BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 12 2012, pages 1612–1618
doi:10.1093/bioinformatics/bts192

Systems biology

Advance Access publication April 17, 2012

Stability of gene rankings from RNAi screens

Juliane Siebourg [1,2], Gunter Merdes [1], Benjamin Misselwitz [3], Wolf-Dietrich Hardt [3] and Niko Beerenwinkel [1,2,*]

[1]Department of Biosystems Science and Engineering, ETH Zurich, Mattenstrasse 26, 4058 Basel, Switzerland, [2]SIB Swiss Institute of Bioinformatics, 4058 Basel, Switzerland and [3]Institute of Microbiology, Department of Biology, ETH Zurich, Wolfgang-Pauli-Strasse 10, 8093 Zurich, Switzerland

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Genome-wide RNA interference (RNAi) experiments are becoming a widely used approach for identifying intracellular molecular pathways of specific functions. However, detecting all relevant genes involved in a biological process is challenging, because typically only few samples per gene knock-down are available and readouts tend to be very noisy. We investigate the reliability of top scoring hit lists obtained from RNAi screens, compare the performance of different ranking methods, and propose a new ranking method to improve the reproducibility of gene selection.

Results: The performance of different ranking methods is assessed by the size of the stable sets they produce, i.e. the subsets of genes which are estimated to be re-selected with high probability in independent validation experiments. Using stability selection, we also define a new ranking method, called stability ranking, to improve the stability of any given base ranking method. Ranking methods based on mean, median, *t*-test and rank-sum test, and their stability-augmented counterparts are compared in simulation studies and on three microscopy image RNAi datasets. We find that the rank-sum test offers the most favorable trade-off between ranking stability and accuracy and that stability ranking improves the reproducibility of all and the accuracy of several ranking methods.

Availability: Stability ranking is freely available as the R/Bioconductor package staRank at http://www.cbg.ethz.ch/software/staRank.

Contact: niko.beerenwinkel@bsse.ethz.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 21, 2011; revised on March 19, 2012; accepted on April 11, 2012

1 INTRODUCTION

Genome-wide gene silencing experiments are important in many fields of biology and medicine as they provide a first overview of which genes might play a role for a specific experimental condition. Many screens have been performed to study signaling in model organisms like *Drosophila melanogaster* (Boutros *et al.*, 2004; Saj *et al.*, 2010). In infection biology, viruses such as *Influenza* and HIV (Cherry, 2009; Hao *et al.*, 2008; Karlas *et al.*, 2010; Zhou *et al.*, 2008) as well as bacteria, including *Salmonella*, *Bartonella* and *Shigella* (Agaisse *et al.*, 2005; Misselwitz *et al.*, 2011; Philips *et al.*, 2005; Rämet *et al.*, 2002; Reiterer *et al.*, 2011; Truttmann *et al.*, 2011), have been analyzed to identify the key host genes involved in pathogen entry into the cell. In cancer research, RNA interference (RNAi) screens have been used to study dysregulated signaling pathways and to identify novel drug targets (Berns *et al.*, 2004; Ngo *et al.*, 2006).

In such high-throughput experiments one faces the problem of detecting the typically few relevant variables from a large, high-dimensional, noisy dataset. We focus here on data from microscopy image-based RNAi screens (Bickle, 2010), where genes are knocked down individually by post-transciptional gene silencing. Small interfering RNAs (siRNAs) of 22 base pairs are introduced into a cell, where they induce cleavage and degradation of a target messenger RNA, complementary to the siRNA and thus to the eventual depletion of the respective protein (Fire *et al.*, 1998; Hannon, 2002; Mello and Conte, 2004).

A typical setup for such a microscopy image-based screen consists of several 384-well plates, where each well contains cells with exactly one gene knocked down. There are different strategies for the knock-down, two of which will be covered by different datasets we analyze in Section 4.2. The first is to take replicates of the same biological experiment, meaning that for each gene, the same siRNA knock-down is performed several times. The second strategy uses biologically different experiments per gene. Here, each well for the same gene contains a different type of siRNA targeting the gene (Echeverri *et al.*, 2006). A third approach is to pool these different siRNAs in one well, each of them in lower concentration (Kittler *et al.*, 2007). This strategy adresses two main problems of siRNAs. The first is inefficient knock-down of a gene, for example due to inefficient siRNA binding. Secondly, an siRNA can have so called off-target effects, arising from limited binding specificity or other often unknown pharmacological effects (Qiu *et al.*, 2005).

After transfection with siRNAs, the cells are put in the experimental condition to be studied and subsequently they are imaged. The images are processed by an image segmentation and analysis software and the final experimental readout consists of one or more phenotypic measures retrieved from fluorescence signals of stained proteins.

In most cases, the goal of a first genome-wide screen is to prioritize genes to select a set of top scoring 'hits' for which a secondary validation experiment is performed. For a 1D readout the usual procedure is to rank the genes by their mean or median readout across replicates. However, to account for the variation among replicates genes can also be ranked according to a test

*To whom correspondence should be addressed.

statistic. For example, redundant siRNA activity analysis (RSA) is a ranking method specifically designed for RNAi screens. It ranks all individual siRNAs by readout and then assigns a p-value to each gene based on the rank distribution of all siRNAs targeting it using a hyper-geometric model (König *et al.*, 2007).

