Published in final edited form as:
Cell Syst. 2021 April 21; 12(4): 338–352.e5. doi:10.1016/j.cels.2021.03.001.

SIGNAL: A web-based iterative analysis platform integrating pathway and network approaches optimizes hit selection from high-throughput assays

Samuel Katz[1,2], **Jian Song**[1], **Kyle P. Webb**[1], **Nicolas W. Lounsbury**[1], **Clare E. Bryant**[2], **Iain D.C. Fraser**[1,*]

[1]NIAID, National Institutes of Health, Laboratory of Immune System Biology, Bethesda, MD, 20892, USA

[2]University of Cambridge, Dept. of Veterinary Medicine, Cambridge, United Kingdom

Summary

Hit selection from high throughput assays remains a critical bottleneck in realizing the potential of omic-scale studies in biology. Widely used methods such as setting of cutoffs, prioritizing pathway enrichments, or incorporating predicted network interactions offer divergent solutions yet are associated with critical analytical tradeoffs. The specific limitations of these individual approaches and the lack of a systematic way by which to integrate their rankings has contributed to limited overlap in the reported results from comparable genome-wide studies and costly inefficiencies in secondary validation efforts. Using comparative analysis of parallel independent studies as a benchmark, we characterize the specific complementary contributions of each approach and demonstrate an optimal framework by which to integrate these methods. We describe Selection by Iterative pathway Group and Network Analysis Looping (SIGNAL), an integrated, iterative approach which uses both pathway and network methods to optimize gene prioritization. SIGNAL is accessible as a rapid user-friendly web-based application (https://signal.niaid.nih.gov).

A record of this paper's Transparent Peer Review process is included in the Supplemental Information.

Keywords

high-throughput hit selection; genomics; genome-wide studies; prioritization; bioinformatics; pathway analysis; enrichment; network analysis; software

[*]Lead contact: fraseri@nih.gov.

Authors Contributions

S.K. and I.D.C.F. developed the initial premise for the software and designed the comparative analysis and validation protocol. S.K., J.S., K.P.W., and N.W.L. wrote the R scripts for SIGNAL analysis. S.K., J.S., and K.P.W. created the server-side scripts, web interfaces, and interactive features. S.K. and J.S. secured the web hosting and security protocols. S.K. formatted the HDF, RNAseq, and CRISPR studies for analysis and performed the comparative and validation studies. S.K. and K.P.W. implemented the statistical methods. C.E.B. provided input and mentorship during software development. S.K. and I.D.C.F. wrote the manuscript. All authors read and approved the final manuscript.

Declaration of Interests

The authors declare no competing interests.

Any additional information required to reproduce this work is available from the Lead Contact.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Introduction

High-throughput approaches - such as RNA and CRISPR-based screens, Next Generation sequencing methods, and proteomic analysis - permit the unbiased measurement of the contribution of each gene in the genome to the outcome of a specific biological process; these methods continue to be some of the most powerful tools in research biology (Heckl and Charpentier, 2015, Gilbert et al., 2014, Moffat et al., 2006, Lee et al., 2003). Yet, critical to the utility of the data they yield,is the ability to translate their results into a constrained list of prioritized candidates that can be rigorously investigated on a feasible scale. This bridging analysis step is significantly constrained by attempts to balance the challenges of analysing large datasets in an unbiased manner to excavate novel insights, with the appropriate recognition of known gene candidates considered as validation hits. This phenomenon underlies what can be described as the “quantum leap” in publications of high-throughput studies, when the analysis leaps -often with sparse analytical justification-from considering statistically prioritized lists of candidates to a handful of hits that are selected for further validation guided by the *a posteriori* knowledge of the authors (Lotterhos et al., 2016). Building on recent advances in statistical normalization methods such as edgeR (McCarthy et al., 2012), DESeq2 (Love et al., 2014), and MAGECK (Li et al., 2014), widely applied bioinformatic approaches following data normalization and candidate ranking can be categorized into three major classes; optimizing the setting of cutoffs, prioritizing based on the representation of pre-set gene groups or pathways, and expanding the list of hits based on predicted interaction networks (Birmingham et al., 2009, Tseng et al., 2012). Though these methods provide differing solutions to the challenge of candidate prioritization, their corrective approaches are often associated with analytical trade-offs relating to error correction, novelty identification, and interpretability. In addition to these challenges, two critical gaps persist; the absence of a systematic way by which these solutions can be collectively utilized such that the greatest additive benefit to hit selection accuracy is accrued, and analysis challenges for experimentalists who may lack the computational expertise required for their implementation.

At the outset of many hit selection pipelines, the setting of a single cutoff for defining an initial set of hits creates an intrinsic compromise between decreasing the false positive rate while increasing the false negative rate or vice-versa (Malo et al., 2006, Boutros and Ahringer, 2008). Whether choosing a cutoff that is stringent or lenient, the artificial rigidity of a single cutoff crucially obscures the more complex reality whereby targets identified by the screen exist on a spectrum of confidences with novel biological insights distributed across the range of assay scores or ranks (Ober, 2016). Since a more lenient cutoff strategy also makes the follow up experiments more labor-intensive (as a high number of candidates are likely to fail secondary screen validation), many studies rely on more stringent cutoffs and preferentially err on the side of a low false positive rate. This approach leaves a large portion of potential candidates unexplored and has been found by many meta-analysis studies to be a critical driver of the limited agreement between related studies (Rosenbluh et al., 2016, Bushman et al., 2009, Hao et al., 2013). These findings emphasize the need for improved hit selection strategies that might better capture lower-scoring potential hits while circumventing the rigidity of a single cutoff.

Pathway analysis is often implemented as an additional way to correct for false positives (Creixell et al., 2015), as randomly selected false-positive hits are less likely to be from the same pathway. Various grouping methods have been proposed as a way to apply the pathway analysis approach (Subramanian et al., 2005, Langfelder and Horvath, 2007, Brunet et al., 2004, Kanehisa and Goto, 2000) as well as different statistical methods to quantify significant enrichment (Beißbarth and Speed, 2004, Goeman and Bühlmann, 2007, Barry et al., 2005, Dutta et al., 2012b, Gu et al., 2012, Rahnenführer et al., 2004b). Irrespective of the chosen analysis method, however, a reliance on pathway databases for high-throughput hit selection limits the number of novel genes that can be identified and overlooks the genes that are not yet annotated within pathway databases, thereby obviating one of the most important rationales for performing unbiased genome-scale screens and contributing to the notorious “streetlight effect” in molecular biology (Haynes et al., 2018, Rodriguez-Esteban and Jiang, 2017).

Complementary to the pathway analysis filtering approach, the network analysis approach utilizes protein-protein interaction (PPI) databases as a way to prioritize lower scoring hits from high-throughput studies and expand the dataset (Dutta et al., 2016, Tu et al., 2009, Wang et al., 2009). Various methods for incorporating the information from interaction databases into a prioritization pipeline for high-throughput studies have been developed (Oti et al., 2006, Cowen et al., 2017, Yu et al., 2013, Wang et al., 2018, Zhang et al., 2017). Expansion of the lists of candidates by network analysis can successfully decrease the rate of false negatives, yet it also intrinsically amplifies the noise in the hit selection set (as false-positive candidates in the original high scoring set of hits also expand to include their predicted interactions).

Taken together, the contributions and associated trade-offs of each class of methods to hit selection illuminates the range of possibilities and challenges that present at the juncture between high throughput experiments and subsequent follow up analysis. The unique trade-offs of each approach also suggest the possibility that some of the analytical blind spots can be offset by combining orthogonal approaches, yet a systematic approach by which to optimally integrate these methods has not been established. Here, we have developed an approach which, through iterative use of pathway and network databases, can harness the benefits while mitigating the drawbacks of these analysis methods. Using a set of comparable genome-scale genetic screens of Human Immunodeficiency Virus (HIV) host response factors (Zhou et al., 2008, Brass et al., 2008, König et al., 2008), we demonstrate improved significance and magnitude of the overlap in screen hits by use of the iterative analysis strategy. We also show that this approach leads to more efficient identification of true-positive CRISPR/Cas9 screen hits as compared to exclusively relying on gene ranking, and identifies relevant biological enrichments missed by focusing on the most differentially expressed genes in RNA-seq data. We name this approach SIGNAL, for Selection by Iterative pathway Group and Network Analysis Looping and describe the development of a web-based platform (https://signal.niaid.nih.gov) for unrestricted access to this analysis resource.

Results

Independent screens for HIV dependency factors provide test datasets for addressing the challenges of hit selection

The challenge of optimal candidate selection from genome-scale screens has been implicated in the modest overlap and limited statistical significance of hit sets across parallel studies (Bhinder and Djaballah, 2013, Ein-Dor et al., 2006). High-throughput gene perturbation studies of similar biological phenomena that have surprisingly limited overlap have been found in host factor screens for Influenza (Watanabe et al., 2010) and HIV (Hirsch, 2010). The three independent studies of essential proteins required for early infection of HIV, also known as HIV Host Dependency Factors (HDFs) are amongst the most frequently cited examples of the high discordance of hit identification between parallel high-throughput assays (Hirsch, 2010, Zhu et al., 2014). Independent work by Brass *et al.* (Brass et al., 2008), König *et al.* (König et al., 2008), and Zhou *et al.* (Zhou et al., 2008) used genome-scale RNAi studies to identify cellular host factors required for effective early stage HIV infection. From approximately 300 hits that passed the validation assays conducted in each study (Post-validation hits) only 2 were shared across all three, with a further 28 shared by at least 2 studies (Fig. 1A). An analysis of the normalized scores of all genes and the validated hits reported by each study shows substantial variation in where the validated hits fell in the initial primary screen score distribution (Fig. 1C, Supplementary Table 1). However, a direct selection of the highest ranking 400 hits from each screen (Fig. 1C (red shading)), did not result in substantial changes in the number of overlapping genes (Fig. 1B).

A closer inspection of the statistical and shared enrichment reveals that all three study comparisons crossed standard thresholds of statistical enrichment of overlap for post validation hits, while only one of the comparisons for high scoring hits passed the significance threshold (Fig. 1D). Since the number of shared hits between studies is very similar for the two hit selection categories (Fig. 1E), this improved statistical enrichment for validated hits is largely driven by the smaller size of the post-validation hit gene sets (Fig. 1F). The improvement in overlap significance also reflects the improvement achievable by the bioinformatic and experimental validation methods used by the three studies beyond the primary ‘high score’ metric (Supplementary Table 1), leading to a reduced number of false positives in the hit selection sets. The lack of increase in magnitude of overlap, however, shows that the approaches used did not reduce the false negative rate.

In addition to demonstrating the challenges of current approaches to hit selection, these studies also provide datasets that can be used to test whether alternative hit selection methods can improve enrichment and error correction. Various attempts have been made at developing benchmarking or synthetic datasets to evaluate the accuracy, sensitivity, and specificity of different hit selection approaches (Geistlinger et al., 2020, Mathur et al., 2018, Nguyen et al., 2019, Roder et al., 2019). The identification of a gold standard dataset by which different prioritization methods can be compared remains one of the critical challenges in bioinformatic analysis of high-throughput data on biological signaling (Khatri et al., 2012, Mathur et al., 2018, Mitrea et al., 2013). As the availability of optimal benchmarking datasets are lacking, the comparative analysis of the three described HIV

studies can serve as a proxy for evaluating the accuracy of new methods. A more sensitive hit selection method applied to all three studies would lead to greater magnitude of overlap, while an approach with higher specificity and precision would lead to improved statistical significance of overlap.

Hit selection and prioritization of medium confidence hits by pathway analysis improves statistical enrichment, but not shared hits, across studies of HDFs

In the single cutoff approach commonly used in high-throughput studies, the list of potential candidates is separated into a binary classification of hits and non-hits based on a chosen threshold (Fig. 2A). This approach precludes the inclusion of lower scoring hits based on downstream analysis. As an alternative, we tested a dual cutoff approach with a stringent threshold for “high confidence” hits, and a more lenient cutoff to define a set of ‘medium confidence’ hits. In this design, the dataset is split into a three-tiered set of high confidence hits, medium confidence hits, and low confidence/non-hits (Fig. 2B). We assigned comparable dual cutoffs to the Z scores of the three screens, with the 400 highest scores assigned as high confidence, the next 1000 assigned as medium confidence, and both groups also required to pass the additional readout score threshold. The rest of the gene candidates were assigned as non-hits (Fig. 2C, Supplementary Table 1).

