HiTSelect: a comprehensive tool for high-complexity-pooled screen analysis

Aaron A. Diaz¹,²,³,†, Han Qin²,⁴,⁵,†, Miguel Ramalho-Santos²,⁴,⁵,* and Jun S. Song¹,²,³,*

¹Institute for Human Genetics, University of California, San Francisco, CA, USA, ²The Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California, San Francisco, CA, USA, ³Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA, ⁴Departments of Obstetrics and Gynecology and Pathology and Center for Reproductive Sciences, University of California, San Francisco, CA, USA and ⁵Diabetes Center, University of California, San Francisco, CA, USA

ABSTRACT

Genetic screens of an unprecedented scale have recently been made possible by the availability of high-complexity libraries of synthetic oligonucleotides designed to mediate either gene knockdown or gene knockout, coupled with next-generation sequencing. However, several sources of random noise and statistical biases complicate the interpretation of the resulting high-throughput data. We developed HiTSelect, a comprehensive analysis pipeline for rigorously selecting screen hits and identifying functionally relevant genes and pathways by addressing off-target effects, controlling for variance in both gene silencing efficiency and sequencing depth of coverage and integrating relevant metadata. We document the superior performance of HiTSelect using data from both genome-wide RNAi and CRISPR/Cas9 screens. HiTSelect is implemented as an open-source package, with a user-friendly interface for data visualization and pathway exploration. Binary executables are available at http://sourceforge.net/projects/hitselect/, and the source code is available at https://github.com/diazlab/HiTSelect.

INTRODUCTION

RNA interference (RNAi) provides a powerful technique for gene knockdown by exploiting a cell’s endogenous machinery for mRNA degradation. RNA-induced silencing complexes target mRNAs via short oligonucleotide guide strands excised from short-hairpin RNA (shRNA). Loss-of-function analysis via high-complexity shRNA screening has successfully identified pathways associated with cancer (1–4), modulators of ricin susceptibility (5), regulators of protein degradation (6) and barriers to cellular reprogramming (7). Additionally, the clustered regularly interspaced palindromic repeats (CRISPR) pathway has been adapted from the bacterial and archaeal immune systems for gene knockout in eukaryotic cells via synthetic single guide-RNA (sgRNA). CRISPR-based screening in human cells has been shown to scale genome-wide effectively (8,9).

A typical approach is to use barcoded lentiviral vectors to express genome-wide libraries of oligonucleotide guide-RNA (shRNA or sgRNA). Cells are infected in a pooled fashion. Subsequently, cells exhibiting a phenotype of interest (often in response to a treatment) are identified and isolated. Barcodes are recovered from both phenotype-positive and phenotype-negative cells via polymerase chain reaction (PCR) amplification and quantified via next-generation sequencing (NGS). The relative abundance of guide-RNA in the phenotype-positive population, as measured by the abundance of reads mapping to the corresponding barcode, allows one to infer the effect of inhibiting the targeted gene on phenotype. Despite the power of genome-wide screens, off-target effects, variance in gene silencing efficiency and variance in sequencing depth of coverage can be significant, making it difficult to distinguish knockdowns/knockouts that truly regulate a given phenotype from background noise (10). Moreover, identifying relevant genes and pathways not only requires rigorous statistical methods for hit selection but also meta-analysis of relevant, secondary genomic data such as gene expression, chromatin state, known physical interactions and functional annotations. Thus, the enormous potential of pooled genome-wide screens is currently hampered by limitations in the methodologies for analysis. To address this roadblock, we developed a computational pipeline for analyzing pooled, high-complexity screens and implemented it into a user-friendly software package called HiTSelect (Supplementary Software). The
HiTSelect pipeline includes comprehensive modules for screen hit selection based on robust statistics, a module for the integration of gene expression data, a module for functional annotation analysis, a module for genetic interaction analysis and a module for gene network visualization (Figure 1). We illustrate the power of HiTSelect using the data from our recent screen for genes that function as barriers to reprogramming somatic cells to induced pluripotent stem (iPS) cells, as well as another recently published data set from a screen for growth factors in ovarian carcinoma (4,7). We compare the gene-ranking module of HiTSelect to both the RNAi Gene Enrichment Ranking (RIGER) and the Redundant siRNA Activity (RSA) algorithms (11,12). We show that HiTSelect’s ranking algorithm is both more sensitive and more specific, HiTSelect’s predictions correlate better with secondary validation assays and HiTSelect is less prone to off-target effects than RSA or RIGER. While RSA and RIGER were designed for multi-well plate-based screens, and/or pooled screens with microarray read-out, HiTSelect is specifically designed for high-complexity shRNA and CRISPR screens with NGS readout. In addition, HiTSelect provides tools for metadata integration, analysis and visualization that are not available in any other software.