After ranking, a threshold is chosen to distinguish between hit and non-hit genes. This threshold can be a fold-change, deviation from the mean, or simply the fraction of top k ranking genes. The number of genes in the final subset will usually be restricted by the capacity of the re-screen and typically contains at most on the order of a few hundred genes. Selecting the optimal genes is a difficult problem, because many data points lie very close to each other and, at the same time, they are subject to considerable noise. Rather than defining hits, we focus here on the gene ranking itself, because (i) we did not find any evidence for two separate groups in the data (such as a bimodal readout distribution), and (ii) in practice, the top k genes will be selected based on available resources. Thus, we assume that each gene has an individual effect and that readout values are drawn from a continuous distribution.

A general problem with rankings is that reproducibility is strongly affected by small perturbations of the data and that different ranking criteria can lead to very different results (Fagin *et al.*, 2003). Since the screens are expensive and time consuming, in a whole-genome setting, only a few samples per gene are available. The analysis is further complicated by high levels of noise resulting, among other factors, from the uncertainty in quantifying image-based readouts and from the above mentioned off-target effects. Thus, the reliability of such gene rankings is a major concern directly affecting the chances of validating primary hits in follow-up experiments.

Gene rankings have been considered in the context of identifying differentially expressed genes from microarray data. To quantify the robustness of a ranking, resampling or subsampling methods are often used (Efron, 1979). For example, to benchmark different statistical tests for their reproducibility in detecting differentially expressed genes, (Qiu *et al.*, 2006) use a subsampling approach. (Pavlidis, 2003) apply a jackknife procedure to investigate the number of replicates per gene in a microarray experiment that are needed to obtain stable results. The R package 'Gene Selector' (Boulesteix and Slawski, 2009) implements several ranking statistics and provides a bootstrap procedure to estimate the robustness of the ranking result.

Another way of generating more stable results is learning the optimal ranking statistic for a given dataset based on resampling (Elo *et al.*, 2008; Mukherjee *et al.*, 2005). The probabilities obtained in this manner can also inform the variable selection procedure. For example, (Mukherjee *et al.*, 2003) have used bootstrapped p-values from t-tests to select genes more robustly. Hall and Miller (2009) discuss the consistency of boostrap estimators for rankings. They also model the variability of rankings which they find to be lower at the extremes (Hall and Miller, 2010). Stability selection is a more general variable selection method based on subsampling to estimate selection probabilities of variables (Meinshausen and Bühlmann, 2010). For this approach, an upper bound on the expected number of false positives has been derived under certain assumptions.

Rank aggregation has also been proposed to improve ranking stability. The 'Gene Selector' package provides aggregation of rankings by, for example, rank averaging or rank product. (Pihur *et al.*, 2009) propose a genetic algorithm to find an aggregated ranking that minimizes the distance to the individual rankings. Their results are quite stable, but a drawback of this method is that it is computationally very expensive and practical only for very small lists of genes.

In the presence of multivariate data, the hit selection problem can also be addressed by multivariate approaches like support vector machines (Guyon *et al.*, 2002) or other classification methods [see (Lai *et al.*, 2006) and (Stiglic and Kokol, 2010) for examples]. However, since we have 1D readouts we only consider univariate methods for the rankings.

In this article, we compare different ranking methods to identify those that produce the most stable gene lists. We analyze mean, median, t-test and rank-sum test, and quantify their reproducibility. The notion of stable sets, as defined in stability selection, is used to assesses the stability of a ranking. However, a ranking should not only be stable but also as accurate as possible. A constant ranking obtained, for example, by sorting genes alphabetically would be perfectly stable, but estimate biological effects very poorly. Finding an optimal trade-off between accuracy and reproducibility is a major goal when selecting hits in RNAi screening. We introduce stability ranking to improve the stability of any given base ranking method, while maintaining and sometimes improving its level of accuracy, and compare it to rank averaging. The performance of rankings is tested on simulated data and on real data from three image-based RNAi screens.

2 STABILITY RANKING

Let G be a set of p genes. For each gene knock-down, an experimental response is measured, called readout. We want to prioritize the genes with significantly altered readouts. For each gene $g \in G$, there are n measurements $g_1, \ldots, g_n$, which could either be replicates or correspond to individual siRNAs. Let $\mathrm{rk}(g)$ be the true rank of a gene g, $S_k = \{g \,|\, \mathrm{rk}(g) \leqslant k\}$ the set of genes that are ranked below a certain cutoff k, and $N_k = \{g \,|\, \mathrm{rk}(g) > k\}$ the set of genes with ranks above the cutoff. Then the goal is to infer the set of true top k genes, S_k, from few noisy observations. As discussed above, we do not optimize the cutoff parameter k, but rather aim at inferring S_k for all $k \in \{1, \ldots, p\}$.

