Plant Molecular Biology & Biotechnology

www.academyjournals.net

Original Article

A Predicted Interactome for Coffee (*Coffea canephora var robusta*)

Elisabeth FITZEK*, Matthew GEISLER*

Department of Plant Biology, Southern Illinois University, 1125 Lincoln Drive Carbondale, IL 62901-6509, USA

Received: 05.07.2011 ***Accepted: 09.09.2011*** ***Published: 11.09.2011***

Abstract

Protein interactions are central to many important cellular processes and give complexity and adaptiveness to biological responses. Comprehensive interactomes have been established for many model organisms using high throughput experimental methods but have yet to be fully explored. Evolutionary conservation of many core biological processes has enabled us to generate a predicted protein interactome for an economically important plant with complex metabolism, *Coffea arabica.* Of the over 12.000 genes identified in coffee by transcript sequencing, only 939 of them were predicted to have 4587 interactions. These include 4126 interactions conserved across all eukaryotic organisms, another 461 that appear to be plant specific, and 29 appear chloroplast specific and cyanobacterial in origin. A confidence value for each interaction was established on the basis of the amount and type of evidence. Small hubs (3-10 partners) make up 30% of the proteins. Using GO (gene ontology) anotation revealed significant enrichment for proteins involved in translation and the cytosolic ribosome, and a depletion of unknown protiens. This was expected, as only conserved interactions would be predicted using our methods, and these are the best studied. However there were some highly conserved interactions in coffee between otherwise unknown proteins. Dividing the entire network into subnetworks (clusters) based on highly interconnected proteins, we identified potential functional modules. The strongest such cluster shows the connections between proteins of the large and small subunits of the ribosome, while other clusters were identified as the proteosome and transcription initiation complexes. Several interconnected subnetworks were identified using cluster analysis. This predicted *Coffea arabica* interactome is not comprehensive, but provides a skeleton of conserved interactions from which to connect together more plant and coffee specific pathways.

Key words: *Coffea arabica, orthologous proteins, predicted protein interactome, protein-protein interaction*

[*]*Corresponding Author: E. Fitzek, E-mail:elfitzek@siu.edu, Phone: +001 618 5293919;*

INTRODUCTION

Model organisms provide a reference for the deduction of gene and protein functions in species which are more difficult to work with in the lab, but are more economically important. With the advent of high throughput sequencing(Wendl et al. 1998), it will become increasingly common to see whole transcriptome (RNA-sequencing), EST-sequencing or whole shotgun genome sequences of non-model organisms (Metzker 2009). With the new generation of sequencing (454 and AB SOLiD) genomes of requested organism can be generated in a short amount of time at low cost. This allows the expandtion into fields which would not consider to generate a genome of their species of interest such as evolutionary biology or domestic plants (Rothberg and Leamon 2008) (Hudson 2008). A recent study, conducted EST-libraries of *Coffea canephora* and *Coffea arabica* to compare their expression profiles based on their ploidy (Vidal et al. 2010). Generated datasets can be contributed into exsisting databases such as Phytozome (http://genome.jgi-psf.org/programs/plants/index.jsf) or Sol genomics network (http://solgenomics.net/) and in addition function as comparative resourses (Mueller 2005) (Paterson et al. 2009). Whereas, functional annotation of genes by experimentation will take considerable time, and in some species will never catch up with gene annotation by inference from model organisms. Only automated homology based annotation can keep up with the speed of genome generation.

Eukaryotic organisms often share the same conserved pathways regarding primary metabolism, DNA repair, vesicular transport and other cellular processes thus it is possible to tap a wide range of model organisms in order to build up annotation for a newly sequenced genome (Curwen 2004). The ensembl database (http://ensembl.org; (Hubbard et al. 2009) provides an extensive amount of experimentally derived data for the proteins of human (*Homo sapiens*), mouse (*Mus musculus*), yeast (*Saccharomyces cerevisiae*), fruitfly (*Drosophila melanogaster*), nematode worm (*Caenorhabditis elegans*), norwegian rat (*Rattus norvegicus*) and cyanobacteria (*Synechosystis* sp.). Orthologs, derived from a single gene in the last common ancestor, are separated only by speciation rather than duplication and divergence. Orthologs thus are more likely to share the same function in both species. However, not all homologous sequences are orthologs, thus it is important to distinguish these from in-paralogs, produced by duplication within one lineage and out-paralogs, which are produced by duplication prior to speciation from the last common ancestor. The program InParanoid 3.0 (http://inparanoid.cgb.ki.se) offers a method to separate these types of homologs, and establish a list of likely one-to-one orthologs (where no duplication has occurred) and clusters based on one-to-many and many-to-many orthology in which several inparalogs are included and ranked based on sequence divergence (O'Brien et al. 2005). The orthologs establish a link between test and reference organisms, and can be used to explore experimentally derived high throughput protein-protein interaction data that is present for many eukaryotic model organisms. The Arabidopsis interactome version 2.0 predicted over 70.000 interactions for 3617 proteins in *A. thaliana*. Of these proteins, 654 also had 1460 experimentally determined protein-protein interactions, which matched 217 of the predicted interactions, a statistically significant overlap (expected overlap by chance =7.41; chi-squared test $P<10^{-250}$) indicating that predictions using this method are accurate (Geisler-Lee et al. 2007). Geisler-Lee et al. have also shown that there is a high degree of gene co-expression between predicted interacting partners and significant enrichment for the likelihood of both interacting proteins being in the same subcellular localization. The Arabidopsis Interactions Viewer (http://bar.utoronto.ca; (Winter et al. 2007)) was developed to provide a user interface with a live database for the Arabidopsis interactome. These interactions construct an enormous map which shows predicted pathways between proteins, including signaling, metabolic pathways, and gene regulation. Since its release, the Arabidopsis interactome has been highly accessed and used in numerous published experimental analyses of interactions in metabolic and regulatory pathways, often used as a starting point for testing new hypotheses (Chan Zhou 2010; Dietz et al. 2010; Liu and Howell 2010). A predicted protein-protein interactome of evolutionarily conserved pathways is a highly useful tool that can be constructed from species with sequenced genomes or transcriptomes, but relatively little molecular experimental data, and is an important addition to initial gene annotation of newly sequenced genomes (Lewis et al. 2010; Lovell and Robertson 2010; Peregrín-Alvarez et al. 2009).

Coffee (*C. canephora var robusta*) belongs to the *Rubiaceae* family a close relative of the *Solanaceae* family which also includes tomato, eggplant, pepper, potato and tobacco. To help overcome the problems presented pests such as *H. hampei* that have a tremendous impact on the growth and development of the coffee fruit(Damon 2000), and to improve economic and environmental costs, an international committee (ICGN) was formed in 2005 to sequence the coffee genome in order to understand the genetic and molecular basis for relevant traits. A large EST library for *C. canephora* generated from different tissues has been assembled into unigenes and annotated, and is publically available through the solanaceae genome network (SGN) (Lin et al. 2005). Coffee is an example of the growing number of economically important species for which genome based resources have arrived prior to the accumulation of the large amounts of direct experimental molecular research typically found in model plants. Functional annotation for most coffee genes has thus far relied on inference from sequence homology to genes and protein domains of model organisms such as arabidopsis, yeast and *E. coli* (Fang et al. 2010; Finn et al. 2009; Flicek et al. 2007). To keep up with the fast approach tools to sequence whole genomes as well as automated annotation are not far behind. The annotations are based on *ab initio* prediction and/or best hit alignments via BLAST with known databases or EST libraries as subject (Wilming and Harrow 2009). Programs such as GENSCAN or MAKER are automated annotation tools which are able to identify splice sites, codon usage as well as cross reference homologs (Cantarel et al. 2007; Chris Burge 1997; Madupu et al. 2010). In this study we extend the prediction of gene function by adding protein-protein interactions derived from interacting orthologs in model organisms.

