Supplemental Methods

Parameter Optimization

We first optimized the algorithm and parameters to minimize selection of "hub" nodes (or genes that are highly connected due to publication bias). We set these as our "grey listed" targets and quantify their representation in 100 runs of the algorithm at various degree penalty values (controlled by the mu parameter; data now shown).

We further optimized the remaining parameters - β, D, and ω. These parameters control the algorithm's inclusion of experimental genes ("terminal nodes") relative to non-experimental genes ("hidden nodes"), network size, and relative density. We ran PCSF over 36 parameter configurations, using all combinations of the following sets: β [1,10,100], ω [0.1, 1, 10, 100], and D [10, 20, 30]. Briefly, β sets the relative cost of capturing gene (node) prizes with adding edges, ω tunes the initial cost associated with connecting the dummy node to all terminal prizes in the input set, and the depth, D, controls how long a given pathway goes in the final network. We looked for a balance of efficiency (ratio of terminal nodes:hidden nodes), network size, and tree composition (rejecting parameter sets that contain any sub-trees with 3 or fewer nodes.

We found the optimal network parameters by looking for a high efficiency (high terminal:hidden node ratio), reasonable network size[1] relative to the input set, and exclusion of grey-listed nodes (Supplementary Figure 1). After vetting the sub-network creation, we selected an optimal parameter setting of β=1, D=30, ω=10, and μ = 0.006 (Figure 3).

We also considered representation of hub nodes, or nodes that have extremely high connectivity in the interactome, and consider how often a parameter set includes these hub nodes in the final solution. Prior to running PCSF simulations, we identified a set of grey-listed nodes by ranking all nodes in our interactome by degree and taking those with a degree > 200 to add to the grey list. At each parameter set, we counted the fractional representation of these grey nodes in 100 runs of PCSF with random input prizes. Each heat map shows this fractional representation across the previously mentioned 36 parameter configurations. The final optimal parameters were β=10, D=10, ω=1, and μ=0.006.

[1] Looking for a reasonable sized network refers to finding a gene target set that is on the order of magnitude with the original input set. A tightly constrained network that is much smaller than the input set would likely contain the highest density of true positives, but would include fewer false negatives. A less constrained, much larger network is more likely to contain find false negatives from the screen, but also contain many false positives.