Research Journal of Biological Sciences 10 (4-5): 44-55, 2015
ISSN: 1815-8846

Gene Comparison Between *Arabidopsis thaliana*, *Prunus mume* and *Prunus persica*

Joanna M. Cross
Department of Horticultural Science, Faculty of Agriculture,
Inonu University (Battalgazi Campus), 44000 Battalgazi, Malatya, Turkey

Abstract: Apricot (*Prunus armeniaca*) is a stone fruit consumed fresh or dried. Turkey is the top world producer with the Malatya area supplying over half of the crop. Market demands are stringent as apricot trees need to be resistant to both heat and cold while producing fruits satisfying the customer. Therefore, crop improvement involves many parameters. The genomes of peach (*Prunus persica*) and Japanese apricot (*Prunus mume*) have been sequenced. Both belong to the same family as apricots. Consequently, around 250 genes were collected for both species along with the reference Arabidopsis. Both targeted and non targeted approaches were applied to diversify the range of protein functions covered. Thus, a set of genes involved in amino acid metabolism was studied along with a second group selected based on the phenotype conferred. Comparison of the three plants shows that gene allocation to a given function is conserved assignment of a clear gene to gene correspondence between organisms is a delicate task and clear gene counterparts do not necessarily share the same cellular compartment.

Key words: Prunus, Arabidopsis, gene comparison, cold response, amino acid synthesis, protein localization

INTRODUCTION

Apricot (*Prunus armeniaca*) is a stone fruit from the same family (Rosaceae) and genus (*Prunus*) as plum, peach, cherry and almond. It is a diploid with a relatively small genome size of 590 Mb (Hagen *et al*., 2002). Trees are cultivated in the Mediterranean Basin, the former Soviet Union countries, Iran, China, South Africa and the United States (Asthma, 2007). Turkey is the top world producer. The Malatya region (38°21'N/38°17'E) contributes over half of the country's apricots with its 10 million trees.

Trees bloom in the middle of March while leaves develop a month after flowering. As a result, the first stages of fruit development rely on reserves accumulated the previous season. Apricot is a climacteric fruit which means that ripening is accompanied by an ethylene burst. Fruits mature at different periods depending on the variety grown. Early ripening fruits are harvested in June while mid and late ripening cultivars are picked in July and end of August, respectively. Hence, harvest season spreads over several months.

Temperatures drop below freezing during the Winter while reaching 40°C in the Summer. Consequently, trees have to be hardy to both cold and heat. In addition, the species is sensitive to diseases with Sharka (Plum Pox Virus) being the major threat (Sochor *et al*., 2012). So far, the disease is absent from the Malatya region though present in Turkey. However, night frosts occur until late April. They are particularly damaging in March at the time of flowering. For instance, in 2014, 3 days of cold, end of March, destroyed 90% of the crop. Since, Spring frosts occur roughly every 3 years, apricot production is quite irregular.

Apricots are appreciated for their taste. Fruit quality depends on sugar and organic acid content (Drogoudi *et al*., 2008). In addition, over 200 volatile compounds have been identified (Gonzalez-Aguero *et al*., 2009). Panel studies have shown that a combination of 18 compounds can mimic apricot taste (Greger and Schieberle, 2007). Apricots harbor high levels of antioxidants including vitamin A. The molecules are suggested to have therapeutic effects on a variety of pathological conditions such as cancer, diabetes and neurodegenerative or cardiovascular diseases. Consequently, there is an interest in identifying apricot varieties with enhanced antioxidant content (Ruiz *et al*., 2005a, b).

Production of new apricot varieties demands a careful balance between farmer satisfaction and customer approval. A regular and optimal ratio between yield and size must be ensured while taste and nutritional content are preserved. Consequently, parameters of importance span a large number of reactions as well as various

signaling pathways. So far, efforts have mainly focused on assessing genetic variation with Random Amplified Polymorphic DNA (Takeda *et al*., 1998), Amplified Fragment Length Polymorphism (Hagen *et al*., 2002; Yuan *et al*., 2007; Krichen *et al*., 2008), Inter-Simple Sequence Repeat markers (Yilmaz *et al*., 2009) and Sequence Characterized Amplified Regions. The results are used for conservation efforts and variety classification (Hagen *et al*., 2002; Bourguiba *et al*., 2012), quantitative trait locus identification (Salazar *et al*., 2014) and as an information basis to breed new varieties (Asthma, 2012). Efforts are under way to develop cultivars with cold and/or Sharka resistance.

At the molecular level, genes differentially expressed during fruit maturation were identified (Geuna *et al*., 2005). Moreover, correlations were made between volatile composition and expression patterns of genes involved in aroma synthesis (Gonzalez-Aguero *et al*., 2009). Recently, the species entered the omics field. Indeed, transcriptional analysis of fruit maturation was performed by collecting expressed sequenced tags (Grimplet *et al*., 2005), designing expressed sequenced tag based microarrays (Li *et al*., 2012) and using chips constructed for peach (Manganaris *et al*., 2011). The data were complemented by a proteomic time-course of fruit development. Moreover, the BGI lists a sequencing project in its early phase. Meanwhile, the genomes of peach (*Prunus persica*) and of Japanese apricot (*Prunus mume*) have been sequenced (Verde *et al*., 2013) and the pathways of peach collected into PeachCyc (Jung *et al*., 2014).

Unfortunately, resources are still incomplete for apricot mainly due to the lack of a sequenced genome. Therefore, it would be beneficial to use data from other species. A logical first choice consists of the two closely related species *Prunus persica* and *Prunus mume*. Those are related to the extent that microarrays designed for *Prunus persica* or *Prunus mume* also function for apricots (Manganaris *et al*., 2011; Li *et al*., 2012). As mentioned, all three species are classified in the genus *Prunus*. However, both *Prunus mume* and *Prunus armeniaca* belong to the same subgenus *Prunus* and section *Armeniaca* while exhibiting similar fruit and tree morphology. Nonetheless, markers clearly separate each plant into different species (Hagen *et al*., 2002). On the other hand, *Prunus persica* is either classified in a separate subgenus *Amygdalus* or listed in the subgenus *Prunus* albeit in a separate section from apricot (Shi *et al*., 2013).

The model plant Arabidopsis represents an appealing second resource with its wealth of experimental data. Thus, a subset of genes was compared between *Prunus persica*, *Prunus mume* and *Arabidopsis* to assess cautions required when translating data from one organism to another. Genes were first assembled for a specific function, i.e., amino acid metabolism. Second, databases of Arabidopsis mutants were screened to identify genes based on phenotype. This ensured that a diversity of gene functions was collected.

MATERIALS AND METHODS

Selection of the amino acid metabolism pathways: The curated pathways Aracyc12.0 (http://pmn.plantcyc.org/) (Mueller *et al*., 2003) and Peachcyc1.0 (http://www.rosaceae.org/) (Jung *et al*., 2014) were used to collect genes for *Prunus persica* and Arabidopsis. They were checked for consistency and redundancies. On the whole, the two databases are consistent except in the following cases. Regarding methionine recycling, Peachcyc harbors two versions of the S-adenosyl-L-methionine cycle while Aracyc contains one version in addition to the Yang cycle. The latter salvages the sulfur of 5'-Methylthioadenosine (MTA) synthesized during the production of ethylene (Sauter *et al*., 2013) and therefore seems quite relevant to developing apricot fruits. Effectively, combination of the early steps of ethylene biosynthesis with the methionine salvage route of Peachcyc produced the Arabidopsis Yang cycle. Moreover, Peachcyc lists two conversions of adenosyl-homocysteine to homocysteine, one direct, one through ribosylhomocysteine. However, the Arabidopsis counterparts of the annotated peach genes potentially converting ribosylhomocysteine were shown to catalyze a Yang cycle reaction with poor or no affinity for their hypothesized substrate (Siu *et al*., 2008). Hence, only the direct conversion of adenosyl-homocysteine to homocysteine was considered.