In this paper, we describe a network of 4586 predicted protein-protein interactions for Coffee (*Coffea canephora var robusta*) using the approach of Geisler-Lee et al. (2007), but using a larger dataset of reference organisms, and including plant and cyanobacterial experimentally determined interactions. Separate datasets are made for one-to-one and many-to-many orthology. We also present lists of protein orthologies between coffee and all model eukaryotic organisms constructed with InParanoid version 3.0 (http://inparnoid.cgb.ki.se; (Remm et al. 2001). Using orthology to *A. thaliana* (TAIR9 release) we constructed a GO annotation grid to determine over represented GO categories in conserved protein-protein interactions and compare to GO annotation distribution of Arabidopsis interactome. The Coffee predicted interactome is freely available to download from (http://sgn.com) and is included as supplemental files as both a database flat file and a pre-constructed network using the popular interactome browser Cytoscape (Shannon et al. 2003).

MATERIAL AND METHODS

The coffee (*Coffea canephora*) EST collection was obtained from SGN (unigene_estscan_pep; release May-2010; http://solgenomics.net). Model organisms were selected on the basis of available experimentally determined protein interactions and included human (*Homo sapiens*), mouse (*Mus musculus*), yeast (*Saccharomyces cerevisiae*) and (*Saccharomyces pombe*), fruitfly (*Drosophila melanogaster*), nematode worm (*Caenorhabditis elegans*), norwegian rat (*Rattus norvegicus*), Escherichia coli K-12 and Arabidopsis (*Arabidopsis thaliana*) and *Synechosystis* sp. PCC6803. Peptide sequences of the model organisms were downloaded from Ensemble (www.ensembl.org; Release 54-May2010), TAIR (http://www.arabidopsis.org; release 2009), Cyanobase (http://genome.kazusa.or.jp; release 2007) and *E.coli* genome database site (http://www.genome.wisc.edu; release Nov2008).

Ortholog construction

Full sets of peptide sequences in FASTA format were used by Inparanoid 3.0 program (O'Brien et al. 2005) pairwise to compare each model organism to coffee. For divergent model organisms such as human, mouse, rat, fruitfly and worm we set the block substitution matrix (BLOSSUM) to 62, while *Arabidopsis thaliana* was performed with a more stringent blossom matrix 80 as this is also a flowering plant and therefore more likely to have similar genes in common. The resulting output was parsed using a small in-house program written in perl (available upon request) to generate a one-one (100% score) ortholog list, and seperately, a many to many ortholog lists with a minium of 40% inparanoid score (to remove clearly divergent sequences). To map gene identifiers used in reference genomes and reference interactomes, we constructed a table of gene identifier aliases. We chose the many to many Inparanoid output to construct the coffee interactome and to evaluate its content (Table 1).

Interactome Construction

References to experimentally determined interactions in model organsims were collected from the Biogrid database (Biogrid-all-singlefile-2.0.53.tab; http://www.thebiogrid.org). Were coffee orthologs to both interacting reference proteins were identified, a predicted interaction was made. The experimental evidence, reference species, publication and authors from the referenced interaction were also recorded. In many cases, the same interaction in coffee was predicted from multiple reference species, or reference interactions using different experimental methods. Cytoscape version 2.6.3 (http://www.cytoscape.org) was used to visualize interactome data. Functional annotation was added for coffee genes with orthologous Arabidopsis proteins, including Gene Ontology (GO slim; ref), and conserved protein family domains from PFAM (Ashburner M 2000; Vidal et al. 2010)(http://pfam.janelia.org/; (Finn et al. 2009).

Confidence Value

For each protein-protein interaction (PPi) a confidence value (CV) was created based on the amount of reference evidence. The CV is calculated from on total number of times a coffee protein-protein interaction was predicted (T), multiplied by the number of different reference species that have orthologous interaction (S), and the number different experimental methods used to demonstrate the interaction in the reference species (E). Thus CV= T*S*E. The interactions were thus seperated into three categories: high confidence value (>10 CV; 282 number interactions; 8.1%), medium confidence (2-10 CV; 1278 number interactions; 17.8%) and low confidence (1 CV; 3027 number interactions; 74%) (Figure1A).

Functional Enrichment Analysis

GOslim and GOfull categories were mapped onto coffee orthologs from *Arabidopsis thaliana* annotation (http://arabidopsis.org). GOslim and GOfull terms were then counted in the proteins in the interactome dataset and compared to the full coffee protein dataset, and to the proteins in Arabidopsis predicted interactome. Statistical significance for enrichment or depletion of terms was determined by chi squared test. The p-value cut off was <1.0E-08 was chosen as the new α value to determine statistical significance.

Table 1 Predicted protein-protein interactions in coffee. PPis were found in dataset of Biogrid, Arabidopsis (total known interactions) and *E.coli*. Interactions found with '1to1' as well as 'many to many' output of Inparanoid. Comparing '1to1' with 'many to many' we see an increase in the interactions found in the different datasets.

Organism	**Sequences**	**INPARANOID 'many to many'**	**'1to1'**	**'Many to many'**	**Increased output for many to many**	**Total known interactions**	**% recovered**
A. thaliana	33410	5101(15%)	143	325	2.1	3881	8.3
C. elegans	27258	1977(7%)	75	84	1.1	6787	1.2
D. melanogaster	20815	1656(8%)	240	287	1.2	32786	0.8
E. coli K-12	4347	290(6%)	86	278	3.2	14253	1.9
H. sapiens	47509	3195(6%)	324	568	1.8	40086	1.4
M. musculus	40732	2442(6%)	2	4	2	1038	0.4
R. norvegicus	32948	2423(7%)	0	11	0	436	2.5
S. cerevisiae	6698	812(12%)	4026	6519	1.6	142657	4.5
S. pombe	5026	805(16%)	228	249	1.1	14008	1.8
***Synechosystis* sp.**	3672	333(9%)	52	53	1	2961	1.6
		total	5176	8378	1.6		
Unique PPis			**3331**	**4587**	1.4		