**MATERIALS AND METHODS**

HiTSelect’s gene ranking algorithms and statistics

Our hit selection method was developed with two concerns in mind: dealing with off-target effects and handling variances in sequencing depth-of-coverage and in knockdown/knockout efficiency across guide-RNA, which can be either shRNA in RNAi screens or sgRNA in CRISPR screens. To address the former, we use multi-objective optimization to identify genes with multiple highly active guide-RNA. To address the latter, we utilize a random effects model to assess sampling variance in the estimation of gene effect size. We focus on a two-group problem and assume that guide-RNA read counts have been tabulated from a phenotype positive sample $S^+$ and a phenotype negative control sample $S^-$. Let $N^+ \text{ and } N^-$ denote the total numbers of sequenced reads from each population, respectively. For clarity, we use Greek letters to index guide-RNA and Roman letters to index genes.

To avoid division by zero in the following formulae, we add a pseudo-count of 1 read to each count. Let $r_{a+}$ denote the read count of the $a^{th}$ guide-RNA in $S^+$. Let $r_{a-}$ denote the read count in $S^-$ normalized by sequencing depth $N^+/N^-$. If the user selects the option to normalize by a user-provided list of control sequences, then we instead normalize by the ratio $M^+/M^-$, where $M^+$ and $M^-$ are the median read counts for the control sequences from $S^+$ and $S^-$, respectively. We model the activity level of the $a^{th}$ guide-RNA in the positive population by the read count log odds ratio (log-odds) $X_a$ that we estimate by the sample log-odds

$$\hat{X}_a = \ln \left[ \frac{p_a / (1-p_a)}{q_a / (1-q_a)} \right].$$

Here $p_a = r_{a+} / N^+$ and $q_a = r_{a-} / N^-$. We designate a guide-RNA as ‘active’ if $\hat{X}_a > 0$ and $r_{a+} > m$, where $m$ is the sample median absolute deviation of the read counts in $S^+$. The latter criterion eliminates undersequenced guide-RNA. We call a guide-RNA active if it has greater odds of being sequenced in the treatment population relative to control. Given a gene $G_j$, targeted by $n$ distinct guide-RNA $\{h_{j,\alpha} \}_{\alpha=1}^n$ with log-odds $\{X_{j,\alpha} \}_{\alpha=1}^n$, we wish to combine them into a single measure of gene effect size. If $\hat{h}_a$ had perfect efficiency in its knockdown/knockout of $G_j$, reads mapping to $\hat{h}_a$ would occur in $S^+$ with some log-odds $Y_a$ determined solely by $G_j$’s effect on phenotype. In practice, there will be less than perfect efficiency and reads will occur with log-odds $X_a = Y_a + Z_a$, for some $Z_a$ to be determined. Thus, var($Z$) gives the variance in knockdown/knockout efficiency across guide-RNA targeting $G_j$. The $Q$-statistic (13) of the residuals $X_a$ allows us to assess the homogeneity of knockdown/knockout efficiency across $\{h_{j,\alpha} \}_{\alpha=1}^n$ as well as compute an estimator of var($Z$). Given that $\hat{X}_a$’s standard error (SE) is $s_a = \sqrt{\frac{1}{r_a+} + \frac{1}{N^- r_a^-} + \frac{1}{r_a+} + \frac{1}{N^- r_a^-}}$ by normal approximation (in the case of sequencing depth normalization; for control sequence normalization, replace this formula with $s_a = \sqrt{\frac{1}{r_a+} + \frac{1}{N^- r_a^-} + \frac{1}{M+ r_a+} + \frac{1}{M^- r_a^-}}$), we have

$$Q = \sum_{a=1}^n u_a \left( \hat{X}_a - \bar{X} \right)^2$$

where $\bar{X} = \sum_{a=1}^n u_a \hat{X}_a / \sum_{a=1}^n u_a$ and $u_a = 1/s_a^2$. Notice that the formula for the log-odds standard error is a decreasing function of increasing depth-of-coverage. From $Q$, DerSimonian and Laird (14) derive estimators of var($Z$), given by

$$\Delta^2 = \frac{Q-k-1}{\sum_{a=1}^n u_a - \sum_{a=1}^n u_a / \sum_{a=1}^n u_a}.$$
and \( Y_j \), given by \( \hat{Y}_j = \frac{1}{n} \sum_{a=1}^{n} w_a \), Here \( w_a = \frac{1}{\sigma^2 + \lambda^2} \) and \( \text{SE}(\hat{Y}_j) = \left( \sum_{a=1}^{n} w_a \right)^{-1/2} \). Thus, we model the effect size for gene \( G_j \) by the summary log-odds \( Y_j \) that we estimate via \( \hat{Y}_j \). This model accounts for variance in knockdown/knockout efficiency across guide-RNA via \( \sigma^2 \) and variance in the sequencing depth of individual guide-RNA via \( \lambda^2 \). We now describe how both the estimate of gene effect size \( \hat{Y}_j \) and its variance estimate \( \text{SE}^2(\hat{Y}_j) \) are used to formulate a probabilistic ranking scheme.

