Supplementary Note 4

Analysis of Tandem Duplications and Gene Families in *Petunia* Species as Compared to Other Solanaceae and *Arabidopsis*

Mitrick A. Johns*, Jennifer Hintzsche

Dept. of Biological Sciences, Northern Illinois University, DeKalb IL 60115, USA
*Corresponding author: e-mail: rjohns@niu.edu

Short title: *Petunia* gene families

Abstract

Polypeptide sequences from *Petunia axillaris* (*PaxiN*) and *P. inflata* (*PinfS6*) as well as tomato, potato, *Nicotiana benthamiana*, and *Arabidopsis thaliana* were subjected to an all-vs-all blast search, then clustered into gene families with OrthoMCL. Of the resulting 27,600 families, 9430 families (39.2% of all genes) contained genes from all six species, and 24.6% of the genes were not clustered into any family. Most multispecies gene families followed the accepted evolutionary lineage, with the *Petunia*, *Solanum*, and Solanaceae clades sharing gene families far more often than other species groupings. Most of the genes unique to a single species or small clade were transposable element-related or had unknown functions, while most genes shared widely among species were involved with cellular and metabolic functions. Gene families with known functions that were almost unique to *Petunia* included several families of F-box proteins, which may function to ubiquitylate specific proteins destined for degradation, and several families of pentatricopeptide repeat genes, which are probably involved with RNA modification in the organelles. *Petunia* also contained several unique families of genes coding for cytochrome P450 and disease resistance genes. Among the families of genes coding for metabolic enzymes that differentially amplified in *Petunia* were genes for 1-aminocyclopropane-1-carboxylate synthase and oxidase, cupredoxin, HXXXD-type acyl-transferase, caffeate O-methyltransferase, and replication protein A. Tandemly duplicated genes fell into a similar set of functions, except that they did not include transposon genes or genes with unknown functions.

INTRODUCTION

Gene duplication is a major way of increasing an organism's diversity of response to specific environmental conditions (Ohno, 1970; Flagel and Wendel, 2009). Genes that have proliferated into multiple copy families can provide important clues to the specialized life habit of a species. For this reason we have studied gene families in the two *Petunia* species, grouping similar genes into families, determining which families have expanded, and analyzing their annotations. We concentrated on three types of comparison: *PinfS6* and *PaxiN* with each other, the two *Petunia* species with other Solanaceae with sequenced genomes, and all Solanaceae with *Arabidopsis thaliana*.

RESULTS

Gene Families Generated by OrthoMCL

For OrthoMCL analysis, genes from the six input species were filtered by requiring them to be at least 20 amino acids long and contain no internal stop codons. Only one representative peptide was used from each gene. A total of 251,612 genes were analyzed, with 35,233 from *PaxiN*, 38,826 from *PinfS6*, 34,727 from *S. lycopersicum*, 39,031 from *S. tuberosum*, 76,379 from *N. benthamiana*, and 27,416 from *A. thaliana*. Table 1 summarizes the number of genes and the number of gene families found in different possible combinations of species. A Venn diagram showing the percentage of genes in families found in all combinations of the five Solanaceae species can be found as Figure 2e in the main text.

Table 1. **Number of genes clustered into families by OrthoMCL.**

Species	Total genes	Genes in families with members of these species only	Families containing members of these species only	Singletons
P. axillaris (PaxiN)	35233	829	314	6615
P. inflata (PinfS6)	38826	1084	428	7782
S. lycopersicum (Sl)	34727	1414	445	8433
S. tuberosum (St)	39031	4243	752	7659
N. benthamiana (Nb)	76379	10848	2430	26425
A. thaliana (At)	27416	3820	1024	4890
Petunia (PaxiN PinfS6)	74059	5948	2402	
Solanum (Sl St)	73758	4469	1175	
All other groups of 2 (mean)	85403	398	94	
All other groups of 2 (range)	62113-115410	27-974	8-268	
Solanaceae (PaxiN PinfS6 Sl St Nb)	224196	23238	3076	
All other groups of 5 (mean)	206773	2041	300	
All other groups of 5 (range)	175233-216885	558-6343	89-946	
All species	251612	98683	9430	61804

OrthoMCL grouped these genes into 27,600 gene families, of which 9430 families contained genes from all six species. These common families had 98,683 genes, or 39.2% of all genes. In contrast, 24.6% of the genes (61,804 genes) were not placed in any gene family (singletons).

The number of gene families unique to specific groups of species was strongly influenced by the evolutionary relatedness of the two *Petunia* species and, to a lesser extent, the two *Solanum* species. There were 314 and 428 families containing only members of *PaxiN* and *PinfS6*, respectively, but 2402 families containing members of both species. Similarly, tomato and potato had 445 and 752 families unique to each species, but 1175 families had genes from both species. As a comparison, all other combinations of two species were found to have between 8 and 268 families (average 94 families) containing members of both species. When extended to all of the Solanaceae (i.e. all species except *Arabidopsis*), 3076 families were found to have genes from all five species. All other combinations of 5 species had between 89 and 946 families in common, with an average of 300.

Of the 27,600 gene families, 2347 families had at least five members from at least one species. These families were used for further analysis. The largest family, consisting of retrotransposon polyproteins found almost exclusively in *S. tuberosum*, contained 1026 members.

