From pathway to regulon in Arabidopsis

by

Wiesława Izabela Mentzen

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Eve Syrkin Wurtele, Co-Major Professor
Xun Gu, Co-Major Professor
David Fernández-Baca
Basil Nikolau
Jonathan F. Wendel

Iowa State University
Ames, Iowa
2006

Graduate College

Iowa State University

This is to certify that the doctoral dissertation of

Wiesława Izabela Mentzen

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Co-Major Professor

Signature was redacted for privacy.

Co-Major Professor

Signature was redacted for privacy.

For the Major Program

TABLE OF CONTENTS

CHAPTER 1. GENERAL INTRODUCTION

The research presented in this dissertation centers around the organization of global gene networks with a focus on the fatty acid biosynthesis pathway in the flowering plant *Arabidopsis thaliana*. Bioinformatic and statistical tools are applied to analyze the data of different kinds (sequence and expression) according to various scopes and angles. My phylogenetic analysis focuses on acetyl-CoA carboxylase, an enzyme catalyzing the first committed reaction in fatty acid biosynthesis pathway, and describes its evolution in the light of the intertwined phylogenies of the multiple members of biotin-dependent enzymes. This work is presented in Chapter 2, "Molecular evolution of acetyl-CoA carboxylase". In Chapter 3, wide array of Arabidopsis transcriptomic data is used to address the question of whether exploitation of this kind of data can reveal the fatty acid biosynthesis pathway, as well as two other pathways (leucine catabolism and starch metabolism). This study led to interesting conclusions pertaining to organization of metabolic processes in plants, and spurred further exploration of this topic. Chapter 4 extends the analysis from Chapter 3 to include the entire Arabidopsis genome and identifies cellular processes that engage the most pronounced groups of coexpressed genes.

Fatty acids biosynthesis

Fatty acids are essential components of all known bacterial and eukaryotic cells. They play critical role in cells as the metabolic precursors for biological membranes and energy reserves. They form the lipid bilayer membrane separating the inside of cell from the outside environment and define the boundaries of intracellular organelles. Fatty acids in a form of glycerolipids are the storage compounds in Eukaryotes. Specialized fatty acids are used for signaling, respiratory, immune system and many other purposes. Those particles must be synthesized de novo in every cell.

Fatty acids are synthesized via a pathway that seems to be conserved throughout the whole life system. The process is taking place in a cytoplasm of bacteria, fungi and animals, in chloroplasts of plants and algae and in apicoplast (a relict plastid) in apicomplexan malaria

parasite *Plasmodium falciparum* (Waller et al, 2003). It is also possible that plants and animals mitochondria contain bacterial type dissociated fatty acid synthase, probably for mitochondrial respiratory purposes (Schneider et al, 1997; Zhang et al, 2003).

The summary reaction of palmitic acid synthesis is:

acetyl-CoA + 7 malonyl-CoA + 14 NADPH + 14 H^+ → palmitic acid + 7 CO_2 + 8 CoA + 14 NADP+ 6 H_2O

Several enzymatic activities, separate in bacteria, and associated in animals constitute the fatty acid synthase. The process starts with acetyl-CoA molecule, which is elongated to malonyl-CoA by acetyl-CoA carboxylase. This compound, shuttled between enzymes by an acyl carrier protein is subsequently iteratively condensed with another malonyl-CoA in a cycle of condensation, reduction, dehydration and reduction, and the resulting growing acyl residue accepts 2 carbon units in the next cycle of elongation. The terminal product, usually C:16 or C:18, is released from ACP by transacylase. The acyl chains can then undergo other modifications: elongation, desaturation or isomeration depending on their cellular fate.

Fatty acid biosynthesis in plants

Plastids are the localization for de novo fatty acid biosynthesis in plants. Enzymes for fatty acid biosynthesis (beta-ketoacyl-[acyl carrier protein] synthase, beta-ketoacyl-[acyl carrier protein] reductase, 3-hydroxyacyl dehydratase, enoyl-[acyl carrier protein] reductase, and acyl carrier protein are present in plastids, often in multiple isoforms. Acetyl-CoA synthase or plastidic pyruvate dehydrogenase are named as possible sources of acetyl-CoA necessary for the first step. Recent studies indicate predominant role of plastidic PyD (Ke et al, 2000, Behal et al., 2002). Fatty acids are elongated in plant cytosol. Some plant mitochondria were shown to synthesize fatty acids: C:8, C:16 and C:18, lipoic acid precursor (Gueguen et al, 2000). Similar sets of enzymes are probably employed in these compartments, however no sequences for cytosolic acyl carrier protein, cytosolic or mitochondrial 3-hydroxyacyl

dehydratase or mitochondrial enoyl-[acyl carrier protein] reductase have been found in Arabidopsis.

Acetyl-CoA carboxylase (ACC)

ACC catalyzes first committed step in fatty acid synthesis: elongation of acetyl-CoA to malonyl-CoA using carbonate anion in reaction:

$$\textbf{ATP + acetyl-CoA + } HCO_3^- \rightarrow \textbf{ADP + orthophosphate + malonyl-CoA}$$

ACC is a member of a family of biotin-dependent carboxylases, along with pyruvate carboxylase, urea amidolyase, propionyl-CoA carboxylase, methyl-crotonyl-CoA carboxylase and sodium – ion pumping oxaloacetate decarboxylase and methylmalonyl decarboxylase. Covalently bound biotin serves as a carboxyl group donor/acceptor (Lane et al., 1974). ACC is composed of four domains: biotin carboxylase, biotin carboxyl carrier, carboxyl transferase alpha and carboxyl transferase beta. Those domains are organized in different ways, from four separate subunits in Bacteria to one heteromeric protein in fungi and animals. Plants can have both heteromeric form of ACC in plastids and homomeric form in cytoplasm and plastids (Konishi and Sasaki, 1994; Alban et al., 1994).

Transcriptional networks

The advances of comparative genomics brought about the results that suggest that the elaborate and diverse functions found in multicellular organisms are achieved through expansion of regulatory systems, in particular regulation of gene expression, rather than by increasing the gene content (Levine and Tjian, 2003). An increasingly active area of system biology focuses on understanding the large scale organization of regulatory networks by extracting and putting together pieces of knowledge contained in single-kind of data. The data from massive experiments is of particular use since it aims at capturing the information describing the whole transcriptome or proteome in the given state. While the proteomics data

for Arabidopsis is still scarce, the importance of meta-analysis of gene expression data has been recognized and led to creation of several web-based analysis services and repositories for microarray data (NASCArrays, GENEVESTIGATOR, PLEXdb etc). It has been found that important cellular processes are organized in modules (also called regulons) that often include many coexpressed genes (Eisen et al., 1998, Hartwell et al., 1999). Analysis of transcriptome allows for inference of these modules. The coexpressed sets of genes corresponding to cellular processes were identified in yeast *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans* and human (Stuart et al., 2003; Tamada et al., 2003; Segal et al., 2003; Magwene and Kim, 2004; Lee et al., 2004). These modules might be important enough to be evolutionarily conserved in eukaryotes (Stuart et al., 2003). Results of several studies indicate that such modules are highly interconnected and hierarchically structured (Babu et al., 2004).

Dissertation organization

This dissertation consists of five chapters. Chapter 1 is the general introduction presenting background on the fatty acid biosynthesis pathway with special attention paid to acetyl-CoA carboxylase and transcriptional networks. Chapters 2 through 4 are the manuscripts prepared for submission to *Comparative Functional Genomics*, *Plant Physiology* and *Plant Cell* journals, respectively. These chapters are formatted according to guidelines established by the publishing journal. All the primary work in these three manuscripts was carried out by myself. Nick Ransom, a computer scientist in our lab, implemented MetaOmGraph, a software for analysis of large "-omics" datasets, that was used in Chapter 3. Chapter 5 contains general conclusions summarizing results from Chapters 2-4.

Literature cited

Alban C, Baldet P, Douce R. (1994). Localization and characterization of two structurally different forms of acetyl-CoA carboxylase in young pea leaves, of which one is sensitive to aryloxyphenoxypropionate herbicides. Biochem J. **300** (Pt 2), 557-65.

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. (2004). Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol. **14**, 283-91.

Behal RH, Lin M, Back S, Oliver DJ. (2002) Role of acetyl-coenzyme A synthetase in leaves of Arabidopsis thaliana. Arch Biochem Biophys. **402**, 259-67.

Eisen MB, Spellman PT, Brown PO, Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. **95**, 14863-8.

Gueguen V, Macherel D, Jaquinod M, Douce R, Bourguignon J. (2000). Fatty acid and lipoic acid biosynthesis in higher plant mitochondria. J Biol Chem. **275**, 5016-25.

Hartwell LH, Hopfield JJ, Leibler S, Murray AW. (1999). From molecular to modular cell biology. Nature **402**, c47-c52.

Ke J et al. (2000). The role of pyruvate dehydrogenase and acetyl-coenzyme A synthetase in fatty acid synthesis in developing Arabidopsis seeds. Plant Physiol. **123**, 497-508.

Konishi T, Sasaki Y. (1994). Compartmentalization of two forms of acetyl-CoA carboxylase in plants and the origin of their tolerance toward herbicides. Proc Natl Acad Sci U S A. **91**, 3598-601.

Lane, M.D., Moss, J.D., Polakis, S.E (1974) Curr. Top. Cell Regul. Acetyl coenzyme A carboxylase **8**, 139-195.

Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res. **14**, 1085-94.

Lei Zhang, Anil K. Joshi, and Stuart Smith. (2003). Cloning, expression, characterization and interaction of two components of a human mitochondrial fatty acid synthase: malonyltransferase and acyl carrier protein. J Biol Chem. **278**, 40067-74

Levine M, Tjian R. (2003).Transcription regulation and animal diversity. Nature **424**, 147-51.

Magwene PM, Kim J. (2004). Estimating genomic coexpression networks using first-order conditional independence. Genome Biol.**5**, R100.

Schneider R, Brors B, Massow M, Weiss H. (1997). Mitochondrial fatty acid synthesis: a relic of endosymbiontic origin and a specialized means for respiration. FEBS Lett. **407**, 249-52

Segal E et al.. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. **34**, 166-76.

Stuart JM, Segal E, Koller D, Kim SK. (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science **302**, 249-55.

Tamada Y, Kim S, Bannai H, Imoto S, Tashiro K, Kuhara S, Miyano S. (2003). Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. Bioinformatics **19,** Suppl 2:II227-II236.

Waller RF et al. (2003). A Type II Fatty Acid Biosynthesis Presents Drug Targets in *Plasmodium falciparum* Antimicrob Agents Chemother. **47**, 297-301.

CHAPTER 2. MOLECULAR EVOLUTION OF ACETYL-COA CARBOXYLASE

Wiesława I. Mentzen, Basil J. Nikolau and Eve Syrkin Wurtele

ABSTRACT

Acetyl-CoA carboxylase is a member of the biotin-dependent enzymes. This family catalyzes the transfer of a carboxyl group to (or from) varied substrates, using covalently-bound biotin as a mediator. Enzymes from this family are widely distributed among prokaryotes and eukaryotes and serve in diverse metabolic pathways. The four essential domains of this enzyme family are found as separate subunits, partially-fused, or completely fused. In order to reconstruct the evolutionary history of this family, we have collected the sequences including a wide variety of forms and organisms, with an emphasis on the fully sequenced prokaryotic and eukaryotic genomes. The resulting trees show a complex evolutionary history. The branching order of the trees is affected by the evolutionary relationship among the organisms, enzyme specificity, and the fusion pattern of the particular enzyme. Acetyl-CoA carboxylases (ACCs) from archaea, bacteria and eukaryotes, which form three branches, polyphyletic to each other, have probably evolved independently, via considerably different structural solutions, to converge on the same function. There are many indications of the recent evolution of biotin-dependent carboxylases. The genome of the hyperthermophilic archaea *Pyrococcus* carries traces of a recent duplication and fusion leading to the formation of the new subunit: oxaloacetate decarboxylase. In plants in which both heteromeric and homomeric forms of ACC are simultaneously present, the homomeric form appears to be replacing the heteromeric one in several species. The evolution of carboxylases illustrates independent paths leading to similar solutions, as well as a great diversity of structures in enzymes sharing the same EC number.

INTRODUCTION

The transfer of single carboxyl groups via biotin provides a broad metabolic function in virtually all living organisms. All biotin-containing carboxylases have a common catalytic mechanism in which a carboxyl group is transferred to or from an acyl group and covalently bound biotin serves as the carboxyl group donor/acceptor (Lane et al, 1974). This mechanism for carboxyl transfer is exemplified by acetyl-CoA carboxylase (ACC). This enzyme catalyzes the biotin-dependent elongation of acetyl-CoA to malonyl-CoA using carbonate anion in the reaction (Fig. 1):

$$ATP + \text{acetyl-CoA} + HCO_3^- \rightarrow ADP + \text{orthophosphate} + \text{malonyl-CoA}$$

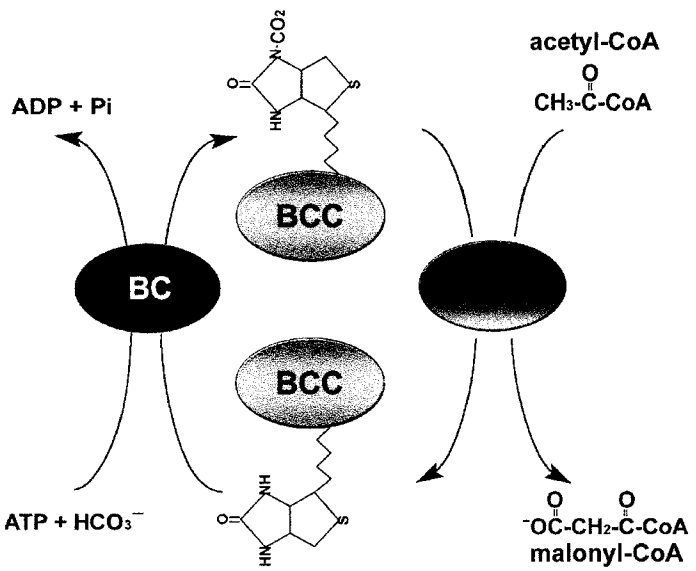


Fig. 1. Reaction catalyzed by acetyl-CoA carboxylase. Biotin prosthetic group of BCC domain serves as carboxyl group carrier and is swung between active centers of BC and CT domains. BC catalyzes carboxylation of biotin with carbonate as a substrate. CT transfers carboxyl group from biotin to acetyl-CoA acceptor. **BC**: biotin carboxylase; **BCC**: biotin carboxyl carrier; **CT**: carboxyltransferase.

Although ACCs catalyze the identical reaction in various species, the metabolic functions of these enzymes are diverse. ACCase has been reported in the Sulfolobaceae *Sulfolobus metallicus, Metallosphaera sedula* and *Acidianus brierleyi* (Burton et al., 1999;

Hugler et al., 2003, Chuakrut et al. 2003) where it appears to function in the hydroxypropionate cycle, an autotrophic CO_2 fixation pathway (Menendez et al, 1999). Malonyl-CoA formation in bacteria and eukaryotes is the first committed step to fatty acid synthesis for membrane lipids, energy storage as triacylglycerides, and signaling. Although at least one archaea (*Pyrococcus furiosus)* contains low levels of fatty acids (Carballeira et al., 1997), archaea do not appear to require malonyl-CoA for fatty acid synthesis for membranes, as they use isoprenoids, rather than fatty-acid based lipids, as primary membrane constituents (Sprott, 1992). In plants, malonyl-CoA serves as a precursor for multiple classes of secondary metabolites such as flavonoids, waxes, and malonated derivatives (Roesler et al., 1994; Nikolau et al., 2003). Mammalian ACCase in mitochondria may function to control the oxidation rate (Ramsay et al., 2001).

ACC is a member of a family of biotin-dependent enzymes including pyruvate carboxylase (PyC, EC 6.4.1.1,), urea carboxylase (UC, EC 6.3.4.6), propionyl-CoA carboxylase (PCC, EC 6.4.1.3), methyl-crotonyl-CoA carboxylase (MCC, EC 6.4.1.4), geranoyl-CoA carboxylase (GCC, EC 6.4.1.5), transcarboxylase (TC, EC 2.1.3.1), and three sodium-ion pumps, oxaloacetate decarboxylase (OxD, EC 4.1.1.3), methylmalonyl-CoA decarboxylase (MMD, EC 4.1.1.41) and glutaconyl-CoA decarboxylase (GCD, EC 4.1.1.70). These enzymes either add, transfer, or remove a carboxyl moiety.

Domains of biotin-dependent enzymes

Carboxylases are composed of three domains: biotin carboxylase (BC), biotin carboxyl carrier (BCC), and carboxyl transferase (CT). Transcarboxylases, which catalyze transfer of carboxyl group between two substrates, contain a BCC and two CT domains with different specificities (Wood, 1979). Decarboxylases contain a substrate-specific CT, biotin decarboxylase, BCC, and typically a membrane anchor and stability-related domains. (Buckel, 2001).

The functional domains of biotin enzymes are organized in different ways (Fig. 2). A heteromeric form of ACC, in which functional domains are located on 4 separate polypeptides (the CT domain is spread across two polypeptides, CT alpha and beta) is found

in bacteria, most plants, algae, and a slime mold *Dictyostelium.* Archaeal ACCases, thus far identifiable solely in autotrophic crenarchaeota lineage *Sulfolobaceae,* differ from their bacterial counterparts in possessing CT alpha and beta activities on one subunit (Fig.2). Their structure more closely resembles corynebacterial ACC/PCC, present also in *P. aeruginosa.* These dual-functioning enzymes are composed of only two subunits, one containing both BC and BCC and the other containing CT. Animals, yeast, and plants contain a homomeric ACCase, in which all the domains are present on a single polypeptide.

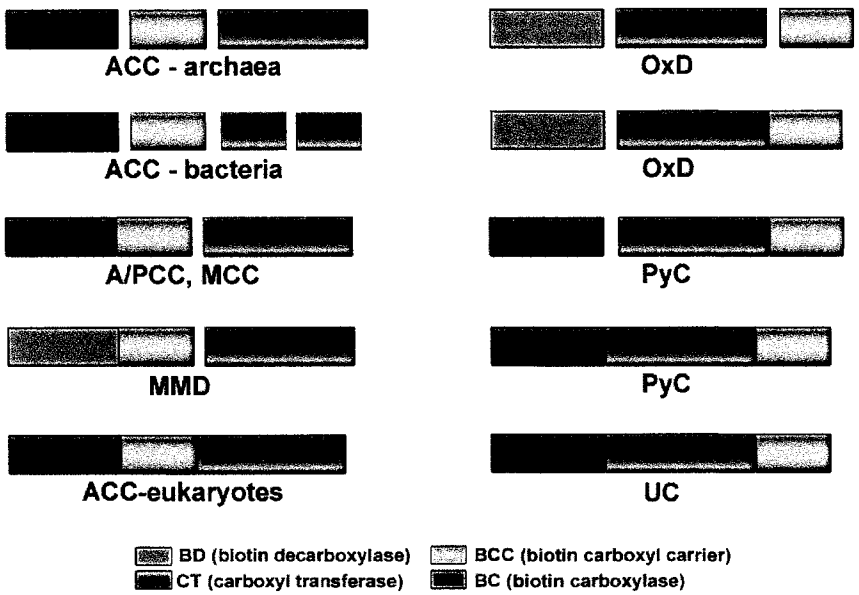


Fig. 2. Organization of functional domains in biotin-dependent enzymes. Individual polypeptides are depicted as rectangles.
ACC, acetyl-CoA carboxylase; **A/PCC**, acetyl/propionyl-CoA carboxylase; **MCC**, methyl-crotonyl-CoA carboxylase; **MMD**, methylmalonyl-CoA decarboxylase; **OxD**, oxaloacetate decarboxylase; **PyC**, pyruvate decarboxylase; **UC**, urea carboxylase.

Phylogenetic relationships

Obermayer and Lynen (1976) hypothesized that biotin-containing carboxylases might have evolved via shuffling, recombination and rearrangement of the primordial genes coding for the functional units. Toh et al. (1993), studying the domain composition of the enzymes whose sequences were available at that time, noted a probable ancient origin of carboxylases, based on the variety of structures, and the organization of functional units and pathways

involved. The similarity in sequence and function of BC with carbamoyl-phosphate synthase, and BCC with lipoyl domain, led them to conclude a common evolutionary origin of these proteins. These conclusions have been reinforced by subsequent biochemical and sequence evidence (Jitrapakdee and Wallace, 2003). Jordan et al., (2003) studied the evolution of carboxylases family, focusing on possible fusions and fissions of the pyruvate-type CT and BCC domains.

ACCases do not form a monophyletic group of enzymes and thus should be analyzed together with other members of carboxylases family. In this paper, we focus on the evolutionary history of ACCases against a background of evolution of other biotin-containing carboxylases. We construct a hypothesis of the history of this family of enzymes based on the phylogenies of constituent domains.

RESULTS AND DISCUSSION

Phylogenetic analysis.

To construct phylogenies for the BC, BCC and CT domains of biotin-dependent carboxylases, sequences spanning various substrate specificities and domain arrangements were collected from diverse organisms. We were especially interested in relationships among forms of acetyl-CoA carboxylases on the kingdom and species level, and among different forms of fusion patterns of the same enzyme. Thus, we used the following criteria for sequence collection: 1) sequences for all biotin-dependent carboxylases encoded in fully-sequenced genomes of archaea; 2) representatives of each enzyme specificity with experimentally-confirmed function; 3) representatives of all fusion patterns for each enzyme specificity and 4) enzymes from representatives of all three superkingoms. Enzymes with experimentally-confirmed and sequence-predicted functions were included.