RESULTS AND DISCUSSION

Generation of a Predicted Interactome Using Orthology

Proteins are routinely annotated according to similarity to known proteins in other organisms either globally (homology based annotation) or due to local similarity in one region (protein domains). We extend this homology-based annotation for coffee (*Coffea robusta*) expressed genes by also predicting protein-protein interactions based on interacting orthologs (interologs) present in other organisms using established methods (Fang et al. 2010) (Peterson et al. 2009). To identify matching proteins we use the software engine INPARANOID (version 3.0; (Remm et al. 2001) (O'Brien et al. 2005)) which distinguishes orthologs from paralogs. It is important to remove inparalogs (which have duplicated since the last common ancestor) as these are likely to have diverged from the original biological role, and probably have not maintained the same interacting partners. A list of orthologous proteins for each species was was generated (Table 1; supplemental table 1) which was then used to identify conserved interacting protein pairs in a large dataset of experimentally determined interactions at BIOGRID (version 2.0.53; (Stark et al. 2006), (Breitkreutz et al. 2008). Two different stringencies were used to establish ortholog matches. 'First, orthologs were selected on a strict 1-1 basis, that is only one ortholog was selected from each cluster, which had a score of 1.0. This approach eliminated all false positives created by divergent gene family members but likely ignores some true positive results. A second approach allowed many to many orthology for cluster members with a score of at least 0.4. This second approach increased the number of predicted interactions at the risk of introducing some false positives. The more stringent approach generated 3337 unique interactions between 800 coffee proteins, while the many-many approach generated an additional 139 proteins and 1250 interactions (for 939 and 4587 total), roughly a 40% increase in interactome size. Perhaps not suprisingly, just a small portion (<10%) of the interactome for each model organism had both orthologs found in coffee. These constitute the conserved proteins common to all eukaryotic organisms, and interactions between these proteins are likely to be conserved as well. The size of the coffee interactome was also considerably smaller than that constructed for Arabidopsis using the same methods (Geisler-Lee et al. 2007). Since the coffee genome is not sequenced, and protein models come from 47,000 sequenced cDNAs, this reduced interactome size indicates the incompleteness of the coffee genome even for conserved genes. Besides *A.thaliana* (8.3%) the recovering rate for the rest of the model organisms between 0.4% and 4.5%. In case of many reference organisms, most notably *A. thaliana, C. elegans, M. musculus* and *R. norvegicus* there are only a small and very incomplete set of protein-protein interactions. *S. cerevisiae* (yeast) on the other hand is a well studied organisms with over 140,000 known interactions in its small genome (6698 proteins)(Breitkreutz et al. 2008). Predicted interactions from orthologous proteins in yeast are thus much more comprehensive than other reference organisms.

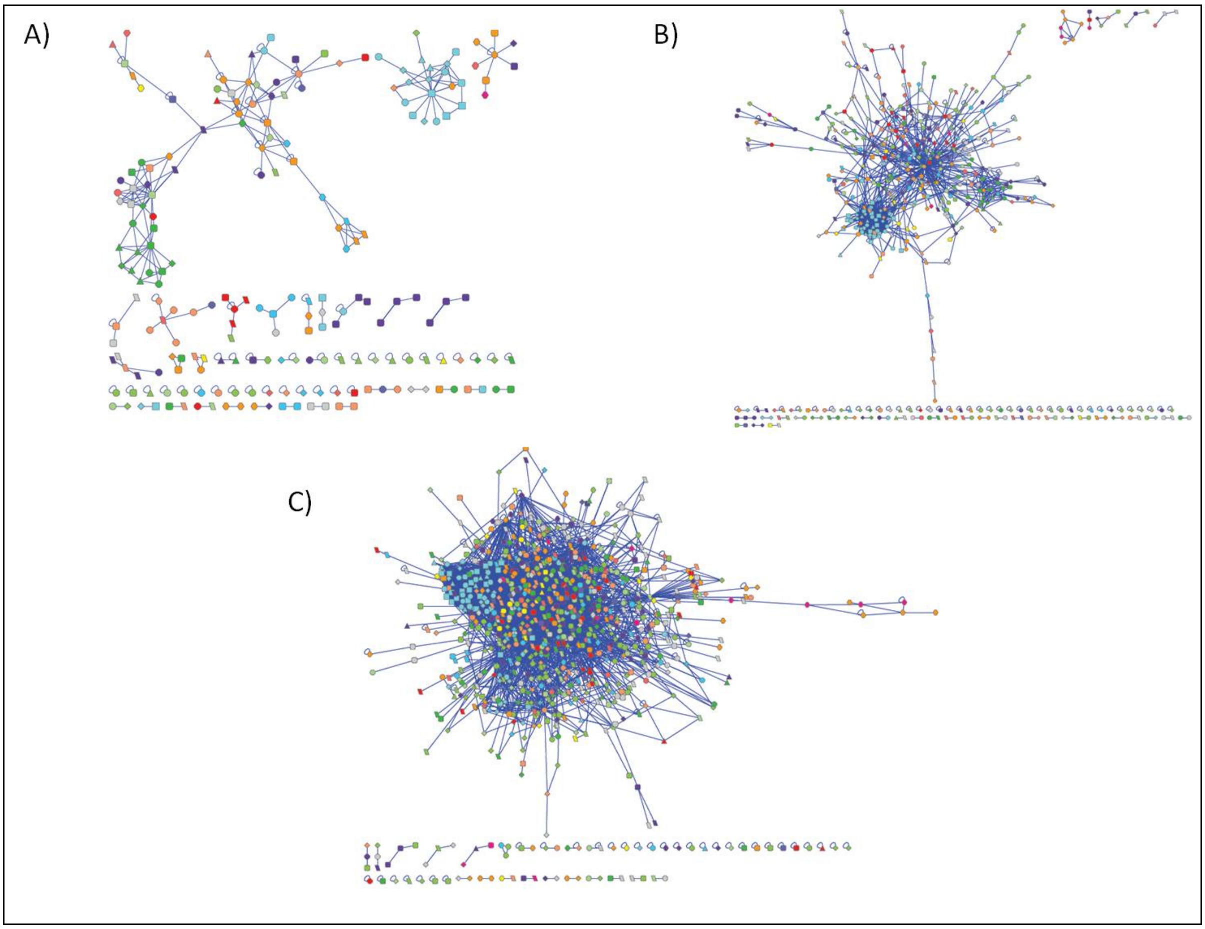

Figure 1 Separation of predicted coffee interactome into three interactomes based on CV value. **A)** ppi with high CV **B)** ppi with medium CV **C)** ppi with low CV. All interactomes are visualized via Cytoscape with organic layout. Node color is based on GOslim molecular function, node shape is based on cellular component and node label is based on coffee id. Edge width is based on CV.

Topology of the Predicted Coffee Interactome

A confidence value (CV) was calculated for each protein-protein interaction. The idea behind asigning CVs is to distinguish interactions predicted from an abundance of experimentally verified reference organisms or multiple experimental methodologies. Based on the confidence value the coffee interactome was seperated into three parts: high (CV>10), medium (CV>2), and low (CV=1) (Figure 1A-C). The majority of the unique ppis have a low CV (3027, about 74% of total interactome) and likely include some false positives. Another 1278 protein interactions have a medium or high CV as they are predicted from multiple sources (Figure 2A). Overall connectivity in the interactome as a whole was determined by examining the frequency distribution of proteins based on the number of predicted partners (Figure 2B). Proteins were subdivided into categories based on connectivity (Figure 2B), free ends (1 interacting partner) were very common (23%), while pipes (2 interacting partners) and small hubs (with 3-10 partners) together make 30% of the total proteins in the dataset, and the overall distribution follows the scale free inverse power law seen in many other biological interactomes, social networks and power distribution grids (R^2 = 0.8773) (supplemental file 2) (Geisler-Lee et al. 2007). Protein hubs with the most interacting partners (50-largest number) are typically highly conserved amongst eukaryotes such as ribosomal proteins, members of the proteosome, or heat shock proteins (Table 2) (Chih-Wen Sun 1997; McIntosh 2009; Oyetunji A. Toogu 2008).

Table 2 The top 20 list of proteins present in super and major hub. List of coffee ID with their corresponding AT id followed by a PFAM description as well the number of interacting partners. Protein sequence identity to yeast was determined via BLASTP.