Gene Families Found in All Species

"Balanced" gene families were defined as having at least 30 members distributed across the six species, with fewer than 80% in any single species or genus. This category included 277 families containing 15,337 genes (data not shown). Most of the gene families in this category contained genes involved in normal cellular metabolism. Only three of these families, with a total of 99 genes, had no known function, and none of the balanced gene families seemed to be associated with transposable elements.

The largest balanced gene family contained subtilase family proteins, with 45-85 genes in each species. The largest 20 families also included six families of protein kinases, four families of membrane transporters, three families of carbohydrate-active enzymes, and two families of pentatricopeptide repeat proteins. Proteins in the latter group are involved with processing organellar RNAs (Barkan, 2011). Other types of family prominent on the list of balanced gene families include transcription factors, disease resistance genes, cytochrome P450 genes, and lipid-active genes.

Gene Families Found Primarily in a Single Species

Gene families that are almost unique to a single species were defined as having at least 30 family members with at least 80% of the members from a single species. There were 108 gene families containing 10,736 genes in this category (Table 2). The number of families varied widely among species, from a low of three families with 155 genes in *PaxiN* to a high of 62 families with 5434 genes in *N. benthamiana*. Families unique to *A. thaliana* were not included. In contrast to the balanced gene families, genes in the almost unique families were overwhelmingly either transposable element genes or genes with no known function. Only four families containing 394 genes are not clearly categorized in one of these two categories: a FAR1-related transcription factor, a 60S ribosomal protein, an F-box protein, and a chromo domain-containing protein.

One group of genes found frequently in species-specific groups is Ulp1 proteases, which regulate the removal of small ubiquitin-like modifier (SUMO) peptides in a cell cycle-dependent manner (Elmore et al., 2011). Ulp1 protease genes are found in several large families that are nearly species-specific in all of the Solanaceae species. Most of these genes are found adjacent to Mutator-like transposase genes. These finding suggest that the Solanaceae contain large numbers of CaMULE or Kaonashi elements (Hoen et al., 2006; van Leeuwen et al., 2007), which consist of an active Ulp1 protease gene coupled to a transposase gene, as seen in cucumber and Arabidopsis.

Table 2. **Near-unique gene families.**

A. *PaxiN* near-unique families: 3 families with 155 genes.[a]

group	total	paxi[b]	pinf	soly	sotu	nben	ath	Family Annotation
ppssna_223	48	43	1	0	2	2	0	RNase H
2 families	107	98	2	1	1	5	0	Unknown protein

B. *PinfS6* near-unique families: 6 families with 349 genes.

group	total	paxi	pinf	soly	sotu	nben	ath	Family Annotation
ppssna_119	67	3	59	0	3	2	0	gag-pol polyprotein
2 families	137	6	128	0	2	1	0	Ulp1 protease
3 families	145	5	137	2	0	1	0	Unknown protein

C. *Solanum lycopersicum* near-unique families: 7 families with 394 genes.

group	total	paxi	pinf	soly	sotu	nben	ath	Family Annotation
ppssna_84	85	2	6	70	4	0	3	Mutator-like transposase
ppssna_91	82	1	1	69	7	2	2	gag-pol polyprotein
ppssna_129	65	0	0	58	7	0	0	Ulp1 protease
ppssna_198	52	1	0	49	2	0	0	retrotransposon zinc finger CCHC-type protein
ppssna_301	40	0	0	40	0	0	0	Unknown protein
ppssna_325	38	0	0	37	0	0	1	transposase-related hAT dimerisation domain
ppssna_427	32	1	0	28	3	0	0	En/Spm-like transposon protein

D. *Solanum tuberosum* near-unique families: 33 families with 4496 genes.

group	total	paxi	pinf	soly	sotu	nben	ath	Family Annotation
ppssna_30	141	0	0	1	140	0	0	'chromo' domain containing protein

ppssna_259	44	0	0	4	40	0	0	RRNA intron-encoded homing endonuclease
15 families	2168	2	2	7	2155	2	0	gag-pol polyprotein
ppssna_11	260	0	0	0	260	0	0	Retrotransposon gag protein
10 families	1480	0	10	4	1466	0	0	Integrase core domain
6 families	403	1	4	5	386	7	0	Unknown Protein

E. *Nicotiana benthamiana* near-unique families: 59 families with 5342 genes.

group	total	paxi	pinf	soly	sotu	nben	ath	Family Annotation
ppssna_95	80	2	3	3	4	67	1	ribosomal protein L37a; Protein synthesis initiation factor
ppssna_186	53	0	0	0	0	53	0	F box protein
ppssna_50	120	3	2	2	0	113	0	FAR1 related sequence
4 families	241	0	0	1	4	236	0	gag-pol polyprotein
ppssna_48	121	0	0	0	0	121	0	Retrotransposon gag protein
ppssna_195	52	2	1	0	1	48	0	Endonuclease/exonuclease/phosphatase
ppssna_169	56	0	0	0	0	56	0	Ribonuclease H
ppssna_151	60	0	0	0	0	60	0	Transposon MuDR mudrA protein
ppssna_422	32	0	2	0	0	30	0	Mutator-like transposase
6 families	730	2	7	34	4	682	1	Ulp1 protease
44 families	3889	26	8	5	7	3843	0	Unknown protein

[a]The families listed here contained at least 30 members with least 80% of the genes in a single species.