We follow the stability selection approach (Meinshausen and Bühlmann, 2010) and let $I \in 2^{\{1,\ldots,n\}}$ be the random variable for data samples of size n that are drawn with replacement from the set of replicates. For a given ranking method and a fixed cutoff k, the probability for a gene $g \in G$ to be in the selected set $\hat{S}_k(I)$ is denoted by $\Pi_k^g = P[g \in \hat{S}_k(I)]$ and estimated from a finite sample $\{i_1, \ldots, i_m\}$ as $\hat{\Pi}_k^g = 1/m \sum_{j=1}^m \mathbb{1}\{g \in S_k(i_j)\}$, where $\mathbb{1}$ is the indicator function, which equals 1 if the argument is true and 0 otherwise, and $S_k(i_j)$ the set of selected genes based on subsample i_j. We regard those genes as stable that are selected with high probability. Formally, for a threshold $\pi \in (0, 1)$, we define the stable gene set

$$\hat{S}_k^{\mathrm{stable}} = \left\{ g \in G \,|\, \hat{\Pi}_k^g \geqslant \pi \right\}.$$

We fix $\pi = 0.9$ throughout the article since the choice of this parameter is not critical, as long as it is not set to very low values (see Supplementary Fig. S3). We will use the size of stable sets as a measure of ranking stability and we now introduce a novel ranking method based on this notion.

Observe that stable sets are nested, $\hat{S}_k^{\mathrm{stable}} \subseteq \hat{S}_{k+1}^{\mathrm{stable}}$, i.e. k-stable genes remain k'-stable for all $k' \geqslant k$. Stability ranking is defined by

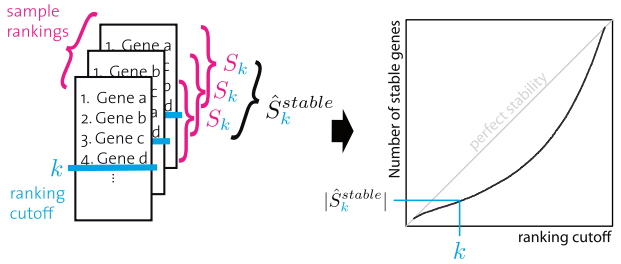

Fig. 1. Illustration of stability selection for a specific ranking cutoff k. The cardinality of the stable set at this cutoff provides an estimate for the number of top k genes, that are expected to be among the top k again, when repeating the experiment under the same conditions

computing the stable sets for all $k \in \{1,\ldots,p\}$ and then ranking the genes by the order in which they enter the stable set:

$$\mathrm{rk}^{\mathrm{stable}}(g) = \left|\hat{S}^{\mathrm{stable}}_{k^*}\right|, \quad \text{where } k^* = \min\left\{k \,|\, \hat{\Pi}^g_k \geqslant \pi\right\}.$$

The cardinality of the stable set provides an estimate of the number of hits that can be expected to be validated with probability π when considering the top k^* genes in the ranking. By validation we mean here that a gene is again among the top k^* genes when repeating the experiment under the same conditions. For noisy datasets, k^* can become much larger than the stable set size (Fig. 1).

Stability ranking is implemented in the R/Bioconductor package staRank. We apply this procedure to several ranking statistics, including mean and median as well as two statistical tests which account for the variation per gene, namely the t-test as a parametric and the rank-sum test as a non-parametric test. The tests are performed as one sided, two-sample tests comparing the replicates of one gene to the total dataset. For datasets generated by different siRNAs per gene, we also apply RSA ranking (König *et al.*, 2007). We analyze the stability and accuracy of the original ranking methods and investigate the improvement due to stability ranking in a simulation study.

3 SIMULATION STUDY

In the absence of evidence for multiple modes in the readout distributions of the RNAi screens we analyzed, in our simulations, we draw knock-down effects for all genes from a unimodal distribution and add individual random noise to it. We simulate datasets from a variety of models. Each model generates datasets of size $p \times n$, where p is the number of genes and n the number of replicates. For each gene, its true effect μ and its observed readouts g_i are drawn in a hierarchical fashion from normal distributions as follows:

$$\begin{aligned}\mu &\sim N(0, s^2)\\ \sigma^2 &\sim \Gamma(\alpha, \beta)\\ g_i &\sim N(\mu, \sigma^2), \quad i = 1,\ldots,n.\end{aligned}$$

Each model is characterized by the variance among gene effects, s^2, and the shape α and rate β of the gamma distribution from which the gene-wise variances among replicates are drawn. The gamma distribution has mean $m = \alpha/\beta$ and variance $v = \alpha/\beta^2$.

We estimated the parameters s, m and v from the effect and replicate distributions observed in the *Drosophila* genome-wide RNAi screen described below (Saj *et al.*, 2010). We then varied the parameters around these estimates, which resulted in 24 different models (Supplementary Fig. S1 and Supplementary Tables S1 and S2). Each model was used to generate datasets of cardinality $n = 2$, 3, 4 and 10.

