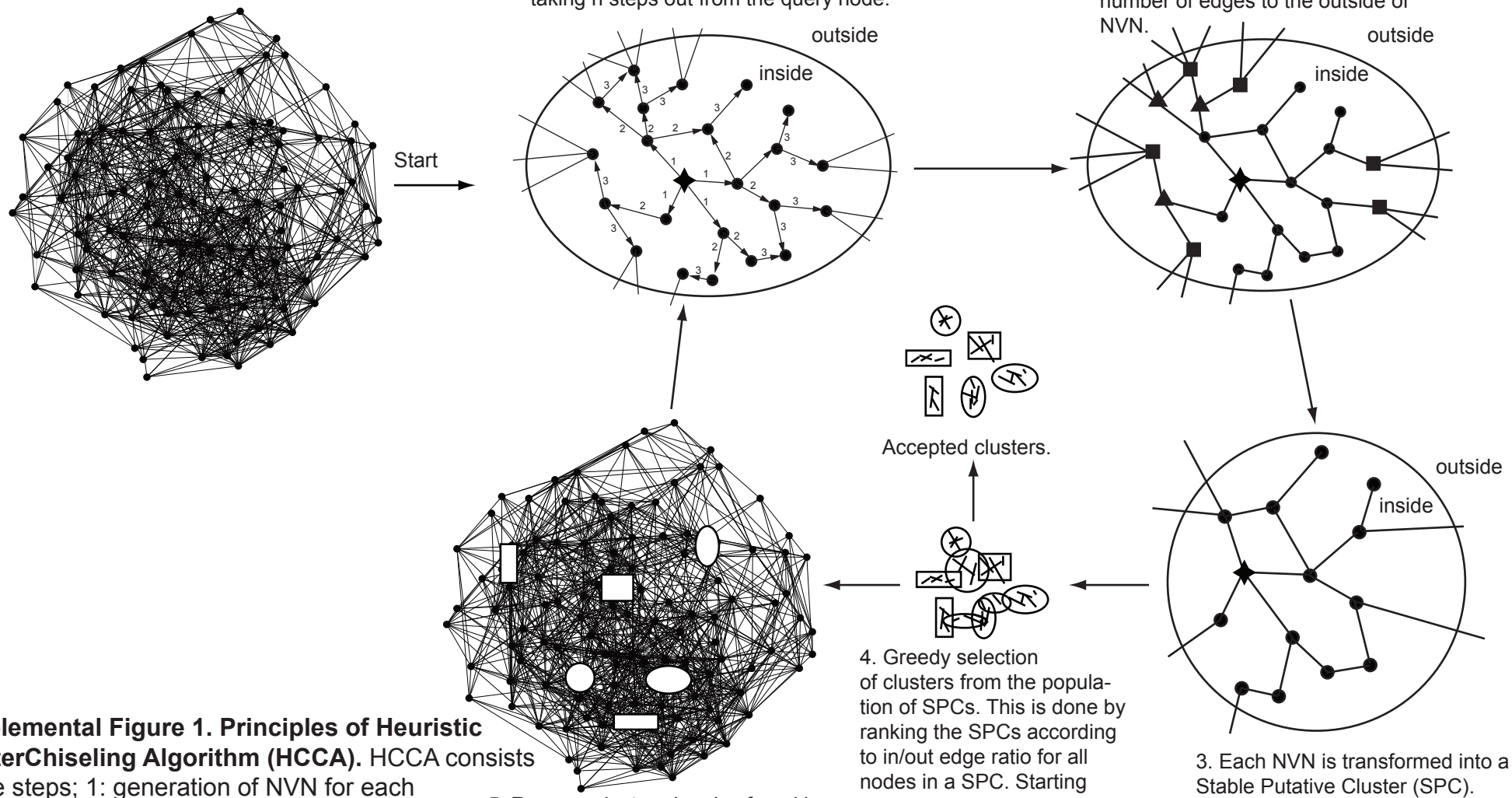
Supplemental Data. Mutwil et al. (2011).
Plant Cell 10.1105/tpc.111.083667.



Supplemental Figure 1. Principles of Heuristic ClusterChiseling Algorithm (HCCA). HCCA consists of five steps; 1: generation of NVN for each node in the network by taking in the n-step neigborhood (3-step neighborhood used in this study), 2: removal of nodes more connected to the nodes outside of NVN which results in, 3: generation of SPCs, 4: selection of clusters from the population of SPCs (preferred cluster size range 40-200 was selected for this study), 5: removal of clustered nodes from the network and return to step 1. A more detailed explanation and comparison of HCCA to other clustering algorithms is available in Mutwil et al. 2010.

1) Enter gene of interest

Probe set ID
Locus ID
Keyword

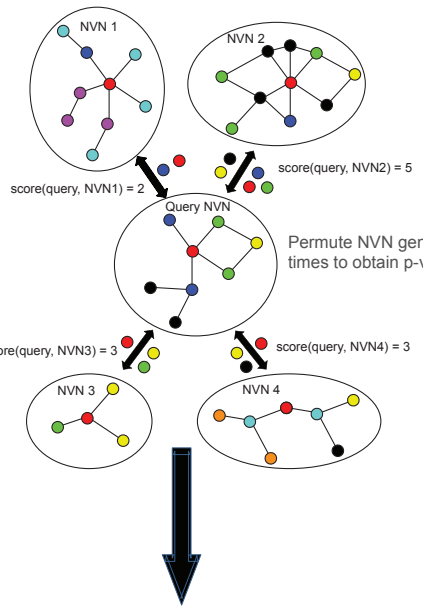
Query gene belongs to "red" PFAM family

2) Generate NVN for each member of the "red" PFAM gene family

NVNs centered around genes belonging to red PFAM families are generated for all species in database.

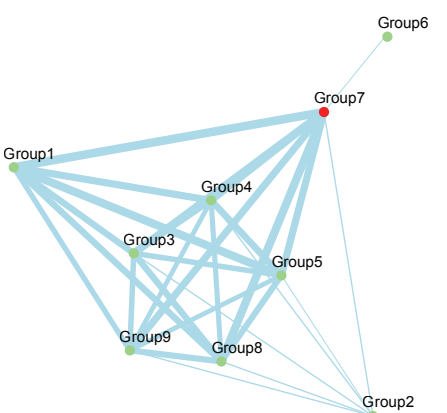
3) NetworkComparer - comparison of query NVN to NVNs centered around red PFAM

Computation of similarity of query NVN to every NVN centered around a gene belonging to red pfam family for all species in the database. Score of 1 is given for each gene family the query NVN and target NVN have in common. Colors of nodes depict different PFAM families.



4) Selection step: Analysis of co-expression group similarities and selection of NVNs for further analysis

The analysis returns three elements:



-A network depicting how similar the analyzed NVNs are to one another. Since certain NVNs are overlapping (i.e. genes belonging to the red PFAM family are co-expressed), they are joined into one group. For example, *AtPSAD1* and *AtPSAD2* (*At4g02770*,*At1g03130*) are co-expressed and therefore represented by one group (group 7). A group containing the query gene is colored red. Any groups which are significantly similar ($p<0.01$) are connected. Thickness of an edge connecting any two groups is proportional to the similarity score those two groups have in common.

The table below shows score of your query gene to all genes found in the family: PsaD

Group:	Network:	Probeset ID	Transcript ID	Score	p-value
Group: 1	ExpMatBar.hrr	Contig4015_at	p36213	23	0.0
Group: 2	ExpMatMedi.hrr	Mtr.34726.1.S1_at	medtr1g132590.1	4	0.0
Group: 3	ExpMatMedi.hrr	Mtr.19267.1.S1_at	medtr5g006220.1	21	0.0
Group: 4	ExpMatRice.hrr	Os.7974.1.S1_at	LOC_Os03g09220.3	20	0.0
Group: 5	ExpMatWheat.hrr	Ta.24304.2.S1_a_at	UniRef90_P36213	20	0.0
Group: 6	ExpMatWheat.hrr	TaAffx.58068.1.S1_at	UniRef90_P36213	1	0.033
Group: 7	ExpMatAra.hrr	263114_at 255457_at	at1g03130 at4g02770	9 31	0.0 0.0
Group: 8	ExpMatPop.hrr	Ptp.5240.1.S1_s_at PtpAffx.71066.5.A1_s_at	gi|poptr1_1|564827|eugene3.00081422 gi|poptr1_1|566261|eugene3.00100819	18 8	0.0 0.0
Group: 9	ExpMatSoy.hrr	Gma.10852.1.S1_a_at Gma.10852.2.S1_at	Glyma10g39460.1 Glyma20g28300.1	19 16	0.0 0.0

-A table showing score of the query gene NVN to all NVNs belonging to the same PFAM family. Red arrows indicate NVNs selected for further analysis.