[b]Species abbreviations as in Table 1.

Gene Families Found in a Single Genus

This study includes two pairs of species from the same genus: the two *Petunia* species, and the two *Solanum* species*, S. lycopersicum* and *S. tuberosum*. To examine genus-level differences, we searched for gene families with at least 30 members that had less than 80% of members from one species, but more than 80% in any two species (Table 3).

Of the eight families that were mainly found in the two *Petunia* species, six families had similar numbers of genes in *PinfS6* and *PaxiN*. The other two families were found at least four times more frequently in one family than the other, but did not reach the cutoff of 80% from one species to be classified as single-species families. The *Petunia*-specific gene families included three categories that were not unknown proteins or related to transposable elements: two families of cytochrome P450 genes, a family of HXXXD-type acyl-transferase family proteins, and a family of replication protein A 70 kDa DNA-binding subunit B genes. The latter two families are quite interesting as they represent potentially significant differences between *Petunia* and other Solanaceae; they are discussed below.

All of the three *Solanum*-specific gene families were primarily found in one species: two from *S. lycopersicum*, which were both transposable element genes, and one from *S. tuberosum*, a family of

genes with unknown function. This finding is consistent with the idea that the two *Solanum* species are more evolutionarily diverged than the two *Petunia* species.

We also found two gene families that reached the criterion of 80% of members from two different species that were not members of the same genus. One of these was a family of disease resistance genes that has many members in the two *Solanum* species as well as in *Arabidopsis*. The other was a family of self-incompatibility proteins that surprisingly was found primarily in *Arabidopsis*.

Table 3. Gene families in which 80% of members came from two different species.

A. Genus *Petunia*.

Group	Total	paxi[a]	pinf	soly	sotu	nben	ath	Family Annotation
ppssna_3	506	218	228	0	4	56	0	Mutator transposase
ppssna_85	85	33	42	0	2	8	0	Unknown protein
ppssna_100	77	6	60	4	1	6	0	Mutator transposase
ppssna_111	70	41	28	0	1	0	0	replication protein A 70 kDa DNA-binding subunit B
ppssna_128	65	48	11	1	5	0	0	RNase H family protein
ppssna_145	61	28	23	3	7	0	0	HXXXD-type acyl-transferase family protein
ppssna_197	52	26	26	0	0	0	0	Cytochrome P450
ppssna_276	42	16	19	2	2	3	0	Cytochrome P450

B. Genus *Solanum*.

Group	Total	paxi	pinf	soly	sotu	nben	ath	Family Annotation
ppssna_203	51	3	3	38	3	1	3	Helitron helicase-like protein
ppssna_230	47	2	2	36	5	2	0	Mutator transposase
ppssna_451	31	2	3	4	21	1	0	Unknown protein
ppssna_477	30	2	0	1	24	3	0	Unknown protein

C. Other Groups.

Group	Total	paxi	pinf	soly	sotu	nben	ath	Family Annotation
ppssna_33	136	2	4	16	36	4	74	Disease resistance protein (TIR-NBS-LRR class)
ppssna_456	30	2	0	3	4	0	21	Plant self-incompatibility protein S1 family

[a]Species abbreviations as in Table 1.

Gene families in Solanacae vs. *Arabidopsis*

We attempted to locate and analyze gene families that are well-distributed among the five Solanaceae species but rare in *Arabidopsis*, or which are almost unique to *Arabidopsis*. Table 4A shows a summary of gene families with at least 30 members that had fewer than 80% of genes in any of the Solanaceae species or in the *Petunia* or the *Solanum* genera, but which had fewer than 2.5% of their genes in *Arabidopsis*. These included 46 gene families containing 3285 genes. The largest annotation categories were disease resistance (749 genes), transposable elements (736 genes), cytochrome P450 (307 genes), and protein degradation (210 genes). In addition, there were 486 genes with no annotation other than

"unknown". At least some of these unknown genes are probably transposon-related: blastp searches using these genes with the NCBI nr database show some weak hits to both DNA transposons and retrotransposons. There were also 797 genes with functions that did not any of the other categories; these included several families of receptor protein kinases, genes involved in the biosynthesis of primary and secondary metabolites, and transporters.

