SUPPLEMENTAL TEXT

Testing and benchmarking Co-expression analysis tools

To identify the most significant co-expression patterns we used three different co-expression software programs, MetaOmGraph, the Botany Array Resource (BAR) expression angler and the Arabidopsis co-expression tool (ACT) based on their large microarray experiment sets and functional ease (Wurtele et al., 2003; Toufighi et al., 2005; Manfield et al., 2006). The Pearson correlation coefficient (PCC) was used to measure co-expression strength in each program, using a threshold of 0.7, which has been considered a rule-of-thumb threshold acceptable for identifying true co-expression pairs, and has been used in many published co-expression analyses e.g. (Cohen et al., 2000; Ren et al., 2005; Conant and Wolfe, 2006; Ren et al., 2007; Tsai et al., 2009). Before we analyzed the co-expression of PG genes, we first tested these three software programs against (i) the ten genes of the plastid-localized linear chlorophyllide biosynthesis pathway downstream of the heme branch and (ii) the nine nuclear genes that encode the subunits of the plastid localized heteromeric ClpPR protease complex and the single gene encoding for the mitochondrial homomeric ClpP2 protease complex.

Because of the unique developmental and environmental controls of the three isoforms of NADPH-protochlorophyllide oxidoreductase (POR) catalyzing the light-dependent, final step of chlorophyllide synthesis, it was assumed that these three genes would not co-express with the rest of the gene set, but could be considered negative controls (Apel, 1981; Armstrong et al., 1995; Holtorf et al., 1995; Runge et al., 1996; Oosawa et al., 2000; Su et al., 2001). The co-expression relationships within this gene set are illustrated in Supp. Fig. 2. Co-expression relationships were found in all three programs for most gene pairs (16/21, 76%), except for the three isoforms of NADPH-protochlorophyllide oxidoreductase (POR). The consistency between the MetaOmGraph and ACT programs was particularly striking (Supp. Fig. 2). Except for the three additional pairs involving PORB, the pattern was identical, and only a relationship between DVR and CHL27 was missing from the expected co-expression set (it was also missing from the BAR analysis). Surprisingly, PORC was found in both programs to demonstrate tight co-expression with each of the other members of the pathway. The PORC isoform is the dominantly expressed isoform in leaf tissue and acts as the primary form of the chl pathway, while PORA and PORB appear to have specialized roles in the greening process. Thus, PORC might be found to co-express with the other chl synthesis genes if a greater reliance on mature leaf tissue microarray experiments is made. As was expected, PORA showed no co-expression relationships with the other members of the pathway and PORB only showed three connections, all found in MetaOmGraph.

To test the programs against a gene set with physical linkages, we analyzed the nuclear-encoded genes of the ClpPR protease complex. The ClpPR complex functions in plastid protein homeostasis and

the protein subunits have always been found as one associated unit, never as free subunits (Peltier et al., 2004; Olinares et al., 2011). ClpP2 is mitochondrial-localized and has not been found to be part of the plastid ClpPR complex, as have the other nuclear-encoded Clp proteins (P3-P6, and R1-R4) and thus we consider it to serve as our negative control. It was found that precisely half of the expected co-expression pairs were found in all three programs (Supp. Fig. 2). Several of the expected co-expression pairs were not found in any of the three programs, indicating that certain members of the complex might demonstrate post-transcriptional control of protein accumulation, e.g. ClpR1. If we considered only the pairs where at least one program identified a co-expression relationship, among those 23 pairs, 61% (14) were found in all three programs. Importantly, none of the programs found a co-expression relationship between genes of the ClpPR complex and the true-negative, ClpP2.

We thus demonstrate the successful identification of (presumed) true-positives of a functionally linked and a physically linked gene set, to the exclusion of true-negatives using three independent software programs. While we would not expect each software program to produce identical co-expression results, we demonstrate that the programs consistently identify the majority of the (presumed) true-positive gene pairs within each gene set.

Leister, et al. previously analyzed the co-expression patterns of nuclear-encoded plastid proteins and found that they distributed into 23 unique co-expression groups, or regulons, suggesting that there are multiple layers of transcriptional control over nuclear-encoded plastid gene expression (Biehl et al., 2005).This indicates that meaningful co-expression patterns can be discerned despite the influence of diurnal fluctuations on plastid gene expression as a whole. We sought to validate that meaningful patterns could be found for PG genes by determining whether the PG genes preferentially co-express with other PG genes or with genes of the plastid as a whole. We tested this by classifying all genes of the genome by their sub-cellular locations using the following 4 groups: I) plastoglobular (25 proteins), II) plastid-localized by experimental data curated in PPDB (http://ppdb.tc.cornell.edu) (1123 proteins), III) plastid predicted by TargetP (Emanuelsson et al., 2000) (3299 proteins), or IV) none of the above classifications (18235 proteins). Pearson correlation coefficients (PCCs) were calculated for all pair-wise combinations between each PG gene and each gene of the microarray. In this way, each of the 21,158 genes could be sorted by their “co-expression tightness” with any given PG gene. If one were to incrementally increase the PCC threshold, fewer and fewer co-expressing genes would exist above the threshold. It was expected that connections to other genes of the PG would preferentially be maintained at higher thresholds. Thus, the fraction of genes from each group above various thresholds was calculated (Supp. Fig. 3). The genes of the PG were indeed consistently maintained to higher PCC values than genes not curated or predicted as plastidic (i.e. Group IV). Also, certain PG genes were found to associate with Group II just as tightly as

Group I, including aldo/keto reductase and FBN 2. This suggests these PG genes are more centrally involved in plastid activities.