The table below shows which pfam families are enriched in vicinity networks of PsaD family in analyzed species.

Family:	Description	Times found in analyzed vicinity networks
PsaD	PsaD. This family consists of PsaD from plants and cyanobacteria.	12
Chloroa_b-bind	Chlorophyll A-B binding protein.	11
PsbP	PsbP. This family consists of the 23 kDa subunit of oxygen evolving system of photosystem II or PsbP from various plants (where it is encoded by the nuclear genome) and Cyanobacteria.	10
PSI_PsaH	Photosystem I reaction centre subunit VI.	9
MSP	Manganese-stabilising protein / photosystem II polypeptide.	9
PSI_PSAK	Photosystem I psaG / psaK.	8
NAD_binding_1	Oxidoreductase NAD-binding domain. Xanthine dehydrogenases, that also bind FAD/NAD, have essentially no similarity.	8
AAA	ATPase family associated with various cellular activities (AAA).	8
PsbY	Photosystem II protein Y (PsbY).	7
PsbX	Photosystem II reaction centre X protein (PsbX).	7
PSI_PsaE	Photosystem I reaction centre subunit IV / PsaE.	7
FBPase	Fructose-1-6-bisphosphatase.	7
Copper-bind	Copper binding proteins, plastocyanin/azurin family.	7
PsbO	Oxygen evolving enhancer protein 3 (PsbQ).	6
PsaN	Photosystem I reaction centre subunit N (PSAN or PSI-N).	6
PsaL	Photosystem I reaction centre subunit XI.	6
PSI_PsaF	Photosystem I reaction centre subunit III.	6

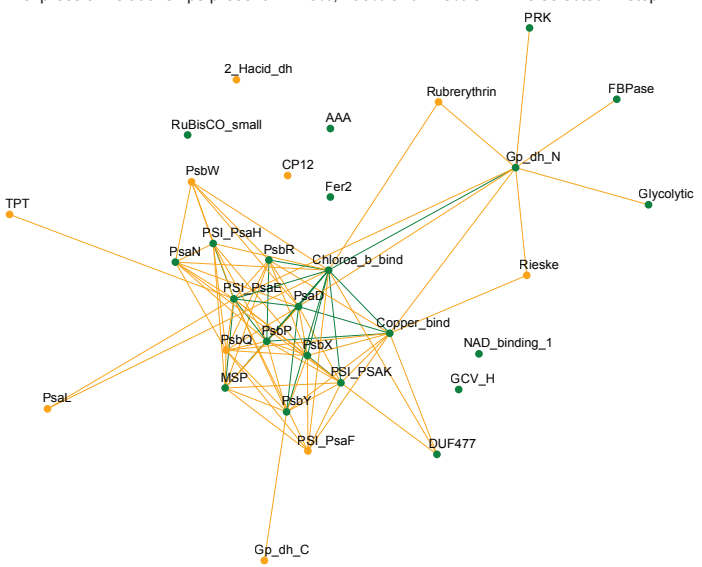
-A table reporting the frequency of specific PFAM families being present in NVNs of the red PFAM family.

5) Generation of ancestral network and revealing transcript indentities

The program displays the ancestral network and table that displays the identity of potential functional homolog.

6) Result

The analysis returns two elements:
-Ancestral network depicting conserved PFAM families and co-expression relationships between them. Green, orange and red nodes and edges represent PFAMs and co-expression relationships present in ≥75%, ≥50% and ≥25% of NVNs selected in step 4.

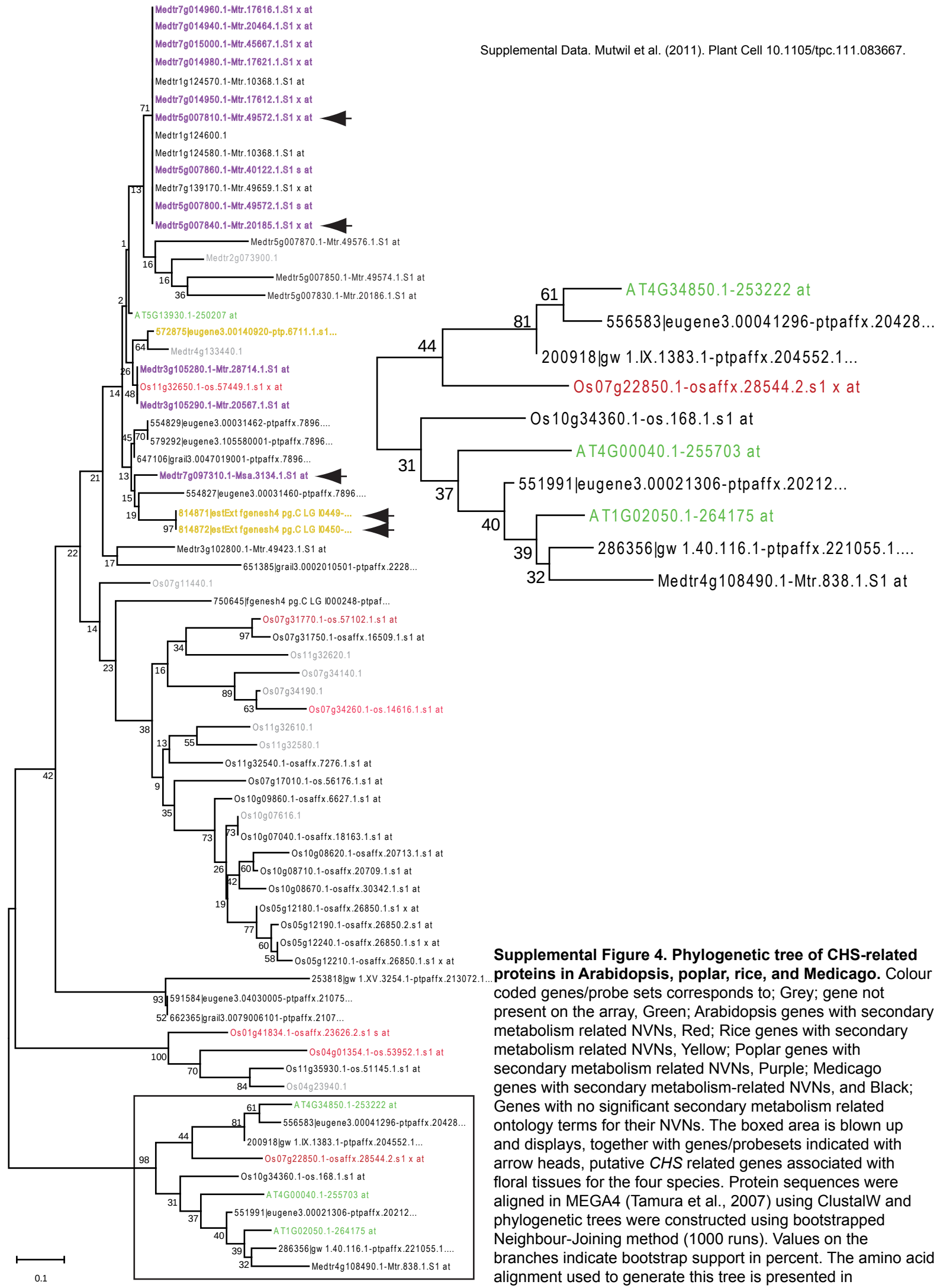


-A table revealing the probesets (genes) constituting the enriched PFAMs.