If we denote the number of active guide-RNA targeting gene \( G_j \) as \( t_j \), then the pair \((t_j, Y_j)\) characterizes the evidence for \( G_j \) being a hit. We want to preferentially rank genes whose knockdown/knockout exhibits a large effect \( \hat{Y}_j \) is high) and for which that effect is reproducible \( \hat{t}_j \) is high). To this end we developed a multi-objective optimization algorithm to maximize both quantities simultaneously. Our method for gene ranking uses the concept of Pareto dominance. A \( k \)-tuple \( A = (A_1, \ldots, A_k) \) of real numbers measuring \( k \) criteria for ranking is said to dominate another \( k \)-tuple \( B = (B_1, \ldots, B_k) \) if: (i) for all \( l \), \( B_l \leq A_l \), and (ii) there exists at least one \( l \in \{1, \ldots, k\} \) such that \( B_l < A_l \). If these conditions hold, then we write \( A \preceq B \). Since this relationship defines only a partial ordering on the set of \( k \)-tuples, it is possible that neither \( A \preceq B \) nor \( B \preceq A \). We will write \( A \equiv B \) in the case that \( A \) and \( B \) are not comparable. The gist of our algorithm will be to allow genes to compete for Pareto dominance in the multi-objective criterion \((t_j, Y_j)\).

A Pareto dominant gene is a superior candidate for a hit, since it shows an increase in summary guide-RNA log-odds, in the number of distinct active guide-RNA, or both. We therefore score genes by the number of other genes that they dominate. In the absence of uncertainty, our method would be to give each gene 1 point for every gene it dominates and 0.5 point for every gene to which it is not comparable. We then rank genes by the number of points they have. In this way, genes compete for dominance in the ranked list. This algorithm can be summarized as:

**ALGORITHM 1: Deterministic multi-objective ranking**

(i) For each gene \( G_j \):
   (a) Set \( \text{score}_j = 0 \)
   (b) For each gene \( G_i \in R, i \neq j \):
      (1) If \( (t_i, Y_i) \prec (t_j, Y_j) \), then \( \text{score}_j = \text{score}_j + 1 \)
      (2) If \( (t_i, Y_i) \equiv (t_j, Y_j) \), then \( \text{score}_j = \text{score}_j + 0.5 \)
   (ii) Rank genes based on their scores.

To make our method more robust, we incorporated uncertainty in the estimation of \((t_j, Y_j)\), using a probabilistic version of this algorithm. For the independent random variables \( A = (A_1, \ldots, A_k) \) and \( B = (B_1, \ldots, B_k) \), we can define \( p(B < A) = \prod_{i=1}^{k} p(E[B_i] < E[A_i])|a_i, b_i| \), where \( a_i \) and \( b_i \) are sampled values of \( A_i \) and \( B_i \), respectively. In order to be able to compute \( p(E[B_i] < E[A_i])|a_i, b_i| \), we need to make additional assumptions about how \( A_i \) and \( B_i \) are distributed. In the case of normal distributions with unknown means and known variance \( \sigma^2_A \) and \( \sigma^2_B \), it can be shown (15) that

\[
p(E[B_i] < E[A_i])|a_i, b_i| = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sigma_A \sqrt{2\pi}} e^{-\frac{(a_i-x)^2}{2\sigma_A^2}} \frac{1}{\sigma_B \sqrt{2\pi}} e^{-\frac{(x-b_i)^2}{2\sigma_B^2}} dy \, dx.
\]

Hughes (16) demonstrates that this formula can be further reduced to

\[
p(E[B_i] < E[A_i])|a_i, b_i| = \frac{1}{\pi} \text{erf}(\frac{b_i-a_i}{\sqrt{\sigma_A^2 + \sigma_B^2}}) \approx \frac{1}{\sigma_A \sqrt{2\pi}} e^{-\frac{(a_i-b_i)^2}{2\sigma_A^2}}.
\]