BCC domain. Biotin carboxyl carrier (BCC) is the only domain common to all biotin-dependent carboxylases, transcarboxylases and decarboxylases. The BCC polypeptide has two sub-domains, one near the N-terminus, involved in formation of a complex with BC (Choi-Rhee and Cronan, 2003), and one near the C-terminus, which contains biotin and is

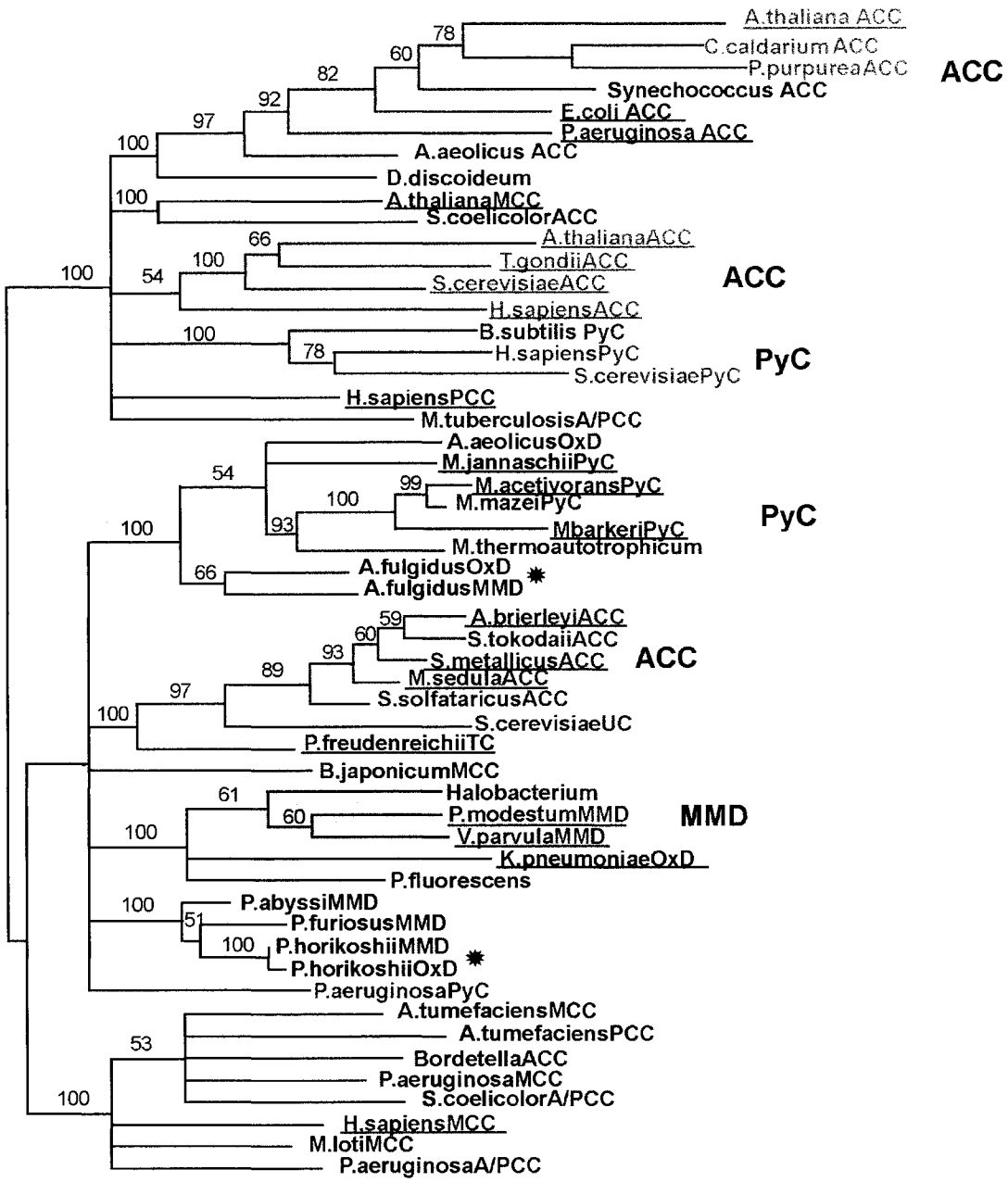


Fig. 3. Phylogeny of BCC subunit of biotin-dependent carboxylases. Species names in green indicate eukaryotes, **blue**: **bacteria**, **red**: **archaea**, entries have experimentally confirmed function. Clades that are homogenous with respect to the enzyme function are shaded and signed at right.
ACC, acetyl-CoA carboxylase; **A/PCC**, acetyl/propionyl-CoA carboxylase; **MCC**, methyl-crotonyl-CoA carboxylase; **MMD**, methylmalonyl-CoA decarboxylase; **OxD**, oxaloacetate decarboxylase; **PyC**, pyruvate decarboxylase; **TC**, transcarboxylase; **UC**, urea carboxylase; no enzyme name means that no annotation for protein is available.

required for catalytic activity (Athappilly and Hendrickson, 1995). The N-terminal sub-domain appears to be unique to BCCs that occur as a separate peptide, and thus was not used for these studies.

Although ACC from *Myxococcus xanthus* has been shown to be biotinylated *in vivo,* and the enzyme has the highest activity toward acetyl-CoA substrate (Kimura et al., 2000), the primary sequence of the BCC domain of this protein is divergent from all other BCC sequences. In fact, the biotin attachment domain has no detectable homology to any other known BCC, so we did not include it in this analysis.

Because of the small size of the biotin-containing domain (~87 amino acids), and because it is composed of a mixture of well-conserved and highly saturated residues, this domain is not particularly suitable for phylogenetic analysis. However, several patterns stay robust under different methods of tree construction (Fig. 3).

BCC domains of ACCs form three major monophyletic branches (green shaded boxes in Fig. 3), although altogether this is a polyphyletic group. Two of these three branches contain members from a single superkingdom: archaea or eukaryota. In the third branch, BCCs from plant heteromeric ACC form a clade with the Cyanobacterium *Synechococcus sp.*, as would be expected for genes derived from Cyanobacteria via plastidial endosymbiosis. Interestingly, BCCs from archaeal decarboxylases form clusters of different enzymes for the same organism (asterisks in Fig. 3) suggesting a recent duplication and divergence within each of these lineages. The BCC tree reflects the universal character of the BCC domains, which do not need to mirror the specificity of the enzymes in which they function.

Biotin carboxylase domain is the most highly conserved of the three domains. The sequence similarity is fairly uniformly distributed along the domain, thus we used a 438-amino acid sequence spanning almost the entire BC domain for our analyses. These extensive regions of similarity allow for well-resolved phylogeny (Fig. 4). The BC domain of ACCases form three separate major branches, which, as for BCC, are not a monophyletic clade. In addition, similar to the case for BCC, two of the three branches contain members of a single superkingdom (archaea (except *Aquifex aeolicus*), and eukaryota), and in the third branch containd bacterial and plastidic ACC. Interestingly, ACC from the hyperthermophilic

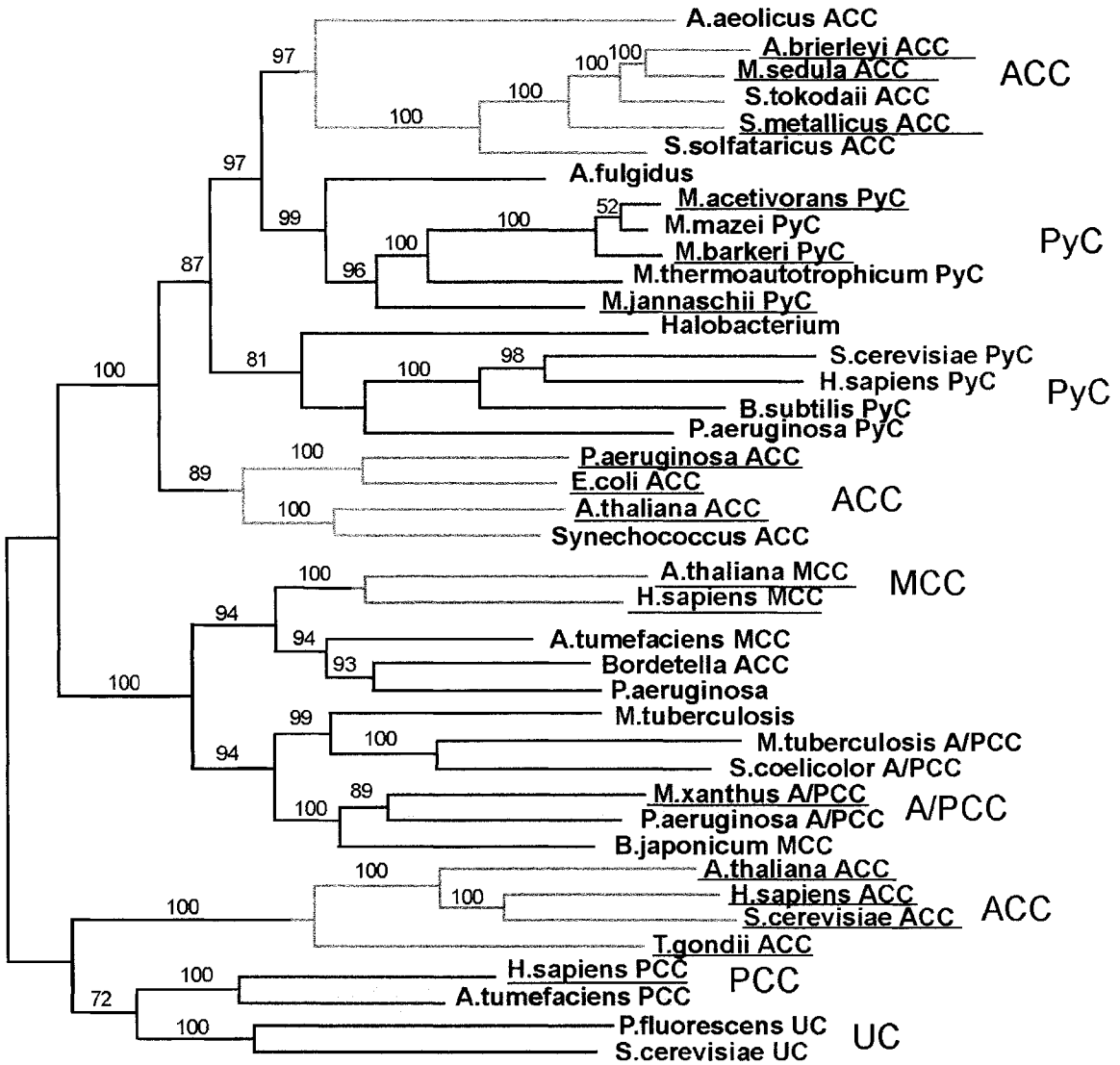


Fig. 4. Phylogeny of BC subunit of biotin-dependent carboxylases. Species names in green indicate eukaryotes, **blue**: **bacteria**, **red**: **archaea**, underlined entries have experimentally confirmed function. Clades that are homogenous with respect to the enzyme function are shaded and signed at right.
ACC, acetyl-CoA carboxylase; **A/PCC**, acetyl/propionyl-CoA carboxylase; **MCC**, methyl-crotonyl-CoA carboxylase; **PyC**, pyruvate decarboxylase; **UC**, urea carboxylase; no enzyme name means that no annotation for protein is available.

bacterium *A. aeolicus* clades with archaeal ACCs; this may be a result of horizontal transfer between these hyperthermophilic organisms.

PyCs form two monophyletic branches: one of archaeal PyC and the other encompassing bacterial two-subunit PyC, eukaryotic homomeric PyC, and bacterial

homomeric PyC. *Archaeoglobus fulgidus* BC of unknown function clades together with archaeal PyC, indicating a possible function for this polypeptide.

The small PCC and UC clades are monophyletic, and each includes both eukaryotic and eubacterial superkindoms. The relationships between clades containing A/PCC and MCC are more complex, and these clades include many proteins with sequence-predicted function.

CT domain. The two families of CT domains, one that uses an acyl-CoA substrate (acetyl-, methylmalonyl-, methylcrotonyl- or propionyl-CoA), and the other specific for pyruvate (OxD and PyC) are not homologous (Samols et al., 1988). We have conducted separated analyses for these two families.

The polypeptides of the acyl-CoA-type (Fig. 5) share low homology but have high structural similarity (Zhang et al., 2003; Hall et al, 2003; Wendt et al, 2003; Diacovich et al., 2004). Therefore we guided the sequence alignment by known and predicted secondary structures of CT domain. The deepest branching reflects the subunit composition of the proteins: there is a split into a group of enzymes with separate CT subunit and two sister ACC clades, corresponding to homomeric enzyme and CT present on two separate subunits, alpha and beta. The group of enzymes in which CT is present as a single subunit shares the highest similarity on the sequence level and this is seen as clades that unite enzymes with mixed functions (exception being the CT domains of glutaconyl-CoA decarboxylases, which form a well-separated clade). However, few of these single-subunit CTs have an experimentally confirmed function, and if only those with confirmed functions are taken under account, the eukaryotic and bacterial A/PCC, PCC and MCC might in fact form monophyletic branches. Archaeal sequences for A/PCC and MMD do not form sister taxa with their bacterial and eukaryotic counterparts. Function of the archaeal *Halobacterium* CT-like sequence, annotated as "putative PCC homolog" could not be deduced from its position in the tree since it is not tightly claded within any branch. In contrast, *Halobacterium* putative MMD clades with human PCC and bacterial A/PCC and thus might have A/PCC function.

The tree of the pyruvate-type CT domain, like that of the acetyl-type CT domain, branches predominantly according to fusion pattern ([CT-BCC] versus [BC-CT-BCC]) rather than according to the substrate specificities of the enzyme (Fig. 6). The homomeric PyC,

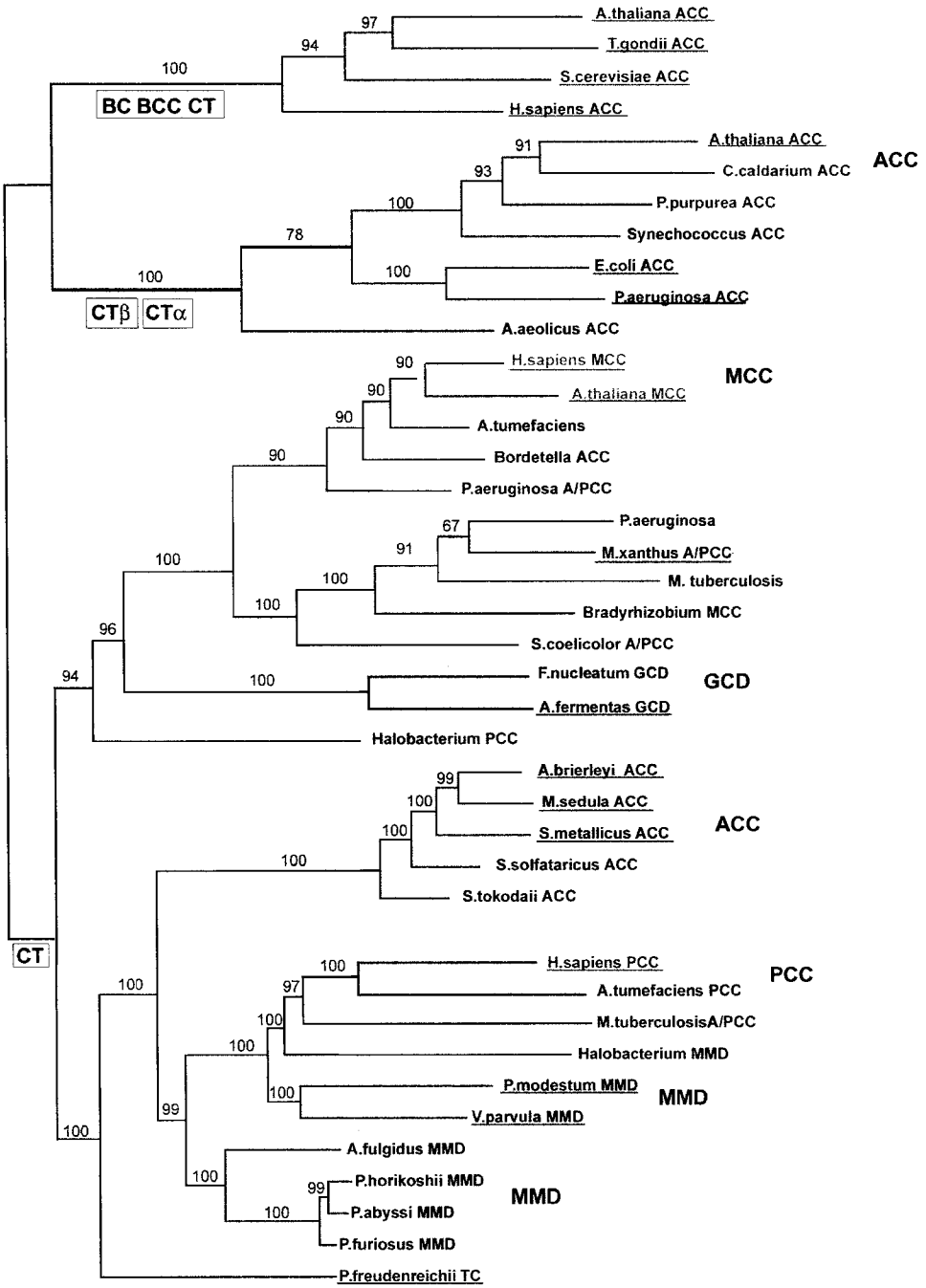


Fig. 5. Phylogeny of acetyl-CoA type carboxyltransferase domain. Species names in green indicate eukaryotes, **blue**: **bacteria**, **red**: **archaea**, underlined entries have experimentally confirmed substrate specificity. Three deepest clades are homogenous with respect to the enzyme fusion pattern, as at left. Clades that are homogenous with respect to the enzyme function are shaded and signed at right.
ACC, acetyl-CoA carboxylase; **A/PCC**, acetyl/propionyl-CoA carboxylase; **MCC**, methyl-crotonyl-CoA carboxylase; **MMD**, methylmalonyl-CoA decarboxylase; **OxD**, oxaloacetate decarboxylase; **PCC**, propionyl-CoA carboxylase; **PyC**, pyruvate decarboxylase; **UC**, urea carboxylase; **TC**, transcarboxylase; **GCD,** glutaconyl-CoA decarboxylase, no enzyme name means that no annotation for protein is available.

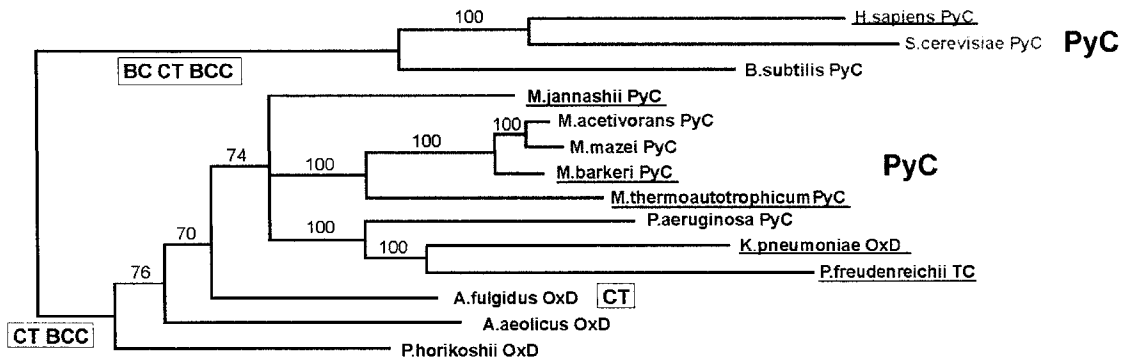


Fig. 6. Phylogeny of pyruvate-type carboxyltransferase domain. Species names in green indicate eukaryotes, **blue**: **bacteria**, **red**: **archaea**, underlined entries have experimentally confirmed function. Clades that are homogenous with respect to the enzyme function are shaded and signed at right. Enzyme fusion pattern ais indicated at left, except *A. fulgidus*, which has unusual pattern (separate CT).
OxD, oxaloacetate decarboxylase; **PyC**, pyruvate decarboxylase; **TC**, transcarboxylase.

both bacterial and eukaryotic, form a well-separated clade. Archaeal and bacterial [CT-BCC] PyC are claded together with OxD of the same subunit structure, rather than with homomeric PyC. OxD from the Archaea *P. horikoshii* does not form sister taxa with *A. fulgidus* OxD, the latter of which has an unusual composition, with the CT domain as a separate subunit.

History of domains of carboxylases

The phylogenies in Figures 3-6 reveal the complex histories of individual domains that evolved as independent units.

Biotin carboxylase occurs as a single subunit in bacterial and plant plastidic heteromeric ACC, and in archaeal ACC and PyC (Fig. 7A). In light of the ancient requirement for ACCase for autotrophic CO_2 fixation (Menedez et al., 1999), the archaeal PyC BC domain may have arisen from the duplication and divergence of the archaeal ACC BC domain, although no archaeal species have been reported to possess both ACC and PyC activities. At some point, the BC domain fused with the BCC domain and this ancient protein diverged into A/PCC- and MCC-specific proteins (Fig. 7B).

The trees for CT domain tend to group enzymes with similar subunit structures together. This suggests that in some cases, the formation of domain combinations preceded the differentiation of protein function. The enzymes with two subunit structures ([BC-BCC] and [CT]), such as A/PCC, PCC and MCC, probably have evolved from a common ancestor

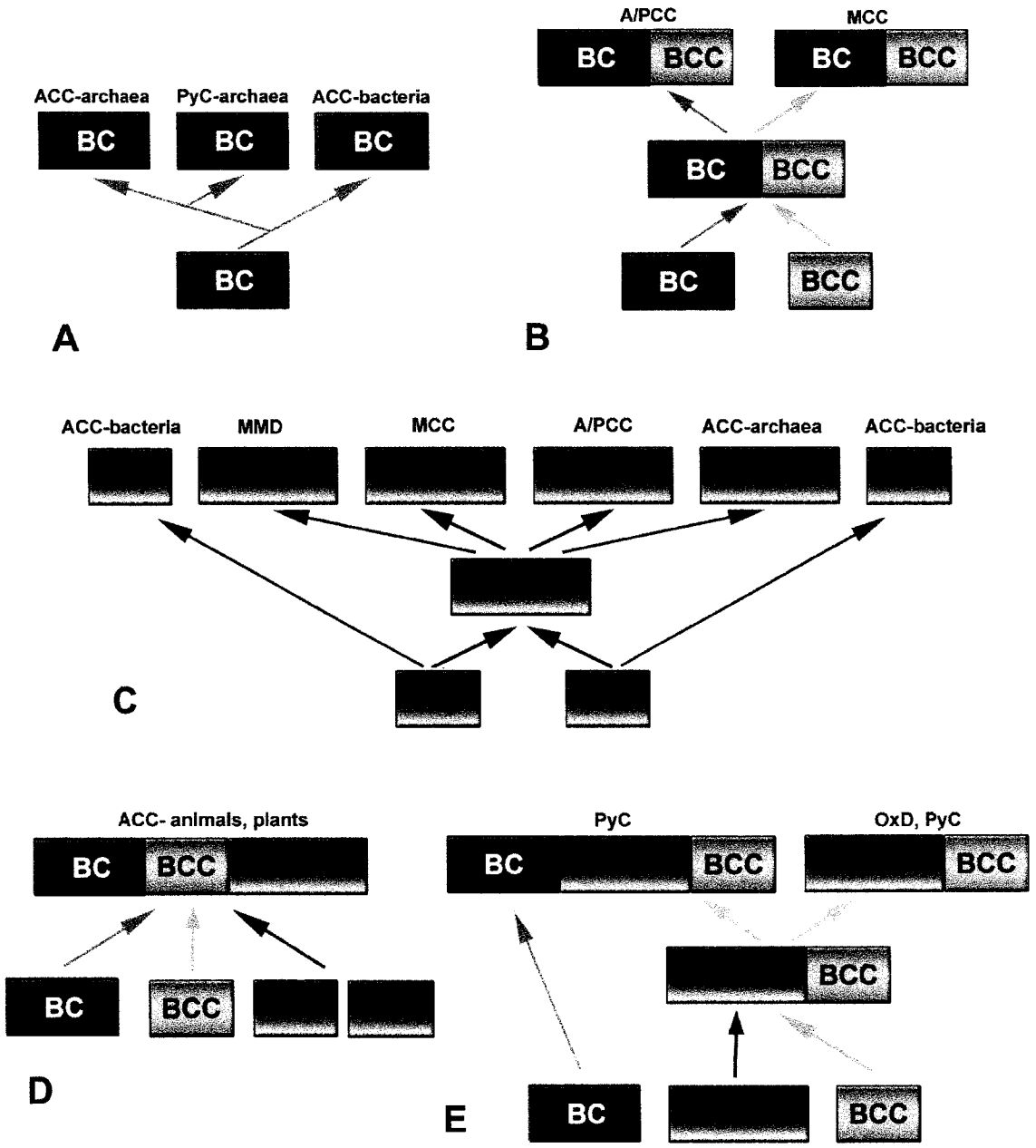


Fig. 7. Schematic representation of phylogenetic history of domains of carboxylases.
A. The ancestral BC domain diverged into bacterial, present now in ACC and into archaeal one, now present in ACC and PyC **B**. Fusion of the ancestral BC and BCC domains followed by divergence of the resulting protein into ACC, PCC and MCC. **C**. CT domain is composed of two subdomains, α and β, which function as separate subunits in bacterial ACCs, while fused CT subunit diverged into MMD, MCC, A/PCC and archaeal ACC forms. **D.** Fusion of BC, BCC and CT domains led to formation of multidomain ACC in early eukaryotes. **E.** Both PyC and OxD originated by fusion of CT and BCC domains, which subsequently diverged into pyruvate- and oxaloacetate-specific ones, and BC domain joined in eukaryotic and some bacterial PyC.

of similar fusion pattern. CT alpha and CT beta domains, still present in bacteria and plants, presumably fused in some lineages and gave rise to MMD, MCC, A/PCC and archaeal ACC (Fig. 7C). The reverse scenario, in which an ancient CT domain is split into CT alpha and CT beta, is also possible. However, the marked similarity of the secondary structure of Ctα and CTβ to each other (Bilder et al., 2006) would rather suggest duplication, divergence and fusion.