Coffee ID	AT ID	PFAM Description	Interacting partner	% identity to yeast
CGN-U121410	At4g05320	Ubiquitin family	182	95
CGN-U123074	At5g52640	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase	169	61
CGN-U120144	At4g38630	Ubiquitin interaction motif	92	49
CGN-U119616	At4g36130	Ribosomal Proteins L2, RNA binding domain	88	67
CGN-U124607	At3g12110	Actin	75	83
CGN-U128328	At4g26840	Ubiquitin family	74	51
CGN-U124246	At4g34670	Ribosomal S3Ae family	68	63
CGN-U120876	At5g59240	Ribosomal protein S8e	65	54
CGN-U119944	At2g34480	Ribosomal L18ae protein family	61	51
CGN-U132587	At4g25630	Fibrillarin	59	72
CGN-U122910	At5g48760	Ribosomal protein L13	59	51
CGN-U119943	At1g48830	Ribosomal protein S7e	59	55
CGN-U123729	At3g11940	Ribosomal protein S7p/S5e	59	66
CGN-U120924	At5g35530	KH domain	58	71
CGN-U124856	At1g74050	Ribosomal protein L6, N-terminal domain	58	55
CGN-U122701	At1g43170	Ribosomal protein L3	58	67
CGN-U121974	At3g09630	Ribosomal protein L4/L1 family	57	59
CGN-U121813	At1g67430	Ribosomal protein L22p/L17e	57	61
CGN-U122609	At2g17360	RS4NT (NUC023) domain	57	66
CGN-U122581	At5g04800	Ribosomal S17	57	63

To verify this observation the % identity of the coffee proteins towards yeast of the top 20 most connected as well as least connected proteins was generated via BLASTP. The average of % identity for most connected proteins was 63 whereas for the least connected proteins was 33.65. A t-test was performed to determine significant difference between 20 most connected to 20 least connected proteins (*p-value* 6.87E-10). This is likely due to the conservation of the pathways connected by proteins with high connectivity across all eukaryotes thus making them easier to predict using our methods. The correlation of increased hub size with conservation may be in part due to the length of evolutionary time for these ancient proteins to have established beneficial interactions with other proteins (conservation leading to increased connectivity), or that the large number of interactions increases the effect of purifying selection (connectivity driving conservation) (Berg et al. 2004). It has been shown that highly interconnected proteins tend to evolve more slowly than proteins with few interacting partners, thus favoring the latter hypothesis (Pereira-Leal et al. 2007).

GO Enrichment Analysis

Gene ontology (GO) is a controlled language annotation of genes based on sequence homology or direct experimental evidence. It is organized into three categories: molecular function, biological process and cellular component which provides the user with a searchable index to better understanding towards the proteins in addition to descriptive annotation and raw sequences. Homology based annotation can accurately assign molecular function and localization (component) data due to the presence of modular domains in protein architecture, but frequently gets biological role wrong as individual gene family members take on specialized roles. An improvement to biological role annotation is to first identify one-to-one orthologs, as these are the most likely to have kept the original role, while inparalogs have undergone specialization or functional divergence. To analyze the predicted coffee interactome, we mapped GO slim and GO full annotation to coffee by using a one-to-one ortholog comparison to Arabidopsis (Tair09; www.arabidopsis.org). This dataset included 171unique genes in the coffee interactome, as well as 4930 genes with no predicted partners (provided in supplemental file 2)

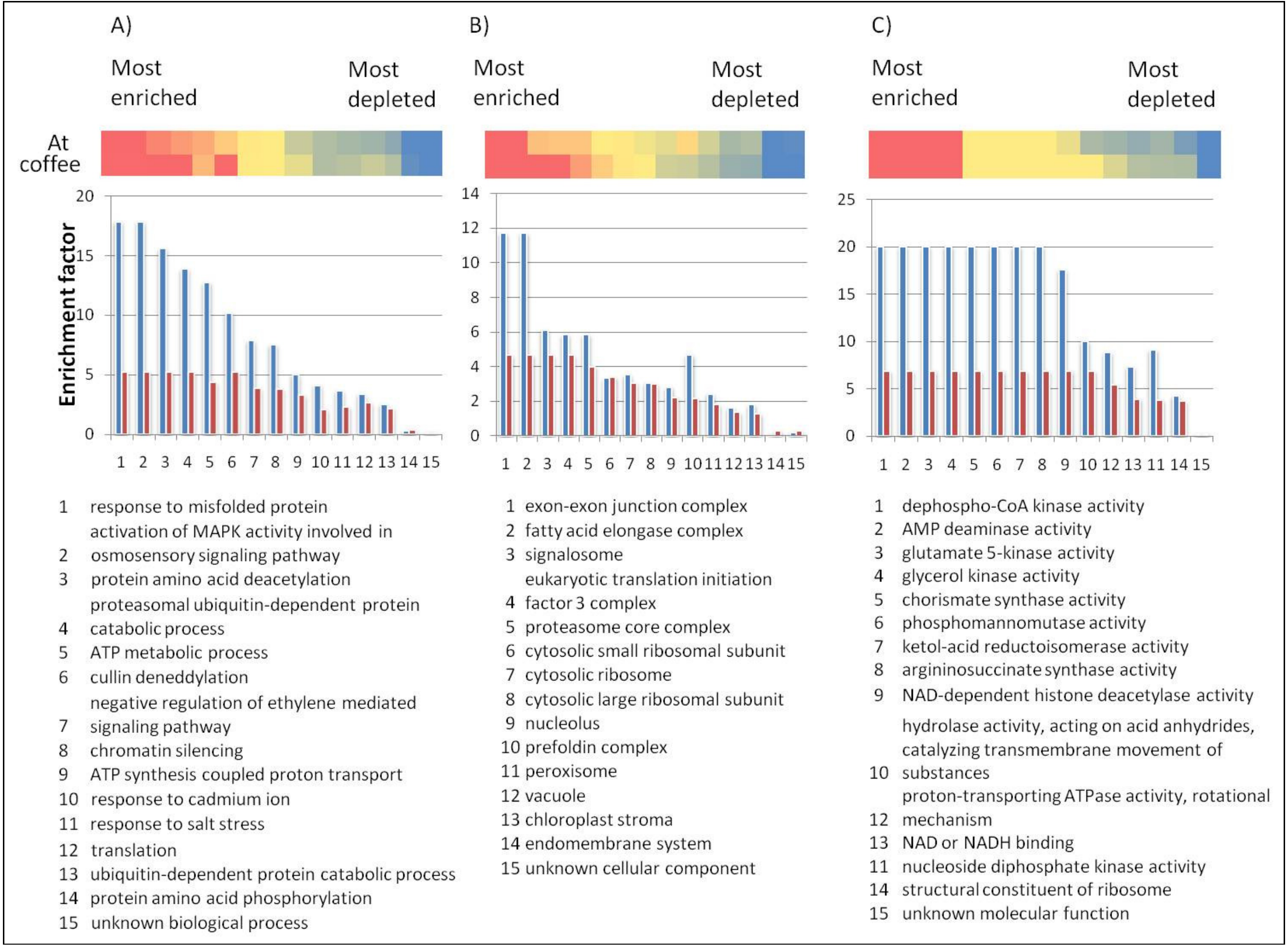

Figure 3 GOfull analysis of predicted coffee interactome. **A)** biological processes, **B)** cellular component and **C)** molecular function. Comparison of predicted coffee interactome to Arabidopsis genome (blue), comparison of predicted coffee interactome to coffee genome (red). On top of each bar is the corresponding enrichment factor (EF) color coded from most enriched (red) to depleted (blue). Each unit was sorted by enrichment factor.