In general, guide-RNA pools targeting different genes will have different knockdown efficiencies. Modeling the binary activities of guide-RNA pools with a single value \( p \) is not sufficient. We can use the above formulas to estimate \( p(G_i > G_j) \equiv p(E[Y_i] < E[Y_j])|\hat{Y}_i, \hat{Y}_j| \) for the gene pair \( G_i \) and \( G_j \). We can compute \( p(E[Y_i] < E[Y_j])|\hat{Y}_i, \hat{Y}_j| \) by using the above formulas. We then rank the number of active guide-RNA, \( t_j \), as Poisson distributed, and \( p \) as a function of Poisson distributed, modeling the binary activities of guide-RNA pools targeting a given gene \( j \) as independent identically distributed Bernoulli random variables (17). Thus, to compute \( p(E[t_i] < E[t_j])|\hat{t}_i, \hat{t}_j| \) we need to compute the probability that \( \hat{t}_i \) is drawn from a Poisson distribution with mean less than that of the Poisson distribution from which \( \hat{t}_j \) was drawn. The test statistic

\[
m' = \frac{\hat{t}_i - \hat{t}_j}{\sqrt{\text{SE}(\hat{t}_i) + \text{SE}(\hat{t}_j)}}
\]

is asymptotically standard normal under the null hypothesis of equal Poisson means (18) and has been shown to be more powerful than an exact test in determining whether two samples were drawn from Poisson distributions with the same rate (19). Thus, we replace \( m' \) above with \( m \) when we compute \( p(E[t_i] < E[t_j])|\hat{t}_i, \hat{t}_j| \). In general, guide-RNA pools targeting different genes will have different complexities, both in the original viral library and in the initial population of infected cells. We apply the Anscombe transform \( x \rightarrow 2\sqrt{\frac{x+3}{8}} \) to \( \hat{t}_i \) and \( \hat{t}_j \), before computing \( m \), to stabilize variance in sampling across genes targeted by guide-RNA pools of varying degrees of complexity (20). This has the practical effect of producing a more uniform Pareto dominance estimate for small differences in the number of observed active guide-RNA (Supplementary Figure S1). Note that once we define \( p(G_i > G_j) \) we can immediately define the probability of non-comparability: \( p(G_i \equiv G_j) = 1 - p(G_i > G_j) - p(G_j > G_i) \). Let \( R \) denote the set of all genes that have at least one active guide-RNA. We compute all pairwise Pareto dominance probabilities between genes in \( R \) and assign a score to gene \( G_j \) via the formula \( \text{score}_j = \sum_{i \in R \cap \{j\}} p(G_i > G_j) + 0.5 \sum_{i \in R \cap \{j\}} p(G_i < G_j) \). We then rank genes by their score. Our algorithm, as implemented in HiSelect, can be summarized as follows:

**ALGORITHM 2: STOCHASTIC MULTI-OBJECTIVE RANKING**

(i) For each gene \( G_j \in R \):
   (a) Set \( \text{score}_j = 0 \)
   (b) For each gene \( G_i \in R; i \neq j \):
      (1) \( \text{score}_j = \text{score}_j + p(G_i > G_j) \)
Joint-spectral network centrality for gene interaction networks

Given a set of genes of interest \( \{G_j\}_{j=1}^k \), we wish to summarize multiple lines of evidence for their interaction into a single metric of a gene’s influence. HiTSelect interfaces with GeneMANIA (http://genemania.org) to aggregate genetic interaction (the perturbation of one gene causing a change in expression of a second gene), physical interaction (protein–protein interaction) and co-localization (genes expressed in the same tissue) experimental data. Suppose we have \( r \) total types of interaction, and for each interaction type, we have a measurement of the interaction strength \( x_{ij}^m \), \( m = 1, \ldots, r \), between genes \( G_i \) and \( G_j \) for the \( m \)th interaction type. We want to summarize these interaction networks. HitSelect computes the interaction strengths, \( x_{ij}^m \), using GeneMANIA. We then replace each \( x_{ij}^m \) with its network-wide quantile \( q_{ij}^m \), i.e., its quantile estimated from the set of numbers \( \{ x_{ij}^m \}_{j=1}^k \). This is done so that interaction strength estimates can be made comparable across networks. In matrix notation we write \( A_m = [a_{ij}^m] \). We formulate a statistic on the set of graphs with common nodes labeled by the genes \( \{G_j\}_{j=1}^k \), and vertex sets that have the \( [A_m]_{m=1}^r \) as their adjacency matrices. We refer to this ensemble of networks as the joint network.

Bonacich first identified the components of the dominant eigenvector of a connected network’s adjacency matrix as a measure of network centrality for the corresponding nodes (21). This eigenvector solves \( A_m x = \lambda_m x \), where \( \lambda_m \) is the largest eigenvalue of \( A_m \) in absolute value. We can extend network centrality to an estimate of joint-network centrality by solving the following optimization problem:

\[
\min_x \sum_{i=1}^r \| A_i x - \lambda_i x \|_2 + \| 1 - \| x \|_2 \|_2.
\]

The first term assures \( x \) is an approximate eigenvector for each of the \( r \) networks. The second term, \( \| 1 - \| x \|_2 \|_2 \), assures convergence of iterative solution algorithms to a non-trivial solution by keeping iterates on the unit sphere and away from the origin.