ACCases of different structure are polyphyletic. None of the four domain-based trees clades eukaryotic homomeric, heteromeric, and archaeal ACCs together. This indicates that eukaryotic ACCases were probably not formed by fusion of bacterial or archaeal genes for multisubunit ACC protein. Instead, eukaryotic ACCases probably evolved independently as a result of independent fusion of ancestral BC, BCC and CT domains in early eukaryotic lineages (Fig. 7D). Ancient CT and BCC domains fused to form subunits of OxD and PyC. BC domain fused at N-terminal end of this peptide forming PyC in fungi, plants and some bacteria (ex. *Bacillus subtilis*), while the biotin decarboxylase and other small structural subunits comprise OxD, as well as other decarboxylases (Fig. 7E).

Archaeal carboxylases may have been acquired from bacterial species, for example Actinobacteria or hyperthermophilic bacteria, and employed in new roles. The universal distribution of ACC in the bacterial and eukaryotic world versus the isolated occurrence of carboxylases in archaeal lineages (ACC only in *Sulfolobaceae*; also other carboxylases are not common: PyC is present only in methanogenes, decarboxylases in *Pyrococcus* and *Archaeoglobus*) and high similarity between these sequences, suggestive of their recent acquisition, are in favor of such a hypothesis. On the other hand, the similarity may reflect close relationships of species, and archaeal ACCs, although highly similar to each other, are very divergent from bacterial and eukaryotic sequences, which would make the identification of their bacterial origin difficult. The crucial function of ACC in primordial CO_2 fixation (Menendez et al., 1999) strongly suggests its archaeal origin and subsequent loss.

Recent rearrangements

Some differentiations and fusions (or fissions) in carboxylase family appear to have occurred quite recently. A good example of such a recent event is seen in the genomes of *Pyrococcus horikoshii* and *Archaeoglobus fulgidus* (Fig. 8B). These Archaea each have two forms of biotin-dependent decarboxylases, methylmalonyl-CoA decarboxylase (MMD), and oxaloacetate decarboxylase (OxD). In *P. horikoshii,* CT and BCC of OxD are located on a single peptide. The two pyrococcal BCCs domains are nearly identical (67 of the 68 residues) and the two archaeoglobal BCCs share high similarity, too. In contrast, CTs domains of *Pyroccus horikoshii* and *Archaeoglobus fulgidus* are of non–homologous, acyl-CoA-specific and pyruvate-specific, types. Thus it appears that in both organisms BCCs were duplicated and diverged, and also that in *P. horikoshii* the CT and BCC subunits of OxD fused recently. This single event, so markedly written in the genomes, delivers the example of: 1) convergence of arrangements' evolution; 2) different histories of domains; 3) universal character of building blocks of carboxylases, and 4) dynamic and continuous evolution of carboxylases.

The evolution of biotin-dependent carboxylases is thus apparently an on-going process. In plants the rearrangement, displacement and substitution of forms of ACC has been noted (Sasaki et al, 1995; Konishi et al, 1996). The ACC subunits are being transferred from plastid to nucleus; apparently this process has been caught "midstream". We have found no chloroplasts coding a BC subunit, while CT beta is still present in most plastids, except Graminae. In Arabidopsis, one of two copies of homomeric ACCases have been tandemly duplicated and the difference in their expression pattern may indicate the differentiation of function leading in the future to a replacement of plastidial ACC genes with a single homomeric nuclear-encoded ACC (Yanai et al., 1995), as it happened in other plants. However, other studies indicate that second copy of ACC might be targeted either to the mitochondrion or coexist with heteromeric ACC in the plastid (Baud et al., 2003).
In rice, a member of Graminae, in which a nuclear-encoded homomeric form of ACC has entirely replaced the heteromeric one, we identified a 74 aa-long remnant of CTβ in the

plastid genome. A similar segment of plastidic-like sequence containing CTβ fragment is present in the nuclear genome of rice.

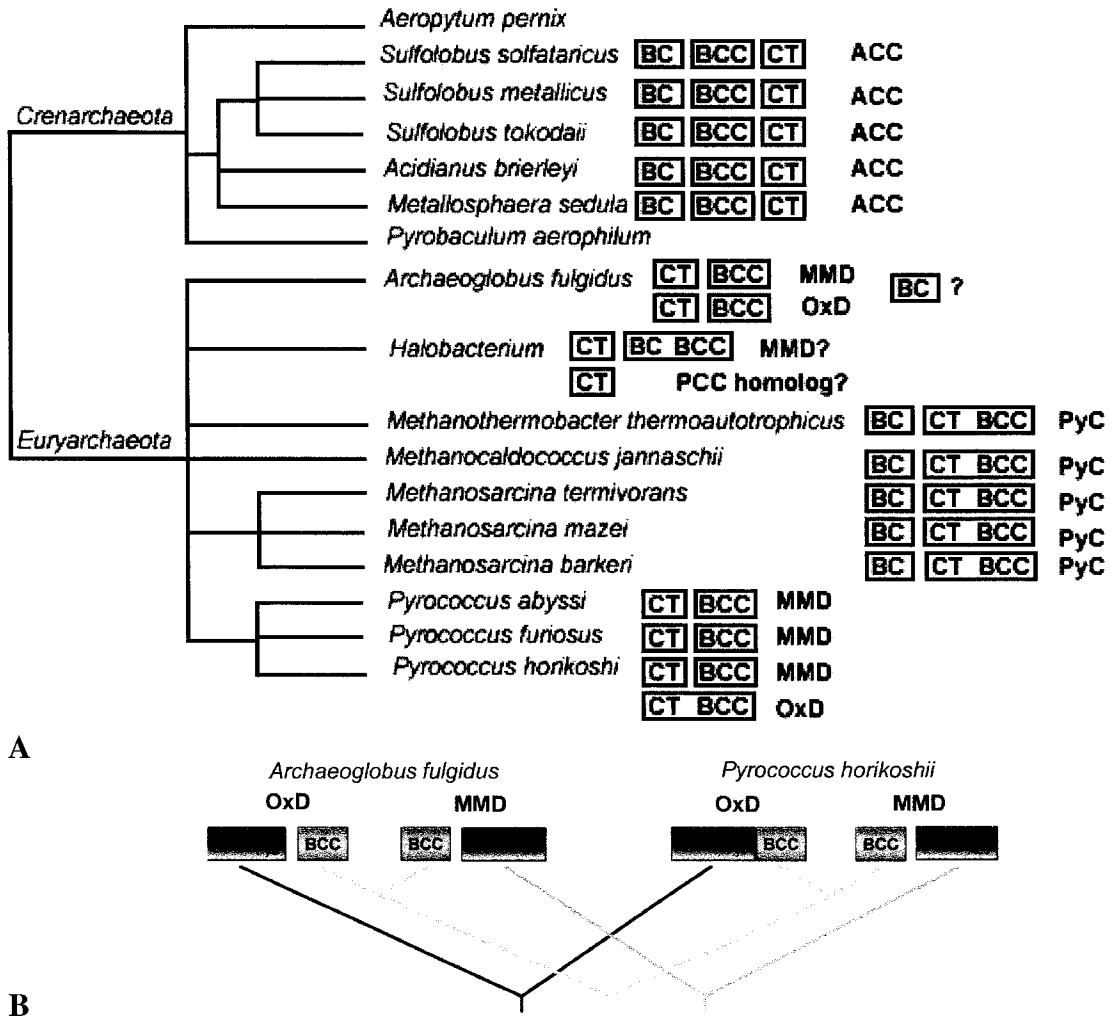


Fig. 8. A. Species tree of fully-sequenced archaean genomes, showing genes encoding biotin-dependent enzymes present in given genome. **B**. BCC domains of two enzymes with non-homologous CT domains (OxD and MMD) share more similarity among proteins within the same species than among proteins with the same enzymatic specificity. The two *Pyrococcus horikoshii* BCC domains, of which one is fused with CT domain of OxD, share 67/68 identical residues.
ACC, acetyl-CoA carboxylase; **MMD**, methylmalonyl-CoA decarboxylase; **OxD**, oxaloacetate decarboxylase; **PyC**, pyruvate decarboxylase.

Fusion patterns

Carboxylases can be classified by two distinct fusion patterns: **BC** (or **BD**)-**BCC-CT**, represented by ACC, PCC, MCC, MMD, and **BC** (or **BD**)-**CT-BCC**, as in PyC, UC and OxD (Fig. 2). Usually, either the three domains are on separate subunits, or all three are fused together, or BCC is fused with one of the remaining two. An exception is transcarboxylase, which contains BCC and two CT domains, one specific for acetyl-CoA and another for pyruvate ($\mathbf{CT_{ACC}}$- $\mathbf{CT_{PC}}$ **–BCC**), all three on separate subunits, as in *Propionibacterium freudenreichii* (Wood HG., 1979) or fused together as in *Giardia intestinalis* (Jordan IK et al, 2003). Because the pattern of gene fusion can be very specific, it provides an indication of possible function of the multidomain protein.

Many forms of multidomain carboxylases arose probably as a result of fusion, in some cases an independent event. The similar fusion patterns of homomeric PyC and UC that have non-homologous CT domains and recent fusion of BC and BCC subunit in archaeal OxD are examples of such independent fusion events.

Substrate specificity versus species specificity

A shift can be observed between the substrate-specificity and the species-specificity from domain to domain. CT is most activity-specific while BCC appears to be much more universal unit. This can be seen in case of decarboxylases from *P. horikoshii* and *A. fulgidus*. An example of universality of BC is the *S. coelicolor* protein containing BC and BCC domains, which is active in combination with either of two CT subunits: one with ACC specificity and another PCC-specific (Diacovich et al., 2004). It seems that while the CT subunit determines the enzyme specificity, the BC and BCC might be to some degree substitutable between the biotin-dependent carboxylases present in a given lineage. This discrepancy is reflected in the different histories of the domains.

Annotation challenges

In public databases, genes for carboxylases are often misannotated in process of sequence-based function predictions, or they lack annotation. The assignment of function for potential carboxylases is especially difficult because of similarities between structurally similar CTs with various specificities and universal character of BC and BCC subunits that often render the sequence homology inconclusive. For example, enzymes from Actinobacteria *Mycobacterium tuberculosis* and *Streptomyces coelicolor* are annotated as ACC, but possess both ACC and significant PCC activities, and have PCC-like structure. The residues responsible for substrate binding can also be dispersed along the primary structure, (as in yeast ACC (Zhang et al., 2003)) or, although chemically identical, may not be in equivalent positions, (as for example in enolases superfamily (Hasson et al., 1998), or chloramphenicol acetyltransferase and UDP-N-acetylglucosamine acyltransferase from hexapeptide repeat protein superfamily (Todd et al., 2001), underscoring the importance of spatial arrangement and not primary sequence. Moreover, very few residues might be responsible for enzyme specificity: Diacovich et al., (2004) were able to completely reverse specificities of the *S. coelicolor* carboxylase between ACC and PCC by introducing a single point mutation. This highlights the limited use of sequence-based functional prediction methods for these enzymes. A gene for CT subunit, which is not in proximity of genes for other potential carboxylases' domains, may deliver clues as to whether the substrate is the fatty acyl or the oxoacid type, but still will not allow inference as to whether that CT is a part of a carboxylase or a decarboxylase. When there is an isolated BC or BCC subunit in the genome, as is the case for *A. fulgidus*, (Fig. 8A), nothing short of biological activity assay would likely answer the question of its function. In such cases other criteria, like fusion pattern, gene neighborhood and genome content, are of assistance. Using these information we can for example classify the CT and BC-BCC peptides (GI:10581017 and GI:10581019) from *Halobacterium* as probable components of PCC. The CT protein was annotated in Genbank as part of putative MMD, however the genome of *Halobacterium* does not contain the biotin decarboxylase necessary to form a functional MMD, and the *Halobacterium* CT protein clusters with PCC enzymes (Fig. 5).

CONCLUSIONS

These results paint the picture of a mix-and-match pool of domains in early stages of life that could be arranged in different combinations. Multiple events of domain duplications, rearrangements, fusions and fissions were involved in the process of carboxylases evolution. These led to a plethora of structures and functions, and some functions seem to be reached in parallel by different structural solutions. Many of those events happened very early in evolutionary history, as suggested by presence of long multifunctional peptides like UC or PyC in bacterial species, while others are relatively recent. The two implementations of ACC – bacterial and eukaryotic- coexist in plants and deliver an example of displacement of bacterial, heteromeric form by the homomeric one, a process that can be observed at many stages. Although ancient events are hard to reconstruct, the near-ubiquitous distribution and simple multisubunit structure of ACC throughout the bacterial kingdom indicates this form of the enzyme might resemble an ancient prototype of the domains that gave rise to a variety of the biotin-dependent enzymes seen nowadays in processes of domain shuffling and fusion.

METHODS

The aminoacid and DNA sequences for this study were obtained from publicly available databases: Genbank (SwissProt (http://us.expasy.org/sprot/), TIGR (http://www.tigr.org), dictyBase (http://dictybase.org/), Chlamydb (http://www.biology.duke.edu/chlamydb) and HAMAP (http://us.expasy.org/sprot/hamap/plastid.html). The representatives of biotin carboxylase family that have experimentally confirmed enzymatic function (Tab.1) were used as queries for BLAST search against all available Genbank data. Those newly found sequences were subjected to another round of BLAST scan to check whether they are not most similar to an enzyme other than the one used as a query. Gene neighborhood was also taken into account for extracting sequences for multisubunit BLAST hits.

All carboxylase-related protein sequences found in archaeal genomes were used in our study. For other organisms, only one of the orthologous proteins is shown.

Amino acid sequences were aligned with ClustalW (Thompson et al., 1994) in BioEdit package (Hall, 1999) or ClustalX program (Thompson et al., 1997) and with T-Coffee program (Notredame et al, 2000) (available at http://www.ch.embnet.org/software/TCoffee.html) and adjusted manually. Alignment of acetyl-specific CT domain was guided by alignment of experimental and predicted structural features (Zhang et al, 2003). Phylogenetic analyses of aligned sequences were performed using neighbour joining and parsimony methods with 100 bootstrap resamplings from PAUP software package (Swofford, 1999) and with MrBayes software (Huelsenbeck and Ronquist, 2001). Protein secondary structure prediction was performed at the PSIPRED Protein Structure Prediction Server (McGuffin et al, 2000).

REFERENCES

Athappilly FK and Hendrickson WA. 1995. Structure of the biotinyl domain of acetyl-coenzyme A carboxylase determined by MAD phasing. *Structure* **3:**1407–1419.

Baud S, Guyon V, Kronenberger J, Wuilleme S, Miquel M, Caboche M, Lepiniec L, Rochat C. 2003. Multifunctional acetyl-CoA carboxylase 1 is essential for very long chain fatty acid elongation and embryo development in Arabidopsis. *Plant J.* **33**:75-86.

Beh M, Strauss G, Huber R, Stetter KO, Fuchs G. 1993. Enzymes of the reductive citric acid cycle in the autotrophic eubacterium Aquifex pyrophilus and in the archaebacterium Thermoproteus neutrophilus. *Arch Microbiol* **160**:306-311.

Bilder P, Lightle S, Bainbridge G, Ohren J, Finzel B, Sun F, Holley S, Al-Kassim L, Spessard C, Melnick M, Newcomer M, Waldrop GL. 2006. The Structure of the Carboxyltransferase Component of Acetyl-CoA Carboxylase Reveals a Zinc-Binding Motif Unique to the Bacterial Enzyme. *Biochemistry.* **45**:1712-1722.

Buckel W. 2001 Sodium ion-translocating decarboxylases. *Biochim Biophys Acta.* **1505**:15-27.

Burton NP, Williams TD, Norris PR. 1999. Carboxylase genes of Sulfolobus metallicus. *Arch Microbiol.* **172**:349-53.

Chapman-Smith A, Cronan JE Jr 1999. Molecular biology of biotin attachment to proteins. *J Nutr* **129:**477S-484S

Choi-Rhee Eunjoo and Cronan John E. 2003. The Biotin Carboxylase-Biotin Carboxyl Carrier Protein Complex of *Escherichia coli* Acetyl-CoA Carboxylase. *J. Biol. Chem.,* **278**:30806-30812.

Chuakrut S, Arai H, Ishii M, Igarashi Y. 2003. Characterization of a bifunctional archaeal acyl coenzyme A carboxylase. *J Bacteriol.* **185**:938-47.

Cronan JE Jr. 2001 The biotinyl domain of Escherichia coli acetyl-CoA carboxylase. Evidence that the "thumb" structure is essential and that the domain functions as a dimer. *J Biol Chem.* **276:**37355-64.

Diacovich L, Mitchell DL, Pham H, Gago G, Melgar MM, Khosla C, Gramajo H, Tsai SC. 2004. Crystal structure of the beta-subunit of acyl-CoA carboxylase: structure-based engineering of substrate specificity. *Biochemistry.* **43**:14027-36.

Egli MA, Gengenbach BG, Gronwald JW, Somers DA, Wyse DL. 1993. Characterization of Maize Acetyl-Coenzyme A Carboxylase. *Plant Physiol.* **101**:499-506

Hall PR, Wang YF, Rivera-Hainaj RE, Zheng X, Pustai-Carey M, Carey PR, Yee VC. 2003. Transcarboxylase 12S crystal structure: hexamer assembly and substrate binding to a multienzyme core. *EMBO J.* **22**:2334-47.

Hasson MS, Schlichting I, Moulai J, Taylor K, Barrett W, Kenyon GL, Babbitt PC, Gerlt JA, Petsko GA, Ringe D. 1998. Evolution of an enzyme active site: the structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase. Proc Natl Acad Sci U S A. **95**:10396-401.

Huber, R. *et al.* 1992. *Aquifex pyrophilus* gen. nov. sp. nov. represents a novel group of marine hyperthermophilic hydrogen oxidizing bacteria. *Arch. Micrtobiol.* **15**, 340-351.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. **17**:754-5.

Hugler M, Krieger RS, Jahn M, Fuchs G. 2003 Characterization of acetyl-CoA/propionyl-CoA carboxylase in Metallosphaera sedula. Carboxylating enzyme in the 3-hydroxypropionate cycle for autotrophic carbon fixation. *Eur J Biochem.* **270**:736-44.

Jitrapakdee S, Wallace JC. 2003. The biotin enzyme family: conserved structural motifs and domain rearrangements. *Curr Protein Peptide Sci.* **4**:217-229.

Jordan IK, Henze K, Fedorova ND, Koonin EV, Galperin MY. 2003. Phylogenomic analysis of the Giardia intestinalis transcarboxylase reveals multiple instances of domain fusion and fission in the evolution of biotin-dependent enzymes. *J Mol Microbiol Biotechnol.* **5**:172-89.

Kimura Y, Miyake R, Tokumasu Y, Sato M. 2000. Molecular cloning and characterization of two genes for the biotin carboxylase and carboxyltransferase subunits of acetyl coenzyme A carboxylase in Myxococcus xanthus. *J Bacteriol.* **182**:5462-9

Konishi T, Shinohara K, Yamada K, Sasaki Y. 1996. Acetyl-CoA carboxylase in higher plants: most plants other than gramineae have both the prokaryotic and the eukaryotic forms of this enzyme *Plant Cell Physiol.* **37**:117-22.

Lane MD, Moss JD, Polakis SE. 1974. Acetyl coenzyme A carboxylase *Curr. Top. Cell Regul* **8**:139-195.

Li SJ, Cronan JE Jr. 1992. The genes encoding the two carboxyltransferase subunits of Escherichia coli acetyl-CoA carboxylase. *J Biol Chem.* **267**:16841-7

Lim F, Morris CP, Occhiodoro F, Wallace JC. 1988. Sequence and domain structure of yeast pyruvate carboxylase *J Biol Chem.* **263**:11493-7.

McGuffin LJ, Bryson K, Jones DT. 2000. The PSIPRED protein structure prediction server. Bioinformatics. **16**:404-405.

Menendez C, Bauer Z, Huber H, Gad'on N, Stetter KO, Fuchs G. 1999. Presence of acetyl coenzyme A (CoA) carboxylase and propionyl-CoA carboxylase in autotrophic

Crenarchaeota and indication for operation of a 3-hydroxypropionate cycle in autotrophic carbon fixation. *J Bacteriol.* **181**:1088-98.

Mukhopadhyay B, Purwantini E, Kreder CL, Wolfe RS. 2001. Oxaloacetate synthesis in the methanarchaeon Methanosarcina barkeri: pyruvate carboxylase genes and a putative Escherichia coli-type bifunctional biotin protein ligase gene (bpl/birA) exhibit a unique organization. *J Bacteriol* **183**:3804-10

Nikolau BJ, Ohlrogge, Wurtele ES. (2003) Plant biotin-containing carboxylases. Arch Biochem Biophys. 2003 Jun 15;414(2):211-22

Notredame C, Higgins D, Heringa J. 2000. T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.* **302**:205-217.

Obermayer M. and Lynen F. 1976. Structure of biotin enzymes *TBIS* **1**: 169-171

Ramsay RR, Gandour RD, van der Leij FR. 2001. Molecular enzymology of carnitine transfer and transport. *Biochim Biophys Acta.* **1546:**21-43.

Roesler KR, Shorrosh BS, Ohlrogge JB. 1994. Structure and expression of an Arabidopsis acetyl-coenzyme A carboxylase gene. *Plant Physiol.* **105:**611-7.

Samols D, Thornton CG, Murtif VL, Kumar GK, Haase FC, Wood HG. 1988. Evolutionary conservation among biotin enzymes. *J Biol Chem.* **263**:6461-4.

Sasaki Y, Konishi T, Nagano Y. 1995. The Compartmentation of Acetyl-Coenzyme A Carboxylase in Plants. *Plant Physiol.* **108**:445-449.

Sprott, GD. 1992. Structures of archaebacterial membrane lipids. *J. Bioenerg. Biomembr.* **24**:555-566.

Thompson JD, Higgins DG, Gibson TJ. 1994. ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position–specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**: 4673-4680.

Thompson JG, Gibson,TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**:4876-4882.

Todd AE, Orengo CA, Thornton JM. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol.* **307**:1113-43.

Toh H, Kondo H, Tanabe T 1993 Molecular evolution of biotin-dependent carboxylases. *Eur J Biochem.* **215**: 687-696.

Wendt KS, Schall I, Huber R, Buckel W, Jacob U. 2003. Crystal structure of the carboxyltransferase subunit of the bacterial sodium ion pump glutaconyl-coenzyme A decarboxylase. *EMBO J.* **22**:3493-502.

Wood, HG. 1979. The anatomy of transcarboxylase and the role of its subunits. *CRC Crit. Rev. Biochem.* **7:**143 – 160.

Yanai Y, Kawasaki T, Shimada H, Wurtele ES, Nikolau BJ, Ichikawa N. 1995. Genomic organization of 251 kDa acetyl-CoA carboxylase genes in Arabidopsis: tandem gene duplication has made two differentially expressed isozymes. *Plant Cell Physiol.* **36**:779-87.

Zhang H, Yang Z, Shen Y, Tong L. 2003. Crystal structure of the carboxyltransferase domain of acetyl-coenzyme A carboxylase. *Science.* **299**:2064-7.

CHAPTER 3. ARTICULATION AND HIERARCHICAL ORGANIZATION OF CORE METABOLIC PROCESSES IN PLANTS

A paper to be submitted to *Plant Physiology*

Wiesława Mentzen, Nick Ransom, Basil J. Nikolau and Eve Syrkin Wurtele

ABSTRACT

Elucidating network structure and function in multicellular eukaryotic organisms is an emerging and important goal. We describe transcriptional modularity of three core metabolic processes in the model plant Arabidopsis. These pathways form transcriptional modules roughly in the frame outlined by the chemical sequence of reactions defining each pathway. The modules incorporate a wider set of transporters, cofactors and substrate-producing enzymes, and regulatory molecules that may represent a common task. These data define a novel hierarchical transcript-level structure, where genes performing smaller, more specific tasks are recruited into higher-order modules with a broad function in catabolism.