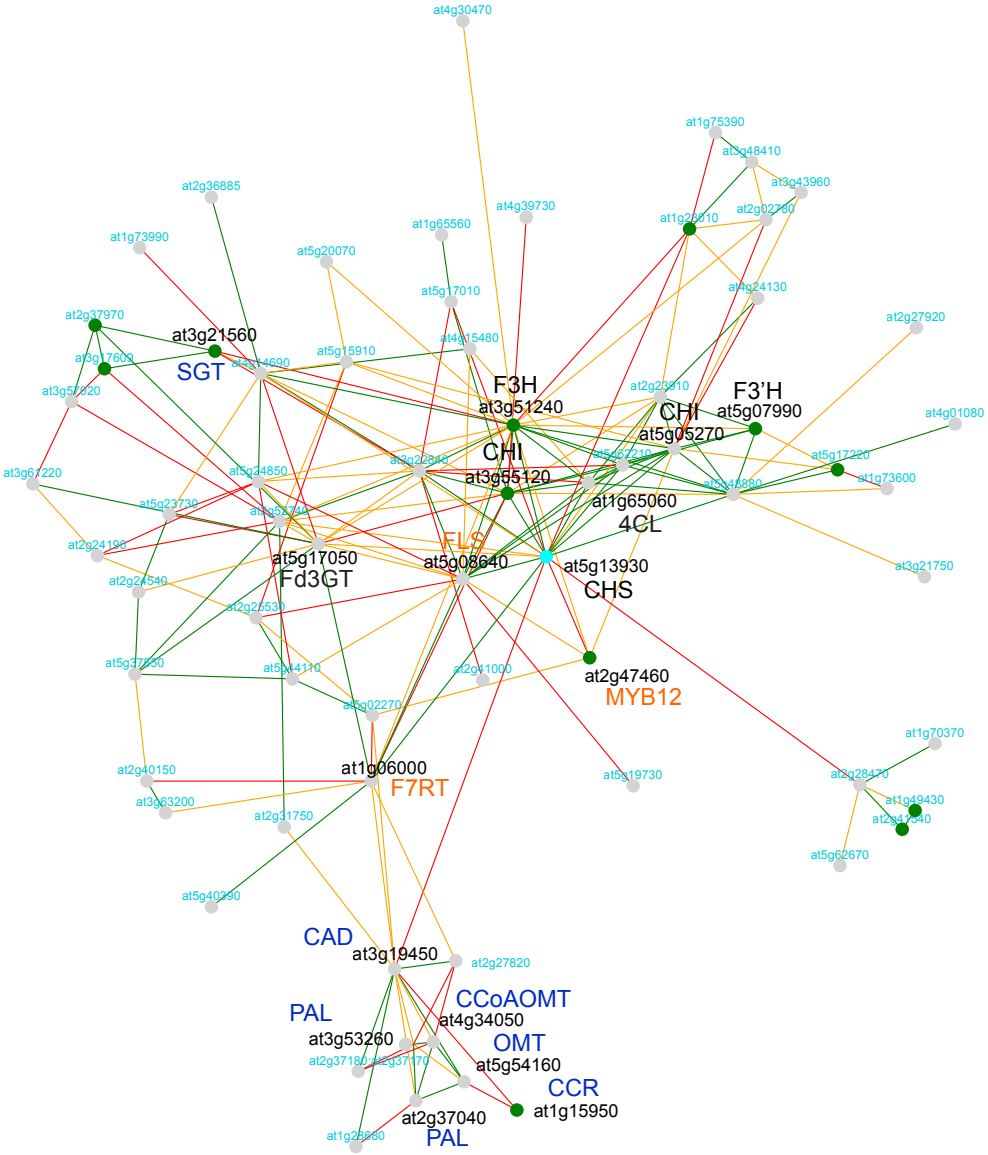
The table below displays the probeset identity of enriched families in selected species

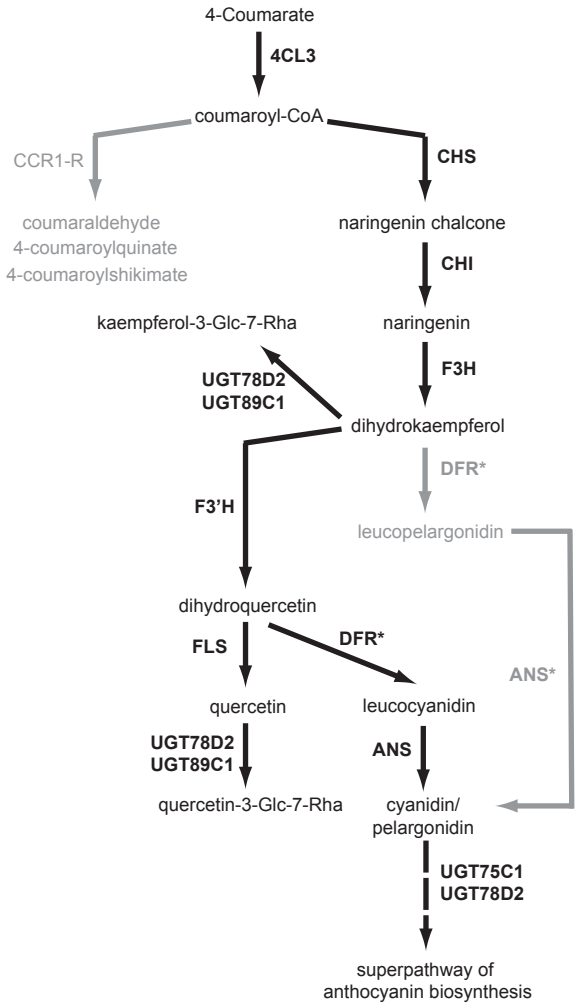
gene family	Description	Mtr.19267.1.S1_at	255457_at	Ptp.5240.1.S1_s_at
PsaD	PsaD. This family consists of PsaD from plants and cyanobacteria. PsaD is an extrinsic polypeptide of photosystem I (PSI) and is required for native assembly of PSI [reaction centers] and is implicated in the electrostatic binding of ferredoxin within the reaction centre. PsaD forms a dimer in solution which is formed by PsaE,however PsaD is monomeric in its native complexed PSI environment.	medtr5g006220.1	at4g02770	gi|poptr1_1|564827|eugene3.00081422
Chloroa_b-bind		medtr2g008680.1 79medtr2g100320.2,Medtr2g100320.1 medtr2g104420.1 medtr2g104980.1 medtr4g034420.1 medtr4g129180.1 medtr2g093540.1 medtr4g127240.1 AC235962_10.1.A_C235657_17.1	at3g47470 at5g01530 at1g44575 at5g03840 at1g44900 at1g61520 at1g61520 at3g47470 at4g10340	gi|poptr1_1|744667|estext_genewise1_v1.c_1340185 gi|poptr1_1|814251|estext_fgenesh4_kg.c_lg_s00033.1|jgi|poptr1_1|744667|estext_genewise1_v1.c_1340185 gi|poptr1_1|653111|grail2.0023650901 gi|poptr1_1|713465|estext_genewise1_v1.c_lg_v12774 gi|poptr1_1|599710|eugene3.00200046 gi|poptr1_1|671313|grail2.0120903701 gi|poptr1_1|749330|eugene3.00010473 gi|poptr1_1|743510|estext_genewise1_v1.c_1250189 gi|poptr1_1|652073|grail2.0002067901 gi|poptr1_1|666834|grail2.0012050701 gi|poptr1_1|850949|estext_fgenesh4_rm.c_lg_s00952.1|jgi|poptr1_1|664073|grail3.0012031401 gi|poptr1_1|724963|estext_genewise1_v1.c_lg_s2607
PsbP	PsbP. This family consists of the 23 kDa subunit of oxygen evolving system of photosystem II or PsbP from various plants (where it is encoded by the nuclear genome) and Cyanobacteria. The 23 KDa PsbP protein is required for PSII to be fully operational in vivo, it increases the affinity of the water oxidation site for Cl- and provides the conditions required for high affinity binding of Ca2+	medtr2g035290.1 79medtr2g103130.2,Medtr2g103830.2	at1g06680	gi|poptr1_1|818844|estext_fgenesh4_pg.c_lg_v1224 gi|poptr1_1|551203|eugene3.00020510
PSI_PsaH	Photosystem I reaction centre subunit VI. Photosystem I (PSI) is an integral membrane protein complex that uses light energy to mediate electron transfer from plastocyanin to ferredoxin.	medtr2g122810.1	at3g16140 at1g52230	gi|poptr1_1|712069|estext_genewise1_v1.c_lg_i11093
MSP	Manganese-stabilising protein / photosystem II polypeptide. This family consists of the 33 KDa photosystem II polypeptide from the oxygen evolving complex (OEC) of plants and cyanobacteria. The protein is also known as the manganese-stabilising protein or it is associated with the manganese complex of the OEC and may provide the ligands for the complex.	AC157372_6.1.A_C233653_4.1	at3e64370	gi|poptr1_1|813999|estext_fgenesh4_kg.c_lg_v10034 gi|poptr1_1|716209|estext_genewise1_v1.c_lg_v13743
PSI_PSAK	Photosystem I psaO / psaK.	medtr2g016960.1	at1g31670 at1g03880	gi|poptr1_1|744651|estext_genewise1_v1.c_1400032
NAD_binding_1	Oxidoreductase NAD-binding domain. Xanthine dehydrogenase, that also bind FAD/NAD, have essentially no similarity.	medtr5g022240.1	at5g66190	gi|poptr1_1|852705|estext_fgenesh4_pg.c_270077 gi|poptr1_1|817271|estext_fgenesh4_pg.c_lg_i10561
AAA	ATPase family associated with various cellular activities (AAA). AAA family proteins often perform chaperone-like functions that assist in the assembly, operation, or disassembly of protein complexes.	medtr5g007700.1	at2g29730	
PsbY	Photosystem II protein Y (PsbY). This family consists of several bacterial and plant photosystem II protein Y (PsbY) sequences. PsbY is a manganese-binding protein that has an L-arginine methicking enzyme activity.	medtr2g102830.1	at1g07740	gi|poptr1_1|726168|estext_genewise1_v1.c_lg_s2024 gi|poptr1_1|826158|estext_fgenesh4_pg.c_lg_v11598
PsbX	Photosystem II reaction centre X protein (PsbX). This family consists of several photosystem II reaction centre X protein (PsbX) sequences from both prokaryotes and eukaryotes.	medtr1g034460.1	at3g05120	gi|poptr1_1|743310|estext_genewise1_v1.c_1450010 gi|poptr1_1|803207|estext_fgenesh4_rm.c_lg_v10043
PSI_PsaE	Photosystem I reaction centre subunit IV / PsaE. PsaE is a 69 amino acid polypeptide from photosystem I present on the stromal side of the thylakoid membrane. The structure is comprised of a well-defined five-stranded beta-sheet similar to SH3 domains	medtr7g117070.1	at4g28750 at5g20360	gi|poptr1_1|711610|estext_genewise1_v1.c_lg_i8393
FBPase	Fructose-1-6-bisphosphatase.	medtr2g093130.1	at3g55800	gi|poptr1_1|826389|estext_fgenesh4_pg.c_lg_v11539 gi|poptr1_1|713685|estext_genewise1_v1.c_lg_m23292