In the absence of this regularization term, the origin would be a basin of attraction. This restriction is acceptable, since we are only interested in the relative importance of genes to a particular pathway. HiTSelect solves this optimization problem using a sequential-quadratic program, with the centroid of the dominant eigenvectors of \( \{A_m\}_{m=1}^r \) (normalized to unit length) as a starting value. As a termination criterion, HiTSelect uses the criterion that the objective function is \( \leq 10^{-8} \) or the algorithm has performed more than 400 iterations.

Comparison of HiTSelect to the RIGER and RSA methods

We compared the gene-ranking module of HiTSelect to RIGER and RSA on readout from two independent high-complexity shRNA screens: (i) Tan et al. screened mesenchymal (HeyA8) and stem-like (Stem-A) ovarian carcinoma cell subpopulations for genes responsible for growth and proliferation (4). Pooled libraries of 80 000 clones targeting 16 000 genes were introduced in the HeyA8 and Stem-A subpopulations separately. After 14 days, shRNA copy numbers were quantified in each population via NGS. The authors used the Stem-A population as a control to assess enrichment in the HeyA8 population. We here refer to HeyA8 as the treatment population. (ii) We co-infected human fibroblasts with lentivirus expressing a genome-wide high-coverage (600 000 shRNAs with \( \sim 30 \) shRNA/gene) shRNA library along with SOX2, KLF4, OCT4 and c-MYC reprogramming factors and an shRNA targeting P53 (7).

Following the appearance of colonies with IPS cell characteristics 28 days post induction, the transduced cells were FACS-purified for TRA-1-81, a marker of fully reprogrammed cells. Then, shRNAs recovered from the TRA-1-81 positive (TRA-1-81+) and TRA-1-81 negative (TRA-1-81−) populations were sequenced (Supplementary Table S1). Using the TRA-1-81− population as a control, shRNA enrichment in the TRA-1-81+ population (which we call the treatment population) was determined. We chose these two screens as test data sets, because both used genome-wide high-complexity libraries, used NGS as readout and performed extensive screen validation assays that we can use to estimate the sensitivity and specificity of hit selection algorithms. RIGER analysis was performed using the RIGER.jar Java archive downloaded from the GENE-E website (http://www.broadinstitute.org/cancer/software/GENE-E/extensions.html). As originally described by Luo et al. (12), this implementation uses a Kolmogorov–Smirnov test for significance assessment. RSA was designed for well-based screens (22), and codes for RSA analysis are available (http://carrier.gnf.org/publications/RSA/).

However, the algorithm itself can be adapted to NGS screen readout by replacing well activity level with barcode read count odds ratio. We implemented RSA for NGS readout in MATLAB, using an odds ratio of 1 for the lower-bound parameter and 2 as an upper-bound parameter.

Following their primary screen, Tan et al. (4) validated 135 genes using short interfering RNA (siRNA). Tan et al., having identified ovarian carcinoma subtypes via microarray and clinicopathological parameters, chose these genes with the intention of specifically targeting pathways enriched in the Stem-A subtype. One siRNA was transfected per well in a 96-well plate format. After 96 h of incubation, a colorimetric cell-proliferation assay was performed to quantify siRNA effect on growth. The assay was performed in quadruplicate, and there were two negative-control wells per plate. We compared siRNA-well to control-well proliferation ratios (obtained from the authors) between the HeyA8 and Stem-A populations using single-tailed \( t \)-tests, corrected for multiple-hypothesis testing via the Benjamini–Hochberg method (23). We found 65 genes in HeyA8 cells and 47 genes in Stem-A cells with differential negative effect on cell growth at a \( t \)-test \( P \)-value cutoff of \( P = 0.05 \). We then tabulated true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN) gene-hit calls for HiTSelect, RIGER and RSA. Hit calls were tabulated in the following fashion: we consider a gene TP if it is ranked within the top 5% of genes in the primary screen of HeyA8 and, in the validation assay, has a mean proliferation ratio which is higher in HeyA8 at a \( t \)-test \( P \leq 0.05 \); we call a
gene FP if it is enriched in the HeyA8 primary screen but not in the validation assay ($P > 0.05$). Similarly, if a gene is ranked within the top 5% in the screen of Stem-A and has a higher proliferation ratio in the Stem-A validation with $P \leq 0.05$, then it is a TN. If it is enriched in Stem-A’s primary screen but fails to validate ($P > 0.05$), then we call it an FN. Sample sizes were unbalanced in the validation assay (65 genes down in HeyA8 cells versus 47 genes down in Stem-A). We chose the Matthews correlation coefficient as an estimator of an algorithm’s predictive value, since it is generally considered to be the most robust estimator of a classifier’s success for unbalanced sample sizes (24). We also chose balanced accuracy (BA) as a second measurement of these algorithms’ performance as classifiers. Balanced accuracy ($BA = \frac{TP + TN}{TP + TN + FP + FN}$) is a renormalized accuracy that avoids overestimation of classifier performance arising when the sample is imbalanced (25).