INTRODUCTION

Biological systems are characterized by their capacity to achieve net metabolic inter-conversions while maintaining homeostasis in the face of environmental and developmental cues. This capacity is hard-wired into the genetic blueprint of an organism, and manifested by the controlled expression of the genetic potential of the organism's genome as it responds to divergent signals and prompts. Mechanisms that control the expression of an organism's genetic potential include those that regulate gene transcription, RNA processing, stability and translation, and polypeptide processing and assembly into complexes. Some of these complexes are enzyme catalysts, whereas others are structural or regulatory. A metabolic network can be defined as encompassing the collection of genes, mRNAs, proteins and metabolites that work in coordination to achieve net metabolic conversions.

Advances made over the last decade in the area of functional genomics have provided an increasing ability to globally profile genome expression at the level of RNAs, proteins and metabolites. These data define the transcriptome, proteome and metabolome, respectively. It is conceptually possible to identify metabolic networks based upon experimental data that reveal correlations in abundance of molecules (mRNAs, proteins/protein complexes, or metabolites) that belong to a common metabolic network. Given that the behavior of an organism can be regulated by multiple mechanisms that impact the transcriptome, proteome and metabolome, it is significant to ask the extent to which the transcriptome can reveal metabolism and its regulation.

In the unicellular eukaryote *Saccharomyces cerevisiae*, which offers the advantage of cell population that is homogenous and a relatively simple genome with only 6,604 genes (*1*), the operation of metabolic networks, such as glycolysis and purine metabolism, has been revealed by the analysis of transcriptomics datasets alone (*2-4*). But can such metabolic networks be detected in organisms with a larger, more complex genome, or in multicellular organisms where evidence for metabolic networks can be swamped by noise associated with cellular differentiation? Preliminary microarray studies show that genes belonging to related pathways can be coexpressed across different tissues, as are, for example, genes for protein synthesis in *C. elegans* (*5*), Krebs cycle and respiratory chain in frog (*6*) or lactose biosynthesis in mouse (*7*).

Because it has become technically possible to profile the entire transcriptome of an organism (it is still a technical challenge to determine the proteome and metabolome), we tested the feasibility of using transcriptomics data to reveal the operation of metabolic networks and their regulation, in the plant model system of Arabidopsis.

Arabidopsis is a multicellular eukaryotic organism that has a large genome (about 31,270 genes (*8*)). Moreover, as with other plant genomes, Arabidopsis has undergone whole-genome duplications (polyploidization), followed by radical gene silencing, gene turnover, and expansion of gene families, the most recent of which is thought to have occurred about 25 million years ago (*9, 10*). Therefore, conducting these analyses in Arabidopsis also offers the potential of addressing the complicating factors associated with polyploidization.

RESULTS AND DISCUSSION

In order to determine how transcriptional networks extend across metabolic processes, we selected three core pathways for analysis: fatty acid biosynthesis, starch metabolism, and mitochondrial leucine catabolism (Fig. 1). Fatty acid biosynthesis delivers acyl components for cellular membranes, signaling molecules, and energy storage (e.g., seed oils). Starch is the principal form of glucose storage in plants. Leucine catabolism provides

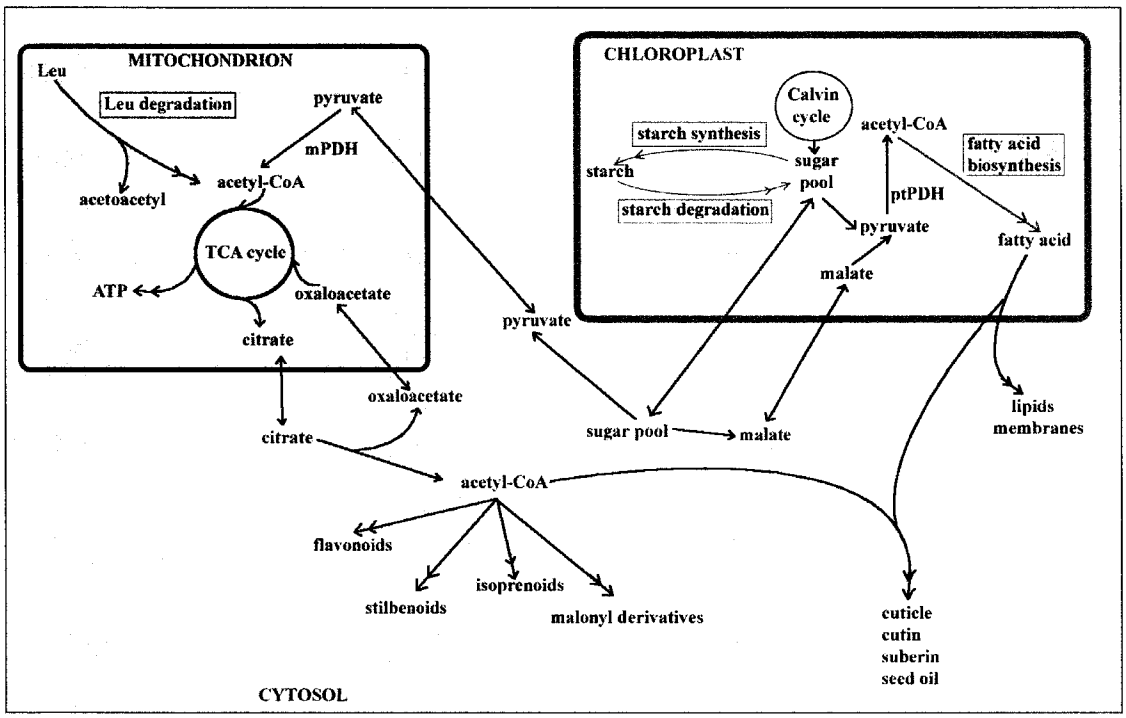


Fig. 1. Metabolic context of Arabidopsis leucine catabolism, starch metabolism and fatty acid biosynthesis pathways (blue rectangles).

an alternate source of acetyl-CoA to sustain respiration and metabolic processes in the absence of photosynthesis (*11*). We included not only the genes encoding enzymes of the "textbook pathway", but also genes encoding enzymes important for synthesis of cofactors and transporters that are thought to be involved in these processes. Specifically, a set of 121 genes associated with these pathways was identified based upon one of three criteria: 1) genes demonstrated by experimental evidence to belong to one of these pathways; 2) genes assigned "putative" functionality in one of the pathways, based on sequence similarity; or 3)

genes encoding transporters and enzymes synthesizing co-factors required for these pathways (table S1).

Public Affymetrix ATH1 chip-transcriptomics datasets comprising of 956 biological samples were used to infer patterns of transcript co-accumulation for the 22,746 Arabidopsis genes that are represented on this chip (see Methods section for details). These data are drawn from 70 experiments that confer a wide range of developmental, and environmental and genetic perturbations on Arabidopsis. A pair-wise correlation matrix for the 121 selected genes was visualized as a graph by placing an edge between every pair of genes whose expression was correlated. Using a Pearson correlation threshold of 0.5, 0.6 and 0.7 yields three transcriptional networks of 105, 76 and 63 genes, connected by 724, 421, and 196 edges, respectively (Fig. 2). As the correlation threshold is increased, three distinct modules emerge, whose member genes closely correspond with the three metabolic pathways that were the focus of this study.

At a 0.6 correlation level the starch metabolism module contains 28 genes encoding all known enzymes of starch metabolism (*12*) (Fig. 2B). The fatty acid synthesis module contains genes encoding all enzymes required for the biosynthesis of 18-carbon fatty acids from pyruvate, as well as genes for biotin and lipoic acid biosynthesis, obligate cofactors for the first two reactions in this pathway (fig. S1). The leucine degradation module contains genes for the 4 known enzymes of mitochondrial leucine catabolism.

To test whether such pathway-specific modules can be revealed in the context of the entire genome, and to determine whether genes other than those encoding enzyme structural genes of a pathway are co-regulated, we calculated Pearson correlations between individual pathway specific hub genes (At2g40840, At4g34030, At2g05990). To test whether such pathway-specific modules can be revealed in the context of the entire genome, and to determine whether genes other than those encoding enzyme structural genes of a pathway are co-regulated, we calculated Pearson correlations between individual pathway specific hub genes (At2g40840, At4g34030, At2g05990) (*13*) and the 22,746 genes represented on the Affymetrix chip (Fig. 3). The results of these analyses demonstrate that despite the complexities associated with a large genome, and the challenge that the plant samples analyzed contain metabolically distinct cellular and tissue compartments, it is still possible to

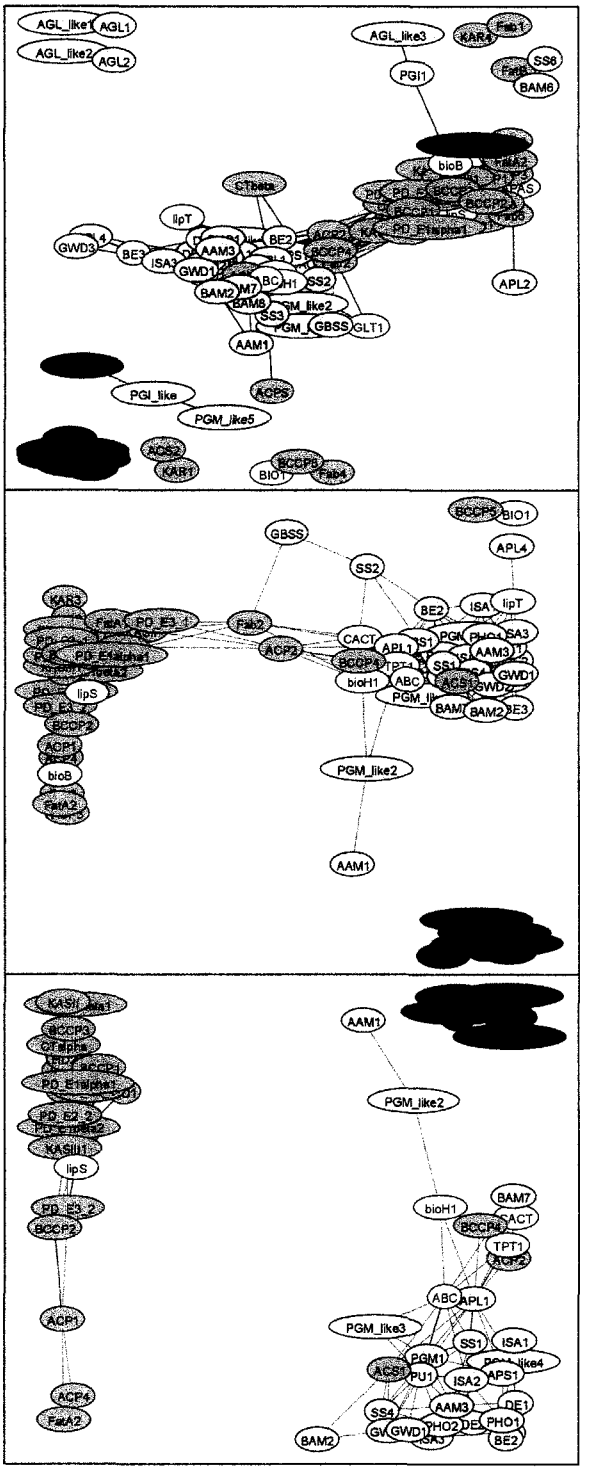


Fig. 2. Coexpression of genes within three core metabolic pathways. The entities with correlation above the threshold are connected with an edge; networks at three thresholds of Pearson correlation are compared: (**A**) 0.5; (**B**) 0.6; (**C**) 0.7. Node colors represent the metabolic function assigned to each gene (blue: fatty acid synthesis, green: starch metabolism, red: leucine catabolism, yellow: transport or cofactor synthesis).

find expression correlations among genes of common metabolic pathways. Specifically, the majority of the genes that show the highest correlation with the fatty acid biosynthesis hub gene, the starch metabolism hub-gene, and the leucine catabolism hub-gene are genes that code for enzyme structural genes required in each of these metabolic pathways (Fig. 3A). In addition to identifying these enzyme structural genes, these analyses reveal genes that indicate additional aspects of each metabolic network. For example, gene At5g52920, which correlates with the fatty acid biosynthesis hub-gene, is one of the fourteen Arabidopsis genes that code for pyruvate kinase (*14, 15*). Pyruvate kinase generates pyruvate, the immediate precursor of the acetyl-CoA used for fatty acid biosynthesis (*15, 16*). These data therefore constrain the complexity associated with the polyploidization of the Arabidopsis genome, and indicate that of the fourteen pyruvate kinase genes that occur in this genome, At5g52920

A

fatty acid bios.		starch metab.		leucine catab.	
ER, At2g05990		DE2, At2g40840		MCC-B, At4g34030	
0.86	At1g34430				
0.86	At5g10160				
0.84	At5g35360	0.85	At3g52180		
0.83	At5g46290			0.82	At2g43400
0.8	At2g30200				
0.8	At3g16950	0.76	At1g06460	0.8	At4g35770
0.79	At1g62640			0.79	At5g21170
0.78	At1g53520			0.78	At1g08630
0.77	At3g25860	0.73	At5g59130	0.78	At1g15040
0.77	At5g52920	0.73	At2g39920		
0.76	At5g16390			0.74	At4g28040
0.75	At5g08415	0.72	At1g67660	0.73	At3g47340
0.73	At2g22230	0.71	At3g18500		
0.72	At1g30120	0.7	At2g31040	0.72	At1g18270
0.72	At5g15530	0.7	At4g33490	0.71	At5g49450
0.71	At1g01090				
0.71	At5g50390			0.7	At3g49790
0.7	At4g12700				

B

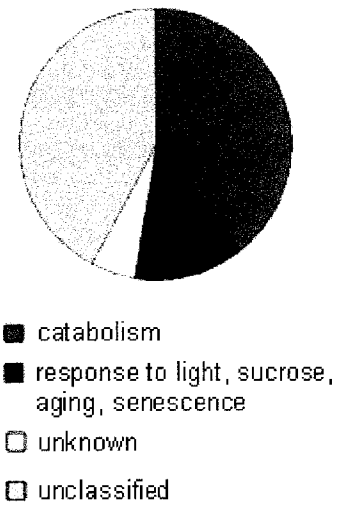


Fig. 3. (**A**) Genes from 22k Arabidopsis ATH1 array that have the highest correlation across 965 chips with the bait genes. Genes used as the bait (at the top of the table) are enoyl reductase (ER), an enzyme from plastidic fatty acid biosynthesis pathway; disproportionating enzyme (DE2), from starch metabolism and methylcrotonyl-CoA carboxylase (MCC-B), active in leucine degradation in mitochondria. Genes that code for proteins predicted to be in the same pathway as the bait gene are highlighted. Potential regulatory genes are in red font. Pearson coefficients (>0.7) were calculated by MetaOmGraph, p-values for enrichment of genes from same pathway as bait are 4.35e-36, 2.02e-15, and 3.1e-23, respectively. (**B**) A higher order catabolic module is revealed from the functional categories of the genes correlated with the leucine catabolism module (Pearson coefficient > 0.5).

is most likely the one required for fatty acid biosynthesis. In addition, these data imply that the functional pathway commonly thought of as "fatty acid synthesis" (acetyl-CoA to fatty acids) should be expanded to begin with pyruvate.

A set of potential regulatory genes whose expression patterns highly correlate with each of the pathway-specific hub-gene was also identified (genes depicted in red font, Fig. 3A). For example, pentatricopeptide repeat-containing protein (encoded by At5g50390) is within the fatty acid biosynthesis module. Because such proteins are thought to participate in modulating the stability of organelle transcripts (*16*), this correlation may indicate a role for this gene in coordinating nuclear-plastidic interactions for regulating fatty acid biosynthesis, via, for example, the control of the plastid-encoded subunit gene of acetyl-CoA carboxylase (*accD*, AtCg00500). Another example is the protein tyrosine phosphatase/kinase gene (At3g52180) that is within the starch metabolic module (Fig. 4A). This gene has recently been shown experimentally to have a regulatory role; knockout alleles of At3g52180 have a starch excess phenotype (*17*). Interestingly, the starch hub-gene shows a high negative correlation with two genes that code for Ras1-like proteins (At5g47200 and At4g17530; Fig. 4B). Consistent with the hypothesis that these genes may have negative regulatory functions in controlling starch metabolism, in humans, RAS1 is a critical component of a phospho-relay pathway important for the negative regulation of many cellular processes (*18*). The leucine catabolism module contains two putative transcriptional regulators, At1g15040 and At5g49450; the former has a sequence-inferred function as a DNA-dependent regulator of transcription, and the latter may code for a bZIP transcription factor. Although these correlations cannot be extrapolated to imply a causative role, identification of candidate regulatory genes delineates hypotheses and suggests experimental approaches for identifying the regulatory function of these genes, as for example has occurred fortuitously for At3g52180.

Further, these transcriptome level analyses provide insights into the hierarchical modularity of Arabidopsis metabolism. This is most evident with the leucine catabolic module. This is a functionally coherent and tightly connected module, nested within a supermodule (Fig. 3B and table S2). The supermodule contains genes that appear to have a

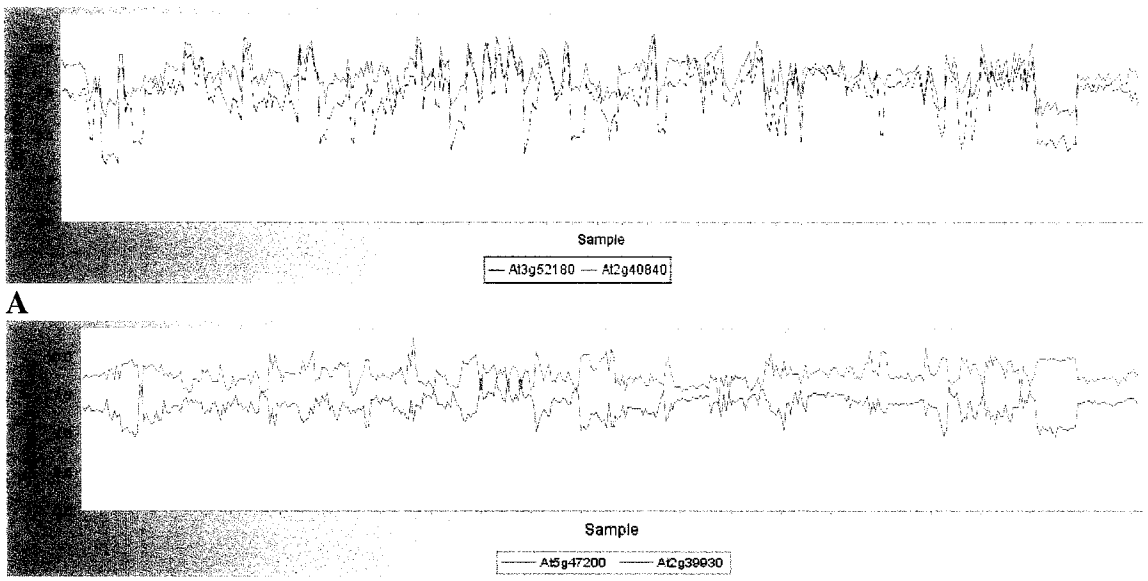


Fig. 4. Correlation of regulatory genes with starch metabolism module. (**A**) Putative positive regulator. green: PTPKIS1, protein tyrosine phosphatase/kinase; blue: DE2, disproportionating enzyme. Correlation = 0.85. (**B**) Putative negative regulator. Red: Ras-related GTP-binding protein, a negative regulator; blue: ISA1, isoamylase, an enzyme in starch catabolism. Correlation = -0.54.

broader but common biological task of maintaining cellular energy balance via catabolism. Specifically, the 28 genes most highly correlated with the leucine catabolic module include genes involved in protein turnover (At1g76410), amino acids catabolism (At3g47340, At1g06570), carbohydrate catabolism (At1g18270, At5g49360, At5g20250), and lipid breakdown (At3g51840). Five genes are annotated as associated with aging, senescence, or respond to light or sucrose stimuli (At3g47340, At5g20250, At2g43400, At4g35770, At4g30270). In plants, which are photoautotrophs, such catabolic processes may be expected to become critical when photosynthesis cannot maintain cellular energy balance, for example during senescence, seed germination, carbon deprivation or autophagy (*11, 19-21*). Consistent with this concept, we find that the genes within this supermodule are co-induced to highest levels of expression under experimental conditions that are expected to limit photosynthetic carbon fixation and induce catabolic processes for maintaining energy balance (oxidative, cold and drought stress, darkness, disruption of plastome-nucleome communication, mutations and illumination conditions that effect the phytochrome response, sucrose starvation, and autophagy (*20, 22*)). Finally, this supermodule contains genes with well-defined or partially defined molecular functions, but undefined physiological functionality (ex. At1g76410, zinc finger protein ; At5g21170, protein kinase; At5g16340,

AMP-binding protein). The inclusion of these genes within this catabolic supermodule may indicate their broad physiological function; a hypothesis that can be experimentally tested by new genetic and biochemical analyses.

This study reveals hitherto unsuspected aspects of metabolic regulation in a complex eukaryotic organism. Each of these core metabolic pathways is structured as a co-expressed module whose transcripts co-accumulate over a wide range of environmental and genetic perturbations and developmental stages. These analyses indicate that modules can be recruited into hierarchical organization (supermodules), hinting at the possibility that supermodules are regulated by common signals that coordinate net metabolic changes within highly interactive networks. The fact that such hierarchical organization is detected from the transcriptome data implies that at least a subset of the higher-order metabolic network may be regulated at the transcript level.

METHODS

Transcriptomic data

The experimental and meta-data was obtained from online microarray depositories: NASCArrays (*30*) and PLEXdb (*31*). The data from 70 experiments comprising of 956 Affymetrix ATH1 microarray slides was normalized to the same range by a scale normalization method described by Yang et al. (*32*). The replicability of experiments was qualitatively assessed on scatter plots and chips with poor biological replicates were discarded. The remaining biological replicates were averaged to yield 424 samples. The normalized data is available online (*33*).

Network of coexpressed genes

The 121 genes used in the initial network construction (Figure 2), including candidates for functions in fatty acid synthesis, starch metabolism and leucine catabolism, are shown in Table S1. A correlation matrix was calculated for these genes. The values of Pearson correlations above 0.3 are statistically highly significant (p-value < 0.00001 after Bonferroni

correction for multiple testing (*34*)). We assessed the validity of correlations higher than 0.6 by performing 10,000 permutations of the corresponding data vectors. Correlations that had a permutation-based p-value<0.0001 (in our case, all pairs of genes examined) were considered for further analysis (fig. S2). The networks in Fig. 2 were constructed by placing an edge between any pair of genes correlated above the threshold.

Enrichment of within-pathway edges in the coexpression network

To test whether there are more correlations between the genes from the same metabolic pathway than could be expected by chance, we compared the coexpression network from experimental data (Fig. 2A) to 10,000 computationally generated random networks. In these random networks the total number of the edges and the number of neighbors for each node was set to be the same as in the original graph. A function for generating random graphs was implemented in R. It constructed graphs by random filling of the adjacency matrix, where nodes' degrees were constrained and self-loops not allowed. Indeed, the proportion of within-pathway edges is significantly higher in our coexpression network compared to such proportions in random graphs (336/421 versus a mean of 152/421, which equals to the difference of ~24 standard deviations, fig. S3).

Correlation with hub genes

The calculation of Pearson correlations between the 3 hub-genes (At2g40840, At4g34030, At2g05990) and all other genes on the chip (Fig. 3A) was conducted in our publicly available software, MetaOmGraph (*4*). MetaOmGraph is able to handle large datasets with an efficient use of memory and is integrated with other publicly available bioinformatics tools in MetNet Exchange suit (http://metnet.vrac.iastate.edu/). The p-values for overrepresentation of pathway genes in gene lists in Fig.3 was calculated by hypergeometric distribution (*35*).