We then analyzed both predicted coffee interactome for enrichment or depletion of GO categories. This was done in order to identify what types of biological pathways were captured by our prediction methods, and what pathways are missing or underrepresented. These will aide users in interpreting the results, account for bias in the predicted interactome and help establish the biological range for which the interactome makes predictions. Examination of GO enrichment also identifies what processes are evolutionarily conserved across plants and eukaryotes. To establish which GO subcategories are significant enriched or depleted, we mapped a total of 4418 Gofull biological process; 5207 GOfull cellular component and 2670 GOfull molecular functions to 939 unique coffee proteins in the predicted interactome. A protein can be assigned more than one GO term hence the large numbers. We used this mapping to further help distinguish proteins in the visualizing tool Cytoscape by asigning them a color and shape based on simplified GO entries (supplemental file 3). To assign enrichment, we compared coffee interacting proteins to all known coffee genes and found significant p-values (after Bonferroni correction for multiple hypothesis testing) in 9 out of 14 biological role subcategories. The coffee known gene set is likely a subset expressed genes in the coffee genome, as it is based on a large EST dataset. However when we compared GO enrichment using the whole Arabidopsis genome as a comparison, we found significant values in 8 out of 14 subcategories (7 of the 9 categories when compared to coffee genome). Among these enriched processes were electron transport/ energy pathways, response to stress, and response to abiotic or biotic stimulus (Figure 3A)

Table 3 Biological processes. Gofull dataset was used to determine best p-values of biological processes (in alphabetical order) of the predicted coffee interactome in comparison to Arabidopsis and coffee genome.

Biological processes	observed in coffee PPI	expected in At genome	Chi2	Enrichment factor	expected in coffee genome	Chi2	Enrichment factor
activation of MAPK	6	0.34	1.4E-22	17.9	21	1.6E-17	5.3
ATP metabolic process	10	0.78	2.1E-25	12.8	6	5.3E-06	5.3
ATP synthesis coupled proton transport	60	11.9	1.6E-44	5.06	60	1.1E-22	3.3
chromatin silencing	32	4.25	2.5E-41	7.53	32	2.8E-16	3.8
cullin deneddylation	12	1.17	1.7E-23	10.2	10	3.2E-07	4.4
regulation of ethylene	22	2.8	1.5E-30	7.87	22	8.4E-12	3.9
proteasomal ubiquitin process	21	1.51	1.2E-56	13.9	6	5.3E-06	5.3
amino acid deacetylation	21	1.34	1.4E-64	15.6	12	1.2E-10	5.3
amino acid phosphorylation	60	207	1.1E-25	0.29	60	4.6E-15	0.4
response to cadmium ion	94	22.7	7.5E-51	4.14	429	1E-100	2.6
response to misfolded protein	6	0.34	1.4E-22	17.9	21	1.6E-17	5.3
response to salt stress	164	44.6	4.2E-72	3.67	164	5.6E-28	2.3
translation	429	127	2E-162	3.37	150	2.7E-23	2.2
ubiquitin-dependent protein process	150	59.2	1.6E-32	2.53	94	8.2E-14	2.1
unknown biological process	155	898	6E-170	0.17	155	1E-219	0.1

There was a depletion of unknown biological processes as expected due to the conservation of the proteins, their orthologs in model organisms tend to be better studied, and annotation by sequence orthology is informative. Out of 939 proteins 614 have known GO categories, whereas 211 have one of the three GO categoris categories noted as unknown. Only 79 proteins have two GO categoroies marked as unknown. However, 3.6% of all unique proteins of the coffee interactome are marked in all three GO categories. What may be surprising is that we did find interactions for several unknown proteins, which highlights the fact that there are still highly conserved genes still have no known biological role. Proteins invoved in phosphorylation of proteins (Tyr and S/T) is highly depleated, indicating that these interactions are likely highly organism specific. There are many inparalogs for these gene families, and only a few one-to-one orthologs. For cellular components we found significant values for 10 out of 16 (coffee genome) and 9 out of 16 (Arabidopsis) subcategories (Figure 3B), including highly enriched cytosol and ribosome localizations for predicted interactors in comparison with Arabidopsis and coffee genome, while the endomembrane system was significantly depleated. Three subcategories of molecuar functions show significant values, most especially structural molecules are highly enriched in the interactome in comparison to Arabidopsis and coffee genomes (Figure 3C).

Table 4 Cellular component. Gofull dataset was used to determine best p-value s of biological processes (in alphabetical order) of the predicted coffee interactome in comparison to Arabidopsis and coffee genome.

Cellular component	**observed in coffee PPI**	**expected in At genome**	**Chi2**	**Enrichment factor**	**expected in coffee genome**	**Chi2**	**Enrichment factor**
chloroplast stroma	372	205	1E-32	1.81	202	7E-06	1.37
large ribosomal subunit	144	46.7	2E-46	3.08	170	9E-28	2.25
cytosolic ribosome	504	143	8E-207	3.53	504	2E-161	3.09
small ribosomal subunit	172	51.5	7E-64	3.34	144	3E-44	3.01
endomembrane system	37	366	5E-71	0.1	326	2E-160	0.29
eukaryotic translation initiation factor 3	10	1.71	2E-10	5.86	24	2E-13	4.01
exon-exon junction	6	0.51	2E-14	11.7	24	8E-17	4.68
fatty acid elongase	6	0.51	2E-14	11.7	6	3E-05	4.68
Nucleolus	170	60.2	6E-46	2.82	18	0.0008	2.16
Peroxisome	138	56.5	1E-27	2.44	138	3E-13	1.83
prefoldin complex	18	3.84	5E-13	4.69	172	3E-66	3.41
proteasome core complex	24	4.09	7E-23	5.86	10	8E-08	4.68
Signalosome	24	3.92	4E-24	6.12	6	3E-05	4.68
unknown cellular component	326	1423	5E-255	0.23	37	8E-16	0.29
Vacuole	202	123	7E-13	1.64	372	1E-06	1.28

We then examined specific annotation sub-categories for enrichment or depletion (Top significant biological roles in Table 3). These were similar for comparisons to the Coffee and Arabidopsis genomes, and included translation (*p-value* 2.3e-162), protein amino acid deacetylation (*p-value* 1.4e-64) and response to salt stress (*p-value* 4.2e-72). Not surprisingly, one of the most conserved set of interactions was for the ribosome (cytosolic small ribosomal subunit *p-value* 7.5e-64, cytosolic large ribosomal subunit (*p-value* 2.45e-46; Table 4, structural constituent of ribosome *p-value* 9.6e-86; Table 5).

Coffee Ribosome Subnetwork

The most conserved densly interconnected subnetwork (cluster) is represented by 42 proteins of the ribosome compartment (Figure 4). This is a highly interconnected subnetwork of 812 interactions. By selecting the first neighbours of the ribosome cluster an extended network of 219 proteins with 2004 interactions is created. Specifically cytosolic ribosome with its cytosolic small and large ribosomal subunits which is a part of the structural constituent of ribosomes.