Table 4. Numbers of genes and gene families found in functional categories: Solanaceae vs. *Arabidopsis*.

	families	paxi genes	pinf genes	soly genes	sotu genes	nben genes	ath genes
A. Solanaceae Gene Families.							
Total number	59	538	794	515	663	775	8
DNA transposon	9	55	172	183	27	111	0
Retrotransposon	3	38	7	22	47	44	0
F-box protein	0	0	0	0	0	0	0
cytochrome P450	4	58	86	40	83	40	0
DNA-interacting	1	6	6	5	8	7	0
Disease resistance	11	139	193	105	204	108	1
Metabolic enzymes	5	38	42	38	59	33	2
Protein degradation	6	68	112	20	50	22	1
Protein kinase	4	24	27	26	49	31	1
Carbohydrate active	4	39	54	33	46	40	0
Lipid active	2	10	13	18	29	13	2
Other	1	12	12	6	4	7	1
Unknown protein	9	51	70	19	57	319	0
B. *Arabidopsis* Gene Families							
Total number	15	11	9	7	8	5	913
DNA transposon	0	0	0	0	0	0	0
Retrotransposon	0	0	0	0	0	0	0
F-box protein	6	6	4	3	5	2	469
cytochrome P450	1	1	2	1	1	1	33
DNA-interacting	0	0	0	0	0	0	0
Disease resistance	0	0	0	0	0	0	0
Metabolic enzymes	0	0	0	0	0	0	0
Protein degradation	0	0	0	0	0	0	0
Protein kinase	2	1	2	2	1	1	80
Carbohydrate active	2	0	0	0	0	0	79
Lipid active	0	0	0	0	0	0	0
Other	3	2	1	1	1	0	216
Unknown protein	1	1	0	0	0	1	36

Table 4B shows 15 gene families (913 genes) that had at least 80% of their members from *Arabidopsis*. The largest single annotation category (469 *Arabidopsis* genes) is proteins containing an F-box. There were also 36 genes with unknown functions and 408 genes in other categories. The unexpected lack of transposable element genes may be attributable to the removal of such genes in the original *Arabidopsis* annotation process.

Functional Categories of Gene families Expanded in *Petunia*

In a further attempt to find interesting gene families that were expanded in *Petunia*, we identified all gene families with at least five members from one of the *Petunia* species and no more than one member from each of the non-*Petunia* species. The families were divided into *PaxiN*-specific (30 families with 261 genes), *PinfS6*-specific (24 families with 198 genes), and *Petunia*-specific (88 families with 1292 genes) categories. The families were then categorized (Table 5).

Table 5. **Gene families with at least 5 members from *Petunia* and no more than 1 member from each of the non-*Petunia* species.**

Family	Total	paxi	pinf	soly	sotu	nben	ath	Family annotation
A. *PaxiN*								
ppssna_3008	11	1	10	0	0	0	0	Blue copper protein; cupredoxin
ppssna_4370	9	9	0	0	0	0	0	AT hook motif DNA-binding family protein
ppssna_5915	8	7	1	0	0	0	0	Cyclin-D-binding Myb-like transcription factor
ppssna_5912	8	8	0	0	0	0	0	1-aminocyclopropane-1-carboxylate synthase
ppssna_5909	8	8	0	0	0	0	0	Cytochrome P450
ppssna_14546	5	5	0	0	0	0	0	Ribonuclease H-like domain
6 families	58	55	3	0	0	0	0	Mutator-like Transposase
3 families	28	27	1	0	0	0	0	Endonuclease/exonuclease/phosphatase
13 families	115	103	12	0	0	0	0	Unknown protein
B. *PinfS6*								
ppssna_3024	11	0	11	0	0	0	0	CCR4-NOT transcription complex subunit
2 families	11	5	6	0	0	0	0	Ribonuclease H
ppssna_5832	8	0	8	0	0	0	0	mitogen-activated protein kinase 12
ppssna_14492	5	0	5	0	0	0	0	protein disulfide isomerase
19 families	0	0	0	0	0	0	0	Unknown protein
C. *Petunia* genus								
11 families	129	52	62	6	5	3	1	F-box family protein
ppssna_111	70	41	28	0	1	0	0	Replication protein A
ppssna_197	52	26	26	0	0	0	0	Cytochrome P450
ppssna_699	24	7	17	0	0	0	0	FAR1-related protein
ppssna_759	23	5	18	0	0	0	0	Disease resistance protein (CC-NBS-LRR class)

ppssna_886	21	7	13	1	0	0	0	FAR1-related protein
ppssna_1900	14	6	7	1	0	0	0	ubiquitin-protein ligase
ppssna_2246	13	5	4	1	1	1	1	cysteine-type endopeptidase inhibitor
ppssna_2248	13	4	6	1	1	1	0	Interferon related developmental regulator
ppssna_2593	12	4	8	0	0	0	0	Zinc finger, RING/FYVE/PHD-type
ppssna_2595	12	6	6	0	0	0	0	Histone superfamily protein
ppssna_2594	12	4	6	1	1	0	0	HXXXD-type acyl-transferase
ppssna_2591	12	10	2	0	0	0	0	1-aminocyclopropane-1-carboxylate oxidase
ppssna_3010	11	3	8	0	0	0	0	Nse4 component of Smc5/6 DNA repair complex
ppssna_3036	11	7	4	0	0	0	0	HXXXD-type acyl-transferase
ppssna_3019	11	7	4	0	0	0	0	GDSL esterase/lipase
ppssna_3020	11	7	4	0	0	0	0	ATP synthase subunit-like protein
ppssna_3027	11	5	3	1	1	1	0	UDP-Glycosyltransferase
ppssna_3531	10	5	4	0	0	1	0	GDSL esterase/lipase
ppssna_3551	10	2	6	0	1	1	0	RNase Phy3 [*Petunia* x hybrida]
ppssna_3544	10	3	7	0	0	0	0	Disease resistance protein (CC-NBS-LRR class)
ppssna_4347	9	5	4	0	0	0	0	HXXXD-type acyl-transferase
ppssna_5824	8	1	5	1	0	1	0	FAR1-related protein
ppssna_5835	8	3	5	0	0	0	0	serpin serine protease inhibitor
ppssna_5846	8	1	5	1	1	0	0	Caffeate O-methyltransferase (COMT) family)
ppssna_5878	8	2	5	0	1	0	0	Cytochrome P450
ppssna_11782	6	6	0	0	0	0	0	Zinc finger, CCHC-type; transcription regulation
13 families	172	103	61	1	2	5	0	Mutator-like Transposase
2 families	93	11	80	1	0	1	0	Ulp1 protease
ppssna_3011	11	0	9	0	1	1	0	Helitron helicase protein
ppssna_1177	18	17	0	0	1	0	0	gag non-LTR retrotransposase
4 families	84	66	10	3	3	2	0	Ribonuclease H
2 families	63	10	53	0	0	0	0	Aminotransferase-like, plant mobile domain
26 families	275	130	124	7	6	8	0	Unknown protein