Regarding cysteine biosynthesis, Peachcyc lists an interconversion route between homocysteine and cysteine which is absent from Aracyc. However, the conversion of homocysteine to cysteine with cystathionine beta synthase seems specific to mammals while plants synthesize homocysteine and then methionine from cysteine (Kushwaha *et al*., 2009). Hence, the interconversion route was not considered.

Finally, Peachcyc lists two routes to convert prephenate to phenylalanine, one through arogenate, another via phenylpyruvate. However, the arogenate pathway appears to be the major route with the phenylpyruvate bypass potentially having a minor role (Tzin and Galili, 2010). Therefore, both pathways were tentatively kept.

Selection of the genes involved in each reaction: Genes were collected as listed in Aracyc and Peachcyc. The final list was determined based on literature, sequence alignments and annotations. Indeed, *Arabidopsis* genes were checked in TAIR (http://www.arabidopsis.org/) Lamesch *et al*., 2012) for experimental evidence regarding their function and localization. They were also blasted (Altschul *et al*., 1990) against a Prunus+*Arabidopsis thaliana* database to verify that the *Prunus persica* counterparts were selected correctly and to identify *Prunus mume* homologues. Peach genes with no Arabidopsis homologues were blasted against the non-redundant database to validate the annotation. All blastings were performed with nBLAST using default settings (Match/Mismatch Scores, 2/-3; Gap costs, existence 5, extension 2). Reactions were associated with a loci rather than a mRNA. Hence, alternative splicing variants were not considered.

Selection of genes based on phenotype: Phenotypes were screened in the Chloroplast 2010 (Ajjawi *et al*., 2010; Lu *et al*., 2008; Lu *et al*., 2011), the RIKEN Phenome (Kuromori *et al*., 2004, 2006) and the Arabidopsis Stress Responsive (Borkotoky *et al*., 2013) databases. The Chloroplast 2010 mutants were screened for altered C/N ratio in seeds or modified starch levels in leaves using default z scores and the option "two siblings with altered parameter". Moreover, hits were checked for consistency of results within two different knock-out lines. Morphological parameters were examined to eliminate plants terminally diseased. The RIKEN Phenome Project was screened for smaller siliques and decreased yields. Plants noted as sterile or small were eliminated to ensure that defects were due to silique development and not to growth problems. The Arabidopsis stress database was screened for cold responses. The literature supplied with each hit was used to identify potential signaling partners. Both metabolic and phenotype base genes were run through the Stanford Interactome database (Jones *et al*., 2014) selecting partners identified in two primary and two confirmation screens.

Gene to gene correspondence assignment: *Arabidopsis* genes identified in the phenotype based search were blasted as described above against a Prunus+*Arabidopsis thaliana* database. Next, gene to gene correspondence was determined relatively. All genes which aligned over more than a conservative 16% of the target were selected. The DNA sequences of *Prunus mume*, *Prunus persica* and Arabidopsis were then aligned using Clustal Omega (Goujon *et al*., 2010; Sievers *et al*., 2011) and default parameters. The identity matrixes and alignments were used to assign the correct Prunus homologues to their Arabidopsis counterparts. Those were considered to be genes aligning preferentially and solely with the Arabidopsis target. Next, *Prunus persica* genes were blasted against Arabidopsis to check that they aligned preferentially with their selected homologue.

Localization: Reactions were assigned to a compartment based on previous literature, TAIR annotations, protein sequence alignment and targeting prediction programs. The prediction programs WoLF PSORT and TargetP (Emanuelsson *et al*., 2007) were used to sort out reactions with multiple cell locations. WoLF PSORT aligns a target with proteins of known compartmentalization while TargetP specializes in detecting mitochondrial, chloroplasticand secretory targeting sequences. Therefore, both approaches are complementary. Default parameters were used with the setting on "plant sequences". NucPred (Brameier *et al*., 2007) and PredPlantPTS1 (Reumann *et al*., 2012) were used to discriminate nuclear and peroxisomal targeted proteins, respectively.

Statistics and probabilities: All calculations were performed in R Version 3.1.2 for Windows 64 bit with the packages gmp, lpSolveAPI and Xnomial. The number of genes per reaction was viewed as a distribution problem of a determined number of isozymes between different functions. The number of possible ways of splitting n objects into k groups of n_1, n_2, n_k elements is given by $n!/n_1!n_2!...n_k!$. This was calculated for observed gene distributions in all three species with package gmp. The latter handles very large integers and decimals. Moreover, it expresses decimals as quotients thus enabling high precision calculations. Therefore, all calculations were done on quotients with a number obtained at the last step for rounding.

Next, it was sought to express results on a relative basis for comparison purposes. A first choice was the Stirling number of second kind which measures the number of ways of splitting n objects into k groups with at least one element. Unfortunately, this could not be calculated reliably due to the size of n ranging from 193-233. Therefore all results were expressed relative to the most frequent distribution in each species. The most frequent combination is obtained by minimizing the denominator in $n!/n_1!n_2!...n_k!$. Moreover, the latter is smaller with reduced n values. Therefore, the assumption was made that a solution could be found with all groups

harboring 1-6 genes. Thus, the denominator could be expressed as $(1!)^{x1}\times(2!)^{x2}\times(3!)^{x3}\times(4!)^{x4}\times(5!)^{x5}\times(6!)^{x6}$. As $\log_{10}(x)$ varies in the same direction as x the $\log_{10}$ form was used, i.e., $x_1\times\log_{10}(1!)+x_2\times\log_{10}(2!)+x_3\times\log_{10}(3!)+x_4\times\log_{10}(4!)+x_5\times\log_{10}(5!)+x_6\times\log_{10}(6!)$.

The function was minimized with package lpSolveAPI using constraints $\sum x_i = 103$ and $\sum ixx_i = 193$ (*Prunus mume*), 201 (*Prunus persica*) and 233 (Arabidopsis) with all x_i being integers between 0 and 103. Identical results were obtained by replacing $\log_{10}(1)$ by a small value. In addition, a few simple solutions were confirmed by manual checking. Additional constraints were added to determine the maximum distributions for a given number of reactions with 2 genes or when the observed number of reactions with >3 genes is kept.

A given random gene distribution was calculated for two species using the observed values. Next, categories were reduced to reactions with 0 or 1, 2, 3 and >3 genes. The small sample size precluded the use of a χ^2-test to determine whether the gene distribution between two species was random or not. Hence, a Multinomial Goodness of Fit test was performed with package XNomial. Unfortunately, the Full Enumeration Method (xmulti) required examining 10^{18} possible combinations. Therefore, Monte Carlo simulations were used (xmonte) on 10000 trials. Simulations with 100000 or even 1000000 trials produced more precise results albeit at a reduced speed. Since, the outcome of the test was identical all calculations were performed based on 10000 trials. Moreover, results were similar for a log-likelyhood ratio, probability or even a χ^2-test.