The genes belonging to catabolic supermodule (Fig. 3B and table S2) were identified by intersection of lists of all genes that are correlated above 0.5 threshold with each of eight genes from leucine catabolism module.

Software

The computations (normalization, correlation matrix, permutation-based p-values, random graphs generation, hypergeometric distribution) were conducted in R software (*36*). The RNA profiles were plotted in MetaOmGraph (*33*). The network of metabolic genes in Fig. 2 was visualized using GraphExplore (*37*). The pathway data is from MetNetDB database. The R code is available upon request.

ACKNOWLEDGEMENTS

We thank D. Cook for assistance with normalizing the microarray data and the MetNet group and D. Nettleton for advice and stimulating discussions. Supported by National Science Foundation grants 0416730, DBI-0520267 and DBI-0209789.

REFERENCES

1. R. Balakrishnan *et al.*, "Saccharomyces Genome Database" http://www.yeastgenome.org/ (February 1st 2006).
2. P. M. Magwene, J. Kim, *Genome Biol.* **5(12)**:R100. (2004).
3. E. Segal, R. Yelensky, D. Koller. *Bioinformatics* **19,** Suppl 1:i273 (2003).
4. E. Segal *et al.*, *Nat Genet.* **34,** 166 (2003).
5. S. Kim *et al.*, *Science* **293**, 2087 (2001).
6. D. Baldessari *et al.*, *Mech Dev.* **122**, 441 (2005).
7. W. Zhang *et al.*, *J Biol.* **3**, 21 (2004).
8. TAIR6 Genome Release, (Nov 11, 2005), http://www.arabidopsis.org/
9. G. Blanc, K.H. Wolfe, *Plant Cell.* **16,** 1679 (2004).
10. K. L. Adams, J.F. Wendel, *Curr Opin Plant Biol.* **8**, 135 (2005).
11. M. D. Anderson *et al*, *Plant Physiol.* **118,** 1127 (1998).
12. Except alpha-glucosidase, which recent experimental evidence indicates may not be part of starch metabolism (*23*).
13. At2g05990 encodes the enoyl-ACP reductase (ER) that is uniquely required in *de novo* fatty acid biosynthesis (*24*); At2g40840 is a starch metabolism hub gene coding for disproportionating enzyme (DE2) (*25*), and the leucine catabolism hub-gene (At4g34030) encodes the β subunit of methylcrotonyl-CoA carboxylase (MCCase) (*11, 19*).
14. W. C. Plaxton, C. R. Smith, V. L. Knowles. *Arch Biochem Biophys.* **400**, 54 (2002).
15. The Arabidopsis Genome Initiative. *Nature* **408**, 796 (2000).
16. T. Nakamura, G. Schuster, M. Sugiura, M. Sugita, *Biochem. Soc. Trans.* **32**, 571 (2004).
17. T. Niittyla *et al.*, *J Biol Chem.* **281**, 11815 (2006).

18. M. Symons, Y. Takai. *Sci STKE.* **2001**, PE1 (2001). http://stke.sciencemag.org/cgi/content/full/OC_sigtrans;2001/68/pe1
19. A. L. McKean *et al., J Biol Chem.* **275**, 5582 (2000).
20. A. L. Contento, S. J. Kim, D. C. Bassham, *Plant Physiol.* **135**, 2330 (2004).
21. P. Che, E. S. Wurtele, B. J. Nikolau, *Plant Physiol.* **129**, 625 (2002).
22. Experiments from NASCArrays (NASCArrays Experiment Reference Numbers): NASCARRAYS-28, NASCARRAYS-89, NASCARRAYS-14, NASCARRAYS-40, NASCARRAYS-124 (http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl).
23. M. A. Taylor *et al., Plant J.* **24**, 305 (2000).
24. Z. Mou, Y. He, Y. Dai, X. Liu, J. Li, *Plant Cell.* **12**, 405 (2000).
25. Y. Lu, T. D. Sharkey, *Planta.* **218**, 466 (2004).
26. D. J. Craigon *et al., Nucleic Acids Res.* **32**: D575-577 (2004).URL: http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl .
27. L. Shen *et al., Nucleic Acids Res.* **33**, Database issue D614-D618 (2005). http://www.plexdb.org/.
28. Y.H. Yang *et al., Nucleic Acids Res.* **30**, e15 (2002).
29. E.S. Wurtele *et al., Comp Funct Genomics.* **4**, 239 (2003). http://www.metnetdb.org/MetNet_MetaOmGraph.htm
30. C.E. Bonferroni, *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3 (1936).
31. R. J. Cho *et al. Nat. Genet.*, **27**, 48 (2001)
32. R Development Core Team.. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 (2004), http://www.R-project.org.
33. Q. Wang, G. Yao, J. Nevins, M. West and A. Dobra, submitted for publication, http://graphexplore.cgt.duke.edu.

Supporting Online Material
Figs. S1 to S4
Tables S1 and S2

APPENDIX. SUPPORTING ONLINE MATERIAL

Table S1. Genes used in the study

#	Name[a]	Locus ID	Function[b]
FATTY ACID BIOSYNTHESIS			
1	ACP1	At3g05020	acyl carrier protein
2	ACP2	At4g25050	acyl carrier protein
3	ACP3	At1g54630	acyl carrier protein
4	ACP4	At5g27200	acyl carrier protein
5	ACPS	At2g02770	holo-acyl carrier protein synthase
6	ACS1	At4g14070	acetyl-CoA synthase
7	ACS2	At3g23790	acetyl-CoA synthase
8	ACS3	At1g77590	acetyl-CoA synthase
9	BC	At5g35360	biotin carboxylase subunit of acetyl-CoA carboxylase
10	BCCP1	At5g16390	biotin carboxyl carrier subunit of acetyl-CoA carboxylase
11	BCCP2	At5g15530	biotin carboxyl carrier subunit of acetyl-CoA carboxylase
12	BCCP3	At1g52670	biotin carboxyl carrier subunit of acetyl-CoA carboxylase
13	BCCP4	At3g15690	biotin carboxyl carrier subunit of acetyl-CoA carboxylase
14	BCCP5	At3g56130	biotin carboxyl carrier subunit of acetyl-CoA carboxylase
15	CTalpha	At2g38040	carboxyl transferase alpha subunit of acetyl-CoA carboxylase
16	CTbeta	AtCg00500	carboxyl transferase beta subunit of acetyl-CoA carboxylase
17	PD_E1alpha1	At1g01090	pyruvate dehydrogenase E1 alpha subunit
18	PD_E1beta1	At2g34590	pyruvate dehydrogenase E1 beta subunit
19	PD_E1beta2	At1g30120	pyruvate dehydrogenase E1 beta subunit
20	ER	At2g05990	enoyl reductase
21	Fab1	At5g16230	fatty acid desaturase
22	Fab2	At2g43710	fatty acid desaturase
23	Fab3	At5g16240	fatty acid desaturase
24	Fab4	At3g02610	fatty acid desaturase
25	Fab5	At3g02630	fatty acid desaturase
26	Fab6	At1g43800	fatty acid desaturase
27	FatA1	At4g13050	thioesterase
28	FatA2	At3g25110	thioesterase
29	FatB	At1g08510	thioesterase
30	HD1	At5g10160	hydroxyacyl dehydrogenase
31	HD2	At2g22230	hydroxyacyl dehydrogenase
32	KAR1	At3g55310	alpha-ketoacyl reductase
33	KAR2	At3g46170	alpha-ketoacyl reductase
34	KAR3	At1g24360	alpha-ketoacyl reductase
35	KAR4	At1g62610	alpha-ketoacyl reductase
36	KASI	At5g46290	alpha-ketoacyl synthase
37	KASII	At1g74960	alpha-ketoacyl synthase
38	KASIII1	At1g62640	alpha-ketoacyl synthase
39	KASIII2	At2g26640	alpha-ketoacyl synthase
40	MAT	At2g30200	malonyl-acetyl-CoA transferase
41	PD_E2_1	At3g25860	pyruvate dehydrogenase acetyltransferase subunit

Table S1. Genes used in the study (continued).

#	Name[a]	Locus ID	Function[b]
42	PD_E2_2	At1g34430	pyruvate dehydrogenase acetyltransferase subunit
43	PD_E3_1	At4g16155	pyruvate dehydrogenase, lipoamide dehydrogenase subunit
44	PD_E3_2	At3g16950	pyruvate dehydrogenase, lipoamide dehydrogenase subunit
STARCH METABOLISM			
45	AAM1	At1g76130	alpha-amylase
46	AAM2	At4g25000	alpha-amylase
47	AAM3	At1g69830	alpha-amylase
48	AGL1	At3g23640	alpha-glucosidase
49	AGL2	At5g63840	alpha-glucosidase
50	AGL3	At3g45940	alpha-glucosidase
51	AGL4	At5g11720	alpha-glucosidase
52	AGL5	At1g68560	alpha-glucosidase
53	AGL-like1	At4g26620	alpha-glucosidase
54	AGL-like2	At1g67490	alpha-glucosidase
55	AGL-like3	At1g24320	alpha-glucosidase
56	APL1	At5g19220	ADP-glucose pyrophosphorylase
57	APL2	At1g27680	ADP-glucose pyrophosphorylase
58	APL3	At2g21590	ADP-glucose pyrophosphorylase
59	APL4	At4g39210	ADP-glucose pyrophosphorylase
60	APS1	At5g48300	ADP-glucose pyrophosphorylase
61	APS2	At1g05610	ADP-glucose pyrophosphorylase
62	BAM1	At4g15210	beta-amylase
63	BAM2	At2g32290	beta-amylase
64	BAM3	At3g23920	beta-amylase
65	BAM4	At2g45880	beta-amylase
66	BAM5	At5g45300	beta-amylase
67	BAM6	At5g55700	beta-amylase
68	BAM7	At4g00490	beta-amylase
69	BAM8	At4g17090	beta-amylase
70	BAM9	At5g18670	beta-amylase
71	BE1	At3g20440	starch branching enzyme
72	BE2	At5g03650	starch branching enzyme
73	BE3	At2g36390	starch branching enzyme
74	DE1	At5g64860	disproportionating enzyme
75	DE2	At2g40840	disproportionating enzyme
76	GBSS	At1g32900	starch synthase
77	GWD1	At1g10760	glucan water dikinase
78	GWD2	At5g26570	glucan water dikinase
79	GWD3	At4g24450	glucan water dikinase
80	ISA1	At2g39930	isoamylase DB
81	ISA2	At1g03310	isoamylase DB
82	ISA3	At4g09020	isoamylase DB
83	PGI1	At4g24620	phosphoglucose isomerase
84	PGI-like	At5g42740	phosphoglucose isomerase

Table S1. Genes used in the study (continued).

#	Name[a]	Locus ID	Function[b]
85	PGM1	At5g51820	phosphoglucomutase
86	PGM-like1	At5g17530	phosphoglucomutase
87	PGM-like2	At4g11570	phosphoglucomutase
88	PGM-like3	At1g70730	phosphoglucomutase
89	PGM-like4	At1g70820	phosphoglucomutase
90	PGM-like5	At1g23190	phosphoglucomutase
91	PHO1	At3g29320	starch phosphorylase
92	PHO2	At3g46970	starch phosphorylase
93	PU1	At5g04360	pullulanase
94	SS1	At5g24300	starch synthase
95	SS2	At3g01180	starch synthase
96	SS3	At1g11720	starch synthase
97	SS4	At4g18240	starch synthase
98	SS6	At5g65685	starch synthase
LEUCINE CATABOLISM			
99	BCAT1	At1g10060	branched-chain amino acid aminotransferase
100	BCAT2	At1g10070	branched-chain amino acid aminotransferase
101	BCAT5	At5g65780	branched-chain amino acid aminotransferase
102	BCKDH_E1a	At5g09300	branched-chain alpha-keto acid dehydrogenase E1 alpha subunit
103	BCKDH_E1a	At1g21400	branched-chain alpha-keto acid dehydrogenase E1 alpha subunit
104	BCKDH_E1b	At3g13450	branched-chain alpha-keto acid dehydrogenase E1 beta subunit
105	BCKDH_E1b	At1g55510	branched-chain alpha-keto acid dehydrogenase E1 beta subunit
106	BCKDH_E2	At3g06850	branched chain alpha-ketoacid dehydrogenase, E2 subunit
107	BCKDH_E3	At3g17240	branched-chain alpha-keto acid dehydrogenase E3 subunit
108	IVD	At3g45300	isovaleryl-CoA dehydrogenase
109	MCC-A	At1g03090	methylcrotonyl-CoA carboxylase subunit A
110	MCC-B	At4g34030	methylcrotonyl-CoA carboxylase subunit B
TRANSPORT AND COFACTOR SYNTHESIS			
111	ABC	At1g54350	acyl transporter
112	BIO1	At5g57590	DAPA synthase
113	bioB	At2g43360	biotin synthase
114	bioH1	At4g36530	pimelic acid synthase
115	bioH2	At4g24160	pimelic acid synthase
116	CACT	At5g46800	mitochondrial transporter
117	GLT1	At5g16150	glucose transporter
118	KAPAS	At5g04620	KAPA synthase
119	lipS	At5g08415	lipoate synthase
120	lipT	At4g31050	lipoate transferase
121	TPT1	At5g46110	triose phosphate translocator

[a] gene name used for this study

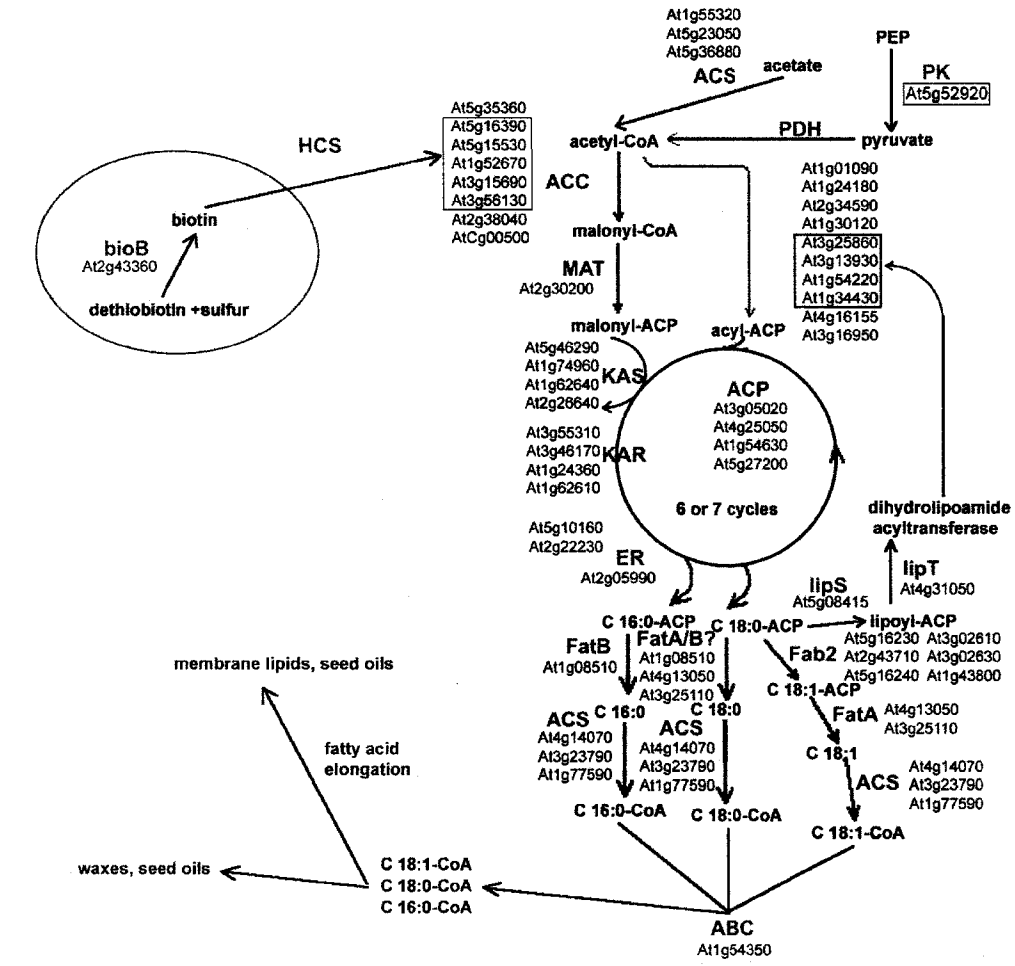
[b] MetnetDB (*29*)

Table S2. Genes in the supermodule containing the leucine catabolism pathway, shown in Fig. 3B, identified as intersection of lists of all genes that are correlated above 0.5 threshold with each of eight genes from leucine catabolism module

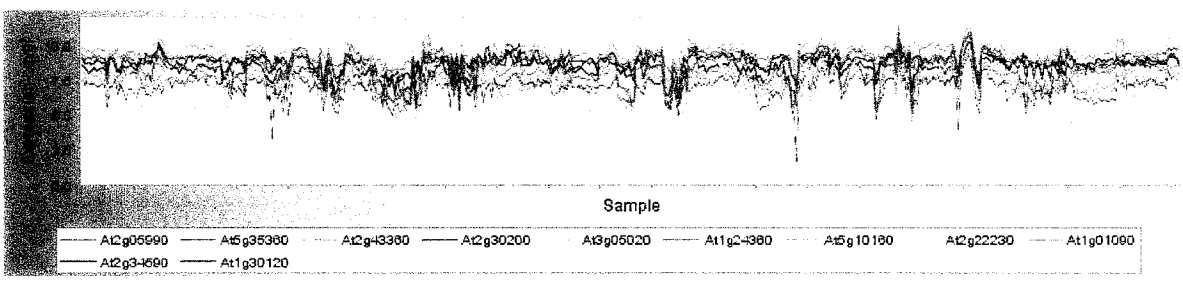
Locus ID	Annotation
At1g03090	3-methylcrotonyl-CoA carboxylase 1 (MCCA)
At1g06570	4-hydroxyphenylpyruvate dioxygenase (HPD)
At1g08630	L-allo-threonine aldolase-related
At1g10070	branched-chain amino acid aminotransferase 2 / branched-chain amino acid transaminase 2 (BCAT2)
At1g15040	expressed protein
At1g15040	glutamine amidotransferase-related
At1g18270	ketose-bisphosphate aldolase class-II family protein
At1g21400	2-oxoisovalerate dehydrogenase, putative / 3-methyl-2-oxobutanoate dehydrogenase, putative / branched-chain alpha-keto acid dehydrogenase E1 alpha subunit, putative
At1g28260	expressed protein
At1g55510	2-oxoisovalerate dehydrogenase, putative / 3-methyl-2-oxobutanoate dehydrogenase, putative / branched-chain alpha-keto acid dehydrogenase E1 beta subunit, putative
At1g55810	uracil phosphoribosyltransferase, putative
At1g58180	carbonic anhydrase family protein
At1g76410	zinc finger (C3HC4-type RING finger) family protein
At1g79690	MutT/nudix family protein
At1g79700	ovule development protein, putative, similar to ovule development protein AINTEGUMENTA
At2g14170	methylmalonate-semialdehyde dehydrogenase; putative
At2g18700	Encodes an enzyme putatively involved in trehalose biosynthesis.
At2g40420	amino acid transporter family protein
At2g43400	Encodes a unique electron-transfer flavoprotein:ubiquinone oxidoreductase that is localized to the mitochondrion.
At3g06850	branched chain alpha-keto acid dehydrogenase E2 subunit (din3)
At3g13450	branched-chain alpha-keto acid dehydrogenase E1 beta subunit (DIN4)
At3g45300	isovaleryl-CoA-dehydrogenase (IVD)
At3g47340	asparagine synthetase 1 (glutamine-hydrolyzing)
At3g51840	short-chain acyl-CoA oxidase
At4g28040	nodulin MtN21 family protein
At4g30270	similar to endo xyloglucan transferase
At4g34030	3-methylcrotonyl-CoA carboxylase 2 (MCCB)

Table S2. Genes in the supermodule containing the leucine catabolism pathway (continued).

Locus ID	Annotation
At4g35770	senescence-associated gene
At4g38470	protein kinase family protein
At5g16340	AMP-binding protein, putative
At5g20250	member of glycosyl hydrolase family
At5g21170	5'-AMP-activated protein kinase beta-2 subunit
At5g41080	glycerophosphoryl diester phosphodiesterase family protein
At5g49360	beta-xylosidase located in the extracellular matrix.
At5g49450	bZIP family transcription factor
At5g63620	oxidoreductase, zinc-binding dehydrogenase family protein, contains PFAM zinc-binding dehydrogenase domain PF00107



A



B

Fig. S1.(**A**) Fatty acid biosynthesis in chloroplast (sequence of reactions depicted by green arrows). Locus IDs (ex. At1g77590) of candidate genes for enzymatic function (enzyme names in blue) are shown; genes whose Locus ID is in red are coexpressed. These genes form connected network in Fig. 2B and encode all enzymes required for the biosynthesis of 18-carbon fatty acids from pyruvate, as well as genes for biotin and lipoic acid biosynthesis, obligate cofactors for the first two reactions in the pathway. Gene for pyruvate kinase (on yellow background), required for production of pyruvate substrate is also coexpressed with genes from the fatty acid biosynthesis pathway. (**B**) Expression profiles of the genes from fatty acid biosynthesis module across 424 samples.

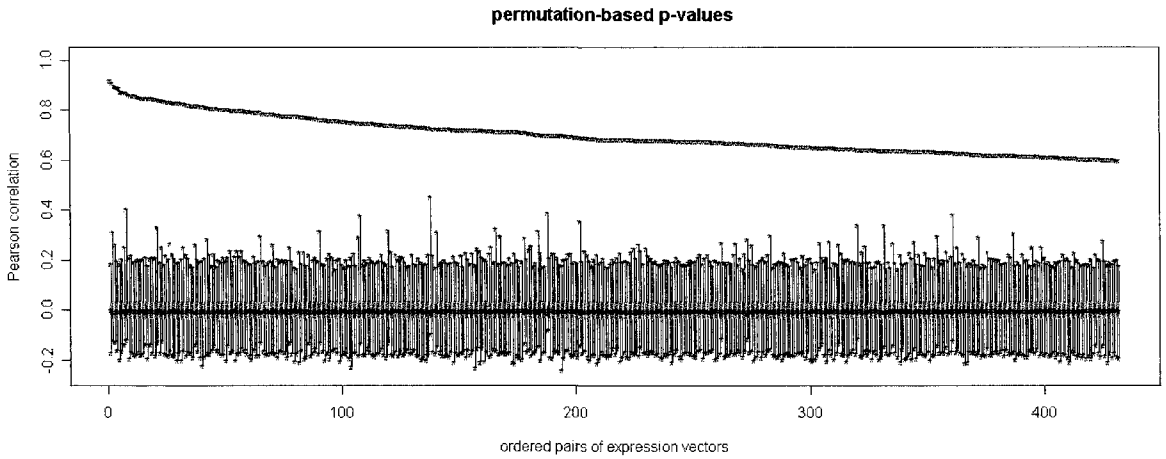


Fig. S2. Permutation-based support for the correlations in the coexpression network of 121 genes in Fig. 2B. For each pair of genes with Pearson correlation above 0.6 this real correlation value (red stars) is compared to the distribution of correlation values obtained in 10,000 permutations of the corresponding expression data vectors (green star, maximum; black star, mean; blue star, minimum, orange bar, values between upper and lower hinges).