Lastly, we assembled lists of negative-control genes and positive-control genes for our iPSC cell screen and compared their rank distributions between algorithms. In the iPSC cell screen, we screened for barriers to reprogramming, i.e. genes whose knockdown enhances reprogramming efficiency. As negative controls, we used the validated targets of known tumor suppressor micro-RNA (miRNA) miR-218 (26), a random gene list and explicit negative-control sequences. As positive control genes, we used a list of 26 barrier genes whose knockdown has been already shown to enhance reprogramming efficiency (Supplementary Table S2). Additionally, we used the experimentally validated targets of miR-17 (whose targets include the tumor suppressor P21) and miR-200 (an inhibitor of the epithelial-to-mesenchymal transition) as positive controls. Both of these miRNA have been shown to enhance reprogramming efficiency (27,28). We also found that HiTSel outperformed RSA and RIGER on simulated data by computing a Receiver Operating Characteristic curve shown in Supplementary Figure S2.

**Single-gene knockdown validation**

BJ fibroblasts were seeded on a 6-well plate and infected with lentivirus expressing shRNA the following day. Each shRNA was cloned into a separate pSicoR-CMV-Puro-T2A-GFP lentiviral vector. To make lentivirus, 293T cells expressing guide-RNA (the hyper-geometric test). Thus, in principle, RSA controls for off-target effects, and it works well for smaller-scale screens on the order of hundreds of wells. However, we find that at the scale of the screens we considered (80 000 and 600 000 shRNA), the hyper-geometric test is not very sensitive. In fact, in both screens, RSA enriches for ‘singletons’, i.e. genes with only one active shRNA. By contrast, HiTSel preferentially ranks genes for which the effect of knockdown is reproducible across multiple shRNA (Figure 2A and B). Compared to HiTSel, RSA’s predictions for the primary screen of Tan et al. do not correlate as well with the results of their siRNA validation assay (Figure 2C). In comparing HiTSel to RSA and RIGER in our iPSC cell screen, we found that the upper-quartiles of the HiTSel and RIGER gene rankings were enriched for positive-control genes to a statistically significant extent (hypergeometric tests: $3.3 \times 10^{-5} \leq P \leq 2.4 \times 10^{-3}$), indicating good sensitivity in hit detection (Figure 2D). However, RSA was not very sensitive and did not rank positive control genes highly ($0.17 \leq P \leq 0.7$). While neither HiTSel nor RSA frequently ranked negative-control genes in the upper quartile range ($0.16 \leq P \leq 0.68$), RIGER ranks control sequences highly ($P = 1.5 \times 10^{-2}$). Thus, RIGER’s false-positive rate is high for some samples. RIGER also enriches for singletons (Figure 2A and B), and RIGER has the least correlation between the primary screen and the siRNA validation assay of Tan et al. (Figure 2C). Furthermore, RIGER ranks negative control sequences highly in our iPSC cell screen (Figure 2D), indicating that RIGER’s FDR is high. Figure 2E shows the overlap of the three algorithms.