As a benchmark, we first tested whether the application of pathway analysis could improve overlap in hits from HDF studies with the standard single cutoff approach. Using the pathway membership list from the Kyoto Encyclopedia for Genes and Genomes (KEGG) database (Kanehisa et al., 2017) we applied pathway analysis exclusively using the high scoring hits of the three screens (Fig. 2D). Statistical significance of overlap increased in two out of the three comparisons as compared to the significance of overlap in the high scoring and post validation hits as described earlier (Fig. 2E). Number of shared hits, however, decreased in all cases as compared to the prior analysis approaches (Fig. 2F). We then ran pathway analysis on the HDF screens using the hit sets prepared by the dual cutoff method. Following analysis of the high confidence set to first identify enriched pathways and the hits therein, medium confidence hits that were members of those pathways were promoted into the hit set. Hits in either the high confidence or medium confidence set that were not part of these enriched pathways were relegated to the non-hit group (Fig. 2G). A comparison of the three studies of HDFs with hit selection by this approach shows that it improves both measures of overlap, significance of enrichment (Fig. 2H) and number of shared hits (Fig. 2I) in all screens as compared to pathway analysis with the single cutoff approach (Fig. 2E, F). Comparison to hit selection by high scores and post validation, however, shows that the strength in pathway analysis is predominantly in the false positive correction (Fig. 2H) and only marginally adds to the sensitivity of the analysis, as measured by the number of shared hits across studies (Fig. 2I).

Hit selection and prioritization of medium confidence hits by network analysis improves shared hits, but not statistical enrichment, across studies of HDFs

Using the tiered dataset approach described in the previous section and the protein-protein interactions curated by the Search Tool for the Retrieval of Interacting Genes (STRING) database (Szklarczyk et al., 2010, Szklarczyk et al., 2019) (see Material and Methods), high

confidence and medium confidence hits were entered into the network and searched for predicted interactions. The interactions were filtered to include only those between a high confidence hit and a medium confidence hit as a means to promote medium confidence hits to the high confidence set (Fig. 2J). Significance of overlap was not improved by network analysis (Fig. 2K) reflecting the expansion of the total number of hits selected and the increase in false positive hits. In contrast, this approach led to a sharp increase in the number of shared hits across studies (Fig. 2L). Testing the significance of the overlap by random permutation also found that the number of shared hits found by network analysis alone was only above a statistical threshold of significance in two out of the three comparisons (Fig. 2K), suggesting that the 'catch-all' approach of network analysis without any false positive correction is prone to the amplification of false positives. This strongly suggests that, complementary to pathway analysis, hit selection by network analysis is a highly sensitive approach but requires additional correction to increase the specificity of the hit selection set.

An integrated serial approach to pathway and network analysis improves both statistical enrichment and number of shared hits

To test a more integrated approach, we designed a combined analysis framework for pathway and network analysis. Using the same three-tiered dataset approach described above, the first step in the analysis identifies enriched pathways from the high confidence hits. Following the identification of significantly enriched pathways, medium confidence hits that are members of the enriched pathways are promoted to high confidence and all high confidence hits that are not part of the enriched pathways are moved to the medium confidence set. Network analysis is then applied to the newly assigned high confidence and medium confidence sets, with medium confidence hits that have a reported interaction with a high confidence hit promoted to the high confidence set. The expanded set of high confidence hits is then assigned as the final hit selection set (Fig. 3A). Applying this integrated serial approach to the three studies of HIV HDFs led to improvements in both significance of overlap and shared hit number across the three studies (Fig. 3B, C). While the improvements in significance or shared hit magnitude were not as strong as with the respective exclusive use of pathway or network analysis alone (Fig. 2H, L), the improvements observed in both metrics when using the integrated serial framework suggests that it can at least partially capture the complementary error correction of the two methods.

An iterative method for the integrated approach further improves hit selection from individual methods

In an attempt to further amplify the combinatorial benefit we observed with the serial approach above, we designed a framework that iteratively applies this strategy. In the iterative design, the same procedure as described for the serial application of pathway then network analysis is applied as a first iteration. The hits selected by the end of this iteration are then reassigned as high confidence hits with the rest reassigned as medium confidence for the second iteration. The same integrated pathway-to-network analysis is repeated to complete the second iteration using the newly assigned high confidence and medium confidence hits as input. When the second iteration ends the new set of high confidence hits are compared to those of the previous iteration. If the two sets are different,

further iterations are applied until the high and medium confidence sets of candidates are no longer changed by further iterations of the cycle (Fig. 3D, Supplementary Figure 1). This approach ensures that the set of selected hits is modified and appended until neither pathway or network analysis can pull it in a different direction, ensuring that the resulting set of hits is at the equilibrium between false positive correction by pathway analysis and false negative correction through network analysis. We applied this iterative approach to the studies of HDFs and found that the three screens required 4 to 5 iterations before the set of high confidence hits no longer changed (Supplementary Figure 2). We then repeated the comparative analysis of the three screens using the hits selected by the iterative approach and observed substantial improvements in both significance of overlap and the number of shared hits (Fig. 3E-F). Of note, the number of shared hits showed a marked improvement over the serial analysis method (Fig. 3C, 3F), suggesting that the repeated iterations are able to capture additional shared hits between screens that could be missed by a less rigorous analysis approach.

To further ascertain whether the above design of the iterative framework is optimized to give the best improvement in hit selection, we designed and tested an iterative pipeline where the order of analysis methods is reversed, with network analysis applied first to the dataset followed by pathway analysis (Supplemental Figure 3A). Reversing the analysis order led to a decrease in both metrics of true positive hit selection (Supplemental Figure 3B-C). The measure of confidence by the random permutation test was also low in two out of the three comparisons, suggesting that the false positive noise was amplified in this hit selection approach. These results suggest that the optimal order for an integrated analysis framework is a false positive correction (such as pathway analysis) followed by a false negative correction (as in network analysis), whereas reversing the order can substantially amplify noise and blunts the power of an integrated and iterative approach.

We also tested the iterative analysis on the post validation hits from the three HDF studies to determine if this framework for prioritization can be applied to hits thresholded by different methods. We assigned as high confidence the post validation hits reported from each screen and as medium confidence hits the top ranking 1000 candidates not selected as hits by each study (Supplemental Figure 4A-B). We observed similar improvements both in the number of shared hits across the screens and in the significance of overlap as we found with the iterative analysis of the highest scoring hits (Supplemental Figure 4C-D).

Taken together, the analysis described above demonstrates that an iterative framework for pathway enrichment followed by network analysis provides the strongest combinatorial benefit of the complementary analysis approaches (Figure 4). By incorporating data that uses two cutoffs, this approach optimizes the ability to identify relevant candidates from segments of the dataset with different levels of noise using a combination of the initial gene rankings from the assay and the known gene characteristics and functions from curated databases. We chose the name SIGNAL for this approach as an acronym for Selection by Iterative pathway Group and Network Analysis Looping.

SIGNAL identifies hits with greater precision, specificity, and mechanistic insight from CRISPR/Cas9 screening and RNA-seq studies

To test our analysis on a broader variety of high-throughput platforms, we used SIGNAL to select hits from a CRISPR/Cas9 screen (Parnas et al., 2015) and an RNAs-seq time course study (Das et al., 2018). Parnas et al. performed a primary genome-wide CRISPR study of bone-marrow derived dendritic cells (BMDC) treated with LPS. The highest ranking 2,569 candidates were then rescreened in a rigorous secondary screen (10 sgRNAs/gene) alongside 1,231 putative positive regulators (Fig. 5A). We assigned the top ranking 25% of the positive regulators as high confidence hits and the rest of the putative candidates as medium confidence hits (Fig.5A). Analysis by SIGNAL selected a subset of 443 genes. We then compared the reported set of primary screen hits versus the more compact number of hits from SIGNAL along a sliding scale of z-score cutoffs in the secondary screen (Fig. 5B). SIGNAL was found to substantially increase the specificity (Fig. 5C) and precision (Fig. 5D) demonstrating that SIGNAL substantially increases efficiency in hit selection for follow up validation from CRISPR/cas9 screens similar to that observed with siRNA studies (Fig. 4).

To assess whether SIGNAL improves insight from pathway enrichments -and to simultaneously demonstrate the application of SIGNAL analysis to RNA-seq datasets - we analyzed a study from Das et al. that reports mRNA expression at four time points between 1 and 24 hours in macrophages treated with LPS and interferon (IFN) γ. Taking the top ranking thousand hits from each time point and splitting it into high confidence and medium confidence hits by a 1:2 ratio (Fig. 5E) the results of each time point were analyzed by SIGNAL. The expected enrichment of immune related pathways was seen in the candidates selected either by highest rank or by SIGNAL. However, SIGNAL-selected also showed an increased enrichment for metabolic related processes over time (Fig. 5F), uncovering the critical role recently established for macrophage metabolic remodeling after extended LPS exposure (Seim et al., 2019, Lampropoulou et al., 2016). Notably, this enrichment is much less evident by looking exclusively at the highest-ranking differentially expressed genes.

signal.niaid.nih.gov is a secure, publicly accessible web-based interface for analysis of high-throughput genomic data

To broaden access to this analysis framework we have made the SIGNAL pipeline available as a secure web-based, user-friendly interface which can be accessed at signal.niaid.nih.gov. The platform is intuitive to use and requires no prior knowledge of computer languages. signal.niaid.nih.gov is hosted by the National Institute of Allergy and Infectious Diseases (NIAID) and uses a secure encrypted HTTPS connection. To increase security and data privacy, once a user's session ends (i.e. close of browser window or move to a new site) the directory with all its files are removed from the SIGNAL server. File names, analysis choices, user IDs, and results are neither collected nor stored.

Uploading a data set to the SIGNAL platform for analysis

To upload a dataset for analysis by SIGNAL the data must be in .csv, .txt, or .xlsx format. The document must also contain one of the following: either a column titled "GeneSymbol" that has HGNC gene symbols in all the rows, or, alternatively, a column titled "EntrezID"

with NCBI EntrezIDs in all the rows. Both ID columns, however, do not need to be included in the upload file. The upload file must also include at least one column with the numeric values to be used for selecting high and medium confidence hits. The name of this column is up to the user. The numeric values can be either continuous values (such as a range of *p* values or a range of Zscores) or assigned values such as assigning a value of 1 to all IDs that should be considered "high confidence", a value of 0.5 to all IDs that should be considered "medium confidence", and a value of 0 for all IDs that should be considered "non-hits" (Supplemental Figure 5). There is also an option to use two columns for setting the cutoffs, such as fold change and false detection rate or Z score and a cell viability readout for example. A panel of dropdown menus on the left side of the browser window allows the user to select the parameters that describe the data and the database settings preferred for the analysis. (Fig. 6A, Supplementary Figure 6).

As many high-throughput assays are now outsourced to core facilities which often supply in return a list of ranked gene candidates and do not include the full range of scores for the entire genome being measured, SIGNAL includes an added feature where the user can add in a "genome background" (Fig. 6A, Supplementary Figure 6). When selected, the *add genome background* feature adds genes to the list that aren't included in the upload file to be used as a background for statistical enrichment analysis. The added background "genes" will not appear as suggested hits by the SIGNAL analysis, the background genes are only used as a means to have more robust statistics on the enrichment of pathways.

SIGNAL also allows for the user to select a list of hits that are exempt from filtering and retained as hits regardless of what the iterative analysis finds. This is recommended for when the study involves genes or a pathway that haven't yet been annotated by pathway databases or when the user has a particular interest or knowledge about the relevance of these gene IDs.

The rapid speed of the analysis processes also allows a user to easily experiment with different cutoffs for the high and medium confidence settings and to compare and contrast outputs. To encourage this approach, a *Reset* action button is included on the platform, which when clicked, resets the input settings and allows the user to easily run a new analysis with different parameters (Fig 6A).

SIGNAL results provide robustly prioritized hits with mapped enrichments of significantly represented pathways

Once the SIGNAL analysis is complete, the window switches to the *Enriched Pathways* tab. This tab provides a list of statistically enriched pathways found in the set of selected hits by SIGNAL analysis. The names of the enriched pathways can also be clicked to open a new tab from the KEGG website showing a schematic of the genes in the pathways with the gene hits from the SIGNAL analysis highlighted. Genes that were marked as high confidence at the input of the analysis are highlighted in blue and those marked as medium confidence are highlighted in red (Fig. 6B-C). This feature makes it possible to further explore if the genes that are driving the enrichment of the pathway are spread across the pathway or concentrated in a particular segment.

