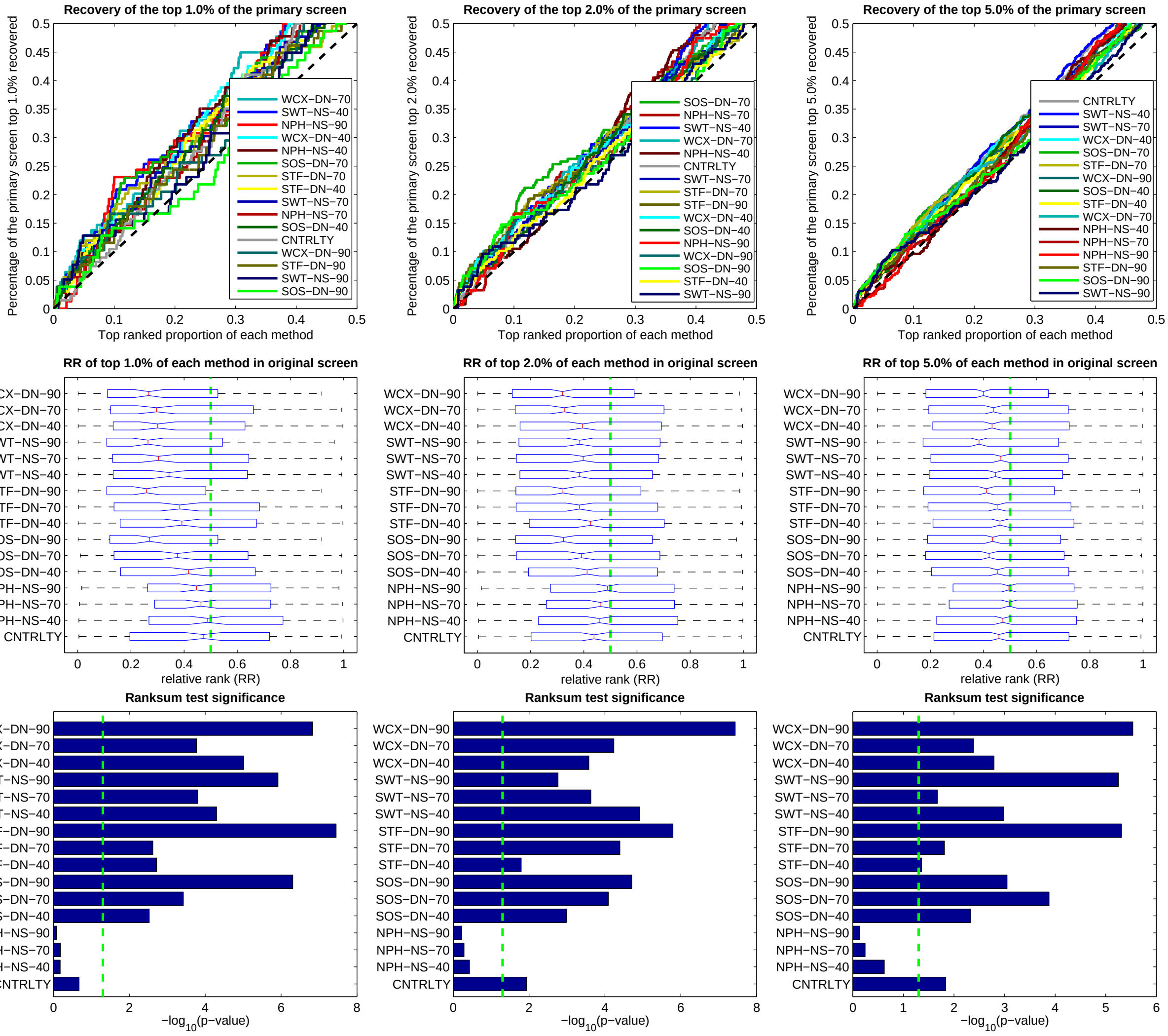


Figure S5: Results for the Tang et al. (2008) cell viability screen (Screen L). Top row: Summarizes the recovery of the top hits in the screen (i.e., genes with the strongest induced phenotype) by each method. The different methods in the legend are ordered according to the corresponding AUC (area under the curve). **Middle row:** Shows the distribution of the top ranked genes for each method in the primary screen (sorted from the strongest to the weakest phenotype). RR stands for relative rank. **Bottom row:** Shows the statistical significance of the corresponding distribution in the middle row.