Supplemental Figure 3. Co-expressed gene vicinity networks for *PSA-D* genes in Arabidopsis (left) and Rice (right). Nodes in the network resemble individual genes and the connecting edges represent co-expressed links. The colouration of nodes and edges are explained in Figure 2. Examples of genes common to the two NVNs are indicated in blue colour with the gene acronyms (see also Supplemental Dataset 2). The expression of the *PSA-D*-related genes across different tissues is displayed above the networks. The interactive networks may be found at http://aranet.mpimp-golm.mpg.de/aranet/a11983, and http://aranet.mpimp-golm.mpg.de/ricenet/r1587, respectively.

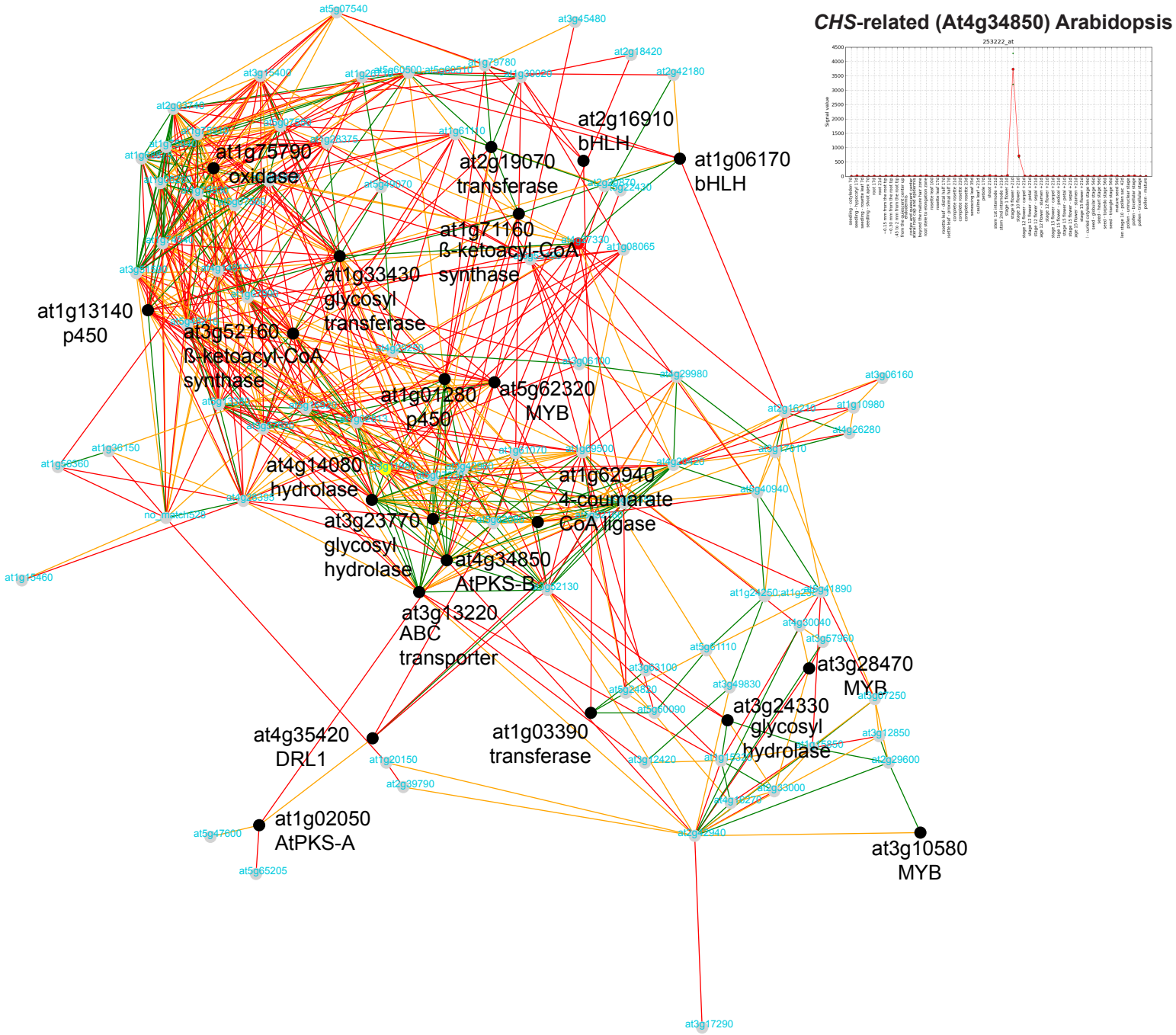


Supplemental Figure 4. Phylogenetic tree of CHS-related proteins in Arabidopsis, poplar, rice, and Medicago. Colour coded genes/probe sets corresponds to; Grey; gene not present on the array, Green; Arabidopsis genes with secondary metabolism related NVNs, Red; Rice genes with secondary metabolism related NVNs, Yellow; Poplar genes with secondary metabolism related NVNs, Purple; Medicago genes with secondary metabolism-related NVNs, and Black; Genes with no significant secondary metabolism related ontology terms for their NVNs. The boxed area is blown up and displays, together with genes/probesets indicated with arrow heads, putative *CHS* related genes associated with floral tissues for the four species. Protein sequences were aligned in MEGA4 (Tamura et al., 2007) using ClustalW and phylogenetic trees were constructed using bootstrapped Neighbour-Joining method (1000 runs). Values on the branches indicate bootstrap support in percent. The amino acid alignment used to generate this tree is presented in Supplemental Dataset 4 online.