When genes involved in other plastidic compartments were tested this pattern did not persist. As an example, the plastidic genes, chlorophyll synthase (chl synthase, AT3G51820), and keto-acyl ACP synthase I (KAS I, AT5G46290), were tested against the *A. thaliana* genome as above (Supp. Fig. 3). As expected, chl synthase demonstrated the tightest association with genes of the curated plastid group and not that of the PG. KAS I demonstrated equal co-expression strength with all four groups. This is likely a reflection of the relevance of the gene to the functioning of the cell as a whole; KAS I is a subunit of the plastid-localized fatty acid synthase complex, responsible for the *de novo* fatty acid metabolism of the cell.

CITED REFERENCES IN SUPPLEMENTAL TEXT

Apel K (1981) The Protochlorophyllide Holochrome of Barley (Hordeum Vulgare L.): Phytochrome-induced decrease of translatable mRNA coding for the NADPH-protochlorophyllide oxidoreductase. European Journal of Biochemistry **120:** 89-93.

Armstrong GA, Runge S, Frick G, Sperling U, Apel K (1995) Identification of NADPH:protochlorophyllide oxidoreductases A and B: a branched pathway for light-dependent chlorophyll biosynthesis in Arabidopsis thaliana. Plant Physiol **108:** 1505-1517.

Biehl A, Richly E, Noutsos C, Salamini F, Leister D (2005) Analysis of 101 nuclear transcriptomes reveals 23 distinct regulons and their relationship to metabolism, chromosomal gene distribution and co-ordination of nuclear and plastid gene expression. Gene **344:** 33-41.

Cohen BA, Mitra RD, Hughes JD, Church GM (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. **26:** 183-186.

Conant GC, Wolfe KH (2006) Functional Partitioning of Yeast Co-Expression Networks after Genome Duplication. PLoS Biol **4:** e109.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol **300:** 1005-1016.

Holtorf H, Reinbothe S, Reinbothe C, Bereza B, Apel K (1995) Two routes of chlorophyllide synthesis that are differentially regulated by light in barley (Hordeum vulgare L.). PNAS **92:** 3254-3258.

Manfield IW, Jen C-H, Pinney JW, Michalopoulos I, Bradford JR, Gilmartin PM, Westhead DR (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. Nucleic Acids Research **34:** W504-W509.

Olinares PD, Kim J, Davis JI, van Wijk KJ (2011) Subunit Stoichiometry, Evolution, and Functional Implications of an Asymmetric Plant Plastid ClpP/R Protease Complex in Arabidopsis. Plant Cell **23:** 2348-2361.

Oosawa N, Masuda T, Awai K, Fusada N, Shimada H, Ohta H, Takamiya K-i (2000) Identification and light-induced expression of a novel gene of NADPH-protochlorophyllide oxidoreductase isoform in Arabidopsis thaliana. FEBS Letters **474:** 133-136.

Peltier JB, Ripoll DR, Friso G, Rudella A, Cai Y, Ytterberg J, Giacomelli L, Pillardy J, Van Wijk KJ (2004) Clp Protease Complexes from Photosynthetic and Non-photosynthetic Plastids and Mitochondria of Plants, Their Predicted Three-dimensional Structures, and Functional Implications. J Biol Chem **279:** 4768-4781.

Ren X-Y, Fiers MWEJ, Stiekema WJ, Nap J-P (2005) Local Coexpression Domains of Two to Four Genes in the Genome of Arabidopsis. Plant Physiology **138:** 923-934.

Ren X-Y, Stiekema W, Nap J-P (2007) Local coexpression domains in the genome of rice show no microsynteny with Arabidopsis domains. Plant Molecular Biology **65:** 205-217.

Runge S, Sperling U, Frick G, Apel K, Armstrong GA (1996) Distinct roles for light-dependent NADPH:protochlorophyllide oxidoreductases (POR) A and B during greening in higher plants. The Plant Journal **9:** 513-523.

Su Q, Frick G, Armstrong G, Apel K (2001) POR C of *Arabidopsis thaliana*: a third light- and NADPH-dependent protochlorophyllide oxidoreductase that is differentially regulated by light. *In* Plant Molecular Biology, Vol 47. Springer Netherlands, pp 805-813.

Sun Q, Zybailov B, Majeran W, Friso G, Olinares PD, van Wijk KJ (2009) PPDB, the Plant Proteomics Database at Cornell. Nucleic Acids Res **37:** D969-974.

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ (2005) The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. The Plant Journal **43:** 153-163.

Tsai H-K, Huang P-Y, Kao C-Y, Wang D (2009) Co-Expression of Neighboring Genes in the Zebrafish (Danio rerio) Genome. International Journal of Molecular Sciences **10:** 3658-3670.

Wurtele ES, Li J, Diao L, Zhang H, Foster CM, Fatland B, Dickerson J, Brown A, Cox Z, Cook D, Lee E-K, Hofmann H (2003) MetNet: software to build and model the biogenetic lattice of Arabidopsis. Comparative and Functional Genomics **4:** 239-245.