The *Gene Hits* tab contains a series of tables that list the genes that were selected by the SIGNAL process and can guide the further prioritization of hits for follow up (Fig. 6C). The tables include a *SIGNAL Gene Hits* table (Supplementary Figure 7) and a *High Confidence Hits not in SIGNAL Hits* table, a list of hits that were assigned as high confidence in the input but were not selected as hits by the SIGNAL analysis. This table is included so that the user can easily review the high-confidence hits that were dropped out by SIGNAL to see if any of those should be manually added back in based on the user's knowledge and judgment. A *Pathway Enrichments* table lists the enriched pathways from the analysis with a range of statistical cutoffs and the gene candidates that drive the enrichment (Supplementary Figure 8). All output tables can be downloaded and saved by the user.

An interactive visualization of pathway hits and network interactions supports hypothesis generation for "missing links" between enriched pathways

Though frequently applied in parallel, there is a need for developing a way of visually representing integrated pathway and network analysis results. To build in a solution within the SIGNAL platform we utilized the Hierarchical Edge Bundling method (Holten, 2006) as a means of grouping network nodes (which in this context would be genes or proteins) that are part of enriched pathways into individual groups. The graph assigns another group as the "additional SIGNAL hits" group to place all the gene hits that are not annotated as part of the selected pathways. For visual clarity the intra-group connections within a given pathway are filtered out, but are shown for the group representing the hits outside the selected pathways. This method allows for clearer visualization of genes driving suggested interactions between pathways and makes it possible to explore putative interactions between pathways through a common 'connecting' gene (Fig. 7A).

Within the SIGNAL platform the user can select up to three pathways from the list of enriched pathways and SIGNAL will generate a graph of all the SIGNAL hits that are part of the selected pathways as well as all the SIGNAL hits that have predicted interactions with the selected 'pathway' members (Fig. 7B). Hovering with a cursor over a specific gene ("node") highlights all the predicted interactions ("edges") for that gene. Clicking on a gene 'fixes' the interactions, so that the user can then click on one of the predicted interactions to observe the interactions from the second node. A panel at the side of the graph provides information about the interaction, such as the evidence source for the interaction and its confidence score from the STRING database (Fig 7B). After clicking through a string of interacting genes the user can click the "Highlight Clicked Pathway" icon and all the genes clicked through in the exploration are highlighted (Fig. 7C). The clicked genes, and the pathways they are members of are tabulated in a separate table that can be downloaded with the rest of the analysis at the download tab.

To demonstrate the utility of this visualization and analysis tool we analyzed the SIGNAL results based on the post-validation hits from the Brass et al. HIV study described earlier (Brass et al., 2008) (Supplementary Figure 4A, center). The N-Glycan biosynthesis and RNA transport pathways were among the strongest enrichments listed in the results (Supplementary Figure 8). A subsequently published meta-analysis of the three studies of HDFs identified the RNA Transport pathway as critical for early HIV infection based on the

combined results from all three studies (Bushman et al., 2009). An additional meta-analysis study using broad network prioritization in all three studies highlighted the Golgi related COG2-4 genes which facilitate vesicular transport and recycling of glycotransferases as essential factors in early HIV infection, along with its interacting gene STX5 (Zhu et al., 2014). The study suggests that these factors potentially modulate HIV infection by regulating glycosylation. Using the visualization platform on SIGNAL we selected the N-Glycan biosynthesis pathway and the RNA transport pathway to generate a network of the related pathway hits and its predicted interactions from the SIGNAL analysis of the Brass et al. data alone (Fig. 6B). Observing the network reveals that the N-Glycan biosynthesis pathway gene MAN2A1 interacts with STX5 and the COG genes. Following the interactions, we can see that the RNA transport pathway gene NUP160 also interacts with STX5. In addition to identifying the essential enrichments previously characterized only after several meta-analysis studies, the SIGNAL generated network also identifies specific gene candidates by which these putative enrichments and regulatory mechanisms can be further investigated. Clicking on the nodes along these interactions and then highlighting the clicked pathway reveals a link between the two enrichments, suggesting that the RNA transport and Golgi factors are both regulated via the N-Glycan maturation enzyme MAN2A1 (Fig. 7C). This analysis is an illustrative example of how the interactive SIGNAL network can be used to identify high-priority candidates for further validation of novel mechanisms.

R script of the SIGNAL framework can be adapted for analysis with bespoke databases and network and enrichment criteria

SIGNAL analysis was developed and tested using a set of parameters and heuristics that are in broad use by researchers. The framework, however, can in principle be applied to any set of complementary pathway and network approaches. Extensive developments in the assessment of pathway enrichments and applications of network analysis theory have enabled more sophisticated approaches to be applied by more computationally skilled investigators. Different analyses also call for databases curated by different criteria to match the specific research question. To broaden the use of the SIGNAL framework to incorporate different methods of pathway and network analysis as well as diverse databases, we also built SIGNAL as an adaptable function in R (https://github.com/niaid/SIGNAL). The SIGNAL function in R can be downloaded and run locally with any user provided databases and network graph. The script can also be adapted for more bespoke analysis approaches using 2nd (Barry et al., 2005, Subramanian et al., 2005, Cowen et al., 2017, Yu et al., 2013) and 3rd (Rahnenführer et al., 2004a, Dutta et al., 2012a, Gu et al., 2012, Wang et al., 2018, Zhang et al., 2017) generation pathway and network approaches or alternative bioinformatic solutions using SIGNAL as a framework for integrating complementary hit selection approaches.

Discussion

In the development of SIGNAL, we focused on the twin challenges of incomplete genome annotation by pathway databases and lack of false positive correction in network analysis approaches. SIGNAL addresses these issues and optimizes the use of these

different database classes to more thoroughly identify biologically significant hits from omic-scale studies. SIGNAL was designed and tested using the first-generation methods of pathway and network analysis (Over Representation Analysis (ORA) (Beißbarth and Speed, 2004, Goeman and Bühlmann, 2007) and Direct Neighbour network (Oti et al., 2006), respectively) both of which remain widely used in the routine reporting of high-throughput studies (Dong et al., 2016). Pathway and network analysis methods have gone through critical evolutions over the past decade. A second generation of enrichment approaches were developed as Functional Class Sorting (FCS) methods which emphasize coordinated changes in the group of genes from the predetermined set (i.e. pathway or functional group) (Barry et al., 2005, Subramanian et al., 2005). A third generation of enrichment analysis approaches were later developed astopology-based (TB) approaches that consider the organization of genes within a pathway and do not weigh all genes in the pathway equally (Rahnenführer et al., 2004a, Dutta et al., 2012a, Gu et al., 2012). Alternative and more discriminating approaches to expanding candidate lists by network analysis have also been created, such as network propagation (Cowen et al., 2017), and methods that incorporate concepts from graph and information theory (Yu et al., 2013). Where more multi-level datasets are available, network prioritization using more sophisticated statistical and machine learning methods such as linear regression and random forest have yielded more discriminating results (Wang et al., 2018, Zhang et al., 2017). It remains to be tested, however, whether the SIGNAL design can be similarly applied to those analysis methods as well as alternative complementary false positive and false negative correction methods.

Some of the intrinsic challenges of relying on curated databases persist even in the SIGNAL design. In the context of using pathway enrichment for hits prioritization, the statistical approach used to assess significant enrichment (a hypergeometric test with FDR) favors a specific range of pathway sizes. Best practices in the use of pathway enrichment statistics suggest that the analysis works best in pathways that contain member genes in the range of 20 to 400 (Ramanan et al., 2012). This range limits the possibility to reliably explore broader pathways such as general metabolic processes which have higher gene membership counts. There is also substantial redundancy and overlap in pathway annotation which can lead to unrelated enrichments being identified, though some bioinformatic solutions for this have already been proposed (Pita-Juárez et al., 2018, Vivar et al., 2013, Simillion et al., 2017).

Network analysis driven data exploration also has a set of persistent challenges. Notably, network analysis databases such as STRING are cell type and treatment agnostic (Ma et al., 2019), making some of the imputed interactions irrelevant or misleading for analysis in different contexts. The latter challenge could be addressed by more cell or disease specific proteinprotein interaction networks being generated, but this will require a substantial investment from the research community to develop and generate such resources. The SIGNAL R code is designed such that it can be adapted to more bespoke analysis pipelines when such datasets are available. Recently developed search engines such as GADGET have used thoroughly sensitive text mining algorithms to map abstracts in PubMed to specific gene IDs, metabolites, and disease keywords (Craven, M. (2015). Gadget. Retrieved from http://gadget.biostat.wisc.edu/), an expansion of these methods to map interactions from the literature to the cell or tissue types and the treatments they were identified in could address

Europe PMC Funders Author Manuscripts

some of the current network database blindspots. Additional creative bioinformatic methods, however, will also be necessary to infer across which cell types and conditions observed protein-protein interactions can be extrapolated to, and in what contexts comparisons are less likely to be informative. Finally, the powerful utility of pathway and network databases and the novel ways in which they continue to be applied further underscores the critical need to support the maintenance of these resources.

Star Methods

Resource Availability

Lead Contact—Further information and requests for source code and data processing protocols should be directed to and will be fulfilled by the Lead Contact, Iain D.C. Fraser (fraseri@nih.gov)

Materials Availability—This study did not generate new unique reagents.

Source data statement—This paper analyzes existing, publicly available data. These datasets' accession numbers are provided in the Key Resource Table.

The datasets analyzed in this study have all been previously published in Brass et al., 2008 (Brass et al., 2008), König et al., 2008 (König et al., 2008), Zhou et al., 2008 (Zhou et al., 2008), Parnas et al., 2015 (Parnas et al., 2015), and Das et al., 2018 (Das et al., 2018)

Code statement: SIGNAL source code, compiled codes corresponding to this manuscript's version of the software, and R code of the analysis generated for this manuscript are available at https://github.com/niaid/SIGNAL.

Scripts statement: The scripts used to generate the figures reported in this paper are available at https://github.com/niaid/SIGNAL/Manuscript/Rcode.

Method Details

Datasets and Databases

Genome-wide studies: The three genome-wide siRNA studies of essential proteins in early HIV infection were published by Brass et al., 2008; König et al., 2008; and Zhou et al., 2008. The complete datasets of scores and metadata of these studies were generously shared by Amy Espeseth (Zhou *et al* screen), Abraham Brass (Brass *et al* screen), and Sumit Chanda (König *et al* screen). Brass et al. and Zhou et al.performed two readouts one at 48 hours post infection and another at a later timepoint. For comparative purposes we only compared the first readout from Brass and Zhou to the study of König et al. to focus on the candidates regulating early infection (Supplementary Table 1). The primary and secondary CRSIPR/cas9 study of LPS treated bone marrow derived dendritic cells was published by Parnas et al. 2015. The RNA-seq time course study of LPS-IFNγ was published by Das et al. 2015 and the data was accessed from the Gene Expression Omnibus (GEO) database, accession number: GSE103958.

KEGG database for pathway enrichment: The KEGG database was downloaded from the KEGG Application Program Interface (API), as described previously (Kanehisa et al., 2017). For the analysis described in this manuscript, the KEGG data was downloaded on May 11, 2019. Pathway lists were filtered for pathways that are related to biological processes (and excluding the ones related to disease) by only selecting pathways with PathwaysIDs of 05000 or less. EntrezIDs were added to the NCBI gene symbols in the KEGG database by the org.Hs.eg.db: R package (Carlson, 2018a) and the org.Mm.eg.db: R package (Carlson, 2018b). The annotated pathway enrichment document was formatted into a matrix of gene IDs and pathway identifiers and subset into 2x2 matrices for competitive enrichment analysis as previously described (Goeman and Bühlmann, 2007).

STRING database for protein network interactions: The STRING database was downloaded from the STRING API as previously described (Szklarczyk et al., 2010). The 9606.protein.links.full.v10.5 was downloaded for human interactions and the 10090.protein.links.full.v10.5 for mouse interactions. Inferred interactions from other species were not included. The network downloads were separated based on the evidence source of their interactions. The evidence source categories followed the STRING database categorizations. The different evidence source network files were then split into three groups based on their evidence scores, 0.15-0.4 as low confidence, 0.4-0.7 as medium confidence, and 0.7-1 as high confidence. The files were then converted into the igraph format using the igraph R package (Csardi and Nepusz, 2005). For the analysis described in this manuscript, the STRING data was downloaded on October 3rd, 2018. Each analysis was performed using a single master igraph that was generated by combining the igraphs of the relevant criteria (evidence source and scores). The networks were used to prioritize lower scoring hits by using the direct neighbor functional approach as previously described (Wang et al., 2009).