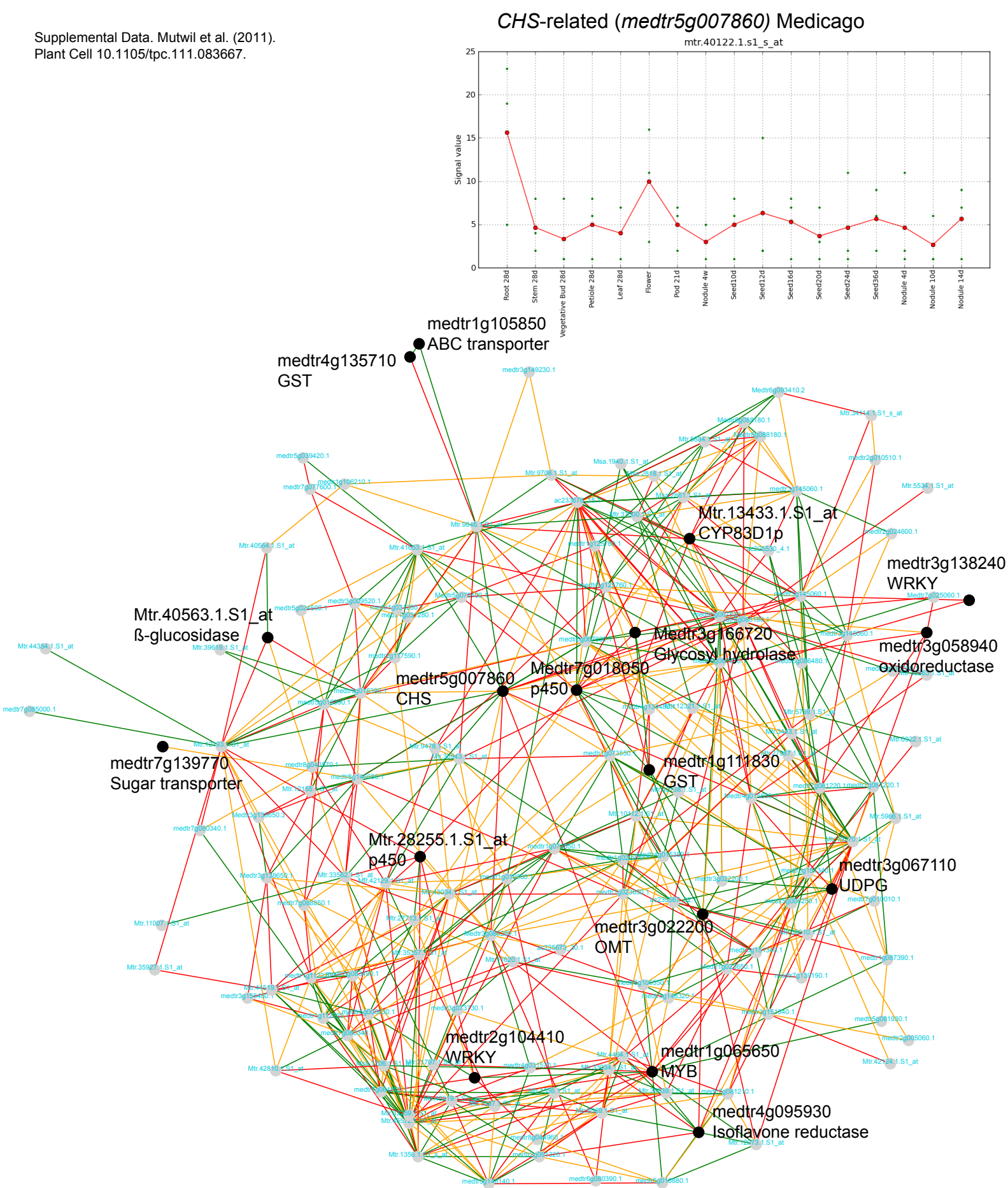


Supplemental Figure 5. Co-expressed gene vicinity network for Chalcone Synthase (*CHS*; *At5g13930*) in Arabidopsis. Many of the genes co-expressed in this NVN (left) participate in the flavonol (orange font), flavonoid (black font) and phenylpropanoid (blue font) pathway (right) leading to anthocyanin production. *The alternative pathway in gray is observed in F3'H (TT7, At5g07990) knockout mutant. The interactive network may be found at http://aranet.mpimp-golm.mpg.de/aranet/a17217.



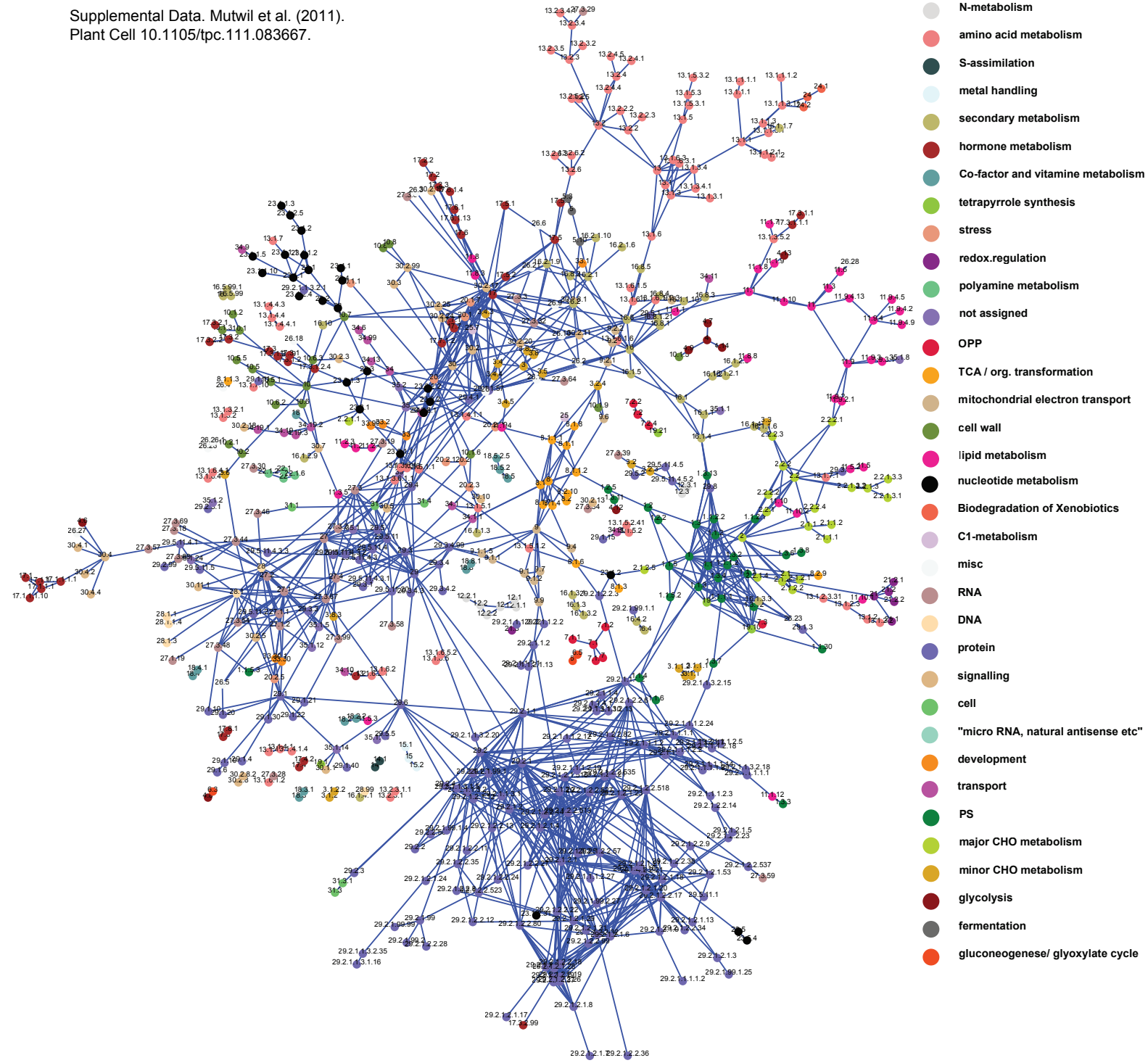
Supplemental Figure 6. NVN for *CHS*-related gene (*At4g34850*) in Arabidopsis.
Nodes in the network resemble individual genes and the connecting edges represent co-expressed links. The coloration of nodes and edges are explained in Figure 2. Expression of the *CHS*-related gene across different tissues is displayed upper left. The interactive network may be found at http://aranet.mpimp-golm.mpg.de/aranet/a14232.