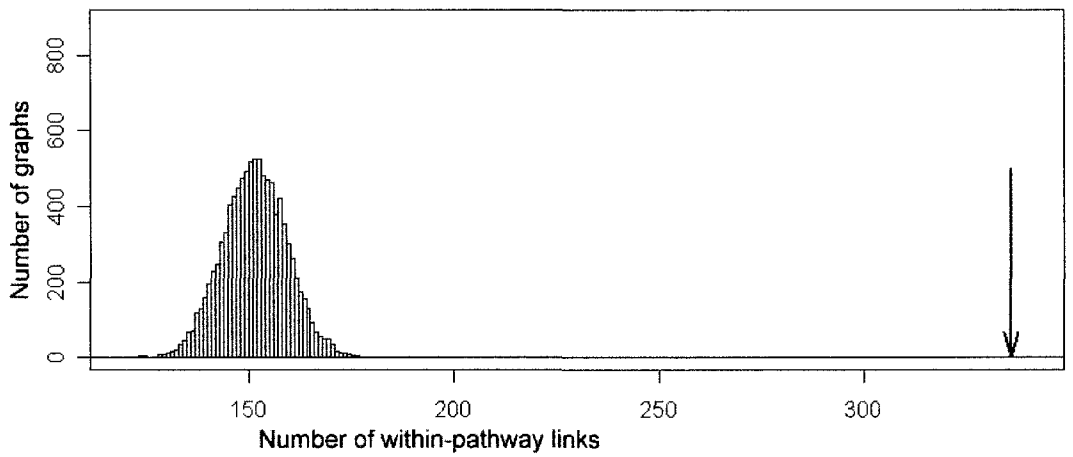


Fig. S3. Enrichment of edges joining genes from the same pathway in the graph in Fig. 2B (red arrow), as compared to random graphs (blue histogram). Histogram shows distribution of numbers of links joining genes from the same pathway in each of 10,000 random graphs with the same links structure as original graph (mean =152; σ=7.8). Red arrow denotes number of within-pathway links (336) based on expression data (from Fig. 2B graph). Total number of links in each graph, regardless of pathway, is 421.

CHAPTER 4. REGULON ORGANIZATION OF ARABIDOPSIS

A paper to be submitted to *The Plant Cell*

Wiesława I. Mentzen and Eve Syrkin Wurtele

ABSTRACT

We identify sets of coexpressed genes in the Arabidopsis genome using graph clustering method. These sets are based on similar expression signature across 926 microarrays that include a variety of environmental and developmental conditions. The 71 largest clusters, comprising totally 9474 genes, correspond to identifiable biological processes, protein complexes, or represent organellar regulation. Photosynthesis, reproduction, and defense response account for processes represented in many major clusters, while specific metabolic processes are less prevalent. By assigning genes to functionally coherent clusters, this analysis extends and delivers support for existing knowledge, and generates new hypotheses about roles of these genes in a broad biological context.

INTRODUCTION

Despite the model plant Arabidopsis genome having been fully sequenced since 2000 the function of many of the 29,000 genes is experimentally undetermined (TIGR). About 9,000 of the genes cannot be ascribed any function, and about 18,000 additional genes have a recognizable domain that represents a general molecular or biochemical function (phosphorylation, glycosylation) but no clear physiological function, such as the time, place and nature of their involvement in cellular processes. Arabidopsis transcriptomic data is now available across a wide range of genetic, environmental and developmental conditions, including mutants, stress, tissue, age, etc. With this wealth of data it is tempting to identify

genes that share common expression signature across a variety of experiments, and examine the composition of such groups of coexpressed genes. Meta-analysis of transcriptome offers a potential for identifying most prevailing cellular processes, associate genes with particular biological processes, and, in particular, assign otherwise unknown genes to the process they are correlated with. Such analyses, sometimes combining also other kinds of data – proteomics, co-precipitation, literature, yeast two hybrid – proved to be valuable for other model organisms, including bacteria, fly, nematode, human and, most of all, yeast. The use of expression data allowed for identification of functionally coherent modules corresponding to major cellular processes in yeast (Stuart et al., 2003; Segal et al., 2003; Magwene and Kim, 2004) and these modules might be important enough to be conserved across the eukaryotic organisms (Stuart et al., 2003). Meta-analysis of Arabidopsis transcriptome receives growing attention and several online repositories for microarray data have been created (NASCArrays, Genevestigator, PLEXdb, etc), even though this organism presents challenges associated with high volume, distribution and incompleteness of biological data. Here, we present a global analysis of coexpression in Arabidopsis genome.

RESULTS

Coexpressed modules of genes

Genes that share a similar expression profile across multiple spatial, temporal, environmental and genetic conditions are likely to be under common transcriptional regulation. Meta-analysis of many experiments facilitates identification of the clusters of coexpressed genes that reflect the most prevailing processes in this plant.

The transcriptomics data for 70 experiments was obtained from public microarray depositories: NASCArrays (http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl; Craigon et el., 2004) and PLEXdb (http://www.plexdb.org/; Shen et al., 2005). The experiments from these databases include a wide variety of environmental, stress, growth, development and mutant conditions. After assessment of the replicate quality, chips with poor replicate quality were removed. The resulting 963 chips were normalized and the

replicates were averaged to yield 424 samples (Methods). Genes with low expression and not highly correlated with others were filtered out (Figure 1).

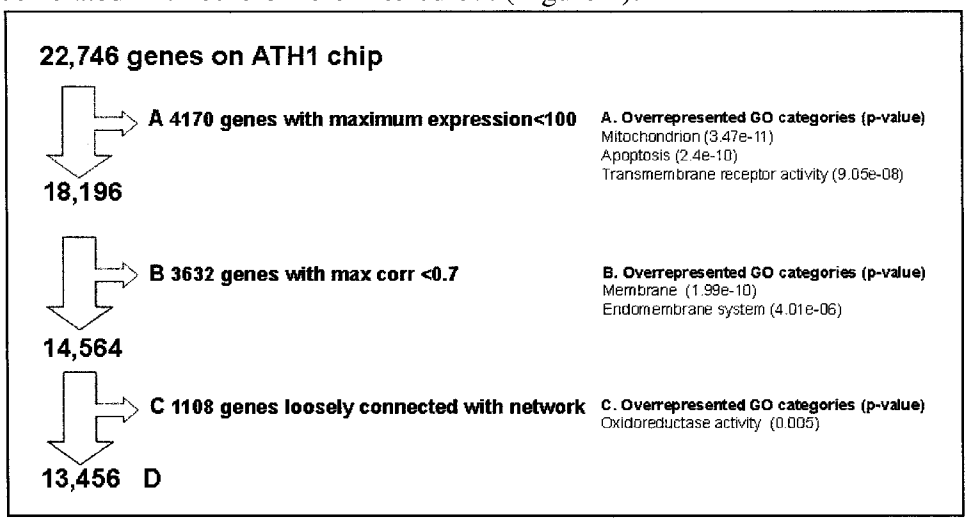


Figure 1. Data processing for the construction of the transcriptional network. Filters were applied to original 22,746 genes on ATH1 chip to remove the genes with low expression (A) and correlation to other genes lower than the threshold of 0.7 (B). The network was then constructed and only biggest connected component of this network was retained; small connected components as well as genes with only one neighbor in the giant connected component were removed (C). The resulting network with 13,456 genes (D) was clustered in subsequent step. Enrichment of GO terms in three groups of removed genes (A-C) is shown.

The expression data for resulting 13,456 genes was viewed as a coexpression network, where genes are represented as nodes, and two nodes are connected by an edge if the correlation between their expression profiles is higher than a threshold of 0.7 (see Methods section for details). In this model, groups of coexpressed genes are represented as the densely-connected regions of the coexpression network. We used the graph clustering method based on flow simulation to identify clusters in this network that correspond to coexpressed genes. This method, developed by Stijn van Dongen (van Dongen, 2000) has been used previously for clustering protein sequence data (Enright et al, 2002) and identifying modules in yeast protein interaction network (Pereira-Leal et al, 2004). The 998 resulting clusters ranged in size from 1 to 1623 genes. We focus our current analysis of cluster functional coherence on the 71 largest clusters, comprising collectively 9474 genes.

The functionalities of the clusters are summarized in Table 1. We used a combination of the automatic analysis of enrichment of GO terms and manual inspection to assign a function to each cluster. Most of the clusters were characterized by a mixture of molecular

Table 1. Summary of clusters' function.

cluster	# of genes	function[a]	FC[b]
1	1629	pollen-specific	ND
2	971	photosynthesis	29
3	869	protein synthesis	65
4	495	cell division	49
5	507	mostly root, mostly membrane proteins	77
6	417	embryonic development	25
7	305	developmental regulation (expressed in leaf apex)	38
8	281	pollen-specific	64
9	234	response to environmental stimula	26
10	223	protein modification, defense response	66
11	215	nuclear, others with very low expression	ND
12	182	development	56
13	154	upregulated in 'response to CO_2 levels' experiment	ND
14	140	regulation of organ development	61
15	138	chloroplast metabolic processes (probably circadian regulation)	56
16	121	information	58
17	115	information	51
18	96	pollen metabolism, transport	52
19	94	regulation of flowering, metabolism, lipid metabolism	61
20	94	information	80
21	92	lipid transport, metabolism, reproduction in flowers	54
22	81	cell wall biosynthesis, carbohydrate metabolism	32
23	77	membrane proteins	69
24	71	defense response	70
25	70	defense response	77
26	68	information	79
27	68	root development	73
28	66	nucleic acid binding, regulation	60
29	63	aerobic respiration in mitochondria	92
30	56	signaling	89
31	52	defense response	25
32	48	nuclear genes, RNA processing, DNA replication	70
33	48	chloroplast organization and biogenesis	62
34	47	mitochondrial genes	96
35	45	kinases, signaling, disease resistance	69
36	43	lipid metabolism in flowers	43
37	42	heat shock response	60
38	40	RNA processing, translation, transcription regulation	82
39	40	catabolic processes deriving energy	51
40	40	transcription, translation, protein folding and transport	86
41	36	regulation, information	83
42	34	regulation	78
43	33	flower/fruit development	44
44	31	metabolic processes in flowers/fruit	22

Table 1. Summary of clusters' function (continued).

cluster	# of genes	function[a]	FC[b]
45	30	proteasome complex	87
46	29	defense response	50
47	29	nuclear, replication, chromosome organization, cell cycle	67
48	28	cell culture and tumor specific	ND
49	27	chloroplast – encoded	100
50	27	signaling	90
51	26	regulation of development, hormone-mediated	74
52	26	endoplasmic reticulum: protein folding and secretion	73
53	25	fatty acid biosynthesis	83
54	23	mixed	ND
55	23	very-long-chain fatty acid metabolism	43
56	22	mixed	ND
57	22	mixed	ND
58	22	transposases, mostly CACTA-type	100
59	21	metabolism of lipids	71
60	21	ubiquitin ligase	53
61	20	catabolism	56
62	20	information, nuclear	78
63	20	stress-induced catabolism, mediated by jasmonic acid	72
64	20	information, nuclear	87
65	20	metabolism	44
66	20	mixed	ND
67	20	signaling, response to pathogen	60
68	20	information	87
69	20	Leucine/glucosinolates synthesis	45
70	19	fruit development	39
71	19	development	31

[a] additional information: **cluster 2**, photosynthesis, chloroplast organization and biogenesis, photorespiration, Calvin cycle; **3**, RNA metabolism, processing, ribosome constituent, protein folding or transport, ribosome genesis, translation, tRNA charging; **5**, (49% membrane proteins) membrane, transport, cell wall modification, biosynthesis, regulation, response; **6**, seed, embryogenesis-related, lipid sequestering, degradation, transport; **7**, regulation (transcription, progression through cell cycle, development), response, pentatricopeptide protein, biosynthesis; **8**, transcription, translation, regulation, response, F-box proteins, transport, carbohydrate metabolism or transport; **9**, response, signaling, regulation of transcription; **10**, response (defense, pathogen, stress, hormone), regulation of transcription, protein modification; **12**, regulation (transcription, development, hormone synthesis), maturation, biosynthesis; **14**, transcription factor, signaling, hormone-responsive, nucleic acid metabolism, organ development; **15**, chloroplast and regulation (56%), probably circadian regulated (6 genes circadian) but also peroxisomal and mitochondrial proteins; **16**, regulation, nucleic acid metabolism, protein synthesis; **17**, regulation of transcription, cyclin, nucleic acid binding, replication, translation, peptidase, protein ubiquitination (not including hypothetical, pseudogenes, unknown and transposons which constitute 39% of all genes in this cluster); **18**, lipid, carbohydrate (metabolism, transport), other transport, signaling; **19**, biosynthesis, TF, auxin-responsive, sexual reproduction, pectinesterase inhibitor, nodulin, protein kinase; **20**, transcription regulation, chromatin assembly and rearrangement, cell cycle, DNA metabolism and repair, RNA processing, helicase, protein import into nucleus, development regulation; **21**, biosynthesis, development, sexual reproduction, lipid (metabolism, transport, sequestering); **24**, defense response, stress response, ABA metabolism and response, signaling, protein kinase, transcription factor; **26**, nucleic acid binding, chromatin structure, transcription factor, meiosis, RNA processing, cell fate, nuclease, signal transduction; **27**,

transcription factor, root, transport, signaling; **33**, including transcription and transcript processing in plastid; **37**, FC 76% if category widened to "response to stress"; **41**, regulatory, transcription factor, nucleic acid binding, protein binding, folding; **42**, nucleic acid binding, chromatin rearrangement, signaling, ubiquitin ligase; **43**, lipid sequestering or metabolism, development; **44**, lipid metabolism, proanthocyanidin synthesis; **46**, response to pathogens, stress, toxins, heavy metals, synthesis of protective compounds ex. glutathione, indol-glucosinolates; **51**, development, transcription factor, hormone signaling, microtubule movement, light-regulated; **55**, VLCFA, wax, cuticle biosynthesis; **59**, terpenoid biosynthesis, sterol biosynthesis, acyltransferase, lipid transfer protein, MD-2-related lipid recognition domain, cytochrome P450; **62**, nuclear, transcription regulation, nucleic acid binding and metabolism, replication, protein folding, kinase; **63**, jasmonic acid biosynthesis or response, defense response, catabolism; **64**, chromosome organization, transcription regulation, translation initiation, transcription, nucleic acid binding, RNA processing, cell cycle; **68**, regulation, signaling, nucleic acid binding; **71**, development, response to hormone

[b] Functional Coherence, calculated as percentage of annotated genes (i.e. other than "expressed" or "hypothetical") whose annotation is consistent with the cluster functional classification

ND – not determined; cluster description is not based on annotation of constituent genes

functions: transporters, transcription factors, signaling molecules, kinases and metabolic genes that work together to achieve common goal. This goal could be for example hormone-mediated development of floral organ, accompanied by metabolic processes, or stress response, leading to rearrangement of chromatin organization. Few clusters contained single gene family or protein complex, for example proteasome complex in cluster 45. Two clusters contained almost exclusively genes from organellar genomes: one mitochondrial and one plastidic. One cluster contained mostly transporters. Even though the genes with low expression were filtered out, three clusters were identified that contained many hypothetical genes, transposons or pseudogenes: cluster nr 11, cluster nr 17 and cluster nr 58 with transposons, mainly of CACTA type, induced by freezing stress. Genes from mixture of categories, expressed in just one condition, are joined in cluster 13. One hundred and twenty six genes were not recruited into any cluster. Many of those are involved in modulation of enzyme activity (protein kinases, protein ubiquitination), response and signal transduction.

The simplified view of the coexpression network formed by genes in these 71 clusters is shown in Figure 2. The link between two clusters means that there are genes in one cluster that are correlated with genes in second cluster. It is interesting to note the grouping of the clusters with similar general functions. This effect is especially pronounced for information-related and environmental response functions.

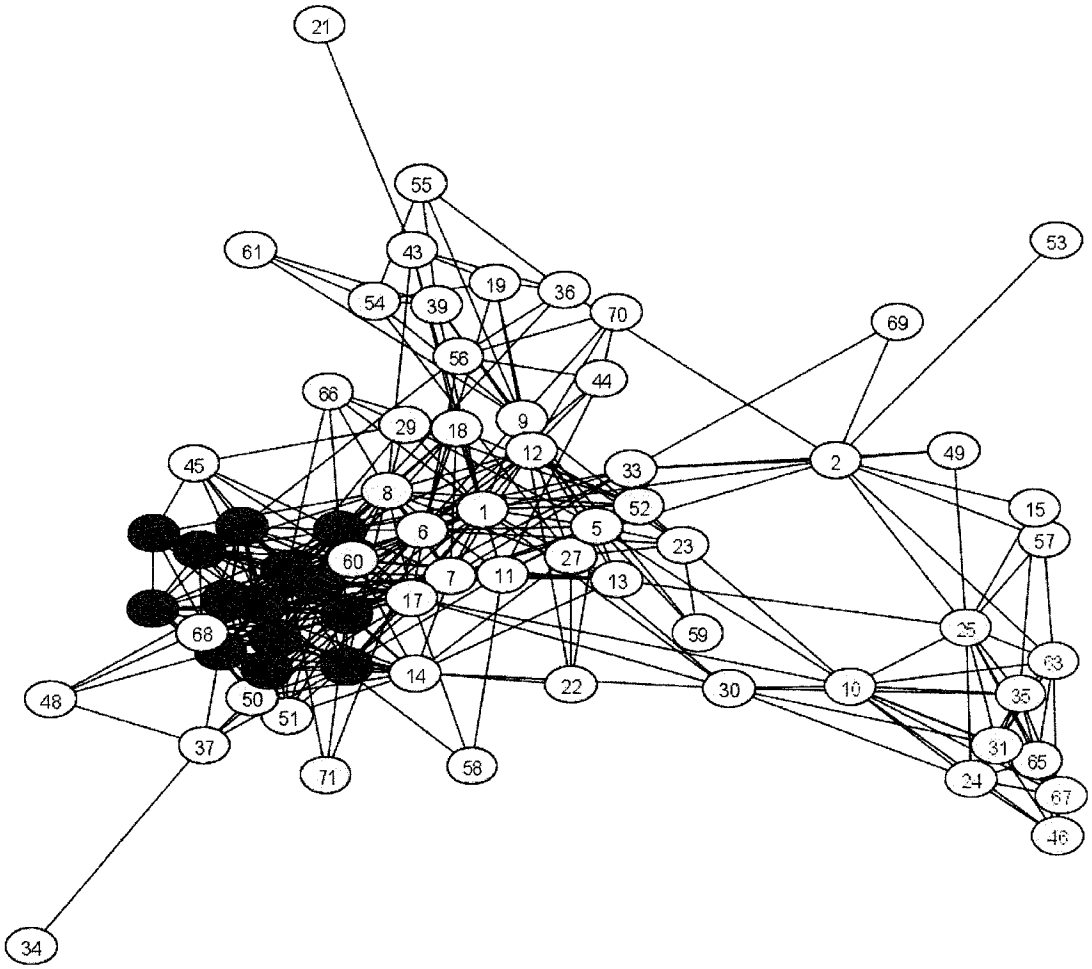
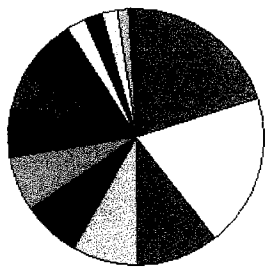


Figure 2. Fragment of coexpression network containing 9,474 genes in 71 largest clusters (represented by gray ovals numbered 1 through 71, as in Table 1.). The link between two clusters means that there are genes in one cluster that are correlated with genes in second cluster. The grouping of information-related clusters (purple) and defense response-related clusters (yellow) may be observed.

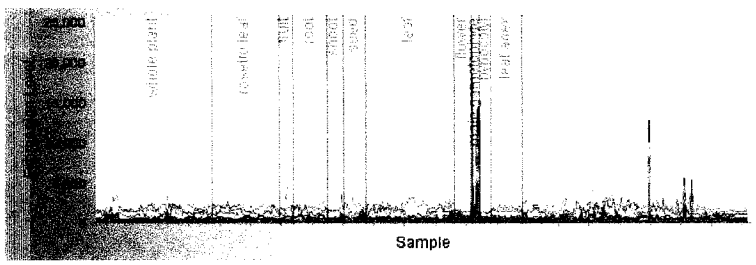
Examples of clusters

Tissue-specific genes form several clusters. One of those is cluster 1, the biggest cluster, composed of 1623 genes, expressed mainly in pollen (Figure 3A). It appears to constitute of genes involved in the regulation of the pollen development, spermatogenesis and pollen tube growth. It is a very dense cluster, containing both the genes with the highest correlation to another gene and genes with highest number of neighbors (genes they are correlated with). The proportion of nuclear regulatory proteins and genes involved in morphogenesis and pollen-specific function is high. The cluster includes 72 protein kinases and 51 transcription

A

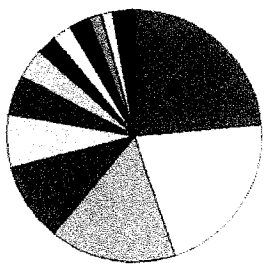


- others
- unknown
- transport
- biosynthesis
- nucleic acid binding/metabolism
- protein kinase/phosphatase
- catabolism
- signaling
- response
- transcription factor
- regulation
- pollen
- development
- tubule, cytoskeleton
- seed, embryo

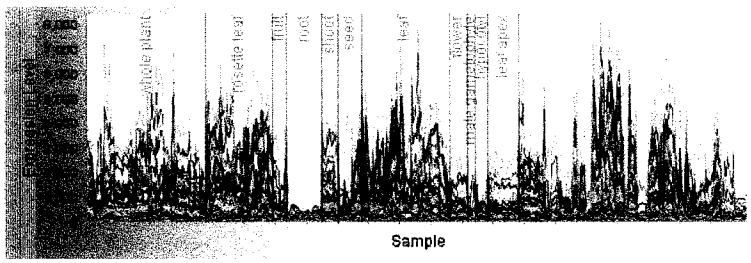


Overrepresented GO categories:	p-value
transcription factor	9.26e-07
carrier	4.36e-06
cytoskeleton	2.7e-06
nucletide phosphorylation	6.01e-05
cytoskeleton organization and biogenesis	0.00299
morphogenesis	0.0436
protein aminoacid phosphorylation	0.0782
regulation of phosphorylation	0.0827

B



- others
- unknown
- photosynthesis
- various metabolic
- protein biosynthesis
- transport
- PPR, TPR
- transcription factor
- other regulation
- protein folding
- chloroplast organization and biogenesis
- photorespiration
- glycolysis
- calvin cycle



Overrepresented GO categories:	p-value
chloroplast	<e-85
thylakoid	1.02e-28
generation of precursor metabolites and energy	5e-26
photosynthesis	1.68e-15
plastid stroma	3.74e-14
protein-chloroplast targeting	1.06e-05
pigment biosynthesis (tetraterpenoid, carotenoid, chlorophyll, porphyrin, terpene)	0.0054
regulation of photosynthesis:	0.0055

Figure 3. Cluster 1, pollen-specific (**A**) and cluster 2, photosynthesis (**B**). Pie charts are based on manual annotation. Representative expression profiles of 200 randomly chosen genes are shown.

factors, among them ARR2 (At4g16110), a two-component response regulator that acts in the cytokinin signaling pathway and binds to the promoters of nuclear genes for several components of mitochondrial respiratory chain complex I (nCI), up-regulated in pollen during spermatogenesis (Lohrmann et al, 2001); AGL 18, also a transcription factor, which represses flowering (Adamczyk et al, 2005) and At3g01330; DEL3, member of the E2F transcription factors that control cell cycle (Kosugi et al, 2002). There are 39 pollen-specific proteins in this group, including two apyrases, AtAPY1 and AtAPY2, that are essential for pollen germination (Steinebrunner et al, 2003); NPG1, calmodulin-binding protein required for pollen development (Golovkin et al, 2003); pollen coat protein; 4 pollen allergens; 4 multi-copper oxidase type I family proteins, nearly identical to pollen-specific BP10 protein; prolifins 3 and 4, similar to pollen specific actin-depolymerizing factor 1 from *Nicotiana tabacum* and 9 genes related to pollen tube growth: SPIK, a potassium channel protein (Mouline et al, 2002); pollen-specific pectin esterases required for enhancing the growth of pollen tube; ROP1, pollen-specific Rop GTPase, a central regulator of tip growth in pollen tubes (Li et al, 1999, Fu et al, 2001); RIC 1 and RIC 6 that interact with ROP1 (Wu et al, 2001) and two 1,3-beta-glucan synthase complex proteins, required for microsporogenesis, pollen tube growth and pollen wall formation.