Most of the gene families that are expanded in one *Petunia* species relative to the other, or in both *Petunia* species relative to the other Solanaceae, are transposable element genes (35 families with 543 genes) or genes with no known function (57 families with 553 genes).

Both cytochrome P450 genes (three families with 60 genes) and disease resistance genes (two families with 33 genes) of the CC-NBS-LRR type (Moffett et al, 2002; McHale et al., 2006) are found in several families that are shared widely among all species, and in other families are almost species-specific (Tables 3 and 7). Both of these categories contain many genes, some of which are needed for common disease and metabolic issues, while others are needed for species-specific problems.

The DNA-interacting proteins include several families of transcription factor as well as some genes that may be involved in chromatin remodeling, DNA repair, or other higher level regulatory functions. Similarly, there are four families of genes that can be categorized as affecting protein degradation, including 11 families of F-box proteins, which are involved with ubiquitylation of proteins targeted for degradation.

The most interesting families in this analysis have relatively specific metabolic functions that may be useful clues to *Petunia* phenotypes; several of these genes are discussed more fully below.

Tandem duplications

We performed a search for tandem duplications that were amplified in *Petunia* relative to tomato. The size distribution of tandem arrays in the three genomes is quite similar, with *PaxiN* having 7732 genes in 2865 tandem arrays, *PinfS6* having 7883 genes in 2967 tandem arrays, and tomato having 8715 genes in 3018 arrays (Figure 1). Despite similar size distributions, the *Petunia* tandem arrays differ strikingly from the tomato arrays in that the average distance between array members in both *Petunia* species is more than twice that of tomato (Table 6).

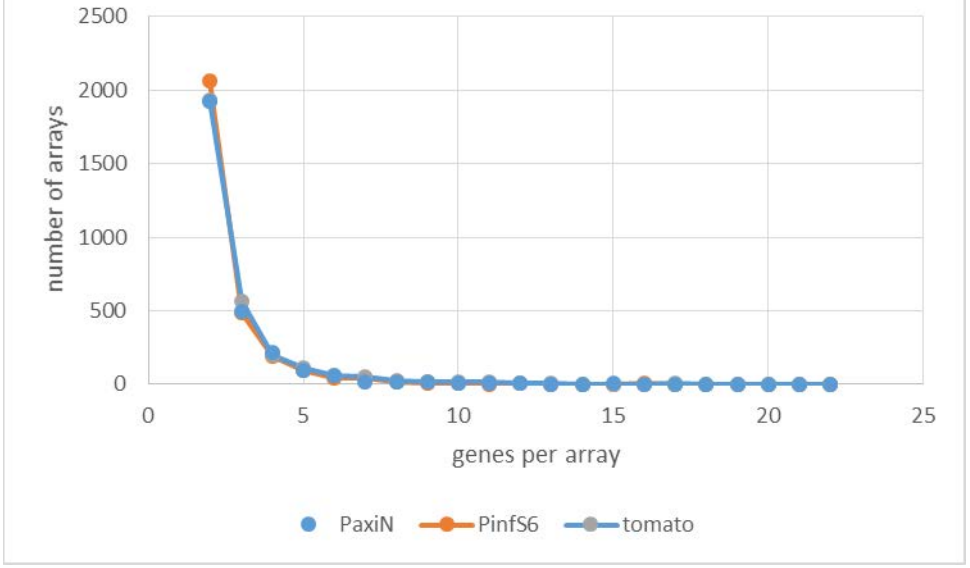

Fig 1. Number of tandem arrays vs. number of genes per array.

Individual tandem arrays were compared between species by examining the best blastp hits and determining which arrays they belonged to. Arrays were considered to be of significantly different sizes if the sum of the array sizes of best blastp hit genes in the target species was less than half the array size in the original species.