Ribosomes are one of the most ancient and necessary machineries in cell essential for cell growth in all organisms (Strunk and Karbstein 2009) (Dunkle and Cate 2010). In bacterial model organsism *Escherischia coli* more than 50 proteins including ribosomal RNA such as 16S ribosomal RNA, S1-S21 in small ribosomal subunit and L1-L36 in the large ribosomal subunit, are involved in the ribosomal translation machinery. In eukaryotic model organsism *Saccharomyces cerevisiae*, so called accessory factors such as ATPases, GTPases as well as exonucleases play an important role in the ribosome assembly. It is not surprising to see protein interactions established between the ribosome subunit proteins to enzymes such as exonucleases as well as S-adenosyl methionine which are essential for utilizing energy from interactions of pre-ribosomes (Strunk and Karbstein 2009). Orthologs of these proteins are present in coffee, and thus we have predicted which of these interactions occur in coffee, and generated a network model of the coffee ribosome/protein synthesis cluster

The distribution of biological processes of the proteins that are first neighbours of coffee ribosome cluster are similarily involved in protein metabolism, but also in ribosome biogenesis, translational termination, histone deacetylation, ubiquitin-dependent protein catabolic process and rRNA modification to name the next significant ones. Interestingly, a DEAD/DEAH box helicase (CGN-U121831) is another protein interacting partner from the first neighbor selection with two interactions worth mentioning. DexH/D proteins are referred to as RNA-dependent ATPases and are involved in disconnecting RNA-protein interactions (Strunk and Karbstein 2009). Eukarotic initiation factor 4E (CGN-U128511) belongs to translational machinery and functions as DNA or RNA binding (CV 24). Elongation factor Tu GTP binding domain (CGN-U 122750) is involved in the stress response and functions as nucleic acid binding site (CV18).

Table 5 Molecular functions. Gofull dataset was used to determine best p-value s of biological processes (in alphabetical order) of the predicted coffee interactome in comparison to Arabidopsis and coffee genome.

Molecular function	**observed in coffee PPI**	**expected in At genome**	**Chi2**	**Enrichment factor**	**expected in coffee genome**	**Chi2**	**Enrichment factor**
AMP deaminase activity	4	0.2	1.6E-17	20.08	154	1.3E-68	3.69
argininosuccinate synthase activity	3	0.15	1.6E-13	20.08	4	7.1E-06	6.9
chorismate synthase activity	3	0.15	1.6E-13	20.08	16	3E-09	3.94
dephospho-CoA kinase activity	4	0.2	1.6E-17	20.08	194	2E-128	0.26
glutamate 5-kinase activity	4	0.2	1.6E-17	20.08	22	5.1E-19	5.42
glycerol kinase activity	4	0.2	1.6E-17	20.08	7	2.8E-09	6.9
hydrolase activity, transmembrane movement of substances	7	0.7	4.4E-14	10.04	10	4.7E-06	3.83
ketol-acid reductoisomerase activity	3	0.15	1.6E-13	20.08	4	7.1E-06	6.9
NAD or NADH binding	16	2.19	1E-20	7.302	3	0.0001	6.9
NAD-dependent histone deacetylase activity	7	0.4	1.3E-25	17.57	4	7.1E-06	6.9
nucleoside diphosphate kinase activity	10	1.1	1.8E-17	9.127	7	2.8E-09	6.9
phosphomannomutase activity	3	0.15	1.6E-13	20.08	4	7.1E-06	6.9
proton-transporting ATPase activity	22	2.49	3.8E-35	8.835	3	0.0001	6.9
constituent of ribosome	154	36.4	9.6E-86	4.23	3	0.0001	6.9
unknown molecular function	194	816	2E-150	0.238	3	0.0001	6.9

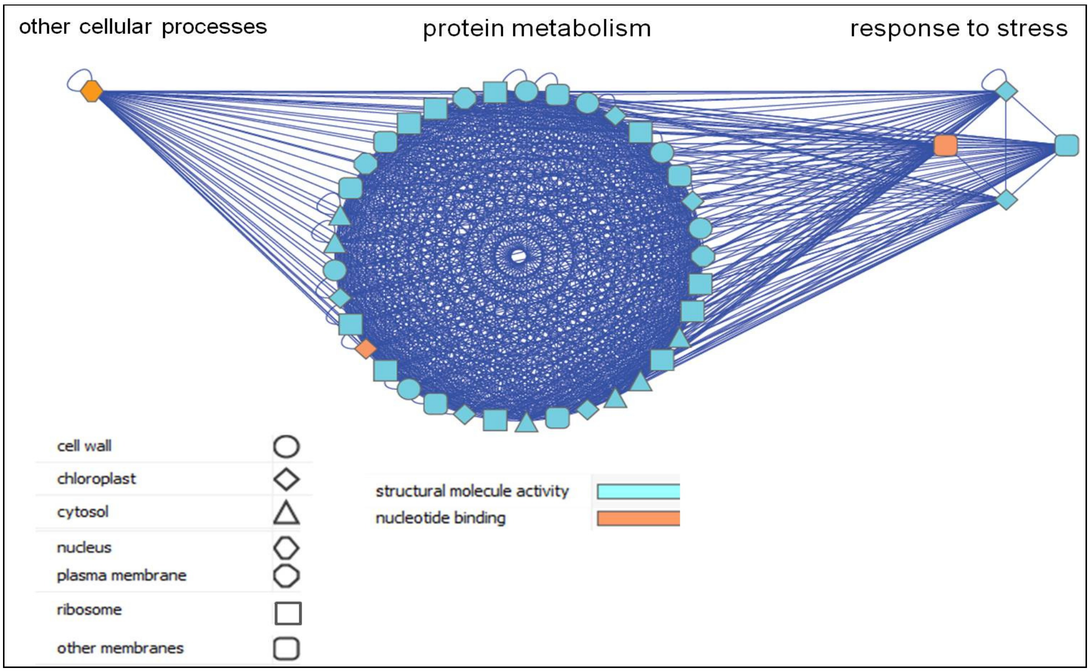

Figure 4 Cluster analysis with MCODE of predicted coffee interactome. Cluster 1 is visualized via Cytoscape. Layout is based on biological processes. Node shape based on cellular component, node color based on molecular function. Node description based on PFAM description.

Coffee Proteasome Subnetwork

The second most conserved subnetwork in coffee is represented by 27 proteins with 191 interactions and reassembles the proteasome machinery in coffee. The proteasome cluster is involved in the ubiquitin protein catabolic process, is part of the proteasome core complex and functions in the peptidase activity (Figure 5). This cluster is reasembled by many proteins which are very important in protein degradation. Interestingly, core histone H2A/H2B/H3/H4 is part of the subnetwork and indicates a nonproteolytic function of the proteasome machinery. Even though proteasomes are primarly involved in protein translocation and degaradation mainly through ubiquitination, it has been reported to have additional roles in processes such as DNA repair and chromatin remodeling (Demartino and Gillette 2007). Three coffee proteins (CGN-U121305, CGN-U120981, CGN-U123237) contain a PCI domain which is thought to be involved as scaffolds of the proteasome lid, COP9 signalosome as well as the eukaryotic translation initioation factor-3 (elF3) (Pick et al. 2009). Proteins with PCI domain interact with each other as well as with proteins or subunits such as SAC3/GANP/Nin1/mts3/eIF-3 p25 family (CGN-U122034), ubiquitin family (CGN-U122810), Mov34/MPN/PAD-1family (CGN-U131698) ATPase family associated with various cellular activities (CGN-U124357) and Ankyrin repeat (CGN-U130477). These interactions have high confidence, with CVs of 108, 12, 108, 36 and 9 respectively corresponding to multiple lines of experimental evidence from different reference organisms. Since proteasome machinery is ATPase dependent process which requires hydrolysis activity the proteasome cluster it is not surprising to find 3 ATPase family associated (AAA) proteins (CGN-U122970, CGN-U124357, CGN-U123280). We reconstructed three out of six ATPases which are part of the 20S proteasomal machinery in eukaryotes (Rabl et al. 2008). These 3 AAA proteins themselves undergo 41 protein interactions including self-interactions and hetero-interactions with proteins such as SAC3 family, proteasome A- and B-type as well as proteins containing PCI domains.