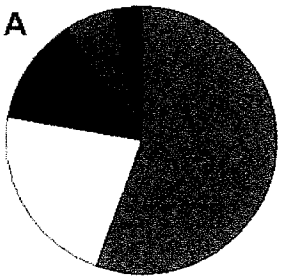
The second biggest cluster (2) contains 971 mainly nucleus-encoded genes involved in chloroplast function: mostly photosynthetic, structural and plastidic protein synthesis activities (overrepresented GO terms: plastid: p-value<e-85, thylakoid: p-value=1.02e-28, photosynthesis: p-value=1.68e-15) (Figure 3B). The cluster contains proteins necessary for plastid biogenesis and organization, formation of thylakoid membrane, biosynthesis of constituents and regulation of the photosynthetic complexes. Eighteen of the genes are involved in chloroplast biogenesis and organization, fission and relocation. An example of these is *thylakoid formation 1* (PSB29), required for thylakoid membrane organization (Wang et al, 2004). One hundred and seventy of these genes have a photosynthesis-related activity. Many are metabolic: eg., biosynthesis of pigments (chlorophyll, carotenoids, porphyrins), aminoacids, tetrahydrofolate, isoprenoids, starch and nucleotides. Metabolism of plastidic proteins is strongly represented by 75 genes; the function of these genes may be

coupled with photosynthesis. For example, PRPL11 is a ribosomal protein that affects photosynthesis (Pesaresi et al, 2001). Fourteen genes from the Calvin cycle, 16 genes from photorespiration, and a subset of 14 genes (some with putative functional assignments) from plastidic glycolysis are also present, reflecting the coupling of metabolic activity during high light conditions with photosynthesis. There are 40 genes encoding proteins with tetratricopeptide (TPR) or pentatricopeptide (PPR) domains. Proteins with these domains have been suggested to be transcript-specific regulators of gene expression in plastids (Nakamura et al, 2004), and have been implicated in plant organelle biogenesis (Lurin et al, 2004). Specifically, HCF107 is a tetratricopeptide that processes the polycistronic chloroplast *psbB-psbT-psbH-petB-petD* operon coding for proteins of the photosystem II and cytochrome b6/f complexes (Sane et al, 2005). A total of 19 plastid-encoded genes are in this cluster, including *psbH* and *petB*. FLU, another TPR containing protein, is a negative regulator of chlorophyll synthesis (Meskauskiene et al, 2001). Out of six nuclear sigma factors that modulate specificity of the plastidic RNA polymerase (SIG1-SIG6), five are present. Protein import is represented by HCF106, translocation pathway component, which imports proteins into thylakoid lumen (Mori et al, 1999); TOC159, a transit sequence receptor required for import of proteins and essential for chloroplast biogenesis (Bauer et al, 2000); and CPFTSY, inferred to be a chloroplast signal-recognition particle receptor protein. Transporters of sodium, calcium and other metals are represented, too. Two hundred and three genes have no known function. The expression of this cluster is light-regulated, and is high in every organ except roots.

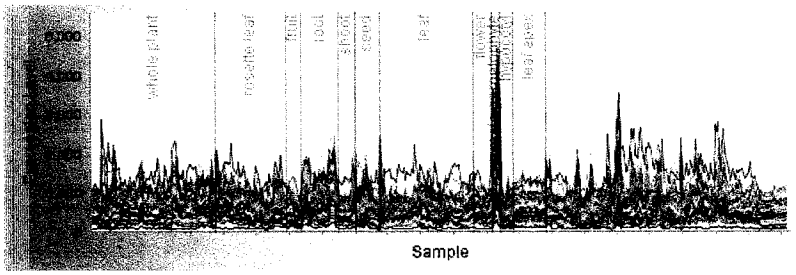
Cluster 29 (63 genes) contains mostly genes involved in mitochondrial aerobic respiration (Figure 4A). There are 36 genes with a function in aerobic respiration, 8 ATP synthase components, 2 genes for pyruvate dehydrogenase and 14 of unknown function. 51 of the proteins encoded are predicted or demonstrated to have mitochondrial localization. The expression is high and well correlated, highest in pollen.

Cluster 20 provides an example of a set of genes involved in nuclear regulation (Figure 4B). Sixty two out of 94 genes have some kind of nucleic acid-associated function: transcription factors, splicing factors, chromatin remodeling, histone deacetylases, RNA helicases, DNA repair, RNA processing. Five of these genes have been implicated in the

A

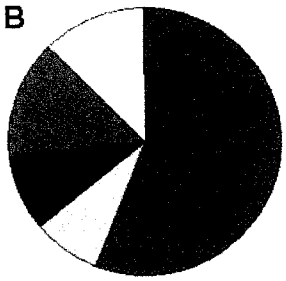


- aerobic respiration
- unknown
- ATP synthase
- others
- pyruvate dehydrogenase

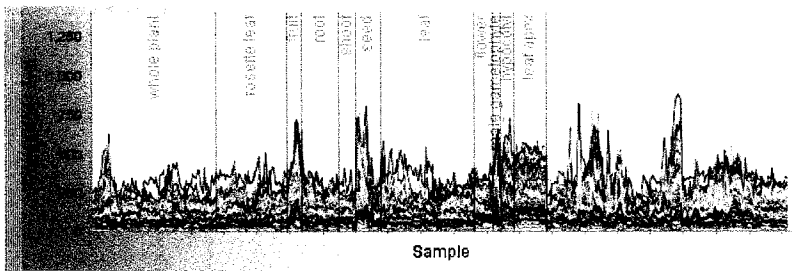


Overrepresented GO categories:	p-value
mitochondrion	1.4e-52
cytoplasm	3.04e-20
hydrogen-transporting ATP-ase	0.00119
mitochondrial inner membrane	0.00863

B

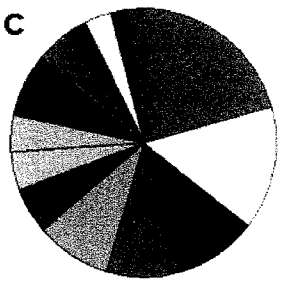


- nuclear
- development, biogenesis
- response
- flower
- others
- unknown

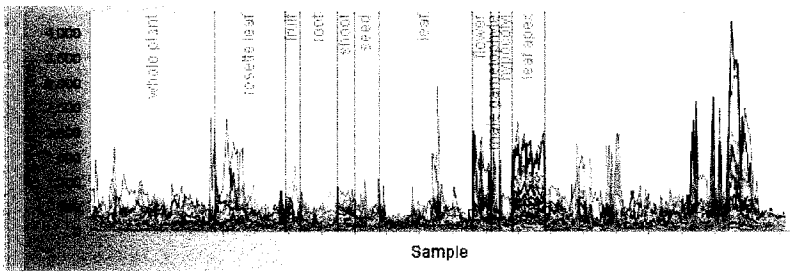


Overrepresented GO categories:	p-value
nucleic acid binding	6.63e-10
nucleus	2.5e-09
RNA helicase activity	9.57e-05
nuclear pore	0.000518
chromatin remodeling complex	0.0101

C



- others
- unknown
- transcription factors
- signal transduction
- protein kinase
- chromosome organization and biogenesis
- microtubule
- development
- cell cycle
- replication
- mitosis
- RNA processing
- regulation
- glycosyl hydrolase
- DNA metabolism, repair



Overrepresented GO categories:	p-value
mitotic cell cycle	1.31e-05
kinase	3.19e-05
interphase	0.00268
cyclin-dependent protein kinase complex	0.00874
nucleus	0.0178
regulation of mitotic cell cycle	0.0314
DNA replication errors control, RNA mediated gene silencing, cyclin binding, mismatch repair, G2/M ransition and G1/S transition of mitotic cell cycle	0.01-0.06
transcriptional activator activity	0.0848

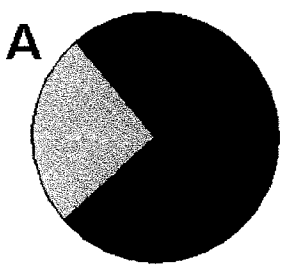
Figure 4. Cluster 29, mitochondrial respiration (**A**), cluster 20, nuclear regulation (**B**), and cluster 4, cell division (**C**). Pie charts are based on manual annotation. Representative expression profiles of 200 randomly chosen genes are shown for cluster 4.

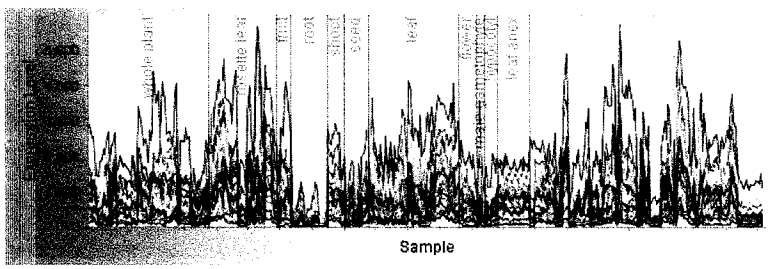
regulation of flower development (At3g12680, HUA1, RNA-binding protein which specifies stamen and carpel identities (Li et al, 2001); At5g04240, ELF6 (early flowering), repressor of the photoperiod pathway (Noh et al, 2004); At2g28290; SYD, which regulates floral homeotic gene expression (Wagner et al, 2002); At5g17690, TFL2 that regulates flowering and floral organ identity by silencing nuclear genes (Nakahigashi et al, 2005), and At4g32551, LUG, a negative regulator of the floral homeotic gene AGAMOUS). No other developmental processes are implicated. The expression pattern of this cluster is relatively low and uniform.

An example of a regulon with a cell-division-related function is cluster 4 (Figure 4C). It contains 495 genes, many with cell division - related function (mitosis, cell cycle, microtubule-related, chromosome organization and biogenesis, replication). Twenty one of these genes are directly involved in mitosis, including cell division control proteins (CDKB1, CDKB2;1, CDKB2;2, CDC2MsF) and cell division cycle protein HBT, as well as cyclins and other cyclin-dependent proteins. Regulatory and signaling molecules are abundant (55 transcription factors, 54 protein kinases, 53 signaling-related genes, 18 other regulatory proteins). Other nuclear functions represented include gene silencing, regulation of organ development, nuclear transport and RNA processing. The expression is highest in leaf apex.

Cluster 49 contains 27 genes, all of which are encoded in plastid genome: 17 encode ribosomal proteins and RNA polymerases, 7 are photosystem-related, and 3 hypothetical (Figure 5A). Three other small clusters are composed exclusively of plastidic genes: 176 (8 genes), 283 (5 genes) and 656 (2 genes). Interestingly, the cluster membership of expressed genes does not necessarily conform to their operon membership. For example, genes from the tricistronic operon, *psaA-psaB-rps14,* each belong to different cluster (numbers 49, 2 and 176, respectively), likewise, the genes from the *accD* operon are scattered among three clusters (number 2, 49 and 283).

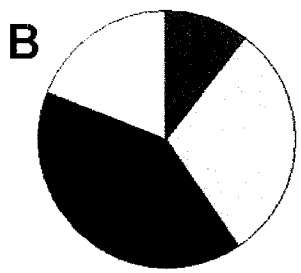
Forty seven genes, 45 of them encoded in the mitochondrion, constitute cluster 34 (Figure 5B). Six of these genes code for respiration-related functions (ATPases, cytochrome, NADPH dehydrogenase) and ribosomal proteins. Notably, most of these genes are hypothetical or unknown, but exhibit high expression. Some proximal mitochondrial genes

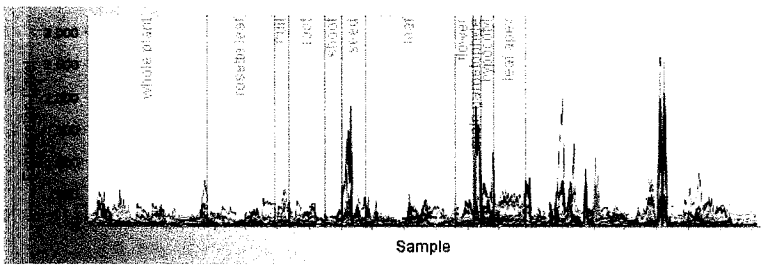




■ ribosome
□ photosynthesis
■ hypothetical

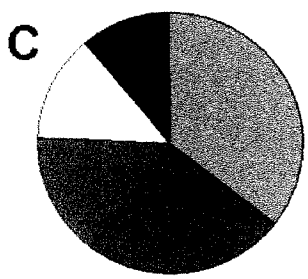
Overrepresented GO categories:	p-value
plastidic ribosome	2.63e-27
photosystem	2.32e-07

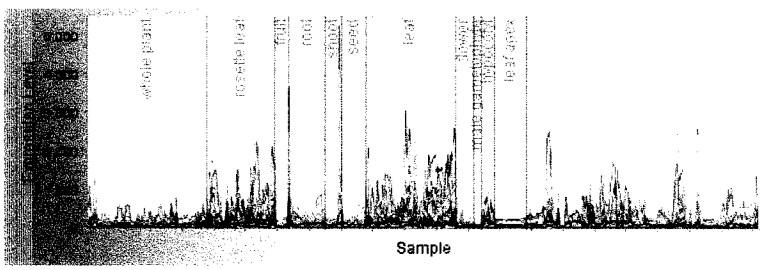




■ ATPase
□ cytochrome, NADH dehydrogenase
■ ribosome
■ hypothetical
□ unknown

Overrepresented GO categories:	p-value
electron transport	1.04e-16
NADH dehydrogenase	1.13e-13
cellular respiration	1.13e-13
mitochondrial membrane	5.28e-08
mitochondrion	2.15e-06
respiratory chain complex	7.81e-06





▨ protein kinase
▨ disease resistance, signaling, response
▨ others
□ unknown
■ heavy metal binding
■ transcription modulator

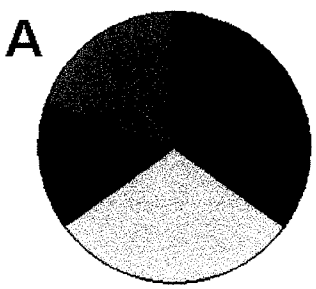
Overrepresented GO categories:	p-value
kinase	0.000155
MAPKKK cascade	0.000697
protein kinase cascade	0.00087
endomembrane	0.00297
response to stress	0.00742
systemic acquired resistance	0.00963
defense response	0.0106

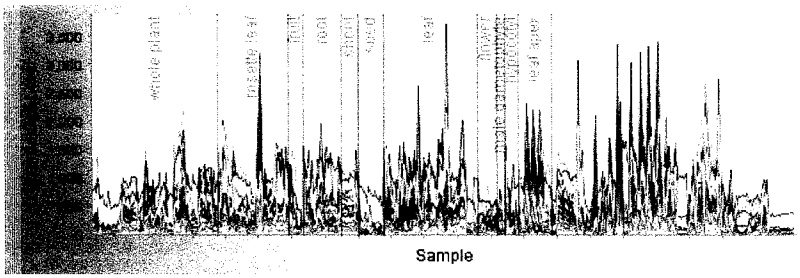
Figure 5. Cluster 49, plastid-encoded (**A**), cluster 34, mitochondrion-encoded (**B**), and cluster 35, protein kinases, signaling and defense response (**C**). Pie charts are based on manual annotation.

have been reported to be co-transcribed (eg., nad3 and rpsL2; rpl5 and cob; nad4L and orf25; atp1 and orf294) (Giege et al, 2000), however, others are from scattered regions of the mitochondrial chromosome. The co-expression of genes from different regions of the mitochondrial genome supports the concept that modulation of RNA stability plays a major role in regulation of gene expression in this organelle (Giege et al, 2000). Clusters 73 and 205 also contain mitochondrial genes. The expression of this cluster is generally high and well-correlated, and is upregulated in seeds, male gametophytes, and starvation.

Cluster 35 (45 genes) provides an example of a signaling cluster, in this case representing signaling events that activate responses to extracellular stimuli, such as pathogens (Figure 5C). It contains 16 protein kinases, some of them linked to response to pathogens (eg., CRK5) and 10 other proteins involved in disease resistance, response and signaling (cellulase, expansin, 6 disease resistance proteins, calmodulin- and cyclic nucleotide binding proteins). Three genes are involved in MAPKKK cascade: MAP kinase MPK3 and 2 MAP kinase kinases, MKK1and MKK2. Twenty three of the encoded proteins have a predicted location in the endomembrane system. At3g56710, SIB1, is a nuclear protein that modulates transcription in chloroplasts (Morikawa et al, 2002) and might coordinate the response of the plastidic genome with the nuclear one. The expression is high in leaves, especially in plants perturbed by pathogen infection or potassium starvation, or during senescence.

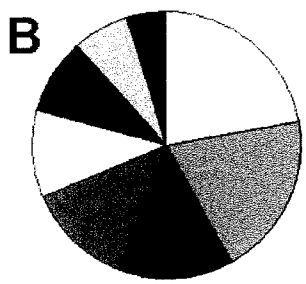
Glucosinolates are defensive compounds against herbivores and may increase activity of protective enzymes, like glutathione transferase (Fahey et al., 2001, and references therein). Cluster 69 contains 20 genes, most of which may be involved in glucosinolate biosynthesis in chloroplasts (Figure 6A). Leucine biosynthesis accounts for eight out of the 13 genes with known biosynthetic functions; others are needed for the biosynthesis of homoserine, lysine, homomethionine, isoleucine, valine, choline and glucosinolate (which are derived from aminoacids (Reintanz et al, 2001)). The enzymes currently annotated in TAIR as being involved in leucine biosynthesis might be active also in the glucosinolate pathway, since those pathways have analogous chemical reactions (Field et al., 2004). Two glutathione transferases, as well as a flavin-containing monooxygenase, another antioxidant involved in glucosinolate production from phenylalanine in rapeseed (Bennett et al. 1993),

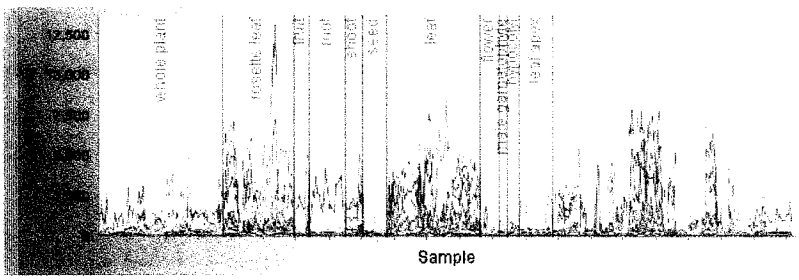




■ Leu biosynthesis
□ biosynthesis
■ toxin catabolism
▨ others

Overrepresented GO categories:	p value
secondary metabolism	0.000223
glycoside biosynthesis	0.000573
glycosinolate metabolism	0.000595
carbohydrate biosynthesis	0.00761
chloroplast	0.0375
other biosynthesis (leucine, aminoacids, branched-chain fatty acids)	0.03-0.1





□ disease resistance protein
▨ protein kinase, often membrane receptor
■ apoptosis
▨ others
□ unknown
■ signaling
□ defense response (not apoptosis)
■ lipid metabolism

Overrepresented GO categories:	p-value
kinase	1.31e-12
membrane	1.25e-09
endomembrane	5.18e-07
response to salicylic acid	0.0312

Figure 6. Cluster #69, glucosinolates biosynthesis (**A**) and cluster # 25, defense response (**B**). Pie charts are based on manual annotation.

and flavonol sulfotransferase that might be involved in pathogen response are also present in the cluster. The expression is spiky and shows light sensitivity; it depends on day length in AtGenExpress Developmental series experiment of the leaf apex (lowest expression: 7 days long-day conditions) (NASCArrays experiment reference number: NASCARRAYS-155), and peaks at four hours of daylight, in a wild-type Arabidopsis, and in a mutant with altered starch metabolism (L. Li, personal communication).

Genes involved in resistance to disease or a pathogen constitute many clusters. Another example is cluster 25. It has 70 genes, 19 of them are disease resistance proteins

(Figure 6B). It also contains other genes involved in pathogen response, among them lectin and lectin kinases, genes related to apoptosis, 17 protein kinases, many of them receptors, and 8 signaling genes. 18 genes are predicted to be integral membrane genes. The spiky expression of this cluster, highest in leaves, is symptomatic for genes responding to environmental stimuli.

Genes not recruited into clusters

One hundred and twenty six genes were not recruited into any cluster. Many of these are involved in protein modification or signal transduction.

Several functional categories have little representation in the 13k genes that formed the defined clusters (Figure 1). Specifically, the genes filtered out because of the low expression were enriched in mitochondrion-located genes (p-value 3.47e-11), apoptosis (2.4e-10), transmembrane receptor activity (9.05e-08) and genes not correlated highly with others are enriched in annotations of membrane location (p-value 1.99e-10) (Figure 1). The genes loosely associated with the main network are not significantly enriched in any functional category, and the second biggest connected component, with 8 genes, contains mostly pseudogenes and transposons.

We looked in detail into functions of the 100 genes with the highest expression, the lowest expression, and the most and the least changing expression. The genes were assigned to functional classes based on the manual curation.

Genes with highest expression were identified as genes with the maximum mean of expression values across all samples (Figure 7A). Photosynthesis–related and protein biosynthesis and modification genes constitute together 58% of this group. Toxin catabolism (5 members) was another overrepresented category.

The majority of the genes found to be expressed at the lowest level (genes with the minimum mean of expression values across all samples) are those for which no expression would be expected: hypothetical genes, transposons and pseudogenes (Figure 7B). The expression signal in this group does not exceed 40 and might be an artifact (eg., of signal processing or normalization). Nucleic acid-binding and disease resistance genes are among

the remaining 23 genes with functional annotations that have the lowest overall level of expression.

The most changing expression profiles belong to the genes that react to various stimuli (Figure 7C). 39 genes had annotation suggesting their involvement in signaling or response to oxidative stress, pathogens or hormones. 12 genes had a function related with

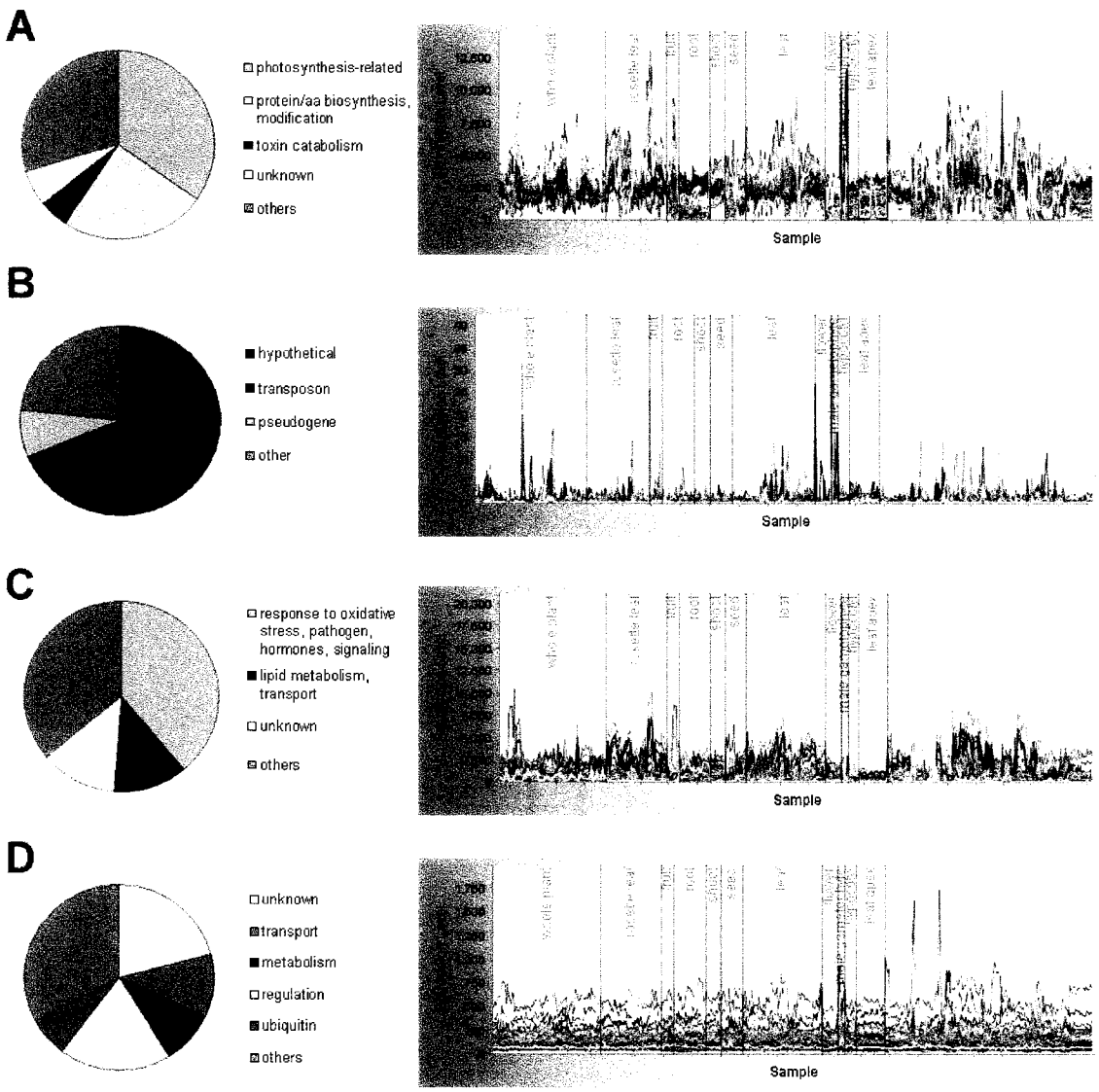


Figure 7. Functional assignments and expression profiles of the 100 genes with **(A)** the highest expression (maximum mean) **(B)** the lowest expression (minimum mean) **(C)** the most changing expression (highest standard deviation of logE) and **(D)** the most steady expression (lowest standard deviation of logE).

lipid metabolism, transport and degradation, possibly reflecting requirements for energy storage that change with the growth phase and the plant part.