RESULTS

Genes were collected for *Arabidopsis thaliana*, *Prunus mume* and *Prunus persica* using both a targeted and non targeted approach: Gene information was gathered for *Arabidopsis thaliana*, *Prunus mume* and *Prunus persica* because *Prunu smume* is the closest species to apricot with a sequenced genome but *Prunus persica* proteins have been curated into pathways and *Arabidopsis thaliana* harbors the most experimental evidence regarding protein function, localization and mutation effects. Moreover, both targeted and non-targeted approaches were used. First, genes encoding proteins involved in amino acid metabolism were collected. The latter are precursors for the synthesis of ethylene, volatile and phenolic compounds. Transcriptomic studies showed significant variations in different biosynthetic pathways (Li *et al*., 2012; Manganaris *et al*., 2011). Both Aracyc12.0 (http://pmn.plantcyc.org/) (Mueller *et al*., 2003) and Peachcyc1.0 (http://www.rosaceae.org/) (Jung *et al*., 2014) were screened and reactions assigned on a metabolic rather than on a gene basis. Hence, catalytic steps with enzymatic evidence are listed even in the absence of corresponding genes. Moreover, several reactions correspond to the same gene list for bifunctional proteins. Enzymes are also placed in their compartments which produces further duplications for multiple locations.

The targeted approach has the drawback of limiting the selected genes to those encoding enzymes. Therefore, a phenotype based search was also completed to identify transcripts important for the early stages of fruit development. The period stands out in several ways. First, as mentioned cold frost damages flowers. In addition, fruits rely on reserves accumulated by the tree the previous season until leaves develop. Hence, three types of phenotypes were searched: those related to cold responses, those suggesting a deficiency in reserve accumulation or metabolic balance and those implicating poor fruit growth.

Several databases were screened to that effect. The Chloroplast 2010 Project (Ajjawi *et al*., 2010; Lu *et al*., 2008, 2011) measures a set of metabolic parametersin Arabidopsis knock-out mutants. It has the advantage of providing data for two different knock-outs for many genes but as the name implies focuses on chloroplastic proteins. Mutants were identified with consistent alterations in morning starch levels of leaves or in seed C/N ratio. The RIKEN Phenome Project morphologically characterizes Arabidopsis mutants obtained by transposon insertion (Kuromori *et al*., 2004, 2006). Mutant with smaller siliques or reduced yield were identified. The Arabidopsis Stress Responsive Database curates genes involved in abiotic stress (Borkotoky *et al*., 2013). Consequently, the database was screened for cold response genes. Some of the references provided listed proteins interacting with the curated product. They were included. Finally, potential interaction partners were searched for all genes in the Stanford Interactome Database (Jones *et al*., 2014). The project uses the split ubiquitin system to identify interactions between all Arabidopsis membrane proteins.

The *Prunus* species show a high conservation of isozyme numbers per reaction: Table 1 and 2 give a summary of the genes collected with literature used. In total, 133 reactions and 48 phenotypic genes were studied. This corresponded to 290 *Arabidopsis* genes or roughly 250

Table 1: Metabolic reactions curated for *Arabidopsis thaliana* (At), *Prunus persica* (Pp) and *Prunus mume* (Pm)

		Nb of genes			Genes perreaction				
Pathway	Rx	At	Pp	Pm	At	Pp	Pm	Local	References
Leu/Val/Ile	19 (10)	23	16	19	2.3	1.6	1.9	Mostly Cl	Binder
Lys/Thr/Met	14 (13)	26	22	22	2.0	1.7	1.7	Mostly Cl	Ravanel *et al.* (1998) Ravanel *et al.* (2004) Jander and Joshi (2009)
Met salvage	13 (10)	28	25	24	2.8	2.5	2.4	Cy	Ravanel *et al.* (1998) Sauter *et al.* (2013)
Phe/Tyr/Trp	21 (16)	55	41	38	3.4	2.6	2.4	Cl	Tzin and Galili (2010)
Ser/Gly/Ala	20 (15)	38	33	31	2.5	2.2	2.1	Cl, Mt, Pe, Cy, Nu	Couturier *et al.* (2013) Liepman and Olsen (2003) Ros *et al.* (2013)
Arg	10 (9)	13	17	16	1.4	1.9	1.8	Cl, Cy	Slocum (2005)
Asn/Asp/Glu/Gln	10 (9)	24	19	20	2.7	2.1	2.2	Cl, Mt, Cy, (Pe)	Coruzzi Liepman and Olsen (2004)
Cys	6 (6)	14	11	11	2.3	1.8	1.8	Cl, Mt, Cy	Novero (2009)
His	10 (8)	11	12	10	1.4	1.5	1.3	Cl	Rajani
Pro	10 (6)	9	12	9	1.5	2.0	1.5	Cl, Mt, Cy	Szabados and Savoure (2010)
Total	133 (102)	241	208	200	2.4	2.0	2.0		

The second column (Rx) provides the total number of reactions for the pathways listed in the first column and in parenthesis the number of stepss with unique genes. Total number of genes involved, number of genes per reaction, pathway localization (Local) and references are provided for each metabolic group. Localization abbreviations are Cl: Chloroplast; Cy: Cytoplasm; Mt: Mitochondria; Nu: Nucleus; Pe: Peroxisome

Table 2: Genes collected based on phenotype conferred. Information was gathered from the Chloroplast 2010 project (Cl2010, Ajjawi *et al.*, 2010; Lu *et al.*, 2008, 2011), the RIKEN Phenome collection (RIKEN, Kuromori *et al.*, 2004, 2006), the Stress Responsive Database (Stress, Borkotoky *et al.*, 2013) and the Stanford Interactomeproject (Interactome, Jones *et al.*, 2014). The table lists the number of genes associated with a given phenotype, the type of proteins encoded with their numbers in parenthesis and in the last column the number of homologues found in Prunus. Unambiguous Prunus counterparts are specified in parenthesis

Database	Phenotype	Genes in At	Proteins encoded	Hits in Pp and Pm
Cl2010	Excess starch am	3	Catalysis (2), kinase/phosphatase (1)	3 (3)
Cl2010	Altered C/N	1	Interactions (1)	1 (1)
RIKEN	Low yield	3	Catalysis (1), interactions (1), transport (1)	2 (2)
RIKEN	Short siliques	7	Catalysis (2), interactions (1), transcription factor (1), other (3)	4 (3)
Stress	Involved in response to cold	25	Catalysis (3), kinase/phosphatase (8), transcription factor (6), other (8)	25 (7)
Interactome	Interact with genes identified above	9	Catalysis (2), kinase/ phosphatase (1), transport (5), other (1)	9 (2)
Total		48	Catalysis (10), kinase/phosphatase (10), interactions (3), transcription factors (7), transport (6), other (12)	44 (18)

counterparts in Prunus. Of note, 102 reactions were encoded by unique genes with the other steps provided as an additional gene function. The database screen enlarged the category of functions studied as 38 genes encoded non catalytic proteins (Table 2). Four *Arabidopsis* genes did not produce hits in either *Prunus* sp. It is important to note that the amino acid metabolism study evaluated genes by family while the phenotypic search concentrated on single units. Therefore, the two lists are studied separately although as will be seen, the conclusions reached are similar.

Perusal of the amino acid metabolism reactions reveals that both *Prunus* sp. harbor 2.0 genes per reaction (Table 1) while Arabidopsis stands at 2.4 isozymes per reaction. The largest variation is observed for the synthesis of aromatic compounds with 3.4 genes per reaction for *Arabidopsis* versus 2.4 or 2.6 for the *Prunus* sp. The number of genes per reaction was counted for all species and listed in Table 3 (columns "obs"). The reaction with zero genes in *Prunus persica* corresponds to the chloroplastic methionine synthase enzyme and will be developed later. Effectively, the number of isozymes per reaction was viewed as a distribution problem of a given number of genes between a certain amount of reactions. Therefore, frequencies of given combinations were calculated relative to the most common distribution. The latter was found to be a large number of reactions encoded by two genes with the complement performed by 3 isozymes (Arabidopsis) or only one (Prunus). Hence, the ratio gene to reaction favors functional redundancy.