To assess reproducibility and accuracy, 300 pairs of datasets are drawn from each model. For each dataset, the different base ranking methods, their stability ranking and their average ranking are computed. Accuracy is assessed by comparing the top k genes from an estimated ranking $\widehat{\mathrm{rk}}$ to the true hits S_k, whereas the reproducibility of a ranking is defined as the overlap in top k gene sets between the two rankings $\widehat{\mathrm{rk}}_1$ and $\widehat{\mathrm{rk}}_2$ estimated from paired datasets (Mukherjee *et al.*, 2005),

$$\begin{aligned}\mathrm{accuracy}(k) &= |S_k(\widehat{\mathrm{rk}}) \cap S_k|/k\\ \mathrm{reproducibility}(k) &= |S_k(\widehat{\mathrm{rk}}_1) \cap S_k(\widehat{\mathrm{rk}}_2)|/k.\end{aligned}$$

Both quality measures take values in $[0,1]$, where 1 indicates complete agreement among top k gene sets and hence perfect accuracy or reproducibility. The final quantities we report are averages across the pairs of datasets (reproducibility) or across all datasets (accuracy), for the top $k/p = 1\%$ or 10% genes.

4 RESULTS AND DISCUSSION

4.1 Simulation results

Using the model described in the previous section, we ran simulations for $p = 1000$ genes and assessed accuracy and reproducibility of the top 10 (1%) genes, because in practice, usually only a small fraction of hits can be selected for follow-up experiments. In Figure 2, reproducibility and accuracy are compared between the original, average and stability rankings for all models. Models with a filled shape showed a significant difference between the two ranking methods (baseline versus stability ranking and aggregated versus stability ranking) at the 0.1% level in a paired t-test after Benjamini–Hochberg correction for multiple testing.

For mean and median ranking, we observe an increase in reproducibility for a large group of models, when using stability ranking (Fig. 2A and B). Remarkably, stability ranking also increases the accuracy of the ranking (Fig. 2F and G). Similarly, for the rank-sum test, reproducibility is improved using stability ranking, although the effect is not as strong, while maintaining the same level of accuracy (Fig. 2C and H). The same behavior can be observed for the RSA rankings (Fig. 2E and J). By contrast, the t-test shows no difference between the two ranking versions, but accuracy is overall very low (Fig. 2D and I). The poor performance might be due to unreliable estimation of the variance from only two to five observations. Thus, non-parametric ranking statistics are preferable for this type of data. Direct comparison of the stability rankings based on mean, median and rank-sum test reveals superior accuracy and reproducibility of the rank-sum test (Supplementary Figs S3 and S4). In general, the performance increases with the width of the effects distribution (see Supplementary Tables S1 and S2, and Supplementary Figure S1 for top 10%). A direct comparison of stability ranking and rank averaging shows similar performance (Fig. 2K–T) with a slight advantage of stability ranking. For all of the four competitive methods, the stability ranking was significantly better in reproducibility for many more models than the average ranking (43 versus 9 for mean, 36 versus 12 for median, 41 versus