Of the 26 tandem arrays with 10 or more members in *PaxiN*, ten were unique to this species, while six of the 20 large arrays were unique to *PinfS6*. Out of 46 total large arrays in the two species combined, 19 were found in *Petunia* but not tomato (Table 7). Prominent among these were three arrays of pentatricopeptide genes unique to *PaxiN*; these genes are involved in RNA editing in the organelles (Nakamura et al., 2012). Three arrays of genes encoding 2-oxoglutarate and Fe(II)-dependent dioxygenase superfamily proteins, two in *PaxiN* only and one in both *Petunia* species, were found; these enzymes play a variety of metabolic roles including hormone and pigment synthesis as well as DNA repair and histone demethylation (Farrow and Facchini, 2014). Each *Petunia* species also has a unique array of cytochrome P450 genes. Many of the *Petunia*-specific tandem arrays encode F box protein

genes, many of which are involved in ubiquitylation of proteins destined to be degraded as well as hormone signaling pathways and self-incompatibility (Smalle and Vierstra, 2004). The arrays that are shared by all three species, including glutaredoxin, protease inhibitors, and genes involved with carbohydrate metabolism.

Table 6. Average distance between tandem array members.

	PaxiN		*PinfS6*		tomato	
Distance	number of arrays	cumulative percentage	number of arrays	cumulative percentage	number of arrays	cumulative percentage
< 10,000 bp	958	33.4	967	32.6	1678	55.6
10,000 - 49,999 bp	1273	77.9	1327	77.3	998	88.7
50,000 - 99,999 bp	327	89.3	393	90.6	167	94.2
100,000 -199,999 bp	215	96.8	201	97.3	88	97.1
200,000 - 499,999 bp	77	99.5	68	99.6	51	98.8
500,000 - 999,999 bp	15	100	8	99.9	21	99.5
> 1,000,000 bp	0	100	3	100	15	100
total number of arrays	2865		2967		3018	
median average dist	19331		20504		8384	

Table 7. Large tandem arrays that are expanded in *Petunia* species.

Array number	size	annotation
***PaxiN* only.**		
paxi_1645	22	Pentatricopeptide repeat-containing protein
paxi_291	15	2-oxoglutarate and Fe(II)-dependent oxygenase superfamily protein
paxi_887	14	Pentatricopeptide repeat-containing protein
paxi_1870	12	LRR receptor-like serine/threonine-protein kinase
paxi_1311	11	F-box protein
paxi_331	11	cytochrome P450
paxi_582	10	Protein phosphatase 2C family protein (signal transduction)
paxi_1014	10	Pentatricopeptide repeat-containing protein
paxi_1881	10	Chaperone DnaJ-domain superfamily protein
paxi_2296	10	2-oxoglutarate and Fe(II)-dependent oxygenase superfamily protein
***PinfS6* only.**		
pinf_1629	16	Unknown protein GO:0003677 (DNA binding)
pinf_2365	16	Disease resistance protein (CC-NBS-LRR class)
pinf_941	13	Cytochrome P450
pinf_848	10	Actin
pinf_891	10	F-box family protein
pinf_970	10	unknown protein, chloroplast related

***PaxiN* and *PinfS6*.**

paxi_2069	17	Polyadenylate-binding protein 2 (=pinf_1516)
paxi_1140	15	F-box/FBD/LRR-repeat protein
paxi_2715	13	Cc-nbs-lrr, resistance protein with an R1 specific domain (=pinf_1603)
paxi_24	12	Protein of Unknown Function (DUF239)
paxi_65	12	2-oxoglutarate and Fe(II)-dependent oxygenase superfamily protein (=pinf_2470)
paxi_558	12	F-box/RNI-like superfamily protein (=pinf_936)
paxi_848	11	F-box/RNI-like superfamily protein
paxi_853	11	F-box/FBD/LRR-repeat protein
paxi_1101	11	Cc-nbs-lrr, resistance protein
paxi_1890	11	F-box/FBD/LRR-repeat protein (=pinf_188)
paxi_2438	10	Unknown protein
pinf_1603	18	Cc-nbs-lrr, resistance protein with an R1 specific domain (=paxi_2715)
pinf_2470	16	2-oxoglutarate and Fe(II)-dependent oxygenase superfamily protein (=paxi_65)
pinf_2209	14	F-box/RNI-like superfamily protein
pinf_1516	12	Polyadenylate-binding protein 2 (=paxi_2069)
pinf_188	11	F-box/FBD/LRR-repeat protein (=paxi_1890)
pinf_323	10	F-box family protein
pinf_936	10	F-box/RNI-like superfamily protein (=paxi_558)
pinf_1451	10	wall-associated kinase

***PaxiN* and *S. lycopersicum*.**

paxi_292	18	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein
paxi_1683	15	SAUR-like auxin-responsive protein family
paxi_38	10	cytochrome P450

***PinfS6* and *S. lycopersicum*.**

pinf_967	13	alpha/beta-Hydrolases superfamily protein

All three species.

paxi_2165	12	subtilisin-like serine protease (=pinf_1903)
paxi_913	10	Glutaredoxin (=pinf_1755)
paxi_1778	10	UDP-glucosyltransferase
pinf_468	12	Serpin (Serine protease inhibitor)
pinf_902	12	cytochrome P450
pinf_1903	11	subtilisin-like serine protease (=paxi_2165)
pinf_526	10	xyloglucan specific endoglucanase inhibitor
pinf_1755	10	Glutaredoxin (=paxi_913)