Observed combinations harbor a majority of reactions encoded by 1-3 genes with a few larger families. The latter explain the low frequencies of the observed gene distributions. Indeed, Arabidopsis harbors a higher number of large gene families than either *Prunus* sp. A reduction of that figure to values observed in Prunus increases the frequency of the combination from 1.3×10^{-11}-2.2×10^{-3} (Table 3). Finally, setting the number of genes with two reactions to that observed, results in

Table 3: Number of reactions with a given count of genes

	Arabidopsis thaliana					*Prunus persica*				*Prunus mume*			
		Max distrib.					Max distrib.				Max distrib.		
Nb genes per Rx	Obs	NA	$x_2 = 36$	x>3	x>3 (Pp)	Obs	NA	$x_2 = 25$	x>3	Obs	NA	$x_2 = 25$	x>3
0	0	0	0	0	0	1	0	0	0	0	0	0	0
1	38	0	20	25	0	48	5	42	31	52	13	46	35
2	36	76	36	62	92	25	98	25	62	25	90	25	59
3	13	27	47	0	1	19	0	35	0	17	0	31	0
4	9	0	0	9	5	5	0	1	5	6	0	1	6
5	3	0	0	3	4	4	0	0	4	2	0	0	2
6	1	0	0	1	1	1	0	0	1	1	0	0	1
>6	3	0	0	3	0	0	0	0	0	0	0	0	0
Rel. Freq.	$4.9\ (10^{-16})$	1	$3.0\ (10^{-4})$	$9.6\ (10^{-14})$	$2.2\ (10^{-3})$	$1.3\ (10^{-11})$	1	$2.3\ (10^{-7})$	$5.8\ (10^{-8})$	$1.1\ (10^{-9})$	1	$1.2\ (10^{-6})$	$1.8\ (10^{-9})$

Observed distributions are given in the columns "Obs" along with the most probable combination (NA) and the most frequent ones for the observed number of reactions encoded by 2 genes ($x_2 = ...$) and for the observed number of reactions encoded by >3 genes (x>3). The distribution with the Pp x>3 was calculated for Arabidopsis as a comparison. Frequencies of all combinations are given relative to the most common distribution. All calculations were performed in R Version 3.1.2 for Windows 64 bit with the packages gmp and lpSolveAPI. The study was performed on the set of metabolic reactions with unique genes. A reaction occurring in two different compartments was considered as two separate entities. The aspartate aminotransferase reactions were discounted due to the uncertainties regarding localization. Moreover, a few metabolic steps involve several protein units each counted as a reaction. Thus, the total number of reactions is 103 versus the previously cited 102

Table 4: Comparison of the number of genes per reaction between At, Pp and Pm

		Arabidopsis thaliana (p-value = $3.8\times10^{-4}\pm6.2\times10^{-5}$)					*Prunus mume* (p-value = 0±0)				
	Genes/Rx	0 or 1	2	3	>3	Tot	0 or 1	2	3	>3	Tot
Prunus persica	0 or 1	27 (18.1)	18 (17.1)	2 (6.2)	2 (7.6)	49 (49)	47 (24.7)	2 (11.9)	0 (8.1)	0 (4.3)	49 (49)
	2	9 (9.2)	11 (8.7)	3 (3.2)	2 (3.9)	25 (25)	5 (12.6)	18 (6.1)	2 (4.1)	0 (2.2)	25 (25)
	3	2 (7.0)	4 (6.7)	6 (2.3)	7 (3.0)	19 (19)	0 (9.6)	4 (4.6)	13 (3.2)	2 (1.6)	19 (19)
	>3	0 (3.7)	3 (3.5)	2 (1.3)	5 (1.5)	10 (10)	0 (5.1)	1 (2.4)	2 (1.6)	7 (0.9)	10 (10)
Tot		38 (38)	36 (36)	13 (13)	16 (16)	49 (30.6)	52 (52)	25 (25)	17 (17)	9 (9)	85 (34.9)

Data are based on the same reaction set as for Table 3. Observed values for each combination are provided with the expected values in parenthesis. Numbers in bold indicate reactions with identical numbers of genes. The bolded total is the sum of all reactions with conserved gene numbers. A Multinomial Goodness of Fit test was used to evaluate whether the observed distribution was different from the calculated one. The resulting p-values are provided. The statistical test was performed in R Version 3.1.2 for Windows 64 bit (R Core Team, 2014) with the package XNomial (Engel *et al*., 2010)

optimal combinations with a mixture of reactions catalyzed by 1-3 (Arabidopsis) or 1-4 (Prunus) enzymes. Both resulting Prunus distributions harbor larger numbers of reactions encoded by single genes then that of Arabidopsis. The real distributions also count a larger number of reactions encoded by single genes in *Prunus* versus in *Arabidopsis*.

Next, the observed combinations were used to calculate the probability of two species to harbor identical or different gene numbers for a given reaction. Four categories were listed namely reactions encoded by 0 or 1, 2, 3 or >3 genes, thus yielding 16 combinations. Those are listed in Table 4 with observed values in the form of Punnett squares for Arabidopsis vs. *Prunus persica* and for *Prunus mume* versus *Prunus persica*. Results for Arabidopsis versus *Prunus mume* are comparable to those of Arabidopsis versus *Prunus persica* which is why they are omitted. A total of 49 or 50 reactions (48 or 49%) have conserved number of genes in *Arabidopsis* and *Prunus persica* or *Prunus mume*, respectively. This is higher than the expected 30.6 or 31.4. The number rises to 85 or 82% when comparing the two *Prunus* sp.

Moreover, extreme combinations such as 0 or 1 gene in one species versus over 3 in the second one also show large differences between observed and calculated values. A Multinomial Goodness of Fit test produced p values in the order of 10^{-4} or less. Consequently, gene numbers per reaction are more conserved between species than would be expected from a random allocation.

In conclusion, the observed gene distributions within a species result from a combination of biological requirements for large gene families and optimal repartition. Moreover, gene allocation is conserved between species particularly so between *Prunus mume* and *Prunus persica*.