In a typical workflow, a researcher may first call hits at the 5% level, in order to have a large enough sample to perform functional annotation clustering and gene network analysis. After obtaining relevant pathways, one may then choose genes from those pathways for validation using a more stringent threshold, such as 1%. At this threshold, the distributions of the number of active guide-RNA are shifted to the left for RSA and RIGER compared to HiTSel in both screens. In particular, RSA and RIGER preferentially rank singletons in the top 1%. HiTSel’s top 1% ranked genes all have at least two (and most have three or more) active shRNA. Thus, HiTSel is less prone to off-target effects since, for all genes ranked in the top 1%, the effect of knockdown is reproducible over multiple guide-RNA. To compare the impact of off-target effects on hit
selection of the three algorithms, we spot-checked a handful of gene hits for ineffective shRNA and also employed the software GESS (29) to identify off-targets attributable to ‘seed effects’. Firstly, we identified the seven most highly ranked singleton genes that were mutually ranked in the top 1% by RIGER and RSA in our iPSC cell screen. None of HiTSellect’s top 1% are singletons. We measured the knockdown efficiency of the single active shRNA for each single gene via reverse transcription quantitative real-time polymerase chain reaction (RT-qPCR), comparing the effect of RNAi to that of a nonsense-sequence control. We then compared these knockdown effect sizes to those observed in knockdown experiments of genes identified in the top 1% by HiTSellect. All of HiTSellect’s top 1% hits have at least two active shRNA, and we performed knockdowns using the two most active shRNA (identified by HiTSellect) separately and averaged the results. All knockdowns were performed in human fibroblasts. The median knockdown efficiency was slightly greater in HiTSellect’s picks (log2 fold-change = −2) than in the RSA/RIGER singletons (log2 fold-change = −1.8), and two of the seven RSA/RIGER singletons did not show significant knockdown efficiency (<2-fold change compared with nonsense-sequence control) (Figure 3A). To explore whether this 29% off-target rate observed in the RSA/RIGER hits was an artifact of the small sample size used, we computed the ‘seed-match frequencies’ for the top iPSC cell screen hits using GESS. GESS identifies shRNAs that may exhibit off-target effects.
by computing the frequency with which the seed sequences in those shRNAs match known transcripts in a database. GESS produces Benjamini–Hochberg-corrected P-values for seed-match frequency (SMF). We ran GESS on the top 1% hits for RSA, RIGER and HiTSelect, using the bottom 1% as negative controls. GESS did not report any HiTSelect hit to have a statistically significant SMF at the GESS P-value cutoff of 0.05. However, 17% of shRNA from RIGER and 35% of shRNA from RSA show a statistically significant SMF (Figure 3B). Taken together, these findings show that HiTSelect is less prone to off-target effects than RSA or RIGER.

HiTSelect provides an integrative analysis suite with visualization tools

While a good ranking algorithm is a prerequisite for robust screen analysis, identifying key pathways also requires going beyond the statistical significance of individual genes and integrating relevant metadata. To this end, we developed modules to automate data aggregation, calculate statistics for data synthesis and visualize the data and results. HiTSelect’s visualization module allows the user to seamlessly browse and search screen results (Supplementary Figure S3A). Genes of interest can be annotated with user-defined gene lists, which can then be exported to a delimiter-separated file or analyzed with other HiTSelect modules. For example, HiTSelect’s functional annotation module interfaces programmatically with DAVID (30) to generate reports presented as interactive visualizations (Supplementary Figure S3B). An example, obtained from clustering the functional annotations of the ips cell screen hits, is available online: http://song.igb.illinois.edu/ipsScreen/docs/david_treemap.html. HiTSelect’s gene network analysis module interfaces with GeneMANIA (http://www.genemania.org), which contains annotations for genetic interactions, physical interactions and tissue co-localization of expression data aggregated from public databases. As a summary statistic for GeneMANIA output, we developed the concept of joint-network centrality. Joint-network centrality generalizes eigenvalue network centrality to multiple networks defined on the same set of genes. Lastly, we developed a Cytoscape interface module to visualize screen hit network interactions with an overlay of metadata from the above modules. Together, these modules provide a powerful and user-friendly system for screen analysis.

HiTSelect is suitable for both shRNA and CRISPR-mediated screens

To demonstrate the versatility of our approach, we use HiTSelect to analyze four screens recently performed via CRISPR knockout and reported by two separate groups (8,31). Wang et al. (8) screened for genes necessary for survival in HL60 and KBM7 cells expressing Cas9 nuclease. Cell populations were infected with barcoded lentivirus expressing a library of 73 151 sgRNA, targeting 7114 genes. After an incubation period of 12 cell doublings, barcodes were recovered via PCR and sequenced. We analyzed these screens’ data using HiTSelect and identified the top 5% of genes most under-represented after the incubation period (Figure 4A and B, left panel). We found that while HiTSelect and RIGER agreed better in the shRNA screen of Tan et al. (4) (Figure 2E), HiTSelect overlapped better with RSA in the CRISPR screen of Wang et al. (8) (Supplementary Figure S4C). Using a combination of HiTSelect’s network analysis module and the MCODE plugin to Cytoscape (32), we then studied interactions between screen hit genes. Strikingly, both screens reproduced four sub-networks comprised of genes associated with the ribosome, proteasome, DNA replication and transcription. In particular, the median joint-network centrality for genes in the ribosome sub-network is in the 97th percentile of joint-network centrality scores overall.