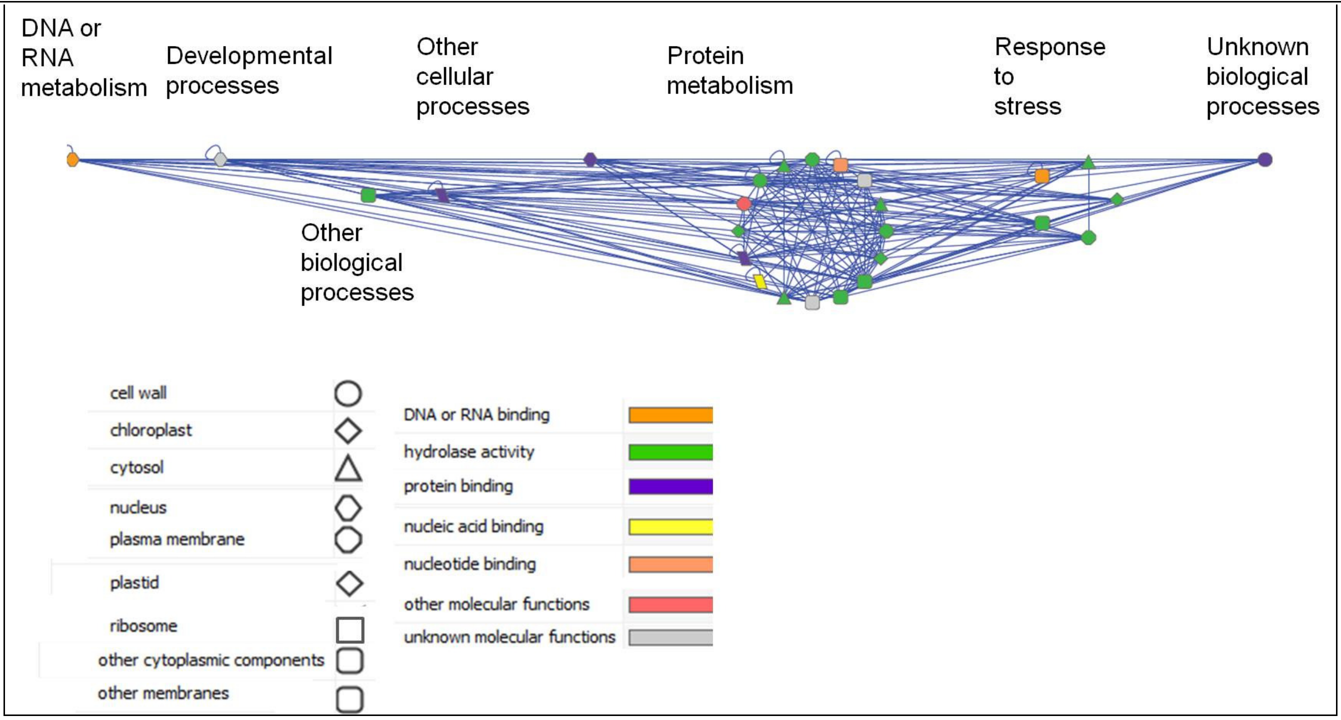

Figure 5 Cluster analysis with MCODE of predicted coffee interactome. Cluster 2 is visualized via Cytoscape. Layout is based on biological processes. Node shape based on cellular component, node color based on molecular function. Node description based on PFAM description.

Coffee Wax Biosynthesis Subnetwork

One interesting and unexpectedly conserved small cluster was MCODE cluster 5. With only 6 proteins and 14 interactions it is a very small subnetwork compared to the previous two cluster groups (Figure 6). This cluster involves malate metabolic processes, and one-carbon metabolism.

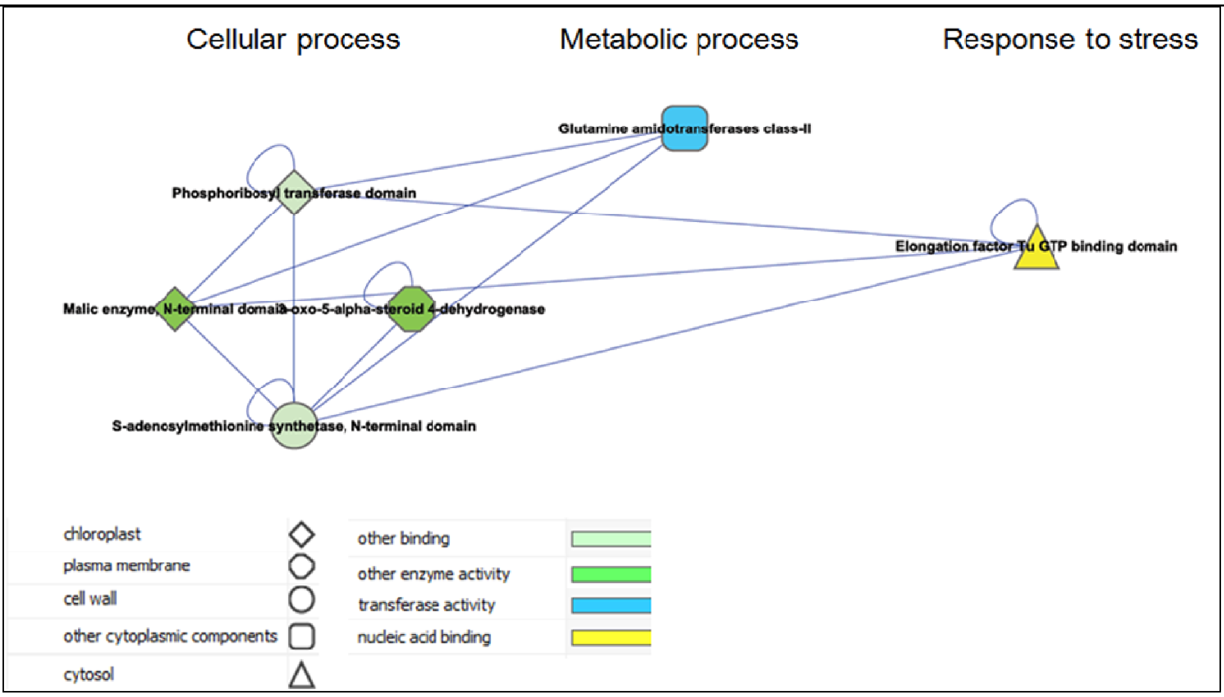
Cellular process
Metabolic process
Response to stress
Glutamine amidotransferases class-II
Phosphoribosyl transferase domain
Elongation factor Tu GTP binding domain
Malic enzyme, N-terminal domain
3-oxo-5-alpha-steroid 4-dehydrogenase
S-adenosylmethionine synthetase, N-terminal domain
chloroplast ◇
plasma membrane ⬡
cell wall ○
other cytoplasmic components □
cytosol △
other binding
other enzyme activity
transferase activity
nucleic acid binding

Figure 6 Cluster analysis with MCODE of predicted coffee interactome. Cluster 5 is visualized via Cytoscape. Layout is based on biological processes. Node shape based on cellular component, node color based on molecular function. Node description based on PFAM description.