Gene and protein ID conversion: Gene to protein ID conversions were done using the biomaRt R package (Durinck et al., 2009) EntrezID to GeneSymbol ID conversions were done using the org.Hs.eg.db: R package (Marc Carlson (2018). *org.Hs.eg.db: Genome wide annotation for Human*. R package version 3.7.0.) and the org.Mm.eg.db: R package (Marc Carlson (2018). *org.Mm.eg.db: Genome wide annotation for Mouse*. R package version 3.7.0.)

Bioinformatics

Iterative analysis in R: Computational analysis was done in the R environment (R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/). Genomic analysis software was supported by the BioConductor platform (Gentleman et al., 2004). To build the SIGNAL analysis pipeline in R, all components were built as separate functions and then integrated together into a master function. Below is a summary of the individual steps taken followed by their integration into an iterative function.

Importing and standardizing databases: The downloads from pathway databases and network databases (KEGG and STRING) were mapped to common IDs (EntrezIDs). The network

database was converted to a set of igraphs and the pathway database was converted to a two-column table of pathway name and pathway members. This enabled efficient mapping between hit datasets and databases.

Pathway enrichment function: A pathway enrichment function was created that creates a contingency matrix for each pathway name in the pathway database and a list of IDs separated into “hits” and “non-hits”. Using a one-sided Fisher’s exact test, the *p*-values, FDR, and Bonferroni correction of enrichment for each pathway name were generated. This analysis loops over all the unique pathway names in the pathway database table. The pathway function is provided with a significance cutoff (<0.055). The function uses the significance to separate the pathway names that passed the threshold of significance for the list of pathway names and creates a list of selected pathways. Using the list of selected pathways, a vector of unique gene IDs that are members of the selected pathways is created. When using a dataset with a single cutoff (hits vs. non-hits), the intersect of the pathway member IDs and hit IDs is sub selected as a new set of hits. When using a dataset with two cutoffs (high confidence hits, medium confidence hits, and non-hits), only the high confidence hits are used as the “hits” for determining pathway enrichment. After the vector of pathway associated genes is created, a vector of the union of high confidence and medium confidence hits is generated. The intersect of the new vector of hits and the vector of pathway genes is taken as the new set of hits.

Network enrichment function: A generated igraph of the selected network database parameters is matched with a list of high confidence hits and medium confidence hits. A new igraph is created based on the intersect of the list of hits with the database igraph. A two-column table of each of the two IDs (“nodes”) from each predicted interaction (“edge”) is created. The list is then matched with a list of the high confidence hits. To find the medium confidence hits that have predicted interactions with high confidence hits, edges that have a match with the high confidence list of hits in at least one of their nodes are kept, while nodes without a match in either of their edges are filtered out. A new vector of unique IDs is generated from the filtered table and the union of the vector and the high confidence list of hits is assigned as the new set of high confidence hits.

Iterative analysis function: The iterative function was built by first creating a pipeline where the pathway enrichment function is applied to the input of a screen containing gene IDs in three groups (high confidence, medium confidence, and non-hits). The output of the pathway analysis steps is then reshaped to match the required input for the network analysis step. Following the network analysis output, the new hit characterizations are assigned as the new input column for pathway analysis. A “confidence category” column keeps track of what confidence level each hit was at the first input while a “proxy score” column updates the level of hit confidence each gene is assigned within each iteration. A separate data frame is created where the selected enrichment pathways from the pathway function are tabulated.

To halt the iterative loop when the iterations converge to the same set of high-confidence hits, the script uses a *while* function relying on a variable nested in an *if* function. Briefly, the variable “counter” is assigned as TRUE at the start of the analysis. An additional variable (“iteration”) counts what iteration of the analysis is currently running. The analysis function

is wrapped within a “while” function that only runs the analysis while counter = TRUE. Following an iteration of pathway and network analysis, an *if* function evaluates if the iteration count is greater than 1. If true, the *if* function evaluates if the table of IDs with associated proxy score of this iteration match the table of IDs with associated proxy scores from the previous iteration. If the condition is true the “counter” variable is assigned as counter = FALSE. This leads to the termination of the function. Otherwise, the counter variable remains TRUE and the condition of the *while* loop is met to commence a new iteration of the analysis. When the analysis is complete a data frame with all the input IDs, the confidence category each ID was assigned at input, the proxy score for each iteration, and whether the ID was assigned as a “hit” by the final iteration is generated. An additional data frame with the pathway enrichments from the final iteration is also generated with each pathway name matched with the intersect of member IDs with the list of hit IDs.

Repeated tests with different datasets have all resulted in the analysis converging to a single set of hits after a finite number of iterations. When testing randomized datasets, however, a number of datasets (out of more than five thousand tested) led to the set oscillating between different sets after a few iterations. To ensure the termination of the iterative analysis even in those rare cases, an additional condition was added to the above described test. The results of each iteration after iteration ≥ 3 are compared to the results of all the previous iterations. If a result is repeated it is indicative of an oscillating pattern. The analysis then finds the iteration within the repeated pattern that has the largest hit set and then terminates the analysis and assigns that iteration as the final output.

Web based interface of SIGNAL (Shiny): The SIGNAL web interface was designed to run on a set of intuitive user inputs and provide the user with the results of SIGNAL analysis and the ability to explore and download the results. Creation of the public facing web page based on R script was done using the Shiny application (Chang et al., 2020). Briefly, the different sets of outputs were separated into different tabs with an additional tab added for input. Inputs required from the user were separated into “selectedInputs” (organism, pathway, network, interaction confidence for network analysis), “conditionalPanel” (selecting interaction network confidence source), “fileInput” (uploading input file), “textInput” (high-conf cutoff value, mid-conf cutoff value), “checkboxInput” (add genome background), and “actionButton” (run analysis, reset analysis). The inputs are assigned to variables that are then matched to variables in the SIGNAL function. A set of warning messages were built in for cases where lists of hits or chosen parameters yield no results in pathway and network analysis.

To create the hyperlinks for each enriched pathway which maps hits onto KEGG pathways, a link2KEGGmapper function is generated following the SIGNAL analysis. The link2KEGGmapper function generates a list of gene names mapped to the organism abbreviation and assigns colors based on the input-provided confidence level. A web path is created for each pathway and added to the end of the https://www.kegg.jp/kegg-bin/show_pathway?%s0%s web address. This generates a unique URL for each pathway based on the list of high confidence and medium confidence hits in its membership to match the URL generated by the KEGG mapper and ID color feature (https://www.genome.jp/kegg/tool/map_pathway3.html).

For the table of pathway enrichment, an enrichment score *(EnrichScore)* for each pathway was calculated. The score is a measure of the robustness of the pathway enrichments by the number of genes represented in the SIGNAL dataset. The EnrichScore also evaluates how many of the genes driving the pathway enrichment were assigned as high confidence in the input. The total EnrichScore is calculated as $\left(\frac{HitGenes}{GeneslnpathWay}+\frac{HighScoresGenes}{HitGenes}\right)/2$

To generate the appended columns of “InteractingGenes” and “NetworkGenePathways” for the SIGNAL gene hits tab, an igraph of the selected hits is generated based on the network input parameters provided by the user and filtered into a sub-igraph for each hit. The interacting genes are then cross-referenced with the pathway input parameters selected by the user and the list of pathway memberships of the interacting genes are tabulated, counted, and added to the “NetworkGenePathways” column. The download tab on the interface was created as a reactive page. As files are added to the directory with additional analysis steps, the download page updates with a list of file names in the current directory. For ease of use the download files are put in a zip file format.

The application is hosted by the National Institute of Allergy and Infectious Disease (NIAID) Office of Cyber Infrastructure and Computational Biology (OCICB) at the following URL: https://signal.niaid.nih.gov. The analysis is run behind two internet security firewalls and all requests are handled using encrypted connections. After a connection ends the directory with the uploaded input file and all the output files generated during the analysis are deleted from the server.

Interactive pathway and network visualization: Interactive visual interfaces were built by integrating the JavaScript language into the R Shiny platform. Communication across the platforms were done by creating JavaScript files in R using the jsonlite R package (Ooms, 2014). and then fed into d3.js file (Bostock et al., 2011).

To create the hierarchical edge bundling maps of selected pathways and SIGNAL hits, an igraph of all the selected hits is generated. A vector of all the selected pathway names and additional group “additional SIGNAL hits” is also created. To filter the network map, first the nodes are filtered based on membership in the selected pathways or interaction with a node in one of the selected pathways. Second, edges are filtered based on having the two nodes in different groups in the vector of selected pathways and novel hits (this removes intra-group nodes). The edges are assigned color grouping based on the node that is in a pathway group. Nodes that appear in more than one group are assigned to a separate group with a different coloring indicating membership in both groups 1 and 2. Visual parameter controls of the graph on the interface are created using the Shiny slider function. A window in the interface maps the node selected by the cursor to the selected network data frame and populates the field with interaction confidence and evidence source information on the selected node.

Running SIGNAL as an R script: To make SIGNAL an adaptable framework for iterative analysis with different datasets and databases beyond the databases and settings used on this platform, an R script version of a standalone SIGNAL function can be downloaded at https://github.com/niaid/SIGNAL/app/www/RscriptsDownload/SIGNAL_R_function.R. The

SIGNAL function relies on calling two separate analysis function, a pathway enrichment function and a network analysis function. The master SIGNAL function applies the pathway and network function iteratively, and the results are tested for when the analysis converges on a single set.

The list of input variables that can be selectively assigned in the adaptable SIGNAL function in R and their required formats are:

screen.datafame: A data frame of the screen.

ID.column: A column within the screen.dataframe for the identifiers of the targets (EntrezID, GeneSymbol, etc.).

criteria.column: A column within the screen.dataframe of the criteria for being considered a hit.

highconf.criteria: A criteria each target has to meet to be considered a “high confidence” hit.

midconf.criteria: A criteria each target has to meet to be considered a “mid confidence” hit.

criteria.setting: Whether the function should be using “equal”, “greater than or equal”, or “less than or equal” when assessing if confidence criteria are met. criteria.setting input should be in the format of “equal”, “greater”, or “less”.

enrichment.dataframe: A data frame to be used for pathway membership in the format of a column of IDs (should be same as ID column in screen.dataframe in ID type and column title) and a column of which group they are part of (each ID~group relationship needs to be in its own separate row).

enrichment.title: Name of the column with the names of the enrichment groups the targets are members of.

stat.test: Name of the statistical test to be used for measuring enrichment confidence. Needs to be in the format of either “pVal”, “FDR”, or “Bonferroni”.

test.cutoff: A numeric value which a less than value in stat.test will be considered a significant enrichment.

network.igraph: an igraph of the network to be used for network analysis (network igraph must use the same ID type as screen.dataframe)

The user provided variables are then used to apply the iterative function as in the previous paragraph. The adaptable version of SIGNAL broadens the scope of its application beyond the use of the specific databases and settings it was designed in.

The SIGNAL function provides an output in the format of a R script list that contains three data frames:

The input data frame with an appended ‘SIGNAL.hit’ column.

A data frame of high confidence and medium confidence designation at each iteration of the analysis.

A data frame of final SIGNAL enrichments from the provided enrichment data frame.

Quantification and Statistical Analysis

Normalization of high-throughput readouts—Normalization of scores from high-throughput HIV studies was performed using the Z score approach described in (Birmingham et al., 2009). Where plate information was available, scores were normalized to plate mean, otherwise scores were normalized to overall mean of the data set. Data from Parnas et al. included already normalized scores. Data from Das et al. were normalized with DESeq2 (Love et al., 2014) using the default parameters.

Cell viability correction—Cell viability correction varied for the different HIV studies and was based on available data as follows:

Zhou et al Screen: The study calculated a Percent Cell Viability measure for each gene target. The population mean and standard deviation of the provided cell viability percentages were estimated and fit to a normal distribution by subtracting the mean and dividing by the standard deviation. Gene candidates with a cell viability score of -2 or less were flagged.

Brass et al Screen: The study included cell count number for each gene target. The counts were log10 normalized and then given a plate-by-plate Z-score normalization. Gene candidates with a cell viability score of -2 or less were flagged.

König et al Screen: Data shared with us for these studies already had a cell viability correction applied to their readout scores.

Hypergeometric test for pathway enrichment—Hypergeometric distributions to calculate the significance of shared enrichments (across screens and for pathway analysis) were done by generating contingency matrices of shared and non-shared hits and then analysing by a one-sided Fishers' Test with the alternative hypothesis set to "greater than" and the null being no shared enrichment. This approach has been previously described as the "competitive enrichment" test or over representation analysis (Khatri et al., 2012).