DISCUSSION

The gene families used in this study were generated by OrthoMCL from an all-vs-all blastp search. OrthoMCL is a clustering method and as such the groups it forms are partly based on unknown internal parameters and vagaries of the input data. However, several intermediate sets of gene families were generated from various combinations of the species, and we found that gene families were relatively stable. On the average, the effect of adding the genes from one additional species was that 96.9% of the genes in the largest 500 gene families remained clustered together in the same family. Also, an average of 434 of these families had all members clustered into the same family when an additional species was added, and an average of 469.5 families had at least 90% of their members clustered into the same family. In general, gene family membership remained intact when new sets of genes were added. For this reason we believe that when genes of the same functional annotation are placed in different gene families, the two families will be detectably different when subjected to phylogenetic analysis and probably functionally differentiated.

The evolution of transposons, as well as that of the Ulp1 protease sequences that are associated with many Mutator-like transposons, is apparently much less constrained than other gene families, as these genes are the largest category of species- and genus-specific genes. The large number of genes with unknown functions in the species-specific categories might also include novel transposons. However, genes with unknown function may also include some with functions that are truly novel to *Petunia*, and they may also include some sequences that are incorrectly labeled as genes.

The most interesting gene families that have been differentially amplified in *Petunia* are those with known metabolic functions that might point to *Petunia*-specific activities. Genes coding for 1-aminocyclopropane-1-carboxylate synthase (ACS) were found in all the species studied, with between 9 and 17 members per genome. However, *PaxiN* has an additional 18 ACS genes in two families, with only two *PinfS6* genes in one family and none in the other. The enzyme coded by these genes is the rate-limiting step in ethylene biosynthesis. It is encoded as a multigene family in all sequenced plant genomes, with different isozymes that have different patterns of gene expression (Yamagami et al., 2003). Similarly, both *Petunia* species share a family of 1-aminocyclopropane-1-carboxylate oxidase genes, not found in the other species. These genes code for the final step in ethylene production (Kende, 1989).

The HXXXD-type acyl-transferase genes are found in both *Petunia* species in relatively equal numbers, but there are very few members from any of the other species. An important sub-category of these enzymes are the hydroxycinnamoyl CoA quinate transferases, which are involved in the production of chlorogenic acids (Sonnante et al., 2010). These secondary metabolites have multiple functions in plants, including lignin biosynthesis, UV light protection and pest resistance.

Caffeate O-methyltransferase (COMT) family enzymes are involved with the biosynthesis of many phenylpropanoids, including lignin, anthocyanin, and a number of defensive and stress-induced compounds (Joshi and Chiang, 1998). There is one member of this family in each of the Solanaceae except *PinfS6*, which has five members.

PaxiN has a family of genes that contain a cupredoxin fold, including blue copper protein and laccase. Genes of this general type are found in all the species studied, with several different families found in each species. In *Petunia*, one family has ten copies in *PinfS6* but only one in *PaxiN*, with no

members in any of the other species. These proteins perform oxidation-reduction reactions in photosynthesis, oxygenations and they have an intense blue color (Nersissian et al., 1998).

FAR1-related sequences are transcription factors that mediate the far-red light responses in higher plants. They have a domain structure similar to Mutator-like transposases. It is possible that the gene families annotated as FAR1-related are transposon-related (Hudson et al., 2003).

Replication protein A is a heterotrimer composed of 70 kDa, 32 kDa, and 14 kDa subunits that is involved with several aspects of DNA metabolism, including replication and repair. Animals and yeast have only a single set of RPA genes, but plants have at least two distinct types of RPA proteins, which may have separated replication functions from repair functions, and which may be specific to organelles or the nucleus (Sakaguchi et al., 2009). Gene family ppssna_111 contains 70 members, 69 of which are from *Petunia*.

METHODS

Peptide sequences from *S. lycopersicon*, *S. tuberosum*, and *N. benthamiana* were downloaded from the Sol Genomics Network web site (http://solgenomics.net). For tomato, the ITAG Release 2.3 genomic annotations were used (Tomato Gene Consortium, 2012). The PGSC version 3.4 was used for *S. tuberosum* (Xu et al., 2011). *N. benthamiana* draft genome sequences from release 0.4.4 (Bombarely et al., 2012). The *Arabidopsis thaliana* peptides and annotations were downloaded from http://www.arabidopsis.org , using version TAIR10.

Blastp from blast+ version 2.2.27 was used to do an all-versus-all comparison of the peptides from the six species (Altschul et al., 1997). The blastp results were used to group genes into families with OrthoMCL version 2.0.8 (Li et al., 2003). The annotations from the genes grouped into families were analyzed using custom software written in Perl. Functional annotations for groups were created by attempting to find a consensus among annotations for individual genes, with a bias against leaving the function as “unknown”.

Tandem duplications in the two *Petunia* species and in tomato were detected using the SynMap function of CoGe (Lyons and Freeling, 2008; Lyons et al., 2008), and were defined as two or more genes with significant sequence homology lying within ten genes of each other on the same contig. Individual tandem arrays were compared between species by examining the best blastp hits and determining which arrays they belonged to. Arrays were considered to be of significantly different sizes if the sum of the array sizes of best blastp hit genes in the target species was less than half the array size in the original species.