Supplemental Data. Mutwil et al. (2011).
Plant Cell 10.1105/tpc.111.083667.



Supplemental Figure 7. Co-expressed gene vicinity for Chalcone Synthase 4 (*CHS4*) in Medicago. Nodes in the network resemble individual genes and the connecting edges represent co-expressed links. The colouration of nodes and edges are explained in Figure 2. Genes potentially associated with flavonol/flavonoid biosynthesis are indicated with acronyms in bold. The interactive network may be found at http://aranet.mpimp-golm.mpg.de/medinet/m19811.

Supplemental Data. Mutwil et al. (2011).
Plant Cell 10.1105/tpc.111.083667.



Supplemental Figure 8. Co-expressed MapMan ontology network for Medicago. Each node indicates a MapMan ontology term, and an edge between two nodes indicates that the two terms are co-expressed with each other. Key to the MapMan ontology is displayed on the right. The network is available at http://aranet.mpimp-golm.mpg.de/medinet/Mapman_network.

Supplemental Figure 9. Pseudocode for Column Selection.

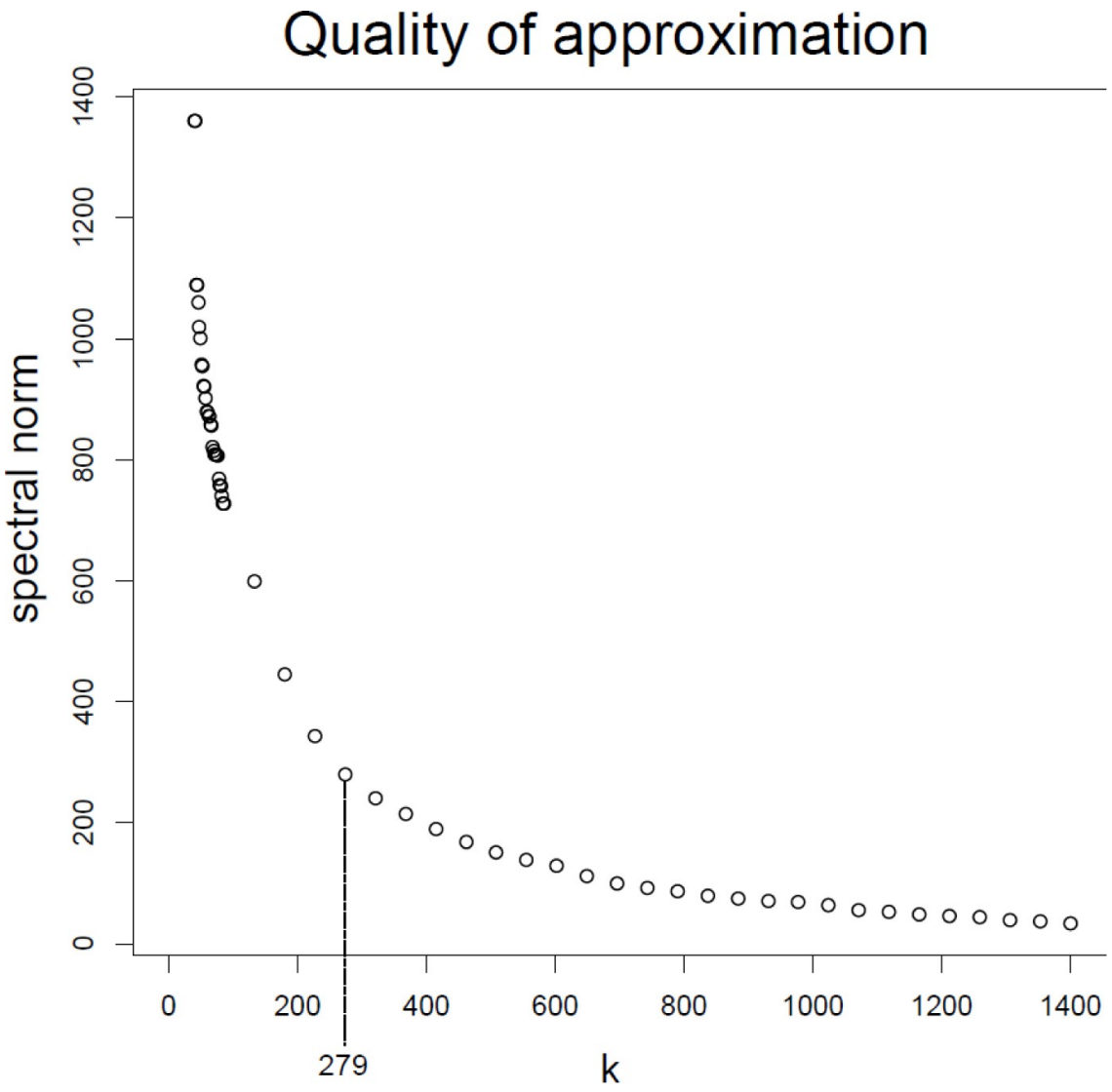
Input: matrix $A_{m \times n}$, with biological experiments as columns, and a range $[t_1, t_2]$ for the number of columns to be considered.

Output: matrix $C_{m \times k\_opt}$, with columns of A

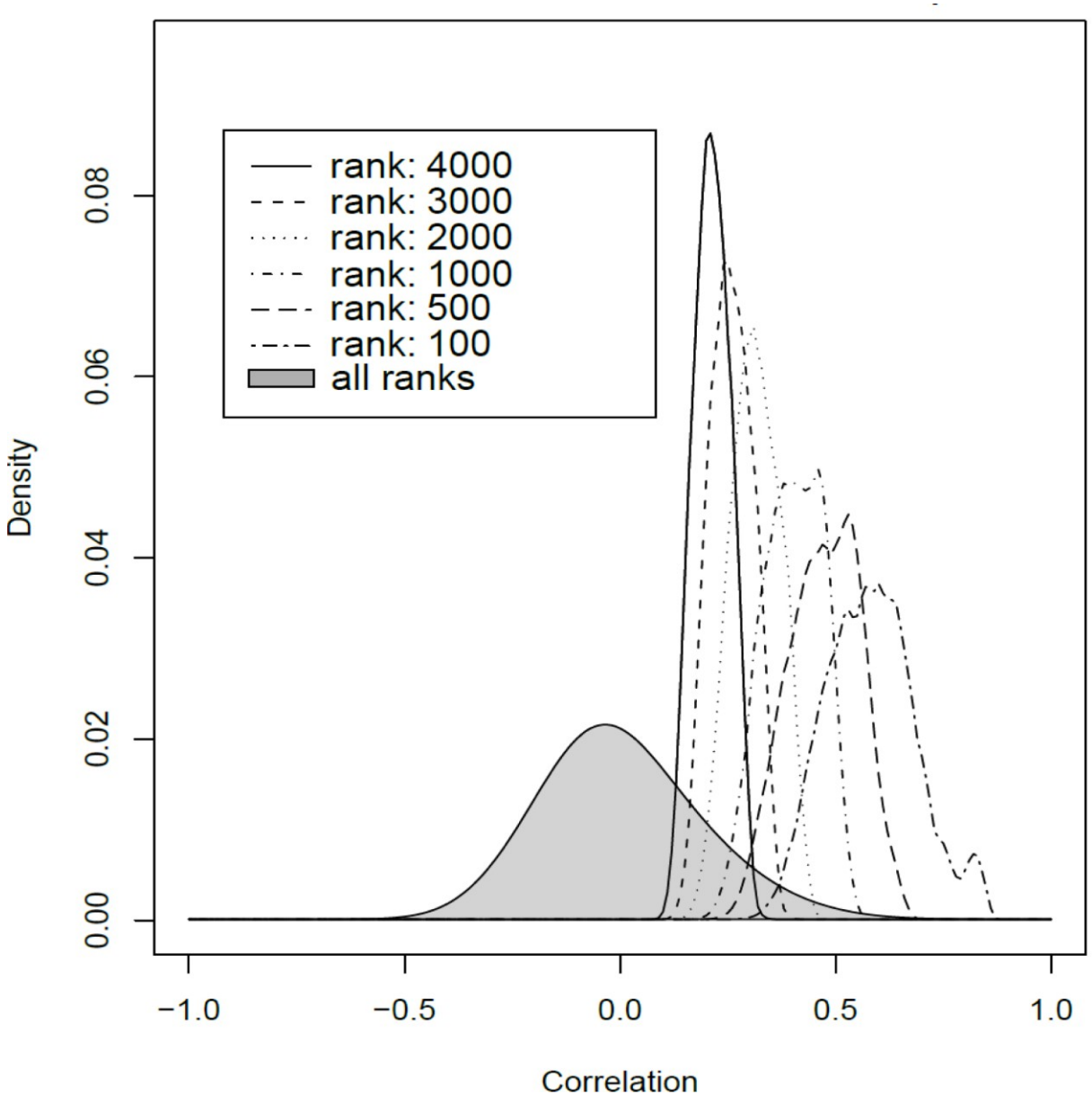
```
1. for k in t1 to t2 do
   //Randomized step
   //randomly select c = k·log(k) columns of A and form a matrix T(m x c)//
2.      for l in 1 to 40 do
3.          for j in 1 to n do
4.              assign a score to the jth column of A: pj = ||Ai||2^2/||A||F^2
5.              select the jth column with probability min(1,c·pj)
6.              if the jth column is chosen
7.                  keep a rescaling factor sqrt(1/min(1, c·pj))
8.          end for
9.          keep the c columns of minimum spectral norm
   //Deterministic step
   //deterministically select k columns from T
   //form the matrix Cm x k using RRQR decomposition with pivoting.
10.     C(k) = RRQR(T)
11. end for
12. apply elbow criterion on the spectral norms of C(k) to determine k_opt
13. return C(kopt)
```