Random permutation testing—Statistical significance of the number of shared hits across studies was calculated using a random permutation test. For each analysis and comparison, 1000 input files were generated having the same size of hits and non-hits (or high confidence hits, medium confidence hits, and non-hits where relevant) with the gene candidates assigned to different confidence groups at random. Each of the randomly generated inputs was run through the same analysis and cross-screen comparison as the non-random input. The number of shared hits found in each analysis of the random input was plotted and compared to a Poisson distribution derived from the maximum likelihood estimator of the non-random hits. A p-value for the random results was computed from its corresponding Poisson distribution. This test was used to increase confidence that the results represented in the findings are driven by the prioritization of biologically relevant candidates

Europe PMC Funders Author Manuscripts

and not by the size of the input or biases in the analysis method and annotation databases used.

Binary classification by secondary screen—A range of 20 Z-scores equally spaced between 0.5 and 2.5 were used as a range of validation cutoffs for the secondary CRISPR study from Parnas et al. Confusion matrices were generated separately for each set of hits at each validation cutoff. Gene candidates from the primary screen that fell above the cutoff in the secondary screen were assigned as validated. Specificity was defined as $\frac{True\ Negative\ (TN)}{TN + False\ Positive\ (FP)}$ and precision was defined as $\frac{True\ positive(TP)}{TP + FP}$.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank colleagues in the Laboratory of Immune System Biology for helpful discussions during the development of this software. We also thank Amy Espeseth, Abraham Brass, and Sumit Chanda for sharing information on the original HDF datasets. We are especially grateful to the NIAID Office of Cyber Infrastructure and Computational Biology (OCICB) for hosting the SIGNAL server, thoroughly testing the application, and providing guidance during the development.

Funding

This work was generously supported by the Intramural Research Program of the National Institute of Allergy and Infectious Diseases (SK, JS, NWL. and IDCF). This work was also supported in part by an appointment to the National Institute of Allergies and Infectious Diseases Emerging Leaders in Data Science Research Participation Program (KPW). This program is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Institutes of Health. The Wellcome Trust Investigator award 108045/Z/15/Z (CEB). This work was also supported by the Fuad and Nancy El-Hibri Foundation, the International Biomedical Research Alliance, and the NIH-Oxford-Cambridge Scholars Programs (SK).

Data Availability

The datasets analyzed in this study have all been previously published in Brass et al., 2008 (Brass et al., 2008), König et al., 2008 (König et al., 2008), and Zhou et al., 2008 (Zhou et al., 2008), Parnas et al., 2015 (Parnas et al., 2015), Das et al., 2018) (Das et al., 2018)

The SIGNAL application can be accessed at https://signal.niaid.nih.gov.

SIGNAL source and compiled codes corresponding to this manuscript's version of the software are available at https://github.com/niaid/SIGNAL.

References

Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics. 2005; 21: 1943–1949. [PubMed: 15647293]

Beißbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics. 2004; 20: 1464–1465. [PubMed: 14962934]

Europe PMC Funders Author Manuscripts

Bhinder B, Djaballah H. Systematic analysis of RNAi reports identifies dismal commonality at gene-level and reveals an unprecedented enrichment in pooled shRNA screens. Comb Chem High Throughput Screen. 2013; 16: 665–81. [PubMed: 23848309]

Birmingham A, Selfors LM, Forster T, Wrobel D, Kennedy CJ, Shanks E, Santoyo-Lopez J, Dunican DJ, Long A, Kelleher D, Smith Q, et al. Statistical methods for analysis of high-throughput RNA interference screens. Nature Methods. 2009; 6: 569. [PubMed: 19644458]

Bostock M, Ogievetsky V, Heer J. D(3): Data-Driven Documents. IEEE Trans Vis Comput Graph. 2011; 17: 2301–9. [PubMed: 22034350]

Boutros M, Ahringer J. The art and design of genetic screens: RNA interference. Nature Reviews Genetics. 2008; 9: 554.

Brass AL, Dykxhoorn DM, Benita Y, Yan N, Engelman A, Xavier RJ, Lieberman J, Elledge SJ. Identification of host proteins required for HIV infection through a functional genomic screen. Science. 2008; 319: 921–6. [PubMed: 18187620]

Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences. 2004; 101: 4164.

Bushman FD, Malani N, Fernandes J, D'Orso I, Cagney G, Diamond TL, Zhou H, Hazuda DJ, Espeseth AS, Konig R, Bandyopadhyay S, et al. Host cell factors in HIV replication: meta-analysis of genome-wide studies. PLoS Pathog. 2009; 5 e1000437 [PubMed: 19478882]

Carlson M. org.Hs.eg.db: Genome wide annotation for Human. R package version 370. 2018a.

Carlson M. org.Mm.eg.db: Genome wide annotation for Mouse. 2018b.

Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web Application Framework for R. R package version 150. 2020.

Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics. 2017; 18: 551–562.

Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, Mustonen V, Gonzalez-Perez A, Pearson J, Sander C, Raphael BJ, et al. Pathway and network analysis of cancer genomes. Nat Methods. 2015; 12: 615–621. [PubMed: 26125594]

Csardi G, Nepusz T. The Igraph Software Package for Complex Network Research. InterJournal, Complex Systems. 2005; 1695

Das A, Yang C-S, Arifuzzaman S, Kim S, Kim SY, Jung KH, Lee YS, Chai YG. High-Resolution Mapping and Dynamics of the Transcriptome, Transcription Factors, and Transcription Co-Factor Networks in Classically and Alternatively Activated Macrophages. Frontiers in immunology. 2018; 9: 22. [PubMed: 29403501]

Dong X, Hao Y, Wang X, Tian W. LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. Scientific Reports. 2016; 6 18871 [PubMed: 26750448]

Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009; 4: 1184–91. [PubMed: 19617889]

Dutta B, Azhir A, Merino L-H, Guo Y, Revanur S, Madhamshettiwar PB, Germain RN, Smith JA, Simpson KJ, Martin SE, Buehler E, et al. An interactive web-based application for Comprehensive Analysis of RNAi-screen Data. Nature Communications. 2016; 7 10578

Dutta B, Wallqvist A, Reifman J. PathNet: a tool for pathway analysis using topological information. Source Code for Biology and Medicine. 2012a; 7: 10. [PubMed: 23006764]

Dutta B, Wallqvist A, Reifman J. PathNet: a tool for pathway analysis using topological information. Source code for biology and medicine. 2012b; 7: 10. [PubMed: 23006764]

Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A. 2006; 103: 5923–8. [PubMed: 16585533]

Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Turaga N, Law C, Davis S, Carey V, Morgan M, Zimmer R, et al. Toward a gold standard for benchmarking gene set enrichment analysis. Briefings in Bioinformatics. 2020.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004; 5: R80. [PubMed: 15461798]

Gilbert, Luke A; Horlbeck, Max A; Adamson, B; Villalta, Jacqueline E; Chen, Y; Whitehead, Evan H; Guimaraes, C; Panning, B; Ploegh, Hidde L; Bassik, Michael C; Qi, Lei S; , et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell. 2014; 159: 647–661. [PubMed: 25307932]

Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007; 23: 980–987. [PubMed: 17303618]

Gu Z, Liu J, Cao K, Zhang J, Wang J. Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. BMC Systems Biology. 2012; 6: 56. [PubMed: 22672776]

Hao L, He Q, Wang Z, Craven M, Newton MA, Ahlquist P. Limited Agreement of Independent RNAi Screens for Virus-Required Host Genes Owes More to False-Negative than False-Positive Factors. PLOS Computational Biology. 2013; 9 e1003235 [PubMed: 24068911]

Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. Scientific reports. 2018; 8: 1362. [PubMed: 29358745]

Heckl D, Charpentier E. Toward Whole-Transcriptome Editing with CRISPR-Cas9. Mol Cell. 2015; 58: 560–2. [PubMed: 26000839]

Hirsch AJ. The use of RNAi-based screens to identify host proteins involved in viral replication. Future microbiology. 2010; 5: 303–311. [PubMed: 20143951]

Holten D. Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. IEEE Transactions on Visualization and Computer Graphics. 2006; 12: 741–748. [PubMed: 17080795]

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017; 45: D353–d361. [PubMed: 27899662]

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000; 28: 27–30. [PubMed: 10592173]

Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLOS Computational Biology. 2012; 8 e1002375 [PubMed: 22383865]

König R, Zhou Y, Elleder D, Diamond TL, Bonamy GMC, Irelan JT, Chiang C-y, Tu BP, De Jesus PD, Lilley CE, Seidel S, et al. Global Analysis of Host-Pathogen Interactions that Regulate Early-Stage HIV-1 Replication. Cell. 2008; 135: 49–60. [PubMed: 18854154]

Lampropoulou V, Sergushichev A, Bambouskova M, Nair S, Vincent Emma E, Loginicheva E, Cervantes-Barragan L, Ma X, Huang Stanley C-C, Griss T, Weinheimer Carla J, et al. Itaconate Links Inhibition of Succinate Dehydrogenase with Macrophage Metabolic Remodeling and Regulation of Inflammation. Cell Metabolism. 2016; 24: 158–166. [PubMed: 27374498]

Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology. 2007; 1: 54. [PubMed: 18031580]

Lee SS, Lee RYN, Fraser AG, Kamath RS, Ahringer J, Ruvkun G. A systematic RNAi screen identifies a critical role for mitochondria in C. elegans longevity. Nature Genetics. 2003; 33: 40–48. [PubMed: 12447374]

Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, Irizarry RA, Liu JS, Brown M, Liu XS. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biology. 2014; 15: 554. [PubMed: 25476604]

Lotterhos KE, François O, Blum MGB. Not just methods: User expertise explains the variability of outcomes of genome-wide studies. bioRxiv. 2016. 055046

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014; 15: 550. [PubMed: 25516281]

Ma J, Wang J, Ghoraie LS, Men X, Haibe-Kains B, Dai P. A Comparative Study of Cluster Detection Algorithms in Protein-Protein Interaction for Drug Target Discovery and Drug Repurposing. Frontiers in pharmacology. 2019; 10: 109. [PubMed: 30837876]

Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R. Statistical practice in high-throughput screening data analysis. Nature Biotechnology. 2006; 24: 167–175.

Mathur R, Rotroff D, Ma J, Shojaie A, Motsinger-Reif A. Gene set analysis methods: a systematic comparison. BioData Mining. 2018; 11: 8. [PubMed: 29881462]

McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research. 2012; 40: 4288–4297. [PubMed: 22287627]

Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, Donato M, Voichita C, Draghici S. Methods and approaches in the topology-based analysis of biological pathways. Frontiers in Physiology. 2013; 4

Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepfer AM, Hinkle G, Piqani B, Eisenhaure TM, Luo B, Grenier JK, Carpenter AE, et al. A Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen. Cell. 2006; 124: 1283–1298. [PubMed: 16564017]

Nguyen T-M, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biology. 2019; 20: 203. [PubMed: 31597578]

Ober C. Asthma Genetics in the Post-GWAS Era. Annals of the American Thoracic Society. 2016; 13 (Suppl 1) S85–S90. [PubMed: 27027959]

Ooms J. The jsonlite package: A practical and consistent mapping between json data and r objects. arXiv preprint. 2014. arXiv:1403.2805

Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. Journal of medical genetics. 2006; 43: 691–698. [PubMed: 16611749]

Parnas O, Jovanovic M, Eisenhaure TM, Herbst RH, Dixit A, Ye CJ, Przybylski D, Platt RJ, Tirosh I, Sanjana NE, Shalem O, et al. A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. Cell. 2015; 162: 675–86. [PubMed: 26189680]

Pita-Juárez Y, Altschuler G, Kariotis S, Wei W, Koler K, Green C, Tanzi RE, Hide W. The Pathway Coexpression Network: Revealing pathway relationships. PLOS Computational Biology. 2018; 14 e1006042 [PubMed: 29554099]

Rahnenführer J, Domingues FS, Maydt J, Lengauer T. Calculating the statistical significance of changes in pathway activity from gene expression data. Stat Appl Genet Mol Biol. 2004a; 3

Rahnenführer J, Domingues FS, Maydt J, Lengauer T. Calculating the statistical significance of changes in pathway activity from gene expression data. Statistical applications in genetics and molecular biology. 2004b; 3: 1–29.

