SUPPLEMENTAL PROTOCOL S1

Biological material growth

Medicago truncatula seeds were grown in aeroponic tanks (caissons) under the following controlled environmental conditions: temperature, 21°C; humidity, 75%; light intensity, 300 µmol m^{-2} s^{-1}; and light/dark photoperiod, 16 h/8 h. Each caisson contained 10 L of low-nitrogen aeroponic medium: *Sinorhizobium medicae* ABS7M (pXLGD4) was grown in TY medium supplemented with 6 mM calcium chloride and 10 µg mL^{-1} tetracycline at 28ºC for 48 h. The culture was washed three times and finally resuspended in 10 mL sterile distilled water to an OD_{600} of 1.0, which was used to inoculate aeroponic caissons containing 10 L of low-nitrogen aeroponic medium (1 mM $CaCl_2$, 0.25 mM MgSO4.7H2O, 1.7 mM KH_2PO4, 3.8 mM K_2HPO4, 0.5 mM K_2SO4, 0.05 mM $FeSO_4.7H_2O$, 0.05 mM Na_2EDTA, 30 µM H_3BO_3, 10 µM $MnSO_4$, 0.7 µM $ZnSO_4$, 1 µM $Na_2MoO_4.2H_2O$, 0.04 µM $CoCl_2$, and 0.2 µM $CuSO_4.5H_2O$).

Prediction of new regulators based on proximity to known symbiosis genes

Based on literature survey we generated a collection of seven high-confidence genes (*NSP2*, *NF-YA1*, *LYK3*, *NIN*, *ERN1*, *NFP* and *RPG*) and eight medium-confidence genes (*LYK4*, *-5*, *ARR8*, *-9*, *NPL*, *LYK10*, *NAP85* and *FLOT2*). To identify new genes that are associated with symbiosis, we ranked all genes based on their proximity on the inferred regulatory connections from module 227 to these input genes. We defined this network by including all genes that were in module 227, all regulators predicted to regulate genes in module 227 and all targets of regulatory genes associated with module 227. To use the network to define a ranking we used a regularized Laplacian graph kernel (Smola and Kondor, 2003). The advantage of using a graph kernel approach, as opposed to a simple graph-based connectivity measure such as shortest paths, is that the derived measures of connectivity are more global and take into account all possible paths that can connect two genes. Because we had two sets of input genes, we weighed them separately, considering the high confidence genes with weight 1, and the medium confidence genes with weight 0.5. We found that this weighting configuration gave the best performance based on leave-one-out cross validation. The graph kernel uses a smoothing parameter to how fast signal diffuses from a node in the graph. The higher the value, the slower the decay. We examined different settings of the smoothing parameter (1, 5, 10, 100) and set this to 100 based on cross-validation and the observation that higher values allowed for the ranking of more genes.

Novel subnetwork identification

We found that several annotated gene sets such as those associated with the cell cycle were highly enriched for interactions both expression-based as well as those in the STRING network. This prompted us to ask whether we could identify novel sub-networks within the network associated with module 227. We applied a spectral clustering algorithm on the largest connected component of the network to obtain 10 clusters. Spectral clustering is meant for clustering entities that are connected by a graph, grouping the entities based on their global network connectivity similarity. We applied an algorithm that first generates the Laplacian of the graph (L= I - D^-0.5 A D^-0.5), computes the first (smallest) m eigenvectors, and clusters it using a k-means algorithm into m clusters (Von Luxburg, 2007). We examined 5, 10 and 20 clusters, however, found that m=10 to given the best results in terms of the number of clusters enriched for a specific term. Six of these clusters were enriched for DNA metabolism, cell cycle, hormone processes, and symbiosis.

REFERENCES

Smola, A.J., and Kondor, R. (2003). Kernels and regularization on graphs. *In*: Learning theory and Kernel machines. Berlin, Germany. Springer.144-158.

Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and computing. **17:** 395-416.