Genes with the least-changing expression have a uniform mixture of metabolism, regulation and transport functions (Figure 7D). There is high proportion of unknown genes in this group. The even level of expression of these genes under all conditions doesn't facilitate the job of ascertaining their function.

Negatively correlated pairs of genes

Cases of negative correlation are far less abundant than positive ones, with the highest value in our dataset about -0.73. An example of a pair of genes highly negatively correlated is At1g06650, 2-oxoglutarate-dependent dioxygenase similar to tomato ethylene synthesis regulatory protein E8 and At5g23430, transducin family protein with nucleotide binding WD-40 repeat, with a corr=-0.734. Another example is At3g11910, DNA binding ubiquitin-specific protease and At4g12800, photosystem I reaction center subunit XI, correlated at -0.72. In 6 out of 8 most negatively correlated pairs of genes, at least one gene was implicated in a regulatory function.

DISCUSSION

The picture of Arabidopsis emerging from this study is that of a plant mainly occupied with gathering energy, reproduction and defense from hostile environment. Genes involved in photosynthesis and photosynthesis-related metabolic processes are those that are expressed at the highest level, and form second biggest cluster. The group of 1623 pollen-specific genes is the most numerous and the most highly correlated. Response and signaling programs are diverse and abundant, reflecting the need for elasticity of the realization of genetic program in changing conditions. 13 of the 71 clusters were assigned function of some signaling/response to external/internal stimuli. These are usually composed of genes of various molecular functions: receptors, kinases, hormone signaling genes and metabolic

genes, (ex for cell wall degradation of a pathogen). Response-related genes are among the most variable with respect to expression level.

Developmental programs account for function of over 10 clusters and include pollen, flower, fruit, root, embryo, and chloroplast development.

The presence of single metabolic pathways as distinct clusters is surprisingly scarce. Exclusively metabolic modules include fatty acid synthesis, aerobic respiration, glucosinolate biosynthesis, and protein biosynthesis. Plastidic glycolysis and Calvin cycle genes are in a cluster together with other genes active in the chloroplast during photosynthesis. Some pathways, like glycolysis, do not form a distinct correlated group probably because enzymes of this pathway have multiple metabolic functions. It is probable that with increasing the annotations of biological function more clusters with metabolic function would appear. However, these data may also reflect a real dichotomy between the textbook metabolic pathways and the Arabidopsis transcriptional network. By identifying those genes that are coregulated with particular metabolic fluxes, the fundamental mechanisms underlying regulation of metabolism can be better understood.

The expression of the plastidic genome is not uniform and is apparently finely tuned by additional level of regulation. Plastidic genes belong to 5 different clusters, sometimes cluster membership differs even within the same operon. Nuclear factors that modulate transcript stability, notably PRP proteins, might be responsible for these differences (Nakamura et al, 2004). Two kinds of RNA polymerases are active in the plastome: plastid-encoded plastid RNA polymerase (PEP), responsible for transcription of mainly photosynthesis I and II genes, and nuclear-encoded plastid RNA polymerase (NEP), controlling chloroplast development genes (ex *accD*). Plastidic genes may have promoters for PEP, NEP, or both types of polymerases (Hajdukiewicz et al, 1997, Ishizaki et al, 2005). The observed clustering into mainly photosynthetic genes including 5 nuclear sigma factors for PEP (cluster 2), mainly ribosomal genes (cluster 49) and additional groups of mixed function (cluster 176, 283 including *accD*, and 656), observed in our data, is likely a reflection of subsets of transcripts from distinct promoters and alternative RNA processing.

There is a grouping of genes into clusters with a similar general function category, observed in the network, particularly for information-related and response categories. This

indicates that these clusters contain genes that participate in multiple response programs and in some subset of conditions are coexpressed. On the other hand, some clusters seem to be isolated in the network (fatty acid synthesis, mitochondrial genes). Those clusters might contain genes that are committed to a distinct process that is carried out in relative independence from other cellular processes.

The genes that code for proteins located in mitochondrion or endomembrane, or involved in apoptosis or regulation of transcription, were among those with the least expression. Many genes for membrane-located proteins, including transporters, had expression profiles unlike any other in the genome. Both these low-expressed and unique-profile classes of genes were not used for network construction. Although many of low-expressed genes are hypothetical and might not be transcribed (76 out of 100 genes with lowest expression in our data were not expressed in the experimental evaluation of the Arabidopsis expression activity by whole genome tiling array by Yamada et al (2003)), some of the hypothetical genes might be active, but in very specific cell types or temporal conditions and thus their expression or ESTs have never been detected. Several well studied genes, for example regulators of flowering CONSTANS and FRI, or myb-type transcription factor CPC, responsible for differentiation of the epidermal cells, are expressed at a very low level in our dataset. Furthermore, genes with flat expression profiles might also have little representation in the network; only 14 out of 100 genes with most steady expression profiles were used for clustering, while 68 were filtered out due to low similarity to any other gene. The profiles of regulatory genes may be flat because the activity of their products is modulated by post-transcript modification – addition of a phosphate group, induced conformation change, binding a cofactor or other subunit. Because regulatory genes are among most comprehensively studied and often guide the annotation of cluster function, their absence in clusters might hinder the identification of developmental programs in our clustered data.

Pairs of negatively correlated genes are potentially interesting. The facts that many such correlations include regulatory protein and that examples of known negative regulators are negatively correlated in our data with their potential targets support biological importance of discovered negative relations. For example, AT2G23430 (ICK1), cyclin-dependent kinase

inhibitor protein, which functions as a negative regulator of cell division and interacts with CYCD3;1 (Zhou et al, 2003) is negatively correlated with this gene (corr=-0.43). ICK1 also negatively correlates at 0.52-0.56 with cyclin-dependent protein kinases CYC2b and CYCA2-similar, cell division control protein CDKB2;1 and other mitosis-related genes. AT1G75950, SKP1, a negative regulator of DNA recombination, is most negatively correlated with DNA polymerase and tubulin-related genes (0.5-0.4). Whether a given correlation translates to negative regulation remains to be experimentally evaluated. Negative correlation might merely reflect the disjoint sets of conditions in which two proteins are active.

The coexpressed clusters identified here might indicate common regulatory program acting upon participating genes. For the groups of genes with expression profiles that are variable and parallel across many diverse conditions such hypothesis is particularly plausible. It has also potential of assigning function to unknown proteins that are strongly connected with such clusters. On the other end of the spectrum there are clusters that unite genes active only in small subset of conditions, even as little as a single sample. In this case there is less ground for the coregulation assumption, but such clusters may also reveal valuable information. For example, genes for increased growth in response to auxin and cytokinin, upregulated only in cell culture and tumors (cluster 48), would be hard to pinpoint if not for the meta-analysis. Data on proteomics or protein interactions would greatly complement the existing expression data and enable a more complete view of the functional programs.

METHODS

Transcriptomics data

Arabidopsis expression data for 963Affymetrix ATH1 chips with probes for 22,746 genes were obtained from Nottingham Arabidopsis Stock Centre microarray database (http://affymetrix.arabidopsis.info/; Craigon et el., 2004) and PLEXdb (http://www.plexdb.org/; Shen et al., 2005). The data represent 70 experiments, including development, stress, mutant, and other studies. All chips in databases were already

individually scaled to the common mean=100, excluding top and bottom 2% signal intensities, using MAS 5.0 algorithm (Affymetrix). The replicability was qualitatively assessed on scatterplots and the inconsistent replicates were removed. The data was normalized to the common range by scale normalization (Yang et al., 2002). Expression values on chips from biological replicates were averaged. All computations in this work, except for graph clustering, were performed in R software (R Development Core Team, 2004). The normalized data is available online at http://www.metnetdb.org/MetNet_MetaOmGraph.htm

Network of coexpressed genes

Pearson correlation matrix was calculated for the 18,195 genes with expression over the chip mean of 100 in at least one chip. 4551 genes whose expression never reached the value of 100 were filtered out because low expression estimations are not reliable and might introduce the noise in the dataset. Only the 14,564 genes that were correlated above the threshold of 0.7 with any other gene were retained for further analysis. The matrix was transformed into a binary matrix by replacing the values of correlation > 0.7 by 1 and 0 otherwise. The resulting matrix induced the adjacency matrix of the coexpression network, in which genes form the nodes and two genes are connected by an edge if they are correlated above 0.7. This value of Pearson correlation in our data generally corresponds to the reliable coexpression across all the microarray slides surveyed.

Clustering the coexpression network

Connected components were identified in the network (*connectedComp* function in R software), yielding one giant connected component with 14,368 nodes and 77 smaller components, ranging from 2 to 8 nodes. Genes connected by a single edge were removed from the biggest connected component, and the resulting network composed of 13,456 genes and nearly 1.5 million edges was clustered by Markov Clustering graph clustering algorithm (http://micans.org/mcl/index.html#source, van Dongen, 2000). 998 clusters were produced;

these were analyzed together with smaller connected components from the previous step. The network in Figure 2 was visualized in GraphExplore tool (Q. Wang, G. Yao, J.R. Nevins and M. West, submitted for publication, http://graphexplore.cgt.duke.edu).

Analysis of clusters' functional coherence

The coherence of functionality of the genes within the clusters was assessed by a combination of automatic analysis of overrepresentation of GO terms (http://gostat.wehi.edu.au/cgi-bin/goStat.pl; Beissbarth and Speed, 2004) and manual inspection of function and expression using tools (AtGeneSearch, MetaOmGraph) in MetNet Exchange toolbox (http://metnet.vrac.iastate.edu/). The pathway data is from MetNetDB database (http://metnet.vrac.iastate.edu/MetNet_db.htm). The RNA profiles were plotted in MetaOmGraph. (http://www.metnetdb.org/MetNet_MetaOmGraph.htm).

ACKNOWLEDGEMENTS

We thank Dr. Di Cook for assistance with data normalization.

REFERENCES

Adamczyk, B.J., Lehti-Shiu, M.D., and Fernandez, D.E. (2005) The MADS domain factors AGL15 and AGL18 act redundantly to repress flowering in short days. 16TH INTERNATIONAL CONFERENCE ON ARABIDOPSIS RESEARCH

Bauer, J., Chen, K., Hiltbunner, A., Wehrli, E., Eugster, M., Schnell, D., and Kessler, F. (2000) The major protein import receptor of plastids is essential for chloroplast biogenesis. Nature **403**, 203-7.

Beissbarth, T., and Speed, T.P. (2004) GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics **20**,1464-1465.

Bennett, R., Donald, A., Dawson, G., Hick, A., and Wallsgrove, R. (1993) Aldoxime-forming microsomal enzyme systems involved in the biosynthesis of glucosinolates in oilseed rape (*Brassica napus*) leaves. Plant Physiol. **102**:1307-1312

Craigon, D. J., James, N., Okyere, J., Higgins, J., Jotham, J., and May, S. (2004). NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res. **32**, D575-577

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research **30**, 1575-1584.

Fahey, J.W., Zalcmann, A.T., and Talalay, P. (2001) The chemical diversity and distribution of glucosinolates and isothiocyanates among plants Phytochemistry, **56,** 5-51.

Field B., Cardon G., Traka M., Botterman J., Vancanneyt G., and Mithen R. (2004) Glucosinolate and Amino Acid Biosynthesis in Arabidopsis, Plant Physiol. **135**, 828-839.

Fu, Y., Wu, G., and Yang, Z. (2001). Rop GTPase-dependent dynamics of tip-localized F-actin controls tip growth in pollen tubes. Journal of Cell Biology **152,** 1019–1032.

Giege, P., Hoffmann, M., Binder, S., and Brennicke, A. (2000) RNA degradation buffers asymmetries of transcription in Arabidopsis mitochondria. EMBO Rep. **1**, 164-70.

Golovkin, M., and Reddy, A.S. (2003) A calmodulin-binding protein from Arabidopsis has an essential role in pollen germination. Proc Natl Acad Sci U S A. **100**, 10558-63.

Gu, Y., Fu, Y., Dowd, P., Li, S., Vernoud, V., Gilroy, S., and Yang, Z. (2005) A Rho family GTPase controls actin dynamics and tip growth via two counteracting downstream pathways in pollen tubes. J. Cell Biol. **169**, 127-138.

Hajdukiewicz, P.T., Allison, L.A., and Maliga, P. (1997) The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. EMBO J. **16,** 4041-8.

Ishizaki, Y, Tsunoyama, Y., Hatano, K., Ando, K., Kato, K., Shinmyo, A., Kobori, M., Takeba, G., Nakahira, Y., and Shiina, T. (2005) A nuclear-encoded sigma factor, Arabidopsis SIG6, recognizes sigma-70 type chloroplast promoters and regulates early chloroplast development in cotyledons. Plant J. **42**, 133-44.

Kosugi, S., and Ohashi, Y. (2002) E2Ls, E2F-like repressors of Arabidopsis that bind to E2F sites in a monomeric form. J Biol Chem. **277**, 16553-8.

Kotera, E., Tasaka, M., and Shikanai, T. (2005) A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. Nature **433**, 326-30.

Li J., Jia, D., and Chen, X. (2001) HUA1, a regulator of stamen and carpel identities in Arabidopsis, codes for a nuclear RNA binding protein. Plant Cell. 13, 2269-81.

Li, H., Lin, Y., Heath, R.M., Zhu, M.X., and Yang, Z. (1999). Control of pollen tube tip growth by a Rop GTPase-dependent pathway that leads to the tip-localized calcium influx. The Plant Cell **11,** 1731–1742.

Lohrmann, J., Sweere, U., Zabaleta, E., Baurle, I., Keitel, C., Kozma-Bognar, L., Brennicke, A., Schafer, E., Kudla, J., and Harter, K. (2001) The response regulator ARR2: a pollen-specific transcription factor involved in the expression of nuclear genes for components of mitochondrial complex I in Arabidopsis. Mol Genet Genomics. **265**, 2-13.

Lurin, C. et al. (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell. **16**, 2089-103.

Magwene, P.M., and Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence. Genome Biol.**5**, R100.

MAPK Group. (2002) Mitogen-activated protein kinase cascades in plants: a new nomenclature. Trends Plant Sci. 7, 301-8.

Meskauskiene, R., Nater, M., Goslings, D., Kessler, F., op den Camp, R, and Apel, K. (2001) FLU: a negative regulator of chlorophyll biosynthesis in Arabidopsis thaliana. Proc Natl Acad Sci USA. **98**, 12826-31.

Mori, H., Summer, E.J., Ma, X., and Cline, K. (1999) Component specificity for the thylakoidal Sec and Delta pH-dependent protein transport pathways. J Cell Biol. **146**, 45-56.

Morikawa, K., Shiina, T., Murakami, S., and Toyoshima, Y. (2002) Novel nuclear-encoded proteins interacting with a plastid sigma factor, Sig1, in Arabidopsis thaliana. FEBS Lett. **514**, 300-4.

Mouline, K., Very, A.A., Gaymard, F., Boucherez, J., Pilot, G., Devic, M., Bouchez, D., Thibaud, J.B., and Sentenac, H. (2002) Pollen tube development and competitive ability are impaired by disruption of a Shaker K(+) channel in Arabidopsis. Genes Dev. **16**, 339-50.

Nakahigashi, K., Jasencakova, Z., Schubert, I., and Goto, K. (2005) The Arabidopsis HETEROCHROMATIN PROTEIN1 Homolog (TERMINAL FLOWER2) Silences Genes within Euchromatic Region but Not Genes Positioned in Heterochromatin. Plant Cell Physiol. **46**, 1747-56.

Nakamura, T., Schuster, G., Sugiura, M., and Sugita, M. (2004) Chloroplast RNA-binding and pentatricopeptide repeat proteins. Biochem Soc Trans. **32**, 571-4.

Noh, B., Lee, S.H., Kim, H.J., Yi, G., Shin, E.A., Lee, M., Jung, K.J., Doyle, M.R., Amasino, R.M., and Noh, Y.S. (2004) Divergent roles of a pair of homologous jumonji/zinc-finger-class transcription factor proteins in the regulation of Arabidopsis flowering time. Plant Cell. **16**, 2601-13.

Pereira-Leal, J.B., Enright, A.J., and Ouzounis, C.A. (2004) Detection of functional modules from protein interaction networks. Proteins. **54**, 49-57.

Pesaresi, P., Varotto, C., Meurer, J., Jahns, P., Salamini, F., and Leister, D. (2001) Knock-out of the plastid ribosomal protein L11 in Arabidopsis: effects on mRNA translation and photosynthesis. Plant J. **27**, 179-89.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ransom, N., Mentzen, W.I., and Wurtele, E.S. MetaOmGraph: A Tool for Plotting and Analyzing Gene Chip Data and Other Large Datasets. In preparation.

Reintanz, B., Lehnen, M., Reichelt, M., Gershenzon, J., Kowalczyk, M., Sandberg, G., Godde, M., Uhl, R., and Palme, K. (2001) *bus*, a Bushy Arabidopsis *CYP79F1* Knockout Mutant with Abolished Synthesis of Short-Chain Aliphatic Glucosinolates Plant Cell **13**, 351-367.

Sane, A.P., Stein, B., and Westhoff, P. (2005) The nuclear gene HCF107 encodes a membrane-associated R-TPR (RNA tetratricopeptide repeat)-containing protein involved in expression of the plastidial psbH gene in Arabidopsis. Plant J. **42,** 720-30.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. **34**, 166-76.

Shen, L., Gong, J., Caldo, R.A., Nettleton D., Cook, D., Wise, R.P. and Dickerson, J.A. (2005) BarleyBase -- An expression profiling database for plant genomics. Nucleic Acids Research. **33**, D614-D618.

Steinebrunner, I., Wu, J., Sun, Y., Corbett, A., and Roux, S. J. (2003) Disruption of Apyrases Inhibits Pollen Germination in Arabidopsis Plant Physiol. **131**, 1638-1647

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science **302**, 249-55.

The Institute for Genomic research (TIGR), TIGR Release 5.0, June 7, 2004. http://www.tigr.org/tdb/e2k1/ath1/ath1.shtml

van Dongen, S. (2000) Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, May 2000.

Wagner, D., and Meyerowitz, E.M. (2002) SPLAYED, a novel SWI/SNF ATPase homolog, controls reproductive development in Arabidopsis. Curr Biol. **12**, 85-94.

Wang, Q., Sullivan, RW., Kight, A., Henry, R.L., Huang, J., Jones, A.M., and Korth, K.L. (2004) Deletion of the chloroplast-localized Thylakoid formation1 gene product in Arabidopsis leads to deficient thylakoid formation and variegated leaves. Plant Physiol. **136**, 3594-604.

Wang, Q., Yao, G., Nevins, J., West, M. and Dobra, A. GraphExplore: a software tool for network visualization; submitted for publication.

Wu, G., Gu, Y., Li, S., and Yang, Z.. (2001) A genome-wide analysis of *Arabidopsis* Rop-interactive CRIB motif-containing proteins that act as Rop GTPase targets, Plant Cell **13**, 2841–2856.

Yamada, K. et al. (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. Science **302**, 842-6.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. (2002) Normalization for CDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. **30**, e15.

Zhou, Y., Wang, H., Gilmer, S., Whitwill, S., and Fowke, L.C. (2003) Effects of co-expressing the plant CDK inhibitor ICK1 and D-type cyclin genes on plant growth, cell size and ploidy in *Arabidopsis thaliana*, Planta **216**, 604 - 613

CHAPTER 5. GENERAL CONCLUSIONS

Fatty acid biosynthesis is an essential biological process and its indispensability is reflected on multiple levels. On the molecular level, the enzyme catalyzing the first committed step of the pathway, acetyl-coA carboxylase is a ubiquitous enzymatic activity, omnipresent in all three superkingdoms. This ancient function appears to have evolved independently in bacteria, archaea and eukaryotes and thus given rise to the variety of structures and functions of biotin-dependent enzymes observed nowadays.

On the pathway level, the complete set of activities necessary to produce 18-carbon fatty acid is transcriptionally coregulated, as revealed by coexpression analysis. This coregulation extends to reactions not regarded traditionally as the part of this pathway: 1) production of the acetyl-CoA substrate from phosphoenolpyruvate in two sequential reactions catalyzed by pyruvate kinase and pyruvate dehydrogenase; 2) production of cofactors (biotin, lipoic acid).

On an even higher level of organization, the system level, there is a separation of fatty acid biosynthesis, achieved by spatial isolation (containment within plastids) and commitment of expression of the genes encoding the participating enzymes. This separation accentuates the relative self-sufficiency and independence of the process. Although all processes within the organism are ultimately coordinated, such completeness and independence lays at the base of biological modularity. Functional modules, which I identified by means of the common expression signature of the involved genes, define developmental programs, several response and metabolic programs, genetic information maintenance and proliferation programs, tissue-specific processes and protein complexes (ribosome, ubiquitin ligase, proteasome). Elucidating these modules is crucial for understanding how complex multicellular organisms such as plants coordinate processes and exert the appropriate controls. These studies demonstrate that the combined use of genomics and transcriptomics data can reveal the higher order organization of multicellular organisms

and, by generating many hypotheses in a single analysis, significantly expedite our understanding of the broad context of gene function.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to many people who made my journey toward PhD possible, bearable and sometimes fun.

My major professor Dr. Eve Wurtele, for being a bottomless source of advice, support, patience and optimistic encouragement throughout this endeavor. She made the work-time as close to play-time as it can be.

Dr. Eve Wurtele and Dr. Basil Nikolau for their generous help during preparation of this dissertation.

My labmates, present and past, who always found the time to share their wisdom, offered help and friendship, and created great, supportive environment that will be missed.

My POS committee members, Dr. Basil Nikolau, Dr. Jonathan Wendel, Dr. David Fernández-Baca and Dr. Xun Gu, for their time serving on my committee and much appreciated advice.

Members of the MetNet group, who delivered valuable critique and expertise.

My friends, for their patient ears, home-made beer and escapes from the world of grad school, those real and those in mind.

And foremost, to my wonderful family.