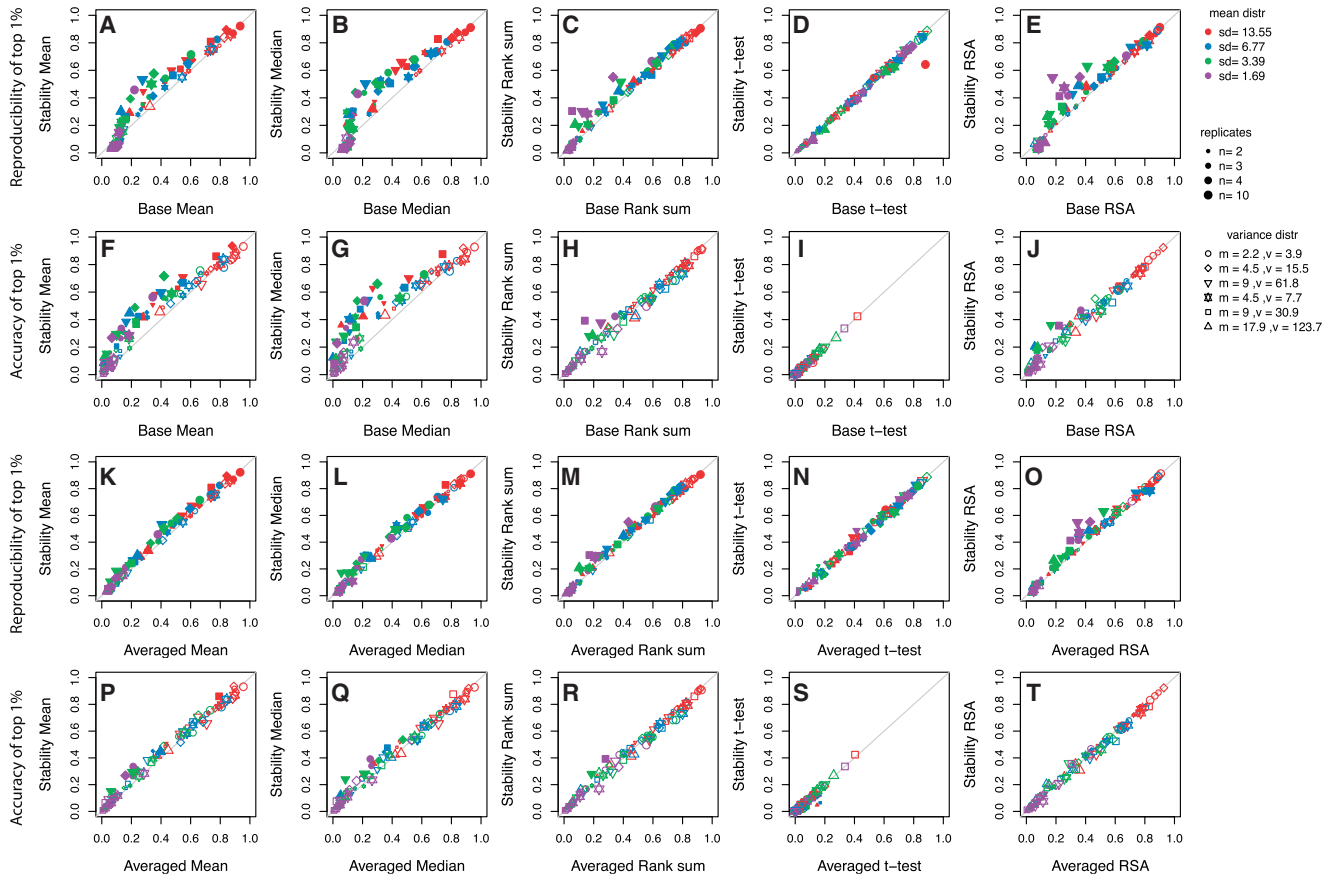

Fig. 2. Reproducibility (top row) and accuracy (second row) of the simulations. Shown are the results for the top 1% genes in base versus stability ranking using median (**A**, **F**), mean (**B**, **G**), rank-sum test (**C**, **H**), *t*-test (**D**, **I**) and RSA (**E**, **J**). The third and forth row (**K**–**T**) show the same plots but for the comparison of the aggregated versus the stability ranking. Each plot shows results for one ranking statistic and each symbol in the plots indicates one model. The colors represent the parameters that were used for the mean effect distribution. The different shapes represent the different gamma distributions for the gene variances and the symbol size indicates the number of replicates used. For models that have a filled shape, the two ranking methods (baseline versus stability ranking) showed a significant difference at the 1% level in a paired *t*-test after Benjamini–Hochberg correction for multiple testing

7 for rank sum and 29 versus 17 for RSA). The accuracy was similar for both aggregated methods, again with a slight advantage for stability ranking (12 versus 0 for mean, 11 versus 0 for median and 2 versus 0 for rank sum).

Figure 3 shows reproducibility and accuracy for one specific model, defined by the parameters $s=1.69$, $m=9$, $v=61.8$ and $n=10$. Interestingly, at the very top of the ranking the *t*-test outperforms the other methods in terms of reproducibility, but at the same time it has the lowest accuracy. The *t*-test base and stability rankings are almost indistinguishable, whereas for the other methods there is a large difference between the two. Especially for the top-ranked genes, stability ranking improves reproducibility by >10%. For median and mean, this also holds true in terms of accuracy up to the top 25% of the ranking. The most accurate ranking is produced by the rank-sum test, slightly outperforming its stability ranking version. Similar effects can be observed for most of the models.

In summary, the rank-sum test offers a good trade-off between accuracy and reproducibility of the ranking. Stability ranking, which can be applied on top of any given ranking method, improves or at least equalizes both accuracy and reproducibility of all ranking methods investigated here. The improvement is the largest if the base ranker does not account for gene-wise variation, such as mean and median ranking, but even the reproducibility of the rank-sum test ranking can be improved and, on average, it is larger then using average ranking.

4.2 Application to RNAi datasets

We apply stability ranking to three RNAi screens, of which one uses four replicates per siRNA, whereas the other two use three to four siRNAs per gene. The first is a whole-genome screen of *D.melanogaster* cells, which was performed to study signaling of the *Notch* receptor (Saj *et al.*, 2010). It consists of 4 identical replicates on ∼12 000 genes each. For each knocked down gene, *Notch* signaling activity was measured based on the ratio of signal to background fluorescence measurements. We refer to it as the Notch screen. The second dataset comprises a screen of ∼7000 human druggable genes (Misselwitz *et al.*, 2011). In this experiment, HeLa cells were infected with *Salmonella* bacteria to study their entry mechanism. For this the infection rate per knock-down is used. This screen was performed using three different siRNAs targeting the same gene. The third dataset is similar to the previous one but was