One member of the cluster is mainly part of the fatty acid elongase complex and endoplasmatic reticulum membrane which are involved in malate dehydrogenase, malic enzyme activity and oxidoreductase activity. The biosynthesis of wax material is part of the cuticle layer and essential for plants to survive on land. Interstingly, all proteins of cluster 6 are mainly found in *S. cerevisiae*, *E. coli*, *C.elegans* or *D. melaongaster*. The hydrophobic layer prevents dehydration as well as acts as an repellant agent of hydrophilic components. The biosynthesis starts out with the Acetyl-Coenzyme A, a product of the Glycolysis/TCA-cycle, to build a pool of fatty acids in the plastids such as leukoplasts via the fatty acid biosynthesis. Once C16 and C18 fatty acids has been created part of them will be translocated to the endoplasmatic reticulum (ER) in order to get additional decoration such as formation of double bonds, addition of ester-groups or hydroxyl-groups (Samuels, Kunst and Jetter 2008). One memeber of the was biosynthesis cluster is ,3-oxo-5-alpha-steroid 4-dehydrogenase (Enoyl-CoA reductase or ECR; CGN-U119801). In Arabidopsis, the ortholog gene *CER10* is responsible of a cuticle phenotype and mutants have a reduced level of all wax components in Arabidopsis (Zheng, Rowland and Kunst 2005). Extending the search of the wax biosynthesis protein (CGN-U119801) we identify a predicted interaction with a β-keto acyl reductase (KCR, CGN-U123973) an enzyme which is known to be involved in wax production during the synthesis of very long-chain fatty acids in the ER. It is an ortholog to *YBR159w* in yeast and a BlastN search revealed a 84% identity with putative 3-ketoacyl-CoA reductase of Brassica napus and 73% identity a predicted protein of Poplus trichocarpa, so is likely conserved throughout angiosperms.

REFERENCES

Ashburner MBC, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25

Berg J, Lassig M, Wagner A 2004. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evolutionary Biology 4: 51

Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bähler J, Wood V, Dolinski K, Tyers M 2008. The BioGRID Interaction Database: 2008 update. Nucleic Acids Research 36: D637-D640

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M 2007. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research 18: 188-196

Chan Zhou YY, Phuongan Dam,Ying Xu 2010. Identification of Novel Proteins Involved in Plant Cell-Wall Synthesis Based on Protein–Protein Interaction Data. journal of proteome research 9: 5025-5037

Chih-Wen Sun SGaJC 1997. A model for the evolution of polyubiquitin genes from the study of Arabidopsis thaliana ecotypes. Plant Molecular Biology 34: 745-758

Chris Burge SK 1997. Prediction of Complete Gene Structures in Human Genomic DNA. Journal of Molecular Biology: 78-94

Curwen V 2004. The Ensembl Automatic Gene Annotation System. Genome Research 14: 942-950

Damon A 2000. A review of the biology and control of the coffee berry borer, Hypothenemus hampei (Coleoptera: Scolytidae). Bulletin of Entomological Research 90: 453-465

Demartino G, Gillette T 2007. Proteasomes: Machines for All Reasons. Cell 129: 659-662

Dietz KJ, Jacquot JP, Harris G 2010. Hubs and bottlenecks in plant molecular signalling networks. New Phytologist 188: 919-938

Dunkle JA, Cate JHD 2010. Ribosome Structure and Dynamics During Translocation and Termination. Annual Review of Biophysics 39: 227-244

Fang G, Bhardwaj N, Robilotto R, Gerstein MB 2010. Getting Started in Gene Orthology and Functional Analysis. PLoS Comput Biol 6: e1000703

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A 2009. The Pfam protein families database. Nucleic Acids Research 38: D211-D222

Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJP, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A, Searle S 2007. Ensembl 2008. Nucleic Acids Research 36: D707-D714

Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH, Geisler M (2007) A Predicted Interactome for Arabidopsis. Plant Physiol. 145: 317-329

Hogue GDBaCW 2003. An automated method for finding molecular complexes in large protein interaction networks. Bmc Bioinformatics 4

Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P 2009. Ensembl 2009. Nucleic Acids Research 37: D690-D697

Hudson ME 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. Molecular Ecology Resources 8: 3-17

Lewis ACF, Saeed R, Deane CM 2010. Predicting protein–protein interactions in the context of protein evolution. Molecular BioSystems 6: 55

Lin CW, Mueller LA, Mc Carthy J, Crouzillat D, Petiard V, Tanksley SD 2005. Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. Theoretical and Applied Genetics. 112: 114-130

Liu JX, Howell SH 2010. bZIP28 and NF-Y Transcription Factors Are Activated by ER Stress and Assemble into a Transcriptional Complex to Regulate Stress Response Genes in Arabidopsis. The Plant Cell Online 22: 782-796

Lovell SC, Robertson DL 2010. An Integrated View of Molecular Coevolution in Protein-Protein Interactions. Molecular Biology and Evolution 27: 2567-2575

Madupu R, Brinkac LM, Harrow J, Wilming LG, Bohme U, Lamesch P, Hannick LI 2010. Meeting report: a workshop on Best Practices in Genome Annotation. Database 2010: baq001-baq001

McIntosh JRWaKB 2009. How Common Are Extraribosomal Functions of Ribosomal Proteins? Molecular Cell 34: 3-11

Metzker ML 2009. Sequencing technologies- the next generation. Nature Reviews Genetics 11: 31-46

Mueller LA 2005. The SOL Genomics Network. A Comparative Resource for Solanaceae Biology and Beyond. Plant Physiology 138: 1310-1317

O'Brien KP, Remm M, Sonnhammer ELL 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Research 33: D476-D480

Oyetunji A. Toogu DCDaBCF 2008. The Hsp90 Molecular Chaperone Modulates MultipleTelomerase Activities. Molecular and Cellular Biology 28: 457-467

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob ur R, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS 2009. The Sorghum bicolor genome and the diversification of grasses. Nature 457: 551-556

Peregrín-Alvarez JM, Sanford C, Parkinson J 2009. The conservation and evolutionary modularity of metabolism. Genome Biology 10: R63

Pereira-Leal J, Levy E, Kamp C, Teichmann S 2007. Evolution of protein complexes by duplication of homomeric interactions. Genome Biology 8: R51

Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Sali A 2009. Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Science 18: 1306-1315

Pick E, Hofmann K, Glickman MH 2009. PCI Complexes: Beyond the Proteasome, CSN, and eIF3 Troika. Molecular Cell 35: 260-264

Rabl J, Smith DM, Yu Y, Chang SC, Goldberg AL, Cheng Y 2008. Mechanism of Gate Opening in the 20S Proteasome by the Proteasomal ATPases. Molecular Cell 30: 360-368

Remm M, Storm CEV, Sonnhammer ELL 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of Molecular Biology 314: 1041-1052

Rothberg JM, Leamon JH 2008. The development and impact of 454 sequencing. Nat. Biotechnol. 26: 1117-1124

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 13: 2498-2504

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M 2006. BioGRID: a general repository for interaction datasets. Nucleic Acids Research 34: D535-D539

Strunk BS, Karbstein K 2009. Powering through ribosome assembly. RNA 15: 2083-2104

Vidal RO, Mondego JMC, Pot D, Ambrosio AB, Andrade AC, Pereira LFP, Colombo CA, Vieira LGE, Carazzolle MF, Pereira GAG 2010. A High-Throughput Data Mining of Single Nucleotide Polymorphisms in Coffea Species Expressed Sequence Tags Suggests Differential Homeologous Gene Expression in the Allotetraploid Coffea arabica. Plant Physiology 154: 1053-1066

Wendl MC, Dear S, Hodgson D, Hillier L 1998. Automated Sequence Preprocessing in a Large-Scale Sequencing Environment. Genome Research 8: 975-984

Wilming L, Harrow J 2009. Gene Annotation Methods. pp 121-136

Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ, Baxter I 2007. An “Electronic Fluorescent Pictograph” Browser for Exploring and Analyzing Large-Scale Biological Data Sets. PLoS ONE 2: e718