Several *Prunus* genes harbor two or more counterparts in Arabidopsis: Homologues in Prunus were searched for all *Arabidopsis* genes identified based on phenotype. Most of the time, BLAST provided several hits with relatively close scores. However, the objective was to determine the most likely Prunus counterpart for a given *Arabidopsis* gene. As a result, a multiple alignment was performed for all selected BLAST hits and identity

Table 5: Percent identity matrix generated for the leucine rich repeat receptor kinase AT4G39400 and similar genes

Identify	ppa000438m	XM_8236051	AT2G01950	ppa022290m	XM_8222737	AT4G39400	ppa000566m	XM_8234124	ppa000552m	XM_8248236	AT3G13380	AT1G55610
ppa000438m	100	-	-	-	-	-	-	-	-	-		-
XM_8236051	99	100	-	-	-	-	-	-	-	-	-	-
AT2G01950	46	46	100	-	-	-	-	-	-	-	-	-
ppa022290m	46	46	66	100	-	-	-	-	-	-	-	-
XM_8222737	46	46	65	99	100	-	-	-	-	-	-	-
AT4G39400	46	45	53	51	51	100	-	-	-	-	-	-
ppa000566m	46	46	50	52	51	67	100	-	-	-	-	-
XM_8234124	46	46	50	52	51	66	98	100	-	-	-	-
ppa000552m	47	46	52	52	52	53	54	54	100	-	-	-
XM_8248236	47	46	52	52	51	52	53	52	98	100	-	-
AT3G13380	46	45	53	51	51	54	54	54	66	66	100	-
AT1G55610	45	45	54	52	52	54	54	53	65	66	79	100

Results were derived by Clustal Omega (Goujon *et al*., 2010; Sievers *et al*., 2011) based on the alignment performed with default parameters. Percent identity numbers are bolded for Pp and Pm homologues as well as for the best aligning At and Pp, Pm homologues. *Prunusmume* gene names (XM_...) were shortened by omitting the two leading zeros and deleting the final ".1"

matrixes obtained. As an example, Table 5 provides the identity matrix obtained for the leucine rich repeat receptor kinase AT4G39400 and hits. As can be seen, AT4G39400 aligns best with ppa000566m and XM_008234124.1 with an identity around 65%. However, both AT3G13380 and AT1G55610 show preferential alignment for the same hypothesized that ppa000566m and XM_008234124.1 intervene in the same signaling network as AT4G39400 while ppa000438m and XM_008236051.1 share roles specific to Prunus. Finally, identity between Prunus homologues is at 98 or 99% showing tremendous gene conservation in the species. The approach is valid provided all genes with a significant alignment to the query are selected. This is why a conservative threshold of 16% alignment was set for the selection of BLAST hits. The number separated random hits from those of potential interest.

Counterparts were determined for all 44 *Arabidopsis* genes with hits. A total of 15 preferentially aligned with single *Prunus* genes. Matches to AT3G26744 were treated as clear counterparts although identity was significant for half of the sequence. Two genes, namely AT2G43790 and AT1G74520 were considered to have clear counterparts. Indeed, though two Arabidopsis sequences aligned with each Prunus hit, identity differed by 8-10%. Finally, a protein alignment identified a preferential counterpart for AT1G50720. The remaining 26 assignments were ambiguous for Arabidopsis, *Prunus persica* and *Prunus mume* 22, nine and ten times, respectively.

Clear counterparts are listed in Table 6 with the identity between Prunus and Arabidopsis as well as that between the *Prunus* genes. Sequence identity between Arabidopsis and Prunus ranges from 60-76% while that between two Prunus counterparts lies at 95-99%. The numbers confirms the high conservation of Prunus sequences. Moreover, genes originate from all databases in similar proportions. This shows the necessity of using several sources for a gene search based on phenotype.

Localization may vary between counterparts: Localization was determined for all genes based on literature, TAIR annotation, protein sequence alignment and prediction. Information was fairly consistent for proteins found in the cytoplasm, the chloroplast, the mitochondria or the nucleus. However, targeting predictions produced contradicting results for proteins located elsewhere. Hence, the compartment for the genes identified based on phenotype remains for the most part unclear. Moreover, a certain number of aspartate aminotransferase proteins could not be clearly assigned. Localization did differ between a few counterparts. Consequently, protein sequences were aligned to pinpoint consistencies between compartment assigned and sequence length. Three examples are provided in Table 7.

Methionine synthase is encoded by three genes in Arabidopsis and two in the *Prunus* species. The enzyme is located in the chloroplast and cytosol. However, targeting programs failed to find a chloroplastic isoform in *Prunus persica*. A protein sequence alignment shows that the results are consistent (Table 7). Indeed, all sequences harbor a methionine at the start of the corresponding cytosolic sequence. Only two proteins exhibit clear localization sequences, namely AT5G20980 and XM-008232732.1. No other sequences were detected for persica. Therefore, *Prunus persica* may harbor purely cytosolic forms of methionine synthase.

The two other examples are branched-chain amino acid transaminase and acetylserine lyase. Once again, Prunus proteins align with Arabidops is isozymes targeted to an organelle but are predicted to be in different compartments. The sequence alignment shows that the results are consistent. Indeed, the protein supposedly located in the cytoplasm displays a shorter sequence than its organelle homologue (Table 7). Surprisingly, the cytoplasmic arabidopsis sequences are slightly longer with an appearance of a signaling peptide. However, the targeting data is based on experimental results.

Table 6: *Arabidopsis* genes with clear counterparts in Pp and Pm. The database used is noted as a letter on the At genes with "a" for the Chloroplast 2010 Project, "b" for the RIKEN phenome collection, "c" for the Stress database and "d" for the Stanford Interactome Project. Counterparts in Pp and Pm are listed with their percentage of identity to *Arabidopsis* genes in parenthesis. The last column displays identity values between Pp and Pm sequences. The alignment between *Pp*, *Pm* and *AT3G26744* genes was significant over half of the sequence. Identity between Pp and Pm over the whole sequence is specified with an*

At gene	Description	Pp counterpart	Pm counterpart	Pp/Pm
AT1G10760[a]	Pyruvate phosphate dikinase	ppa000209m (71)	XM_008247257.1 (69)	98
AT2G40840[a]	Disproportionating enzyme 2	ppa000782m (72)	XM_008245862.1 (72)	98
AT3G52180[a]	Protein phosphatase	ppa007299m (67)	XM_008233843.1 (67)	98
AT1G34790[a]	Zinc finger protein	ppa016755m (64)	XM_008221233.1 (63)	99
AT5G23630[b]	ATPase cation pumps	ppa000424m (74)	XM_008220607.1 (74)	99
AT1G69180[b]	Transcription factor	ppa014900m (65)	XM_008245598.1 (64)	99
AT1G68560[b]	α-l-arabinofuranosidase/β-D-xylosidase	ppa001168m (70)	XM_008244999.1 (70)	99
AT5G49360[b]	α-l-arabinofuranosidase/β-D-xylosidase	ppa001718m (68)	XM_008225083.1 (67)	98
AT2G01390[b]	(TPR)-like protein	ppa004294m (62)	XM_008243399.1 (61)	98
AT4G39400[c]	Leucine-rich repeat receptor kinase	ppa000566m (67)	XM_008234124.1 (66)	98
AT3G26744[c]	Transcription activator	ppa005038m (75)	XM_008241330.1 (74)	9899*
AT2G39810[a, c]	Novel protein	ppa000974m (65)	XM_008240210.1 (64)	97
AT3G59770[c]	Phosphoinositide phosphatase	ppa000157m (72)	XM_008244816.1 (71)	99
AT4G04920[c]	Nuclear targeted protein	ppa000947m (67)	XM_008226829.1 (74)	86
AT5G13650[a,c]	Srv3	ppa002327m (76)	XM_008242662.1 (76)	99
AT5G50720[c]	AtHVA22e	ppa013097m (68)	XM_008241283.1 (68)	95
AT1G74520[c]	AtHVA22a	ppa012417m (71)	XM_008241585.1 (71)	97
AT2G43790[c]	MPK6	ppa006536m (76)	XM_008244822.1 (76)	99
AT3G54300[d]	Synaptobrevin-likeproteinfamily	ppa010737m (76)	XM_008246668.1 (76)	98
AT5G47910[d]	Respiratory burst oxidase protein D	ppa000883m (67)	XM_008224070.1 (67)	97

Table 7: Examples of counterparts with potentially different alignments. Homologue names and methionines are bolded. Arabidopsis sequences were included for all compartments as a reference. Localization abbreviations are as for Table 1. All alignments were performed with Clustal Omega (Goujon *et al*., 2010; Sievers *et al*., 2011) using default parameters