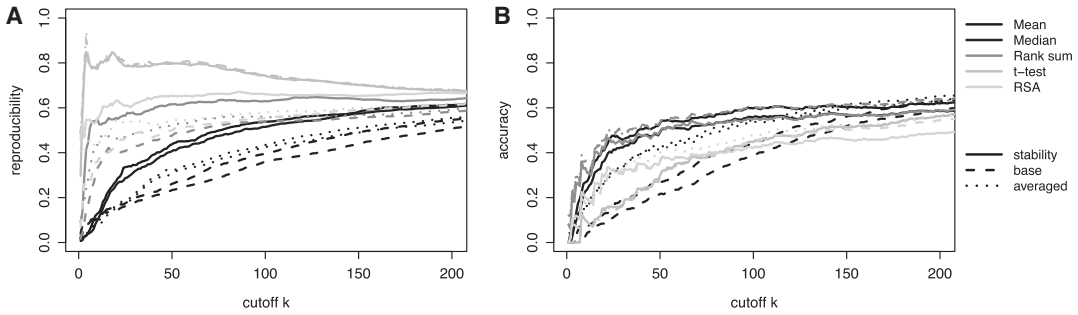

Fig. 3. Reproducibility (**A**) and accuracy (**B**) for the model ($s=1.69, m=9, v=61.8, n=10$). Dashed lines represent the base rankings, solid lines the stability versions

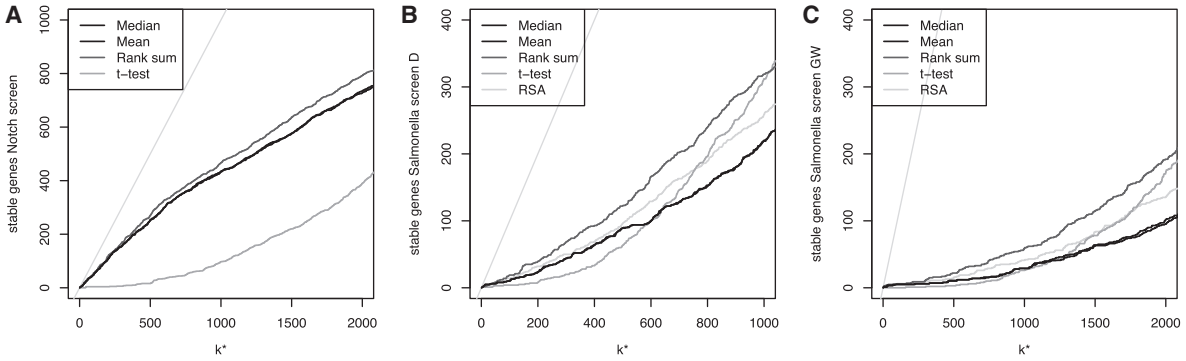

Fig. 4. Growth of the stable set of genes as a function of the ranking cutoff k^* for the Notch screen (**A**), the drugable Salmonella screen (**B**) and the genome-wide Salmonella screen (**C**). For each of the different ranking statistics, stability selection was performed. The diagonal (gray line) indicates perfect stability

performed on a genome-wide scale. Here we used all 14 837 genes for which 4 siRNA values were available. We call these screens the druggable (SalD) and the genome-wide (SalGW) Salmonella screen, respectively. For a more detailed description of the datasets, see Datasets section in the Supplementary Material.

For all datasets, median, mean, *t*-test and rank-sum test were used to calculate base and stability rankings. Since they use different siRNAs per gene, for the Salmonella screens, RSA ranking was also performed. The rankings were directed toward down regulation of the Notch receptor and decrease in infection, respectively. In all screens, the rank-sum test produces the most stable rankings, followed by RSA ranking (Fig. 4). The *t*-test has initially the lowest stability. This changes throughout the ranking, but since the top part is the most relevant one, this method appears impractical. As expected, the stability of the Notch screen, which uses replicates, is much higher than for the Salmonella screens, which use different siRNAs per gene. Table 1 summarizes the stable set sizes for the top 1% and top 10% resulting from the rank-sum rankings for each of the datasets (for the other methods see Supplementary Tables S3–S6).

To compare the reproducibility of rankings on the real data, we employed a bootstrap analysis (Efron, 1979) and resampled the data for each gene with replacement. For each bootstrap run, we used as many values per gene as the original dataset had. Figure 5A shows the bootstrapped reproducibility values of the Notch screen for the top 20% of median, *t*-test and rank-sum test rankings. Overall the reproducibility is very high for most rankings. In particular, the first two ranks show perfect reproducibility for the stability median and stability rank sum. Generally, the stability median rankings and both rank-sum test versions are ∼10% more reproducible than the base median rankings. Above the top 1% the original version is slightly more reproducible. The *t*-test again fails to recover a stable ranking.