Roder J, Linstid B, Oliveira C. Improving the power of gene set enrichment analyses. BMC Bioinformatics. 2019; 20: 257. [PubMed: 31101008]

Rodriguez-Esteban R, Jiang X. Differential gene expression in disease: a comparison between high-throughput studies and the literature. BMC Med Genomics. 2017; 10: 59. [PubMed: 29020950]

Rosenbluh J, Mercer J, Shrestha Y, Oliver R, Tamayo P, Doench John G, Tirosh I, Piccioni F, Hartenian E, Horn H, Fagbami L, et al. Genetic and Proteomic Interrogation of Lower Confidence Candidate Genes Reveals Signaling Networks in β-Catenin-Active Cancers. Cell Systems. 2016; 3: 302–316. e4 [PubMed: 27684187]

Seim GL, Britt EC, John SV, Yeo FJ, Johnson AR, Eisenstein RS, Pagliarini DJ, Fan J. Two-stage metabolic remodelling in macrophages in response to lipopolysaccharide and interferon-γ stimulation. Nature Metabolism. 2019; 1: 731–742.

Simillion C, Liechti R, Lischer HEL, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics. 2017; 18: 151. [PubMed: 28259142]

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102: 15545–15550.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic acids research. 2010; 39: D561–D568. [PubMed: 21045058]

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019; 47: D607–d613. [PubMed: 30476243]

Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. 2012; 40: 3785–99. [PubMed: 22262733]

Tu Z, Argmann C, Wong KK, Mitnaul LJ, Edwards S, Sach IC, Zhu J, Schadt EE. Integrating siRNA and protein–protein interaction data to identify an expanded insulin signaling network. Genome research. 2009; 19: 1057–1067. [PubMed: 19261841]

Vivar JC, Pemu P, McPherson R, Ghosh S. Redundancy Control in Pathway Databases (ReCiPa): An Application for Improving Gene-Set Enrichment Analysis in Omics Studies and "Big Data" Biology. OMICS: A Journal of Integrative Biology. 2013; 17: 414–422. [PubMed: 23758478]

Wang L, Tu Z, Sun F. A network-based integrative approach to prioritize reliable hits from multiple genome-wide RNAi screens in Drosophila. BMC Genomics. 2009; 10: 220. [PubMed: 19435510]

Wang L, Xi Y, Sung S, Qiao H. RNA-seq assistant: machine learning based methods to identify more transcriptional regulated genes. BMC Genomics. 2018; 19: 546. [PubMed: 30029596]

Watanabe T, Watanabe S, Kawaoka Y. Cellular Networks Involved in the Influenza Virus Life Cycle. Cell Host & Microbe. 2010; 7: 427–439. [PubMed: 20542247]

Yu D, Kim M, Xiao G, Hwang TH. Review of biological network data and its applications. Genomics & informatics. 2013; 11: 200–210. [PubMed: 24465231]

Zhang W, Chien J, Yong J, Kuang R. Network-based machine learning and graph theory algorithms for precision oncology. npj Precision Oncology. 2017; 1: 25. [PubMed: 29872707]

Zhou H, Xu M, Huang Q, Gates AT, Zhang XD, Castle JC, Stec E, Ferrer M, Strulovici B, Hazuda DJ, Espeseth AS. Genome-scale RNAi screen for host factors required for HIV replication. Cell Host Microbe. 2008; 4: 495–504. [PubMed: 18976975]

Zhu J, Davoli T, Perriera Jill M, Chin Christopher R, Gaiha Gaurav D, John Sinu P, Sigiollot Frederic D, Gao G, Xu Q, Qu H, Pertel T, et al. Comprehensive Identification of Host Modulators of HIV-1 Replication using Multiple Orthologous RNAi Reagents. Cell Reports. 2014; 9: 752–766. [PubMed: 25373910]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

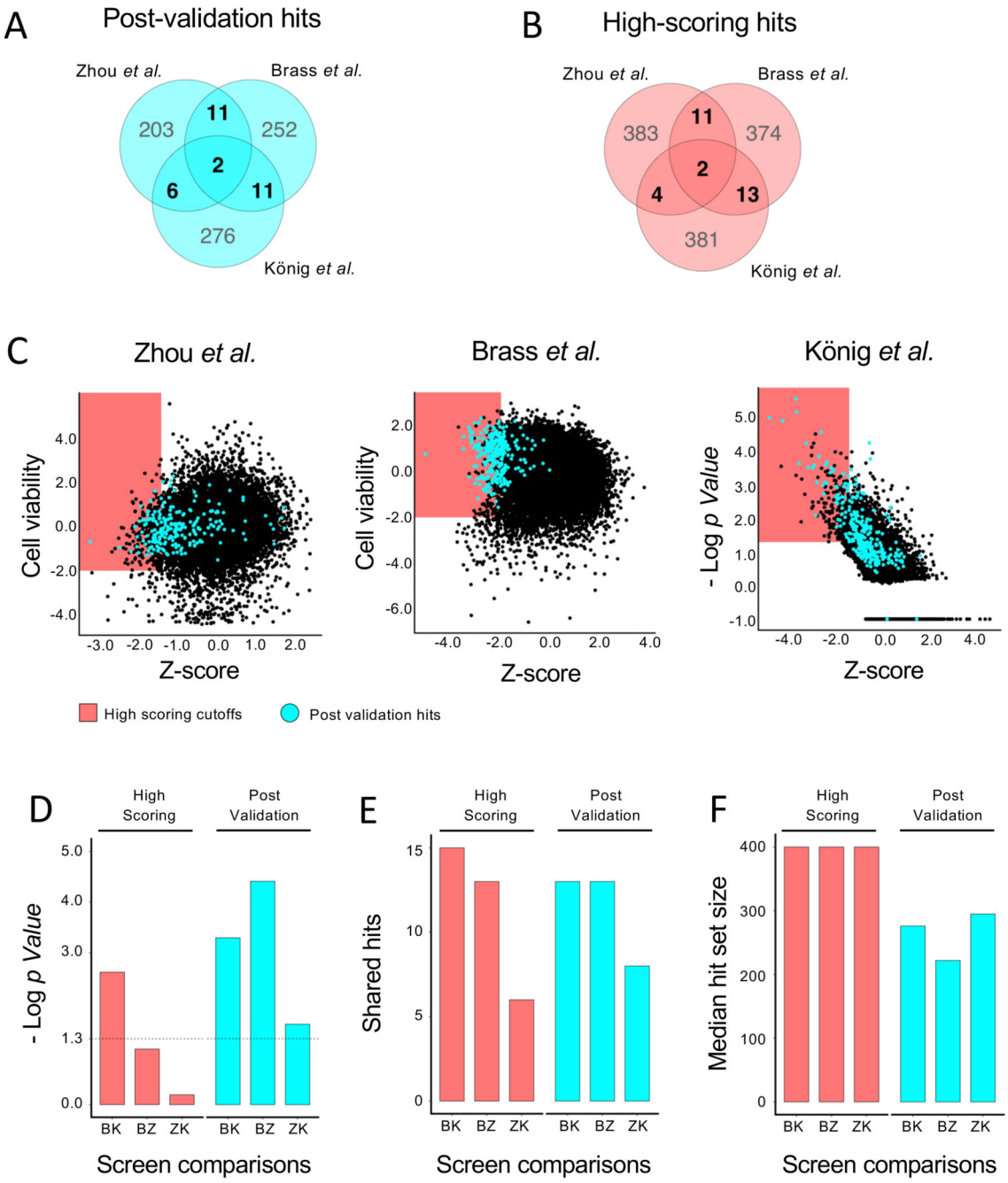


Figure 1. Parallel genome-wide siRNA studies have limited overlap in hits selected by highest score or post-validation.

(A, B) Venn diagram of shared and unshared hits for the three studies of HIV HDFs from the post validation sets (A) or selected by highest score (B). (C) Normalized data from the three siRNA studies of HDFs. Red highlighted area indicates highest scoring hits. Genes highlighted in light blue were selected as post validation hits by the respective studies. (D) Negative Log p Values of the statistical enrichment between the hit sets of the three siRNA HDF studies for hits selected by high score cutoff (red) and post-validation hits selected by analysis and secondary screening (blue). Threshold is set at -Log(0.05). (E) Number of

shared hits in two screen comparisons of the three siRNA studies of HDFs. (F) Median size of the hit selection sets for each two-way comparison of the three siRNA studies of HDFs.

Europe PMC Funders Author Manuscripts

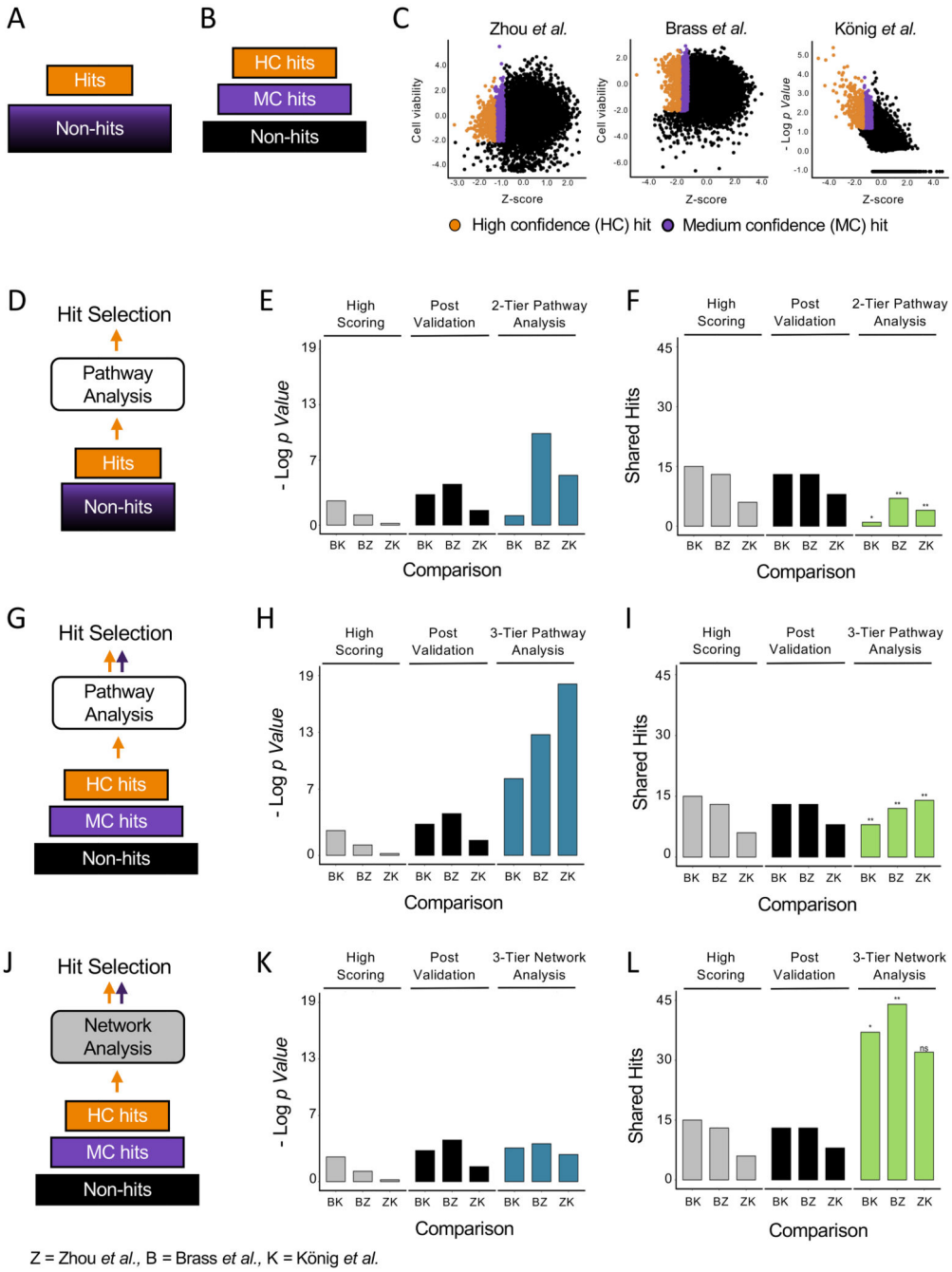


Figure 2. Prioritization of candidates by pathway enrichment or network analysis are complementary in hit selection solutions.