ACKNOWLEDGEMENTS

The authors thank the NIU Department of Computer Science for use of its Gaea computing cluster.

AUTHOR CONTRIBUTIONS

Both authors performed the research and analyzed the data. M.A.J. wrote the article.

REFERENCES

Altschul, S.F., Madden,T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller,W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389-3402.

Barkan, A. (2011). Expression of plastid genes: Organelle-specific elaborations on a prokaryotic scaffold. Plant Physiol. **155**: 1520–1532.

Bombarely, A., Rosli, H.G., Vrebalov, J. , Moffett, P. , Mueller, L.A. , and Martin, G.B. (2012). A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. Molecular Plant-Microbe Interactions **25**: 1523-1530.

Elmore, Z.C., Donaher, M., **Matson, B.C.**, **Murphy, H.**, **Westerbeck, J.W.,** and **Kerscher, O**. (2011). Sumo-dependent substrate targeting of the SUMO protease Ulp1. BMC Biology **9**: 74.

Farrow, S.C. and Facchini, P.J. (2014). Functional diversity of 2-oxoglutarate/Fe(II)-dependent dioxygenases in plant metabolism. Front. Plant Sci. **5**: 524.

Flagel, L.E. and Wendel, J.F. (2009). Gene duplication and evolutionary novelty in plants. New Phytol. **183**: 557-564.

Hillwig, M.S., Liu, X., Liu, G., Thornburg, R.W., and MacIntosh, G.C. (2010). Petunia nectar proteins have ribonuclease activity. J Exp. Bot. **61**: 2951–2965.

Hoen, D.R., Park, K.C., Elrouby, N., Yu, Z., Mohabir, N., Cowan, R.K., and Bureau, T.E. (2006). Transposon-mediated expansion and diversification of a family of ULP-like genes. Mol. Biol. Evol. **23**: 1254-1268.

Hudson, M.E., Lisch, D.R., and Quail, P.H. (2003). The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. Plant J. **34**: 453–471.

Joshi, C.P. and Chiang, V.L. (1998). Conserved sequence motifs in plant S-adenosyl-L-methionine-dependent methyltransferases. Plant Mol. Biol. **37**: 663–674.

Kende, H. (1989). Enzymes of ethylene biosynthesis. Plant Physiol. **91**: 1-4.

Li, L, Stoekert, C. J., and Roos, D.S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. **13**: 2178-2189.

Lyons, E., and Freeling M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. **53**: 661-673.

Lyons E, Pedersen B, Kane J, and Freeling M (2008). The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates rosids. Tropical Plant Biol. **1**: 181-190**.**

McHale, L., Tan, X., Koehl, P., and Michelmore, R.W. (2006). Plant NBS-LRR proteins: adaptable guards. Genome Biology **7**: 212.

Moffett, P., Farnham, G., Peart, J., and Baulcombe, D.C. (2002). Interaction between domains of a plant NBS–LRR protein in disease resistance-related cell death. EMBO J. **21**: 4511–4519.

Nakamura, T., Yagi, Y., and Kobayashi, K. (2012). Mechanistic insights into pentatricopeptide repeat proteins as sequence-specific RNA-binding proteins for organellar RNAs in plants. Plant Cell. Physiol. **53**: 1171-1179.

Nersissian, A.M., Immoos, C., Hill, M.G., Hart, P.J., Williams, G., Herrmann, R.G., and Valentine, J.S. (1998). Uclacyanins, stellacyanins, and plantacyanins are distinct subfamilies of phytocyanins: Plant-specific mononuclear blue copper proteins. J. Biol. Chem. **278**: 49102-49112.

Ohno, S. (1970). *Evolution by Gene Duplication.* Springer-Verlag, New York.

Sakaguchi, K., Ishibashi, T., Uchiyama, Y., and Iwabata, K. (2009). The multi-replication protein A (RPA) system – a new perspective. FEBS J. **276**: 943–963.

Smalle, J. and Vierstra, R.D. (2004). The ubiquitin 26S proteasome proteolytic pathway. Annu. Rev. Plant Biol. **55**: 555-590.

Sonnante, G., D'Amore,R., Blanco, E., Pierri, C.L., De Palma, M., Luo, J., Tucci, M., and Martin, C. (2010). Novel hydroxycinnamoyl-coenzyme A quinate transferase genes from artichoke are involved in the synthesis of chlorogenic acid. Plant Physiol. **153**: 1224-1238.

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**: 635-641.

Van Leeuwen, H., Monfort, A., and Puigdomenech, P. (2007). Mutator-like elements identified in melon, Arabidopsis and rice contain ULP1 protease domains. Mol. Genet. Genomics **277**: 357-364.

Xu, X. et al. (2011). Genome sequence and analysis of the tuber crop potato. Nature **475**: 189-195.

Yamagami, T., Tsuchisaka, A., Yamada,K., Haddon, W.F., Harden, L.A. and Theologis, A. (2003). Biochemical diversity among the 1-amino-cyclopropane-1-carboxylate synthase isozymes encoded by the Arabidopsis gene family. J. Biol. Chem **278**: 49102-49112.