For all screens, subsets of genes had been selected, which were followed up on with validation experiments. Selection of genes was based on a combination of the outcome of the primary screen as well as biological expert knowledge. For the rank-sum test, Table 1

Table 1. Screening and validation results for the top 1% and top 10% of the rank-sum test ranking for each of the three RNAi screens

Gene set	Total	Stable	Re-screened	Validated, *n* (%)
Notch top 1%	129	70	67	37 (55.2)
Notch top 10%	1281	556	225	141 (62.7)
SalD top 1%	69	12	9	7 (77.8)
SalD top 10%	686	198	74	28 (37.8)
SalGW top 1%	148	6	5	2 (40)
SalGW top 10%	1478	111	51	6 (11.8)

Second column indicates the absolute number of genes, third column indicates how many of these were stable. Fourth column indicates the part of the stable genes that was used in the re-screening experiments and the last column shows how many of them were validated. Notch indicates the Notch screen, SalD and SalGW refer to the drugable and the genome-wide Salmonella screens.

Table 2. Comparison of the base and stability version of the rank-sum test ranking

Gene set	Total	Overlap	Rank sum		Stability rank sum	
			Re-screened	Validated, *n* (%)	Re-screened	Validated, *n* (%)
Notch top 1%	129	124	120	68 (56.7)	119	70 (58.8)
Notch top 10%	1281	1160	235	148 (63)	233	147 (63.1)
SalD top 1%	69	64	47	22 (46.8)	43	19 (44.2)
SalD top 10%	686	618	125	41 (32.8)	113	38 (33.6)
SalGW top 1%	148	118	67	7 (10.4)	60	7 (11.7)
SalGW top 10%	1478	1162	113	12 (10.6)	102	11 (10.8)

For each of the three RNAi screens, the top 1% genes and top 10% genes are shown. The second column indicates the total number of genes contained in the top 1% and top 10% of the rankings. The third column shows how many genes both rankings had in common. For both versions then the number of re-screened and validated genes are shown. For description of the row names see caption of Table 1.

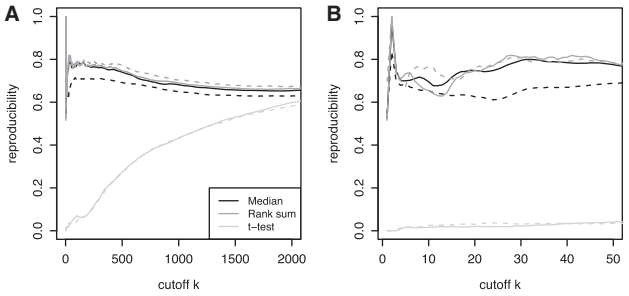

Fig. 5. Bootstrapped reproducibility for the Notch screen as a function of the cutoff k. (**A**) Shows the top 20% genes and (B) a zoom into the top 50 genes. Dashed lines indicate the base rankings and solid lines the stability versions

summarizes how many of the stable top 1% and top 10% genes were selected for re-screening and how many of these were finally validated (see Supplementary Tables S3–S6 for the other rankings). However, assessing the significance of these results is difficult for two reasons. Firstly, the sets of re-screened genes do not represent i.i.d. random samples, because they are biased by the way they were selected. Secondly, the re-screening experiments were not carried out under the same experimental conditions and therefore may lead to different conclusions. In case of the Notch screen, the primary *in vitro* screen was validated *in vivo*. In case of Salmonella, screens were validated using different or only partially overlapping siRNA libraries as compared with the primary screens. For the Notch screen, a total of 233 down regulating genes were re-screened, whereas for Salmonella, 164-infection decreasing genes of the drugable and 119 infection-decreasing genes of the genome-wide screen were chosen for validation.

Overall, most of the stable top 1% genes were re-screened. For the Notch screen the stable top 10% genes contained almost all of the re-screened genes, whereas for the other experiments this fraction is reduced to ∼50%. Validation rates vary considerably but tend to be the higher the more stable a screen is.

Comparing the top 1% and top 10% of the rank-sum test ranking and its stability counterpart we find that for the base ranker always a few more genes had been chosen for re-screening. Yet, in five out of six cases validation rates were higher when using stability ranking (Table 2). The rank-sum test and *t*-test rankings showed the highest overlaps between base and stability ranking, but for the less similar rankings, the stability rankings had also higher validation rates in most of the cases (Supplementary Tables S7–S10).

5 CONCLUSION

We have applied the concept of stability selection to gene rankings to generate more reproducible ordered hit lists for data generated from phenotypic RNAi experiments. We have shown that the robustness of different ranking methods can be very different and that the stable set size can be used as a measure of reproducibility. Since image-based RNAi screening data tend to be very noisy and sparse, the use of stability ranking can improve stability, especially in the top part of the rankings which is of main interest. In the present study, the rank-sum test ranking and its stability ranking version have resulted in the most reproducible hit lists. Stability ranking is very flexible and can be applied to any gene ranking method. It does not only improve ranking statistics that ignore the gene-wise variance, such as mean or median, but it also improved the reproducibility of a statistic like the rank-sum test. Thus, irrespective of the chosen ranking statistic, it appears beneficial to complement the selection of top scoring genes with stable genes to increase validation rates in secondary screens. In principle, the stable sets could also hint at a reasonable cutoffs for hit selection. Analyzing the growth curve of the stable sets for the datasets used in the present study, no such cutoff was be found, but this may be investigated further in future work on different datasets.