(A) Schematic of a single cutoff, two tier data approach. (B) Schematic of a dual cutoff, three tier data approach. (C) Scores from three genome-wide studies of HDF. Normalized scores are plotted on the x-axis and secondary scores that were considered (such as cell viability and assigned p-values) are on the y-axis. Genes with both scores above the cutoff are in orange and genes with Z scores above the secondary cutoff are in purple. (D) Schematic of the pathway analysis approach for hit selection. Candidates are divided by a single cutoff. (E, F) statistical significance of the overlap (E) and number of shared hits (F)

Europe PMC Funders Author Manuscripts

selected by pathway analysis of 2-tiered data versus highest scoring and post validation hits. (G) Schematic of the pathway analysis approach for hit selection from a three-tiered dataset. (H, I) statistical significance of the overlap (H) and number of shared hits (I) selected by pathway analysis of 3-tiered data versus highest scoring and post validation hits. (J) Schematic of the network analysis approach for hit selection. (K, L) statistical significance of the overlap (K) and number of shared hits (L) selected by network analysis of 3-tiered data versus highest scoring and post validation hits. Random permutation test scores: ns = $p > 0.05$, * = $p \leq 0.05$, ** = $p \leq 0.01$.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

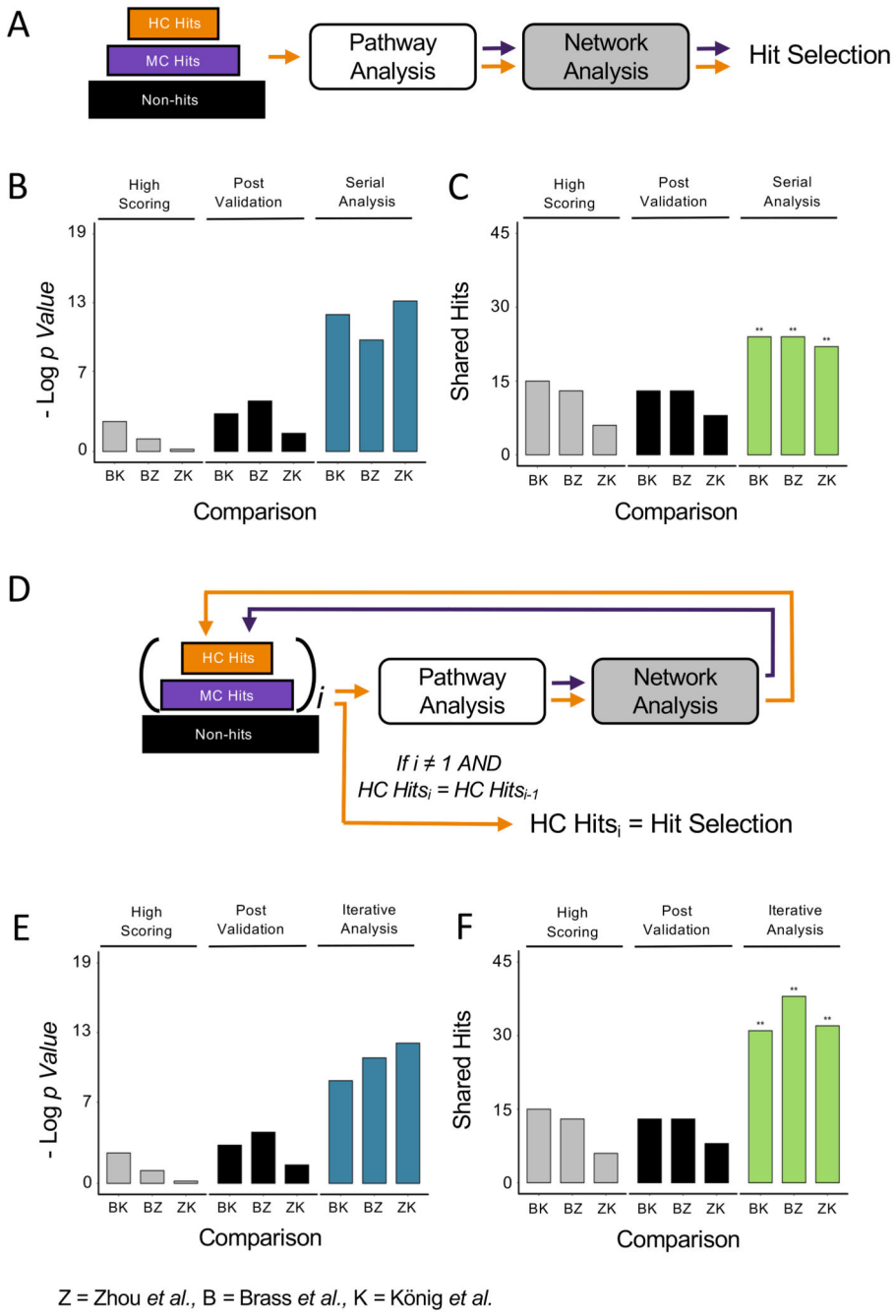


Figure 3. Integrated and iterative approaches to pathway and network analysis improve overlap by multiple measures.

(A) Schematic of the serial analysis approach for hit selection. Pathway analysis is applied to high confidence hits. High confidence and medium confidence hits from enriched pathways are assigned as high confidence hits. Network analysis is applied to the set of high confidence hits. Medium confidence hits that have predicted interactions with high confidence hits are added to the final hit set. (B, C) statistical significance of the overlap (B) and number of shared hits (C) selected by serial analysis of 3-tiered data versus highest scoring and post validation hits. (D) Schematic of the iterative analysis approach for hit

selection. i = number of iterations. Pathway and network analysis are sequentially applied as in the integrated approach. When $i > 1$, if the set of high confidence hits at the end of the current iteration (HC_i) is the same as the set of high confidence hits from the end of the previous iteration (HC_{i-1}) high confidence hits are used as the final hit set from the study. If high confidence set of hits are different, another iteration of integrated analysis is applied. (E, F) statistical significance of the overlap (E) and number of shared hits (F) selected by iterative analysis of 3-tiered data versus highest scoring and post validation hits. Random permutation test scores: ns = $p > 0.05$, * = $p \leq 0.05$, ** = $p \leq 0.01$.

Europe PMC Funders Author Manuscripts

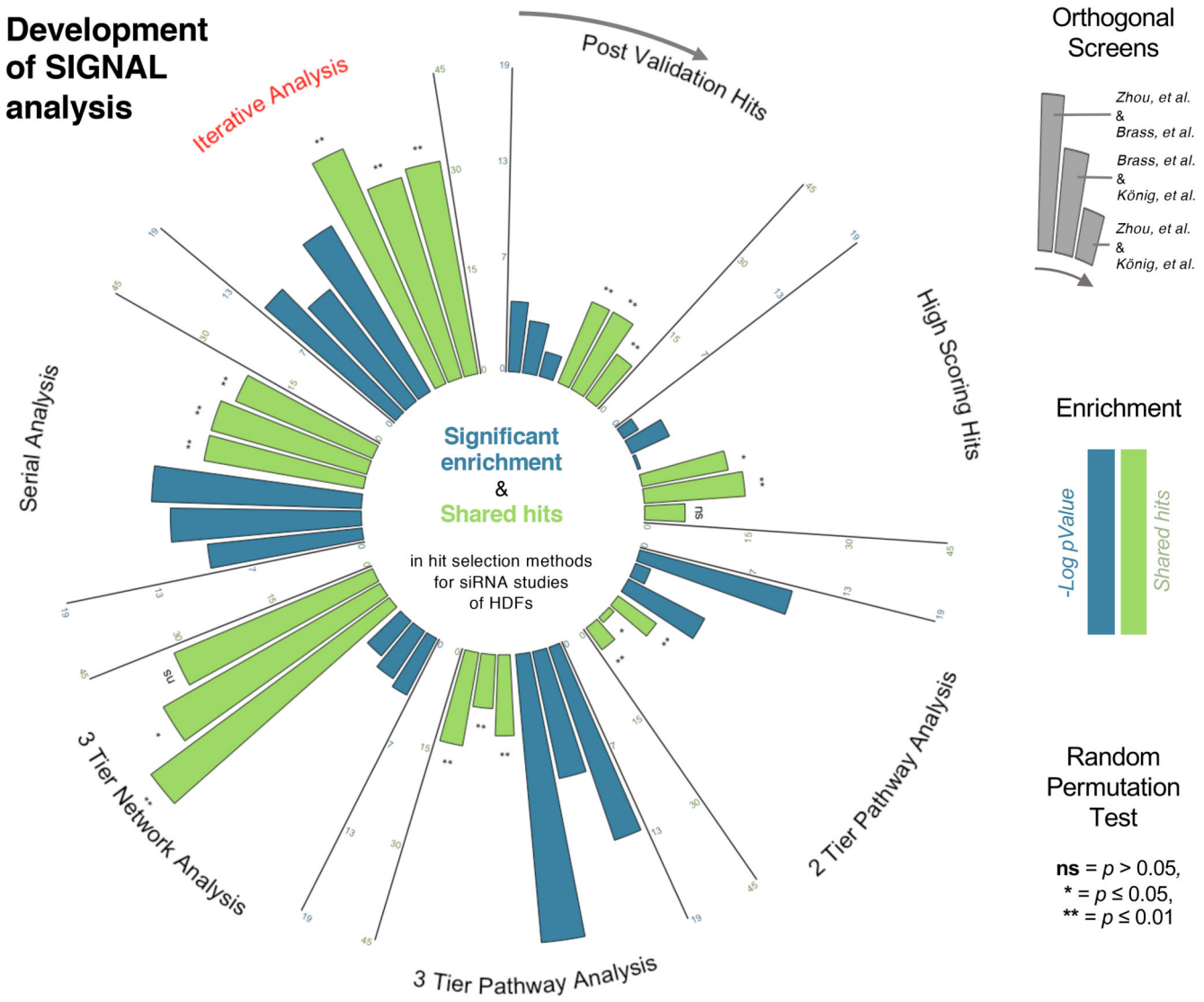


Figure 4. Throughput Ranking by Iterative Analysis of Genomic Enrichment (SIGNAL).

Summary and comparison of the multiple approaches to hit selection evaluated by false positive correction (measured by significance of overlap (blue)) and false negative correction (measured by number of shared hits (green)). Hit selection methods: (clockwise from top center): post validation hits, high scoring hits, pathway analysis using a two-tiered dataset, pathway analysis using a three-tiered dataset, network analysis, serial integration of pathway and network analysis, iterative integration of pathway and network analysis.

Europe PMC Funders Author Manuscripts

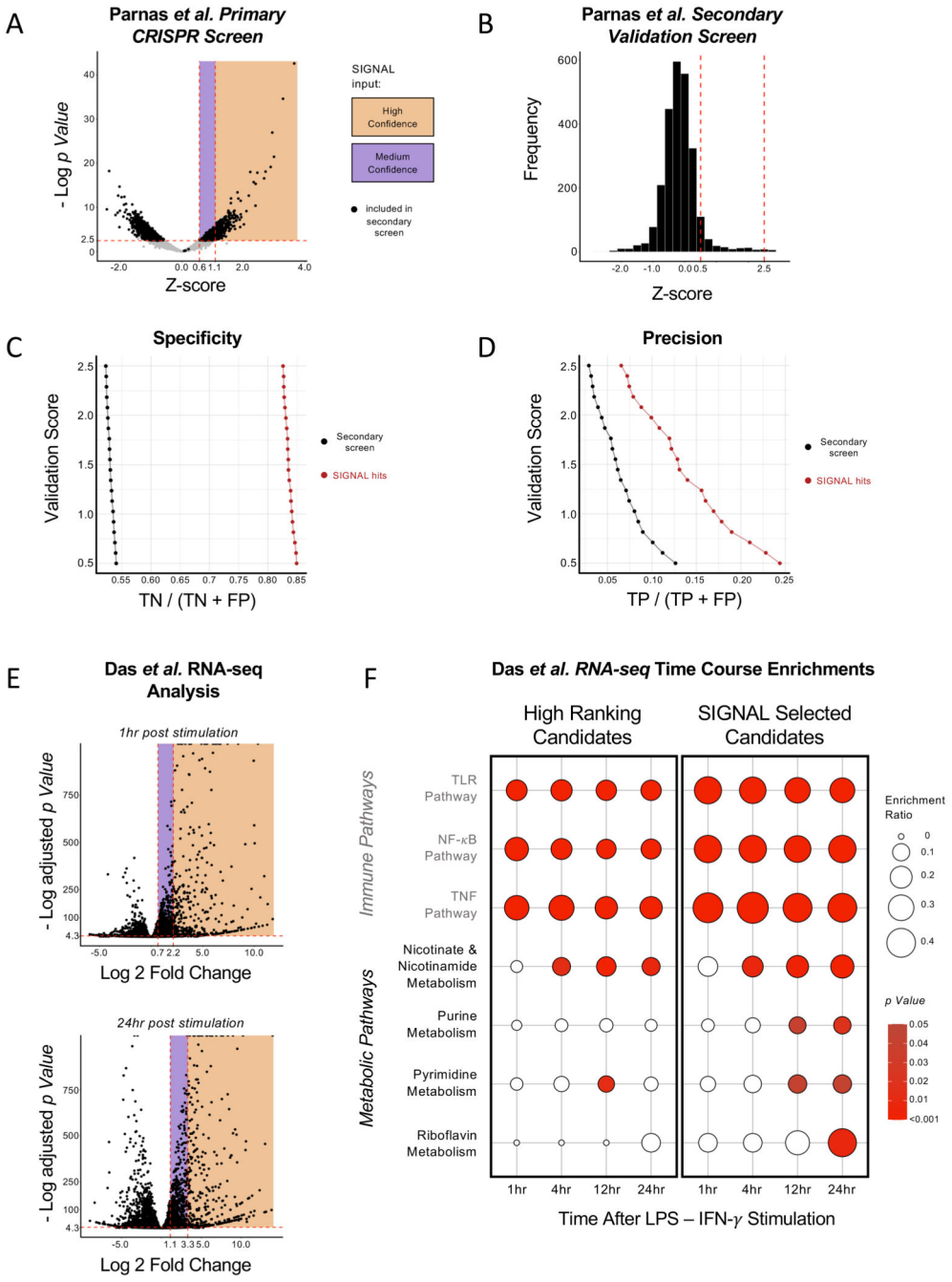


Figure 5. Increased specificity, precision, and elucidation of pathway enrichment by SIGNAL in CRISPR and RNA-seq studies.