In the first step, for a given matrix A and a parameter k, one performs randomized sampling of size k·log(k) columns with probabilities given by pj = ||Ai||22/||A||F2. After repeating the randomized sampling 40 times, suggested and tested in (Boutsidis et al. (2008)), one selects those c columns which minimize the spectral norm of the submatrix T. In the second step, we rely on the deterministic rank revealing QR (RRQR) decomposition. Given a matrix A, its QR-decomposition is a representation of A as Q*R, where R is an upper triangular matrix (all entries below the principle diagonal are zero) and Q is an orthogonal matrix (i.e., a matrix satisfying Q*Q^T = I with Q^T being the transpose of Q and I is the identity matrix). We applied the RRQR decomposition to find the k columns, from the c preselected in T, which are most mutually independent. Note that in our computations, we employ the RRQR decomposition with pivoting, as implemented in LAPACK (Anderson et al., 1999). The randomized step can be avoided to arrive at the column selection problem of Golub (1965) with lower bounds on the spectral norm, still acceptable for applications,. Due to the large size of the analyzed data compendia, here we rely on the deterministic step in order to avoid calculations of singular value decomposition necessary for establishing the probabilities for selecting columns in the randomized step. Here we provide an extension to this method to allow automatic selection of the number of most mutually independent columns. To this end, we first determine the minimum spectral norm for a number k of columns in the range [t1, t2] from A. thaliana data compendium,

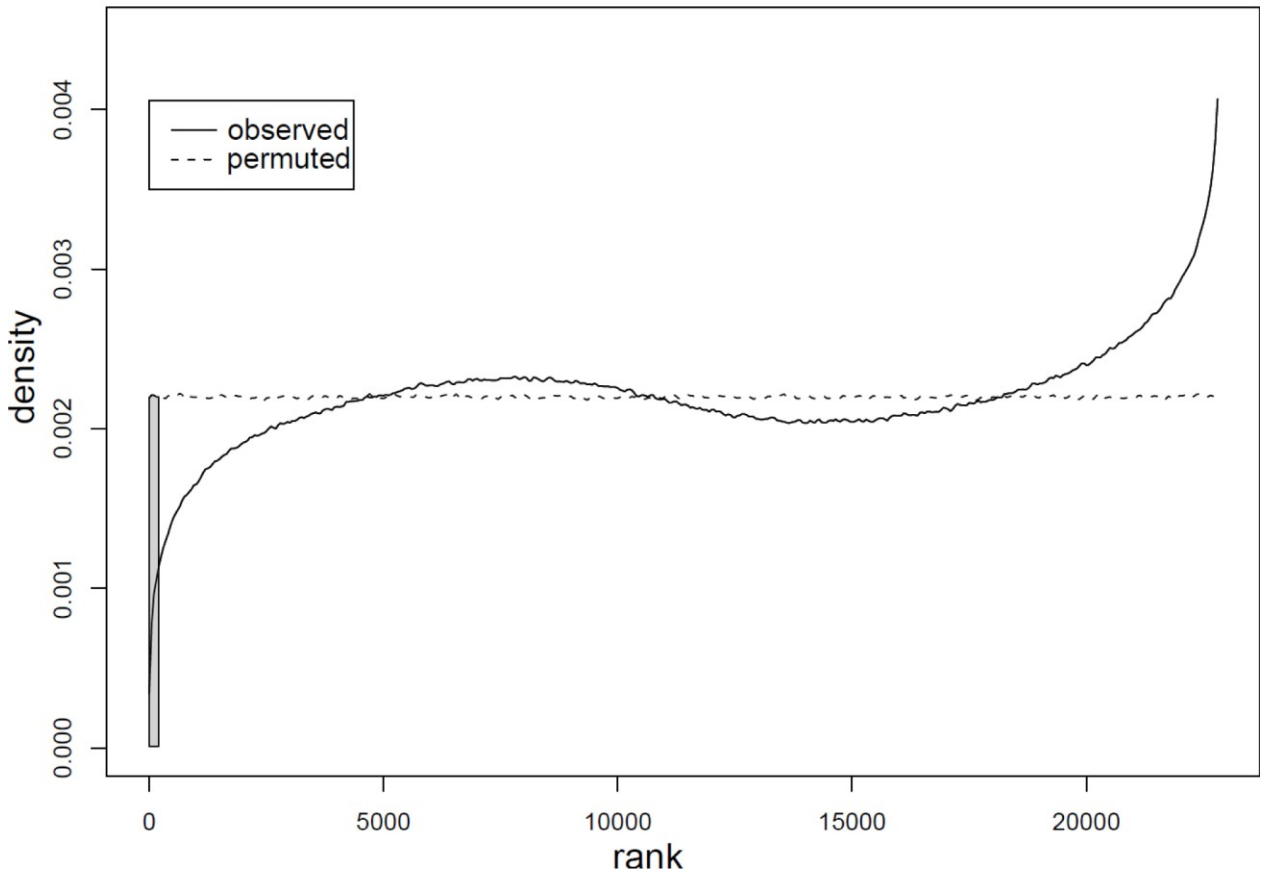
with t1 = 20 and t2 = 1400. The corresponding matrix will be termed optimal with respect to the spectral norm. We define the optimum k as that which appears at the elbow of the curve, determined by the number of columns and the spectral norm, above which only a negligible improvement is observed. By inspecting this curve for A. thaliana data compendium (Supplemental Figure 10), we observe that kopt = 279, and its corresponding spectral norm is 15% of the spectral norm of the optimal 20-column submatrix. A computational speed-up can be achieved by testing from right to left and using previous results for k independent columns. To provide fair comparison with the data compendia from the other species used in the analysis, we rely on the values obtained from applying our method on *A. thaliana* data compendium: We seek that kopt in another species, whose corresponding spectral norm is also 15% of its corresponding optimal 20-column submatrix. This resulted in the following values for kopt = 181, 156, 163, 83, 165, 171, corresponding to barley, rice, Medicago, poplar, wheat and soybean.