Funding: SystemsX.ch, the Swiss initiative in systems biology, under IPhd [2009/025] and RTD [2009/005] (InfectX), evaluated by the Swiss National Science Foundation.

Conflict of Interest: none declared.

REFERENCES

Agaisse,H. *et al.* (2005) Genome-wide RNAi screen for host factors required for intracellular bacterial infection. *Science*, **309**, 1248–1251.

Berns,K. *et al.* (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature*, **428**, 431–437.

Bickle,M. (2010) The beautiful cell: high-content screening in drug discovery. *Anal. Bioanal. Chem.*, **398**, 219–226.

Boulesteix,A.-L. and Slawski,M. (2009) Stability and aggregation of ranked gene lists. *Brief. Bioinformatics*, **10**, 556–568.

Boutros,M. *et al.* (2004) Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science*, **303**, 832–835.

Cherry,S. (2009) What have RNAi screens taught us about viral-host interactions? *Curr. Opin. Microbiol.*, **12**, 446–452.
Echeverri,C.J. *et al.* (2006) Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nat. Meth.*, **3**, 777–779.
Efron,B. (1979) Bootstrap methods: another look at the Jackknife. *Ann. Stat.*, **7**, 1–26.
Elo,L.L. *et al.* (2008) Reproducibility-optimized test statistic for ranking genes in microarray studies. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **5**, 423–431.
Fagin,R. *et al.* (2003) Comparing top k lists. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, pp. 36.
Fire,A. *et al.* (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806–811.
Guyon,I. *et al.* (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.
Hall,P. and Miller,H. (2009). Using the bootstrap to quantify the authority of an empirical ranking. *The Annals of Statistics*, **37**, 3929–3959.
Hall,P. and Miller,H. (2010). Modeling the variability of rankings. *The Annals of Statistics*, **38**, 2652–2677.
Hannon,G.J. (2002) RNA interference. *Nature*, **418**, 244–251.
Hao,L. *et al.* (2008) Drosophila RNAi screen identifies host genes important for influenza virus replication. *Nature*, **454**, 890–893.
Karlas,A. *et al.* (2010) Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature*, **463**, 818–822.
Kittler,R. *et al.* (2007) Genome-scale RNAi profiling of cell division in human tissue culture cells. *Nat. Cell Biol.*, **9**, 1401–1412.
König,R. *et al.* (2007) A probability-based approach for the analysis of large-scale RNAi screens. *Nat. Meth.*, **4**, 847–849.
Lai,C. *et al.* (2006) A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics*, **7**, 235.
Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. Roy. Stat. Soc. B*, **72**, 417–473.
Mello,C.C. and Conte,D. (2004) Revealing the world of RNA interference. *Nature*, **431**, 338–342.
Misselwitz,B. *et al.* (2011) RNAi screen of Salmonella invasion shows role of COPI in membrane targeting of cholesterol and Cdc42. *Mol. Syst. Biol.*, **7**:474.
Mukherjee,S.N. *et al.* (2003) Gene ranking using bootstrapped *p*-values. *ACM SIGKDD Explor. Newslett.*, **5**, 6.
Mukherjee,S.N. *et al.* (2005) Data-adaptive test statistics for microarray data. *Bioinformatics*, **21**, 108–114.
Ngo,V.N. *et al.* (2006) A loss-of-function RNA interference screen for molecular targets in cancer. *Nature*, **441**, 106–110.
Pavlidis,P. (2003) The effect of replication on gene expression microarray experiments. *Bioinformatics*, **19**, 1620–1627.
Philips,J.A. *et al.* (2005) Drosophila RNAi screen reveals CD36 family member required for mycobacterial infection. *Science*, **309**, 1251–1253.
Pihur,V. *et al.* (2009) RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics*, **10**, 62.
Qiu,S. *et al.* (2005) A computational study of off-target effects of RNA interference. *Nucleic Acids Res.*, **33**, 1834–1847.
Qiu,X. *et al.* (2006) Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*, **7**, 12.
Rämet,M. *et al.* (2002) Functional genomic analysis of phagocytosis and identification of a Drosophila receptor for E. coli. *Nature*, **416**, 644–648.
Reiterer,V. *et al.* (2011) Shigella flexneri type III secreted effector OspF reveals new crosstalks of proinflammatory signaling pathways during bacterial infection. *Cell. Signal.*, **23**, 1188–1196.
Saj,A. *et al.* (2010) A combined ex vivo and in vivo RNAi screen for Notch regulators in Drosophila reveals an extensive Notch interaction network. *Dev. Cell*, **18**, 862–876.
Stiglic,G. and Kokol,P. (2010) Stability of ranked gene lists in large microarray analysis studies. *J. Biomed. Biotechnol.*, **2010**, 616358.
Truttmann,M.C. *et al.* (2011) Bartonella henselae engages inside-out and outside-in signaling by integrin β1 and talin1 during invasome-mediated bacterial uptake. *J. Cell Sci.*, **124**(Pt 21), 3591–602.
Zhou,H. *et al.* (2008) Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe*, **4**, 495–504.