(A) Primary screen data from Parnas et al. Hits selected in the initial study for the secondary screen are in black, hits in the area highlighted in orange were assigned as high confidence hits for SIGNAL analysis and hits in the area highlighted in purple were assigned as medium confidence hits. (B) Histogram of Z-scores from the secondary screen by Parnas et al., the limits of scores used for validation cutoffs to test SIGNAL are indicated by the red dotted line. (C) Specificity of hits selected for secondary analysis (black) and hits selected by SIGNAL (red) across a range of 20 validation cutoffs. (D) Precision of hits selected

Europe PMC Funders Author Manuscripts

for secondary analysis (black) and hits selected by SIGNAL (red) across a range of 20 validation cutoffs. (E) RNA-seq data from Das et al. at 1 hour (top) and 24 hours (bottom) after stimulation by LPS and IFN-γ, hits in the area highlighted in orange were assigned as high confidence hits for SIGNAL analysis and hits in the area highlighted in purple were assigned as medium confidence hits. (F) Pathway enrichment of the highest-ranking candidates from Das et al. RNA-seq (left) and SIGNAL selected candidates (right).

Europe PMC Funders Author Manuscripts

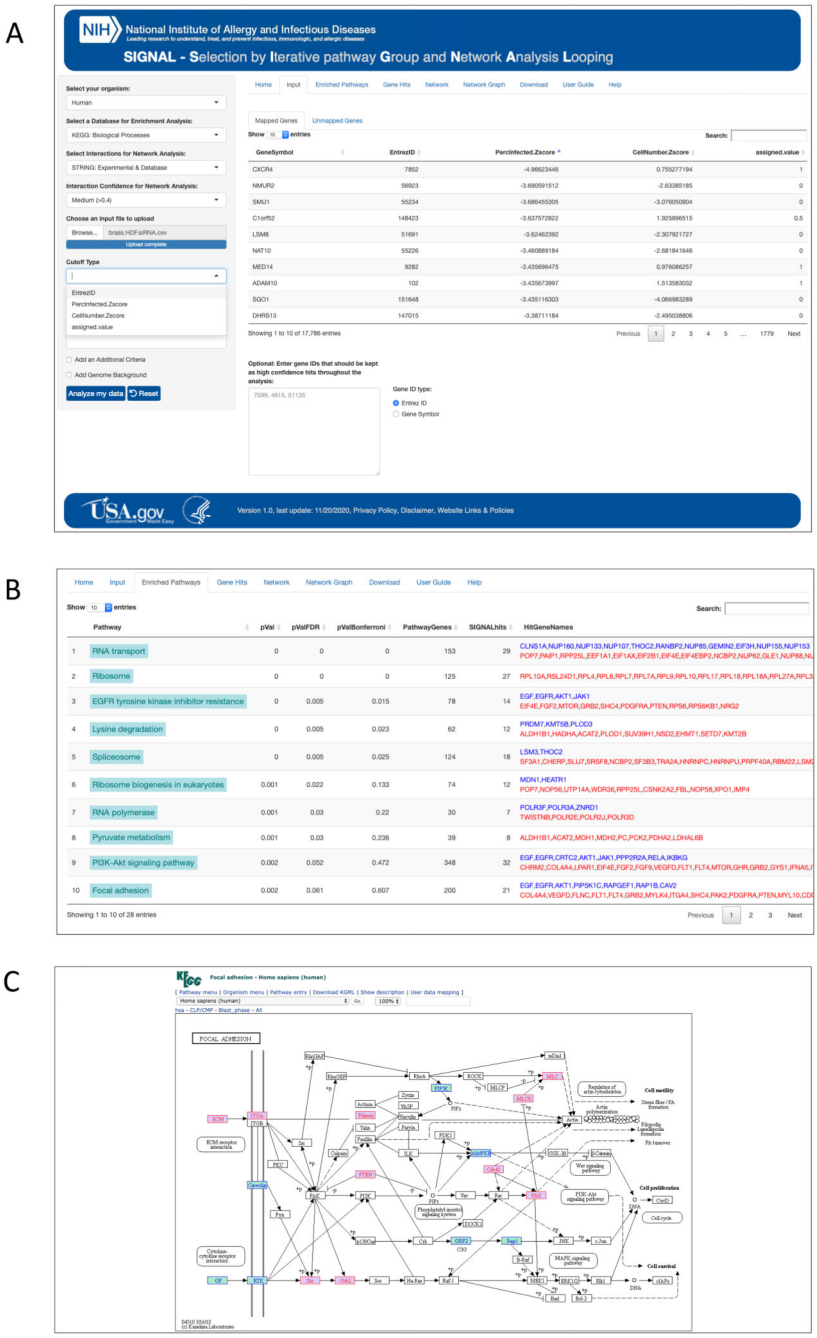
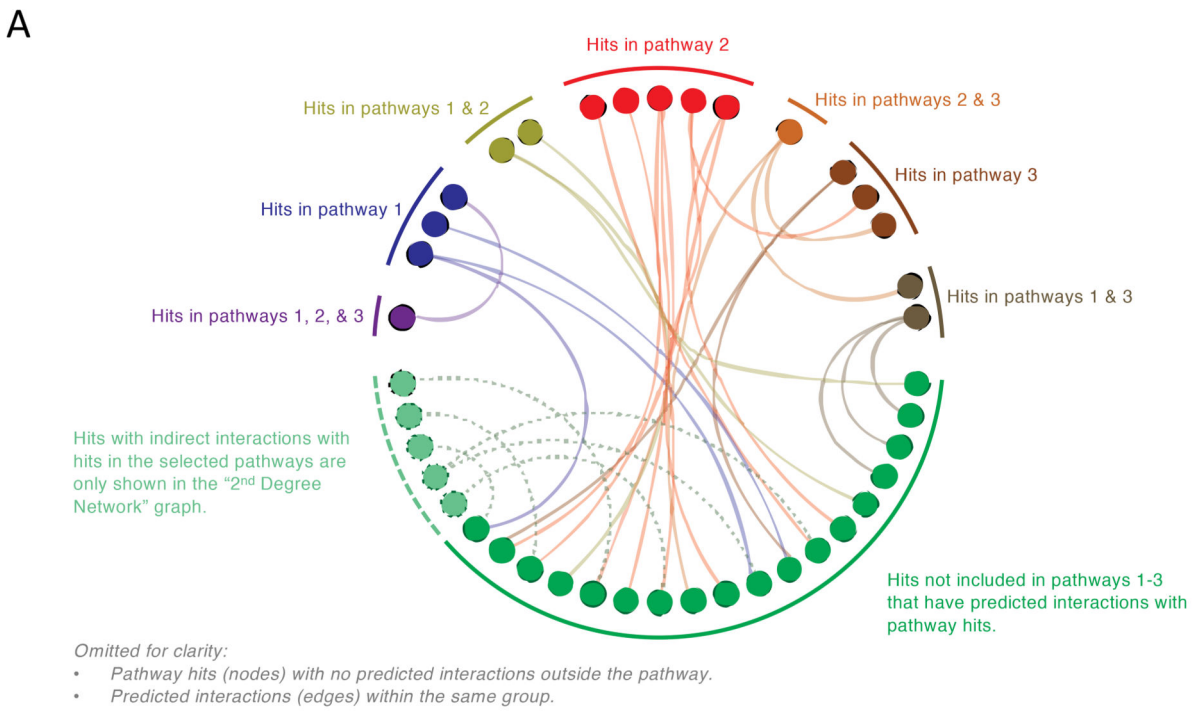


Figure 6. Uploading and analyzing data on signal.niaid.nih.gov.

(A) The landing page of signal.niaid.nih.gov after a gene list has been uploaded. (B) Pathway enrichment tab on SIGNAL. Genes listed on right from enriched pathways are color-coded based on their assignment in the uploaded gene list; High confidence (blue) and medium confidence (red). (C) A mapped KEGG pathway linked from a SIGNAL-output with candidates ranked as high confidence in the input highlighted in blue and candidates ranked as medium confidence highlighted in red.

Europe PMC Funders Author Manuscripts



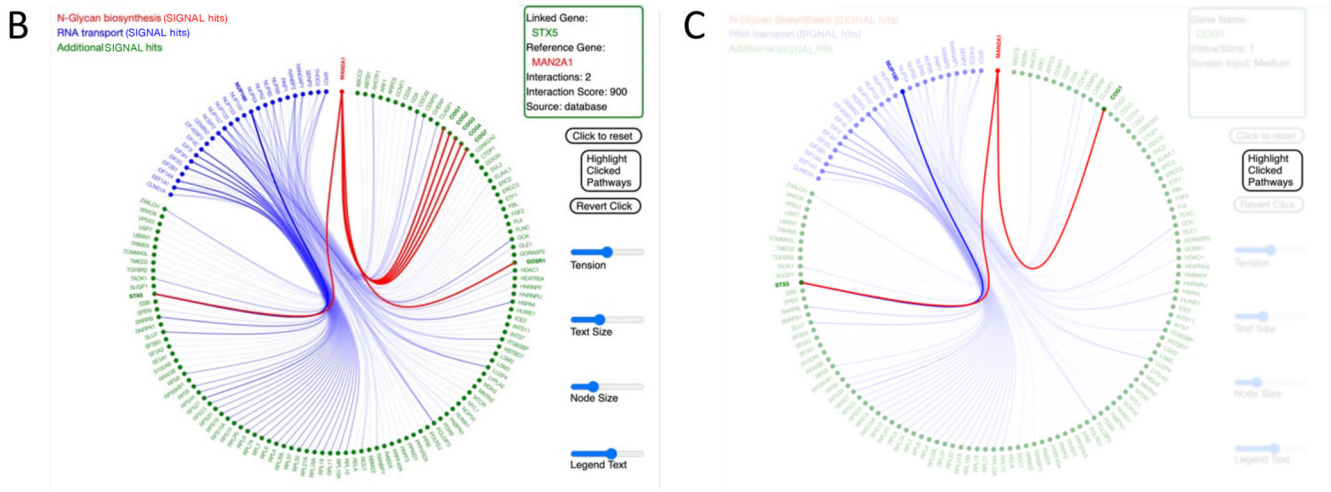


Figure 7. An interactive version of pathway and gene networks enables exploration of putative missing links.

(A) Structure of a SIGNAL network visualization map integrating pathway and network information utilizing hierarchical edge bundling. (B) An interactive version of the pathway and gene network graph in SIGNAL. This example shows the results following the selection of the "RNA transport" and "N-Glycan biosynthesis" pathways in the SIGNAL analysis of the Brass et al. study (Brass et al., 2008) of essential factors for HIV infection. Information

about different nodes and edges appear in the window at top right. (C) “Highlight Clicked Pathway” option in SIGNAL highlights the clicked-on genes and their interactions.

Europe PMC Funders Author Manuscripts