Protein	Alignment
Methionine synthase	AT5G20980 (Cl) MGQLALQRLQPLASLPRRPPSLPPPSSATPSLPCATASRRPRFYVARAMSSHIVGYPRIG
	ppa021650m (Cy) ----------MASHIVGLPRIG
	XM_8232732 (Cl) MKQ----------VSSITFGP-CYGSLCFSAKRPTLLRFTHHFKFHSTRAMASHIVGLPRIG
	ppa001783m (Cy) ----------MASHIVGYPRMG
	XM_8239029 (Cy) ----------MASHIVGYPRMG
	AT5G17920 (Cy) ----------MASHIVGYPRMG
	AT3G03780 (Cy) ----------MASHIVGYPRMG
	*:***** **:*
Branched-chain amino acid transaminase	AT1G50110 (Cy) ----------MAPSSSPLRTTSETDEK
	AT1G10060 (Mt) ----------MALRRCLPQYSTTSSYLSKIWGFRMH---------GTKAAASVVEEHVSGAEREDEE
	AT1G10070 (Cl) MIKTITSLRKTL---------VL---------PLHLHIRTLQTFAKYNAQAASALREERKKPLYQNGDDV
	ppa008826m (Cy) ----------
	XM_8222198 (Mt) MIQRTTRLHKLVRSIGVGSSLSSSSKQLRVHRCFSSVAASNA-EQACEQSVESYNVKKNE
	AT1G50110 (Cy) YANVKWEELGFALTPIDY**M**YVAKCRQGESFTQGKIVPYGDISISPCSPILNYGQGLFEGL
	AT1G10060 (Mt) YADVDWDNLGFSLVRTDF**M**FATKSCRDGNFEQGYLSRYGNIELNPAAGILNYGQGLIEGM
	AT1G10070 (Cl) YADLDWDNLGFGLNPADY**M**YVMKCSKDGEFTQGELSPYGNIQLSPSAGVLNYGQAIYEGT
	ppa008826m (Cy)----------**M**YVMKCSNNGTFEKGQLNRYGNIELNPAAGVLNYGQGLYEGT
	XM_8222198 (Mt) YADVDWDNLGFGLTPTDY**M**YVMKCSNNGTFEQGQLNRYGNIELNPAAGVLNYGQGLYEGT
	:... * .* : **:*:..*:.:*****.: **
Acetylserinelyase	AT3G04940 (Cy) ----------
	AT3G03630 (Cl) MAFASPSLRLLPQSPLGRITSKLHRFSTAKLSLFSFHHDSSSSLAVRTPVSSFVVGAISG
	ppa007201m (Cl) MAILSAPLLLSLPHPF-SFPSKRHRFGTFKVSSSILS----------
	XM_8242347 (Cy) ----------
	AT3G04940 (Cy) ----------MEEDRCSIKDDATQLIGNTP**M**VYLNNIV
	AT3G03630 (Cl) KSSTGTKS-KSKTKRKPPPPPPVTTVAEEQHIAESETVNIAEDVTQLIGSTP**M**VYLNRVT
	ppa007201m (Cl) ------TNGALLRRQFTQRYPLVFAK---ASSVYATREDLDTVNIAEDVTQLIGSTP**M**VYLNKVT
	XM_8242347 (Cy) ----------**M**VYLNKVT
	*****.:

DISCUSSION

Prunus genes show a strong degree of conservation: All species show a similar gene distribution per function with the majority of catalytic steps being performed by 1-3 genes while a few reactions require larger numbers. Moreover, a given reaction is more likely to harbor the same number of genes in a second species than would be

expected from a random distribution. Hence, there is a general conservation in gene distribution patterns between organisms. However, the degree of conservation is particularly pronounced between the two *Prunus* sp. with over 80% of the reactions harboring similar gene numbers. The results are confirmed at the sequence level as gene identities between Prunus counterparts are over 90%. This reflects previous data as microarrays designed for *Prunus persica* function with apricot extracts (Manganaris *et al*., 2011). The surprise comes from the extent of conservation.

The results have to be contrasted with the diversity observed within a species. As mentioned in the introduction, polymorphism studies identified significant variation in apricot varieties (Takeda *et al*., 1998; Hagen *et al*., 2002; Yuan *et al*., 2007; Krichen *et al*., 2008; Yilmaz *et al*., 2009). Moreover, close to a million informative SNPs were found in *Prunus persica* accessions (Verde *et al*., 2013). The polymorphism is reflected at the phenotypic level with significant variations in phenolic compounds, carotenoids, sugar and acid levels (Ruiz *et al*., 2005a, b; Drogoudi *et al*., 2008; Engel *et al*., 2010; Schmitzer *et al*., 2011; Gundogdu *et al*., 2013) as well as fruit ripening times (Asthma, 2012). Consequently, a potential for diversity exists in the *Prunus* sp. in spite of the strong genome conservation.

Variation is also observed in protein length. At least three cases are linked to differences in localization. For instance, *Prunus persica* does not seem to harbor a plastidic methionine synthase. Results are based on predictions which assume a single compartment while several proteins have dual locations (Carrie *et al*., 2009). Hence, the chloroplastic location may have been overlooked. In fact, the prediction programs suggest a second targeting to the mitochondria. The cytosolic methionine synthase is definitely essential in regenerating the methyl group of S-adenosyl-L-methionine, an intermediate in the methionine salvage pathway (Ravanel *et al*., 1998). Conversely, the chloroplastic form ensures autonomy of the organelle for methionine synthesis (Ravanel *et al*., 2004). The single mutant presented in the Chloroplast 2010 database looks smaller but otherwise fairly normal. Consequently, the peculiar localization distribution in *Prunus persica* appears to be possible. The other two cases merely change the repartition of proteins between compartments. The three cases are an underestimate of the potential localization differences between counterparts. Moreover, when related to the number of proteins with several compartments, the ratio appears quite significant. Hence, at the genome scale a noteworthy number of counterparts in different species may encode proteins with different localizations.

In conclusion, gene to gene correspondence seems fairly straight forward to pinpoint in the *Prunus* sp. This opens many possibilities in terms of combining transcriptomic data from different organisms. However, polymorphic variations and potential differences in localization suggest that caution must still be exercised when assessing the fine details of a gene function.

Arabidopsis data is potentially applicable to Prunus: While Arabidopsis is clearly different from a tree, a surprising number of functional elements appear to be conserved. Thus, Prunus homologues were identified for 44/48 *Arabidopsis* genes. Moreover, an Arabidopsis pseudogene listed in amino acid metabolism matched sequences in both *Prunus* sp. Genes are potentially conserved for at least three reasons: a similar role, a similar function but with different applications in each organism, a similar gene origin. Clearly, differentiation between these mechanisms is essential in correctly transferring data from Arabidopsis to Prunus. Precious information is provided by sequence alignments. As an example, identity matrixes show clear examples of *Prunus* genes sharing similar conservation with all genes of the corresponding Arabidopsis family. This would point to a shared function or an evolutionary relic rather than to a similar role.