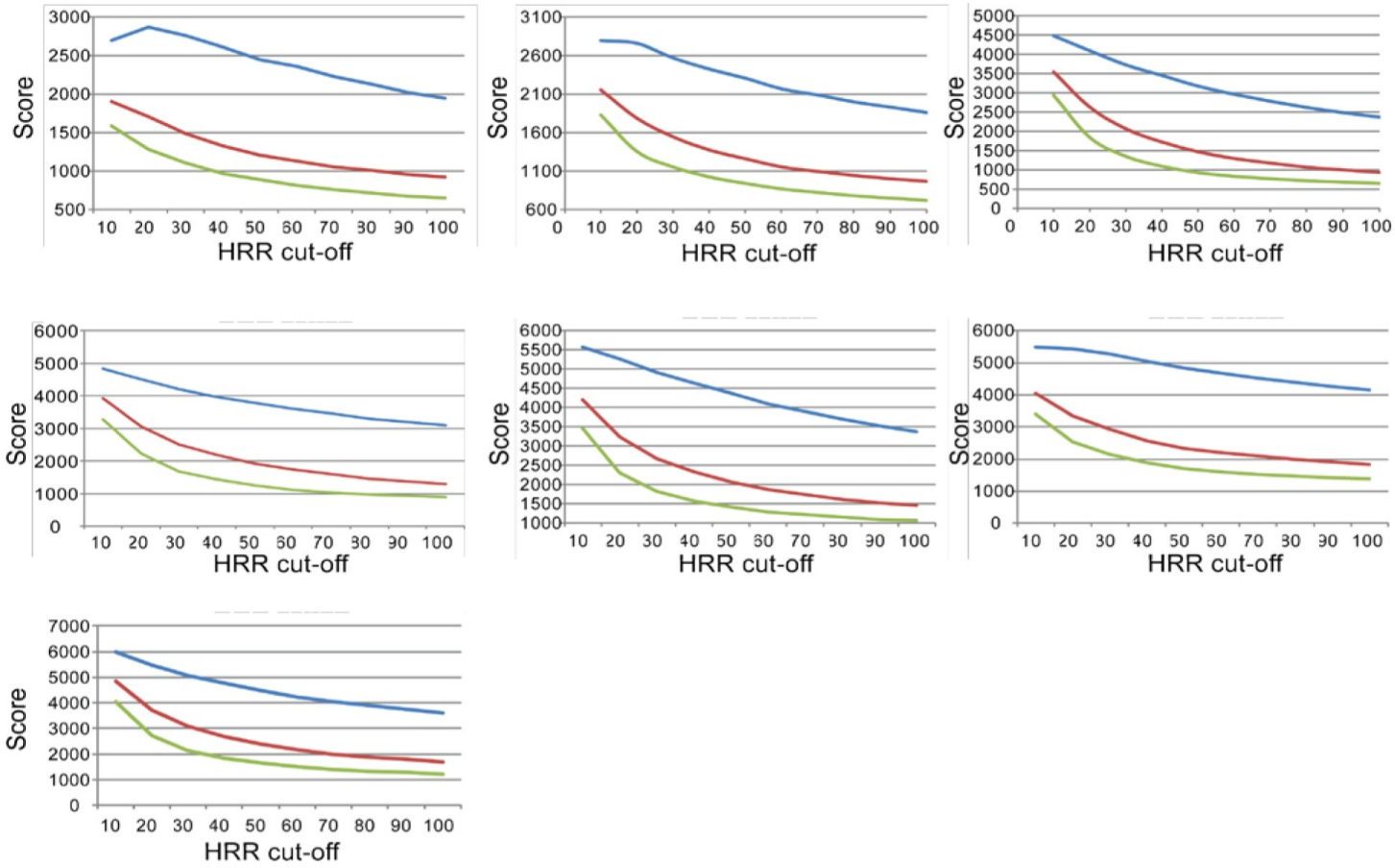
Supplemental Figure 10. Spectral norm of optimal k-column submatrices for Arabidopsis in the range of 20 to 1400.



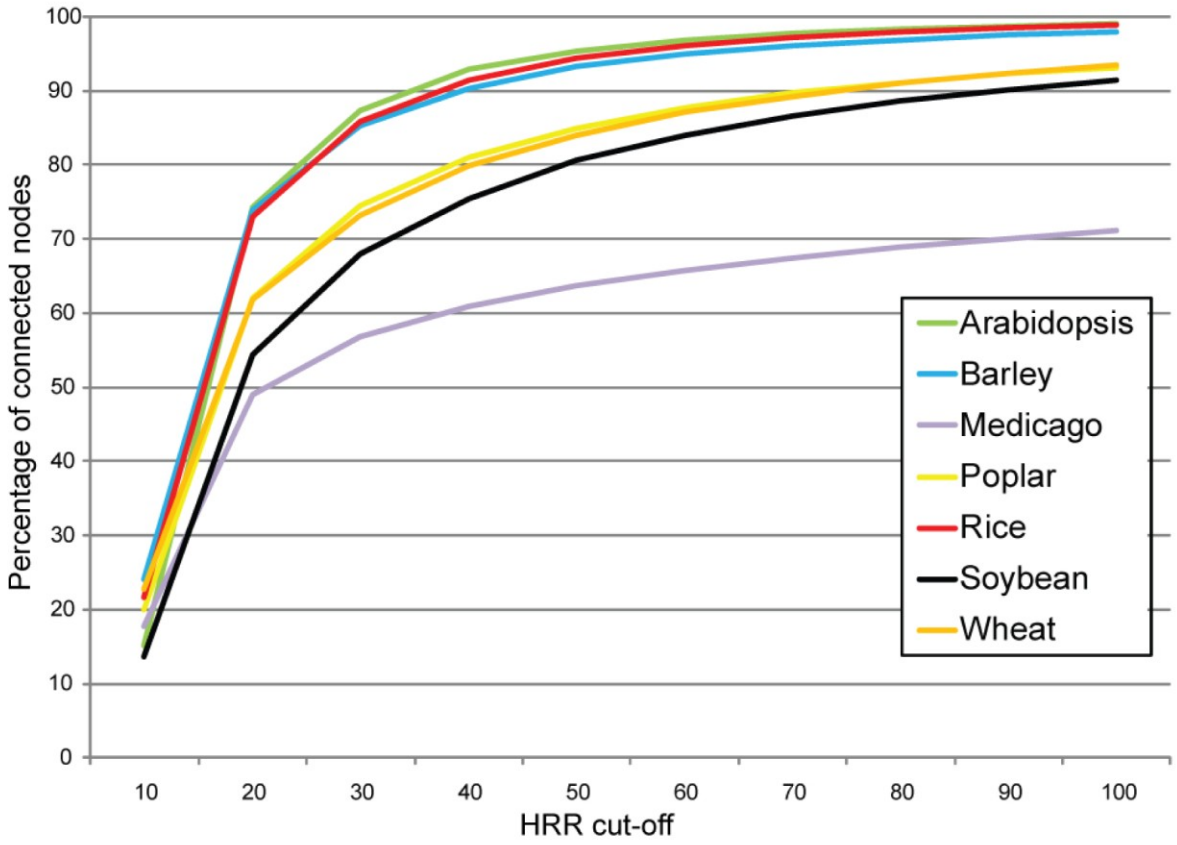
Supplemental Figure 11. Distribution of correlation coefficients observed for specific ranks on the Arabidopsis data collection.



Supplemental Figure 12. Figure 4: Distribution of reciprocal ranks on the Arabidopsis data collection. Permuted ranks are distributed uniformly. The area under the curve shaded in grey determines 1% of the overall area of the permuted ranks and coincides with a rank cut-off of 228.



Supplemental Figure 13. Influence of HRR cut-off and NVN step size on the biological relevance for the seven analyzed species. The x-axis represents HRR value cut-off which was used to construct a co-expression network, while the y-axis represents the score. Blue, red and green lines represent step sizes of 1,2 and 3 used to generate NVNs.



Supplemental Figure 14. Influence of HRR cut-off on the percentage of connected nodes in the networks of the seven analyzed species. The different species are color coded, as shown in the legend.