Figure 4. HiTSelect analysis of CRISPR-mediated negative selection screens in HL60 and KBM7 cells, respectively. Genes ranked in the top 5% by HiTSelect’s gene ranking algorithm (A, B: left panels) are enriched for functional annotations for fundamental biological processes, such as transcription, translation, DNA replication and cell cycle (C). These annotations were discovered using HiTSelect’s functional annotation analysis module. In both screens, analysis using a combination of HiTSelect’s gene-interaction module and MCODE identifies dense clusters of genes associated with the ribosome, proteasome, DNA replication and transcription. In particular, the median joint-network centrality for genes in the ribosome sub-network is in the 97th percentile of joint-network centrality scores overall.
tackling ribosomal proteins. Although Wang et al. also identified enrichment for ribosomal, proteasomal and DNA replication genes, the sub-network of transcription-related genes, which contains mediator complex components and other genes essential for cell survival, was not reported to be statistically significant. Also not previously reported, but enriched in our HiTSelect analysis, are the clusters of genes that are known to be essential for cell survival: cell cycle genes (e.g. components of the anaphase-promoting complex and cyclin-dependent kinases), components of the cytoskeleton and elements of the ubiquitination pathway (Figure 4C). Identification of these additional, essential pathways further validates the sensitivity of HiTSelect.

Likewise, Zhou et al. (31) created a library of 873 sgRNA targeting 291 genes. Two screens were then performed in OCT1 and Cas9 expressing HeLa cells (HeLaOC) to identify targeting 291 genes. Two screens were then performed in OCT1 and Cas9 expressing HeLa cells (HeLaOC) to identify genes necessary for intoxication by chimeric anthrax (PA/LFnDTA) and diphtheria toxin (DT). HeLaOC cells were infected with barcoded lentivirus expressing the aforementioned sgRNA library and then treated with each toxin. The treatment was followed by PCR amplification and sequencing of the vectors’ barcodes in surviving cells. We analyzed these screens using HiTSelect and identified 14 and 16 genes with a non-zero effect size (log-odds >0) in the PA/LFnDTA and DT screens, respectively (Supplementary Figure S4A and B). All seven of the genes validated by Zhou et al. (via a cell viability assay) were identified in HiTSelect’s picks. In particular, anthrax receptor (ANTXR1) and diphtheria toxin receptor (HBEGF) were the number 1 genes in HiTSelect’s ranking of the PA/LFnDTA and DT screens, respectively. Moreover, knockout of these two genes in Zhou et al.’s validation assay produced the greatest increases in cell viability over control, following treatment with PA/LFnDTA or DT, respectively. Due to the small number of gene hits in the screens of Zhou et al. (14–16 genes), DAVID did not identify any over-represented functional annotation terms, and GeneMANIA also did not identify significant interactions among these genes, limiting our re-analysis of the screens. HiTSelect thus provides a rigorous analysis method for selecting hits that correlate well with validation. In addition to these cell-surface receptors, HiTSelect also identifies adenosine triphosphate (ATP) binding proteins in the PA/LFnDTA screen: CFTR and PECR. PECR knockout was validated by Zhou et al. as conferring resistance to PA/LFnDTA, but CFTR was not. Another bacterial toxin, cholera toxin, induces the secretion of intestinal fluid via CFTR upregulation (in an ATP-mediated fashion); thus, HiTSelect revealed that the role of CFTR in anthrax toxicity is a potential avenue for further investigation (33).

**DISCUSSION**

High-throughput, genome-wide screening is a powerful technique to identify genes and pathways regulating a phenotype of interest. However, statistical approaches designed for smaller-scale screens, performed at low throughput in multi-well plates and/or screens using microarrays as readout, do not completely control for variation in knockdown/knockout efficiency or variation in sequencing depth-of-coverage. Thus, analysis techniques that are robust to these random effects are currently lacking. Moreover, there are no bioinformatics tools available to perform integrative analysis of screen metadata. To address these needs, we have developed HiTSelect. In addition to providing a rigorous hit selection method that compares very favorably with existing approaches, HiTSelect is a comprehensive tool with gene network analysis, gene ontology analysis and visualization modules. HiTSelect is free, open source and available as a stand-alone software package for Windows, MacOS and Linux. As exemplified by the algorithm comparisons and data analyses reported here, HiTSelect is expected to enable a high degree of confidence in the interpretation of genome-wide screen data and maximize the depth of biological insights arising from these kinds of screens.

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

**FUNDING**

CIRM [RB4-06028 to M.R.-S.]; National Institutes of Health (NIH) Director’s New Innovator Award [DP2OD4698 to M.R.-S.]; Sontag Foundation Distinguished Scientist Award [to J.S.S.]; NIH [R01CA163336 to J.S.S.]. Funding for open access charge: NIH [R01CA163336].

**Conflict of interest statement.** None declared.

**REFERENCES**