A second issue is the preferential alignment of a given gene with several counterparts. For instance, only 18 *Arabidopsis* genes harbor a clear Prunus counterpart with correspondence assignment being ambiguous in the other cases. Three *Arabidopsis* genes involved in cold response aligned with a single Prunus counterpart thus causing complexities in the comparison of the cold response signaling pathway. A potential answer may be found in the studies on gene duplication. Genome comparison shows clear examples of the enlargement of specific gene families in given organisms. For instance, *Prunus persica* harbors a large number of genes devoted to fruit quality (Verde *et al*., 2013). Moreover, several families important for the production of a lignified seed are enlarged. The increase is attributed to gene duplication as an adaptive process to deal with a lignified stone. In Arabidopsis over 80% of the genome represents duplicated regions (Briggs *et al*., 2006). Duplicates are often maintained albeit with a reduced function. Therefore, the evolutionary information provided by sequencing data is essential for understanding gene to gene correspondence between species.

Databases are a key tool for the modern biologist: As mentioned, moving from model plants to crops requires a sensible perusal of the massive data currently being

generated. As a result, databases of all type are an essential tool for the modern biologist. Unfortunately, progress in sequencing and other data production has exponentially increased the cost and manpower necessary to maintain databases. The funding situation encountered by TAIR a few years ago shows that even major players are at risk. Thankfully, they are now "thriving" thanks to donations. But what is the status of less frequently used databases? Sadly, the issue is not easy to resolve. Indeed, this type of resource needs to generate biological data and publications to justify academic funding. This is usually possible in the early stages of the project but less so in the maintenance phase. Hence, there is a risk of information being lost through lack of funding. Moreover, as the number of available databases grows, it becomes increasingly important to maintain their low cost. Otherwise, the average researcher will be deprived of essential resources in case of insufficient funding. In short, data management is an important and difficult issue to resolve. Currently, the massive amounts of data produced are underused.

CONCLUSION

The results point to a remarkable conservation of the Prunus genes both in terms of number of enzymes per family and of sequence alignment. Second, the basis of comparison across species lies in the existence and correct assignment of homologues. The data show that though homologous genes are detected most of the time the assignment of a clear gene to gene correspondence is often ambiguous between Arabidopsis and Prunus. Finally, clear protein counterparts do not necessarily share the same cellular localization. In short, Prunus data will likely apply to *Prunus armeniaca* thanks to the remarkable conservation of the genus. However, Arabidopsis information may be valid to some extent.

ACKNOWLEDGEMENT

I would like to thank Professors Bayram Murat Asma, HikmetGeckil and ErgunDogan for critical reading of the manuscript.

REFERENCES

Ajjawi, I., Y. Lu, L.J. Savage, S.M. Bell and R.L. Last, 2010. Large-scale reverse genetics in Arabidopsis: Case studies from the Chloroplast 2010 project. Plant Physiol., 152: 529-540.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman, 1990. Basic local alignment search tool. J. Mol. Biol., 215: 403-410.

Asthma, B.M., 2007. Malatya: World's capital of apricot culture. Chronica. Hort, 47: 20-24.

Asthma, B.M., 2012. A new early-ripening apricot, Dilbay. Hort Sci., 47: 1367-1368.

Borkotoky, S., V. Saravanan, A. Jaiswal, B. Das and S. Selvaraj *et al*., 2013. The arabidopsis stress responsive gene database. Int. J. Plant Genomics.

Bourguiba, H., J.M. Audergon, L. Krichen, N. Trifi-Farah and A. Mamouni *et al*., 2012. Loss of genetic diversity as a signature of apricot domestication and diffusion into the Mediterranean Basin. BMC Plant Biol., Vol. 12.

Brameier, M., A. Krings and R.M. MacCallum, 2007. NucPred-predicting nuclear localization of proteins. Bioinformatics, 23: 1159-1160.

Briggs, G.C., K.S. Osmont, C. Shindo, R. Sibout and C.S. Hardtke, 2006. Unequal genetic redundancies in Arabidopsis: A neglected phenomenon?. Trends Plant Sci., 11: 492-498.

Carrie, C., K. Kuhn, M.W. Murcha, O. Duncan and I.D. Small *et al*., 2009. Approaches to defining dual-targeted proteins in arabidopsis. Plant J., 57: 1128-1139.

Couturier, J., B. Touraine, J.F. Briat, F. Gaymard and N. Rouhier, 2013. The iron-sulfur cluster assembly machineries in plants: current knowledge and open questions. Front. Plant Sci., Vol. 4.

Drogoudi, P.D., S. Vemmos, G. Pantelidis, E. Petri, C. Tzoutzoukou and I. Karayiannis, 2008. Physical characters and antioxidant, sugar and mineral nutrient contents in fruit from 29 apricot (*Prunus armeniaca* L.) cultivars and hybrids. J. Agric. Food Chem., 56: 10754-10760.

Emanuelsson, O., S. Brunak, G. von Heijne and H. Nielsen, 2007. Locating proteins in the cell using TargetP, SignalP, and related tools. Nat. Protocols, 2: 953-971.

Engel, R., L. Abranko, E. Balogh, A. Blazovics and R. Herman *et al*., 2010. Antioxidant and antiradical capacities in apricot (*Prunus armeniaca* L.) fruits: variations from genotypes, years and analytical methods. J. Food Sci., 75: 722-730.

Geuna, F., R. Banfi and D. Bassi, 2005. Identification and characterization of transcripts differentially expressed during development of apricot (*Prunus armeniaca* L.) fruit. Tree Genet. Genomes, 1: 69-78.

Gonzalez-Aguero, M., S. Troncoso, O. Gudenschwager, R. Campos-Vargas, M.A. Moya-Leon and B.G. Defilippi, 2009. Differential expression levels of aroma-related genes during ripening of apricot (*Prunus armeniaca* L.). Plant Physiol. Bioch., 47: 435-440.

Goujon, M., H. Mcwilliam, W. Li, F. Valentin and S. Squizzato *et al*., 2010. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res., 38: 695-699.

Greger, V. and P. Schieberle, 2007. Characterization of the key aroma compounds in apricots (*Prunus armeniaca*) by application of the molecular sensory science concept. J. Agric. Food Chem., 55: 5221-5228.

Grimplet, J., C. Romieu, J.M. Audergon, I. Marty and G. Albagnac *et al*., 2005. Transcriptomic study of apricot fruit (Prunus armeniaca) ripening among 13 006 expressed sequence tags. Physiol.Plantarum, 125: 281-292.

Gundogdu, M., T. Kan and M.K. Gecer, 2013. Vitamins, flavonoids and phenolic acid levels in early-and late-ripening apricot (*Prunus armeniaca* L.) cultivars from Turkey. Hort Sci., 48: 696-700.

Hagen, L., B. Khadari, P. Lambert and J.M. Audergon, 2002. Genetic diversity in apricot revealed by AFLP markers: Species and cultivar comparisons. Theor. Applied Genet., 105: 298-305.

Jander, G. and V. Joshi, 2009. Aspartate-derived amino acid biosynthesis in arabidopsis thaliana. Arabidopsis book/Am. Soc. Plant Biol., Vol. 7.

Jones, A.M., Y. Xuan, M. Xu, R.S. Wang and C.H. Ho *et al*., 2014. Border control: A membrane-linked interactome of arabidopsis. Sci., 344: 711-716.

Jung, S., S.P. Ficklin, T. Lee, C.H. Cheng and A. Blenda *et al*., 2014. The Genome Database for Rosaceae (GDR): Year 10 update. Nucleic Acids Res., 42: 1237-1244.

Krichen, L., J.M. Martins, P. Lambert, A. Daaloul and N. Trifi-Farah *et al*., 2008. Using AFLP markers for the analysis of the genetic diversity of apricot cultivars in Tunisia. J. Am. Soc. Hortic. Sci., 133: 204-212.

