Stability of gene rankings from RNAi screens

Juliane Siebourg 1,2, Gunter Merdes 1, Benjamin Misselfitz 3, Wolf-Dietrich Hardt 3 and Niko Beerenwinkel 1,2, 1

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

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@sbsse.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), S. pyogenes (Saj et al. 2000). In infection biology, viruses such as Influenza and HIV (Boutros et al. 2005; Hao et al. 2009; Karlas et al. 2011; Zhou et al. 2008) as well as bacteria, including Salmonella, Bartonella and Shigella (Aogisse et al. 2005; Misselfitz et al. 2011), Philips et al. 2005; Ramelet et al. 2005; Reitter et al. 2011; Truttmann et al. 2011; Truttmann et al. 2014), 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 (Harms et al. 2004; Nes et al. 2005).

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 (Hicks et al. 2011), where genes are knocked down individually by post-transcriptional 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 et al. 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. 2009). A third approach is to pool these different siRNAs in one well, each of them in lower concentration (Kutler et al. 2005). This strategy addresses 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
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 [Genes et al. 2009].

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 [Quin et al. 2007]. 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 1997]. For example, to benchmark different statistical tests for their reproducibility in detecting differentially expressed genes, Quin et al. 2009 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 [Ehr et al. 2008; Mukherjee et al. 2009]. The probabilities obtained in this manner can also inform the variable selection procedure. For example, Mukherjee et al. 2009 have used bootstrapped p-values from t-tests to select genes more robustly. Hall and Miller (2009) discuss the consistency of bootstrap estimators for rankings. They also model the variability of rankings which they find to be lower by, for example, rank averaging or rank products [Pihur et al. 2007] or other classification methods [see [Genes et al. 2009] 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 assess 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 \( r(k) \) be the true rank of a gene \( g \), \( \hat{S}_k = [g | r(k) \leq k] \) the set of genes that are ranked below a certain cutoff \( k \), and \( N_k = [g | r(k) > k] \) 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 \( I = \{1, \ldots, m\} \) 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 \( P[I = I | \hat{S}_k(I)] \) and estimated from a finite sample \( I = \{1, \ldots, m\} \) as \( \hat{P}_k(I) = \frac{1}{m} \sum_{j=1}^m I[g \in \hat{S}_k(I_j)] \), where \( I \) 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

\[
\hat{S}_k \subset \{ g \in G | \hat{P}_k(I) \geq \pi \}.
\]

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, \( \hat{S}_k \subset \hat{S}_{k'} \) for all \( k' \geq k \). Stability ranking is defined by

\[
\hat{S}_k = \bigcap_{k' \geq k} \hat{S}_{k'}.
\]
We analyze the stability and accuracy of the original ranking 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 ∈ [1, . . . , p] and then ranking the genes by the order in which they enter the stable set: 

\[ R_{\text{stable}}(\alpha) = \{ g \mid k = \min \{ k \mid g \in | S_k \} \geq \alpha \} \]

The cardinality of the stable set provides an estimate of the number of hits that can be expected to be validated with probability \( \pi \) 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, \( k^\ast \) 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 models, the ranking is performed as one sided, two-sample tests comparing the replicates and the rank-sum test as a non-parametric test. The tests are accounted for the variation per gene, namely the mean and median as well as two statistical tests which are drawn in a hierarchical fashion from normal distributions as a stability ranking method 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. We simulate datasets of the RNAi screen described below for different models (Supplementary Fig. 1 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 \( R_k \) 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 \( R_{k1} \) and \( R_{k2} \) estimated from paired datasets (Mukherjee et al. 2005).

- **Accuracy**:
  \[ \text{accuracy}(k) = |S_k(R_k) \cap S_k| / k \]

- **Reproducibility**:
  \[ \text{reproducibility}(k) = |S_k(R_{k1}) \cap S_k(R_{k2})| / k \]

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 3, 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. 4 and B). Remarkably, stability ranking also increases the accuracy of the ranking (Fig. 4 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. 4 and H). The same behavior can be observed for the RAS rankings (Fig. 4 and J). By contrast, the \( t \)-test shows no difference between the two ranking versions, but accuracy is overall very low (Fig. 4 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. 4 and I) 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 10 for the top 10\%).
Stability ranking

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 (rank sum and 7 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 \( t = 1.69, m = 9, \gamma = 61.8, n = 10 \). 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 druggable 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 3B). 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 \( \text{Bootstrapped [95]} \) 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 5 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 \( \sim 10\% \) 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 summarizes the screening and validation results for the top 1% and top 10% of the rank-sum test ranking for each of the three RNAi screens.

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

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

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 included 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 druggable and the genome-wide Salmonella screens.

1616
Comparison of the base and stability version of the rank-sum test ranking

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

<table>
<thead>
<tr>
<th>Gene set</th>
<th>Total</th>
<th>Overlap</th>
<th>Rank sum</th>
<th>Stability rank sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Re-screened</td>
<td>Validated, n (%)</td>
</tr>
<tr>
<td>Notch top 1%</td>
<td>129</td>
<td>124</td>
<td>120</td>
<td>68 (56.7)</td>
</tr>
<tr>
<td>Notch top 10%</td>
<td>1281</td>
<td>1160</td>
<td>235</td>
<td>148 (63)</td>
</tr>
<tr>
<td>SalD top 1%</td>
<td>69</td>
<td>64</td>
<td>47</td>
<td>22 (46.8)</td>
</tr>
<tr>
<td>SalD top 10%</td>
<td>686</td>
<td>618</td>
<td>125</td>
<td>41 (32.8)</td>
</tr>
<tr>
<td>SalGW top 1%</td>
<td>148</td>
<td>118</td>
<td>67</td>
<td>7 (10.4)</td>
</tr>
<tr>
<td>SalGW top 10%</td>
<td>1478</td>
<td>1162</td>
<td>113</td>
<td>12 (10.6)</td>
</tr>
</tbody>
</table>

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 a 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 4). 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 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 IPPhd [2009/025] and RTD [2009/005] (InfectX), evaluated by the Swiss National Science Foundation.

**Conflict of Interest**: none declared.

**REFERENCES**