Kuromori, T., T. Hirayama, Y. Kiyosue, H. Takabe and S. Mizukado *et al*., 2004. A collection of 11 800 single-copy Ds transposon insertion lines in Arabidopsis. Plant J., 37: 897-905.

Kuromori, T., T. Wada, A. Kamiya, M. Yuguchi and T. Yokouchi *et al*., 2006. A trial of phenome analysis using 4000 Ds-insertional mutants in gene-coding regions of Arabidopsis. Plant J., 47: 640-651.

Kushwaha, H.R., A.K. Singh, S.K. Sopory, S.L. Singla-Pareek and A. Pareek, 2009. Genome wide expression analysis of CBS domain containing proteins in Arabidopsis thaliana (L.) Heynh and Oryza sativa L. reveals their developmental and stress regulation. BMC Genomics, Vol. 10 10.1186/1471-2164-10-200.

Lamesch, P., T.Z. Berardini, D.H. Li, D. Swarbreck and C. Wilks *et al*., 2012. The arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. Nucleic Acids Res., 40: 1202-1210.

Li, X., N.K. Korir, L. Liu, L. Shangguan and Y. Wang *et al*., 2012. Microarray analysis of differentially expressed genes engaged in fruit development between Prunus mume and Prunus armeniaca. J. Plant Physiol., 169: 1776-1788.

Liepman, A.H. and L.J. Olsen, 2003. Alanine aminotransferase homologs catalyze the glutamate: glyoxylate aminotransferase reaction in peroxisomes of Arabidopsis. Plant Physiol., 131: 215-227.

Liepman, A.H. and L.J. Olsen, 2004. Genomic analysis of aminotransferases in arabidopsis thaliana. Crit. Rev. Plant Sci., 23: 73-89.

Lu, Y., L.J. Savage, I. Ajjawi, K.M. Imre and D.W. Yoder *et al*., 2008. New connections across pathways and cellular processes: Industrialized mutant screening reveals novel associations between diverse phenotypes in arabidopsis. Plant Physiol., 146: 1482-1500.

Lu, Y., L.J. Savage, M.D. Larson, C.G. Wilkerson and R.L. Last, 2011. Chloroplast 2010: A database for large-scale phenotypic screening of arabidopsis mutants. Plant Physiol., 155: 1589-1600.

Manganaris, G.A., A. Rasori, D. Bassi, F. Geuna and A. Ramina *et al*., 2011. Comparative transcript profiling of apricot (*Prunus armeniaca* L.) fruit development and on-tree ripening. Tree Genet. Genomes, 7: 609-616.

Mueller, L.A., P. Zhang, S.Y. Rhee, 2003. AraCyc: A biochemical pathway database for arabidopsis. Plant Physiol., 132: 453-460.

Novero, A.U., 2009. Current status of research on o-acetylserine (thiol) lyase and β-cyanoalanine synthase, two enzymes of plant cysteine biosynthesis: A review. Plant Omics, 2: 181-189.

Ravanel, S., B. Gakiere, D. Job and R. Douce, 1998. The specific features of methionine biosynthesis and metabolism in plants. Proc. Natl. Acad. Sci. USA., 95: 7805-7812.

Ravanel, S., M.A. Block, P. Rippert, S. Jabrin and G. Curien *et al*., 2004. Methionine metabolism in plants: Chloroplasts are autonomous for de novo methionine synthesis and can import S-adenosylmethionine from the cytosol. J. Biol. Chem., 279: 22548-22557.

Reumann, S., D. Buchwald and T. Lingner, 2012. PredPlantPTS1: A web server for the prediction of plant peroxisomal proteins. Front Plant. Sci., Vol. 3.

Ros, R., B. Cascales-Minana, J. Segura, A.D. Anoman and W. Toujani *et al*., 2013. Serine biosynthesis by photorespiratory and non-photorespiratory pathways: An interesting interplay with unknown regulatory networks. Plant Biol., 15: 707-712.

Ruiz, D., J. Egea, F.A. Tomas-Barberan and M.I. Gil, 2005a. Carotenoids from new apricot (*Prunus armeniaca* L.) varieties and their relationship with flesh and skin color. J. Agric. Food Chem., 53: 6368-6374.

Ruiz, D., J. Egea, M.I. Gil and F.A. Tomas-Barberan, 2005b. Characterization and quantitation of phenolic compounds in new apricot (*Prunus armeniaca* L.) varieties. J. Agric. Food Chem., 53: 9544-9552.

Salazar, J.A., D. Ruiz, J.A. Campoy, R. Sanchez-Perez and C.H. Crisosto *et al*., 2014. Quantitative Trait Loci (QTL) and Mendelian Trait Loci (MTL) analysis in Prunus: A breeding perspective and beyond. Plant Mol. Biol. Rep., 32: 1-18.

Sauter, M., B. Moffatt, M.C. Saechao, R. Hell and M. Wirtz, 2013. Methionine salvage and S-adenosylmethionine: essential links between sulfur, ethylene and polyamine biosynthesis. Biochem. J., 451: 145-154.

Schmitzer, V., A. Slatnar, M. Mikulic-Petkovsek, R. Veberic and B. Krska *et al*., 2011. Comparative study of primary and secondary metabolites in apricot (*Prunus armeniaca* L.) cultivars. J. Sci. Food Agric., 91: 860-866.

Shi, S., J. Li, J Sun, J. Yu and S. Zhou, 2013. Phylogeny and classification of Prunus sensu lato (Rosaceae). J. Int. Plant Biol., 55: 1069-1079.

Sievers, F., A. Wilm, D. Dineen, T.J. Gibson and K. Karplus *et al*., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Syst. Biol., 7: 539-544.

Siu, K.K.W., J.E. Lee, J.R. Sufrin, B.A. Moffatt and M. McMillan *et al*., 2008. Molecular determinants of substrate specificity in plant 5'-methylthioadenosine nucleosidases. J. Mol. Biol., 378: 112-128.

Slocum, R.D., 2005. Genes, enzymes and regulation of arginine biosynthesis in plants. Plant Physiol. Biochem., 43: 729-745.

Sochor, J., P. Babula, V. Adam, B. Krska and R. Kizek, 2012. Sharka: The past, the present and the future. Virus, 4: 2853-2901.

Szabados, L. and A. Savoure, 2010. Proline: A multifunctional amino acid. Trends Plant Sci., 15: 89-97.

Takeda, T., T. Shimada, K. Nomura, T. Ozaki, T. Haji, M. Yamaguchi and M. Yoshida, 1998. Classification of apricot varieties by RAPD analysis. J. Jap. Soc. Hortic. Sci., 67: 21-27.

Tzin, V. and G. Galili, 2010. The biosynthetic pathways for shikimate and aromatic amino acids in Arabidopsis thaliana. Arabidopsis book/Am. Soc. Plant Biol., Vol. 8.

Verde, I., A.G. Abbott, S. Scalabrin, S. Jung and S. Shu *et al*., 2013. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat. Genet., 45: 487-494.

Yilmaz, K.U., S. Ercisli, B.M. Asthma, Y. Dogan and S. Kafkas, 2009. Genetic relatedness in prunus genus revealed by inter-simple sequence repeat markers. Hort Sci., 44: 293-297.

Yuan, Z., X. Chen, T. He, J. Feng, T. Feng and C. Zhang, 2007. Population genetic structure in apricot (*Prunus armeniaca* L.) cultivars revealed by fluorescent-AFLP markers in Southern Xinjiang, China. J. Gen. Genomics, 34: 1037-1047.