BMC Plant Biology

BioMed Central

Research article

The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*

Steven B Cannon*[1,2], Arvind Mitra[3], Andrew Baumgarten[1,4], Nevin D Young[1,2] and Georgiana May[1,4]

Address: [1]Plant Biology Department, University of Minnesota, St. Paul, MN 55108, USA, [2]Plant Pathology Department, University of Minnesota, St. Paul, MN 55108, USA, [3]Adam Ave 532, Ithaca, NY 14850, USA and [4]Ecology, Evolution, and Behavior Department, University of Minnesota, St. Paul, MN 55108, USA

Email: Steven B Cannon* - cann0010@umn.edu; Arvind Mitra - amitra55@yahoo.com; Andrew Baumgarten - baum0217@umn.edu; Nevin D Young - neviny@umn.edu; Georgiana May - gmay@umn.edu

* Corresponding author

Published: 01 June 2004

BMC Plant Biology 2004, **4**:10 doi:10.1186/1471-2229-4-10

Received: 01 November 2003
Accepted: 01 June 2004

This article is available from: http://www.biomedcentral.com/1471-2229/4/10

Abstract

Background: Most genes in *Arabidopsis thaliana* are members of gene families. How do the members of gene families arise, and how are gene family copy numbers maintained? Some gene families may evolve primarily through tandem duplication and high rates of birth and death in clusters, and others through infrequent polyploidy or large-scale segmental duplications and subsequent losses.

Results: Our approach to understanding the mechanisms of gene family evolution was to construct phylogenies for 50 large gene families in *Arabidopsis thaliana*, identify large internal segmental duplications in *Arabidopsis*, map gene duplications onto the segmental duplications, and use this information to identify which nodes in each phylogeny arose due to segmental or tandem duplication. Examples of six gene families exemplifying characteristic modes are described. Distributions of gene family sizes and patterns of duplication by genomic distance are also described in order to characterize patterns of local duplication and copy number for large gene families. Both gene family size and duplication by distance closely follow power-law distributions.

Conclusions: Combining information about genomic segmental duplications, gene family phylogenies, and gene positions provides a method to evaluate contributions of tandem duplication and segmental genome duplication in the generation and maintenance of gene families. These differences appear to correspond meaningfully to differences in functional roles of the members of the gene families.

Background

Most genes in *Arabidopsis thaliana* are members of gene families. Similarity searches between all predicted proteins show that 65 – 85% of all *Arabidopsis* genes are homologous to at least one other gene in the genome, depending on similarity thresholds ([1] and analysis in this paper). There is a wide range in gene family sizes, with more than 400 receptor kinase genes [2,3], ~270 – 285 cytochrome P450 genes [1,4,5], and many small families or unique genes. The dramatic variation we observe in gene family size and distribution may be affected by many processes, including tandem duplication with high rates

of birth and death and gene duplication resulting from larger scale genome events such as polyploidy or duplications of large chromosomal regions (referred to in this paper as "segmental duplications"). We provide a quantitative characterization of the gene duplication patterns evident in the evolution of 50 large gene families in *A. thaliana*.

The complete sequencing of the *A. thaliana* genome revealed numerous large-scale segmental duplications [1,6-10]. Several studies have concluded that at least two rounds of duplications have probably occurred in the *A. thaliana* genome, with many losses and rearrangements leaving a mosaic of "segmental duplications" or "duplication blocks" [7,10-14]. Most duplication blocks appear to have come from one round of polyploidy, estimated by various methods to have occurred 20 – 40 Mya, before the evolution of the genus *Brassica* but after the separation of Brassicaceae from other close eudicot families [7,10-12]. The portion of the genome that exists in duplicate regions serves as a baseline for evaluating whether genes in a given gene family have been lost or retained at a rate higher than expected for the genome as a whole. If most duplication blocks did in fact originate during one round of polyploidy, this duplication could also be used to provide an internal reference point to use in comparing the rates of amino acid substitutions in members of different gene families.

While polyploidy is one mechanism by which gene family copy numbers expand, tandem or local duplication is the most commonly evaluated mechanism for gene family expansion. Tandem duplication often results from unequal crossing-over [15] and multiple episodes of unequal crossovers might lead to increasing or decreasing copy numbers in gene families, or to simple cycling of genes without large changes in gene family size. Though not investigated in this paper, transposable elements may also have played an important role in gene duplications and genome rearrangements in *Arabidopsis* [16].

To determine the relative importance of segmental and local duplications in the evolution of large gene families, we developed software to identify clades in gene family phylogenies that have arisen either through segmental or local duplications. In 50 large gene families in *A. thaliana*, we find that contributions made by these two processes differ greatly from gene family to gene family. We discuss the possible biological significance of these differences in gene family evolution.

Results

Strategy

Our general approach consisted of the following steps. Details, parameters, and software are described in the Methods section.

1) Choose initial gene families and preliminary sequence membership. We began with 2001 *Arabidopsis* PIR superfamilies, available at MIPS [17], and refined family membership in the subsequent steps.

2) Narrow the gene family selection on the basis of domain arrangements. We determined the Pfam [18] domains of all sequences in each gene family, assessed the consistency of domain arrangements in each family, and excluded families with particularly complex domain arrangements, such as those in several kinase families.

3) Iteratively construct and refine gene family alignments. We constructed T-Coffee [19] alignments using a maximum of 30 genes from each family, then generated hidden Markov models (HMMs), realigned all proteins in each family to the model, used the model to re-search the full set of predicted *Arabidopsis* proteins, retrieve sequences with expectation values less than 10^{-10}, and re-align those to the HMM.

4) Trim alignments for use in phylogenetic analyses. This involved removing indel regions, first by removing residues falling outside of the "match states" in the HMM, and then by visually inspecting and in some cases removing other poorly aligned or indel regions.

5) Calculate phylogenies. We generated parsimony and bootstrapped neighbor joining trees, and also calculated maximum likelihood branch lengths for the parsimony topologies.

6) Predict segmental duplications in the *Arabidopsis* genome, using DiagHunter [20,21]. In a two-dimensional dot plot of amino acid similarity "hits" between chromosomes, segmental duplications appear as diagonal features. The sets of homologous gene pairs that contribute to such features were used in the next step. Similarity is at a BLASTP bit score threshold of 500, with other parameters described in [21].

7) Determine gene pairs in a gene family having the same coordinates as found in a pair of duplication blocks. Any such gene pair likely duplicated at the same time as the pair's duplication block. We carried out this and the next three steps using OrthoParaMap software developed for this purpose, and described in detail at [22,23].

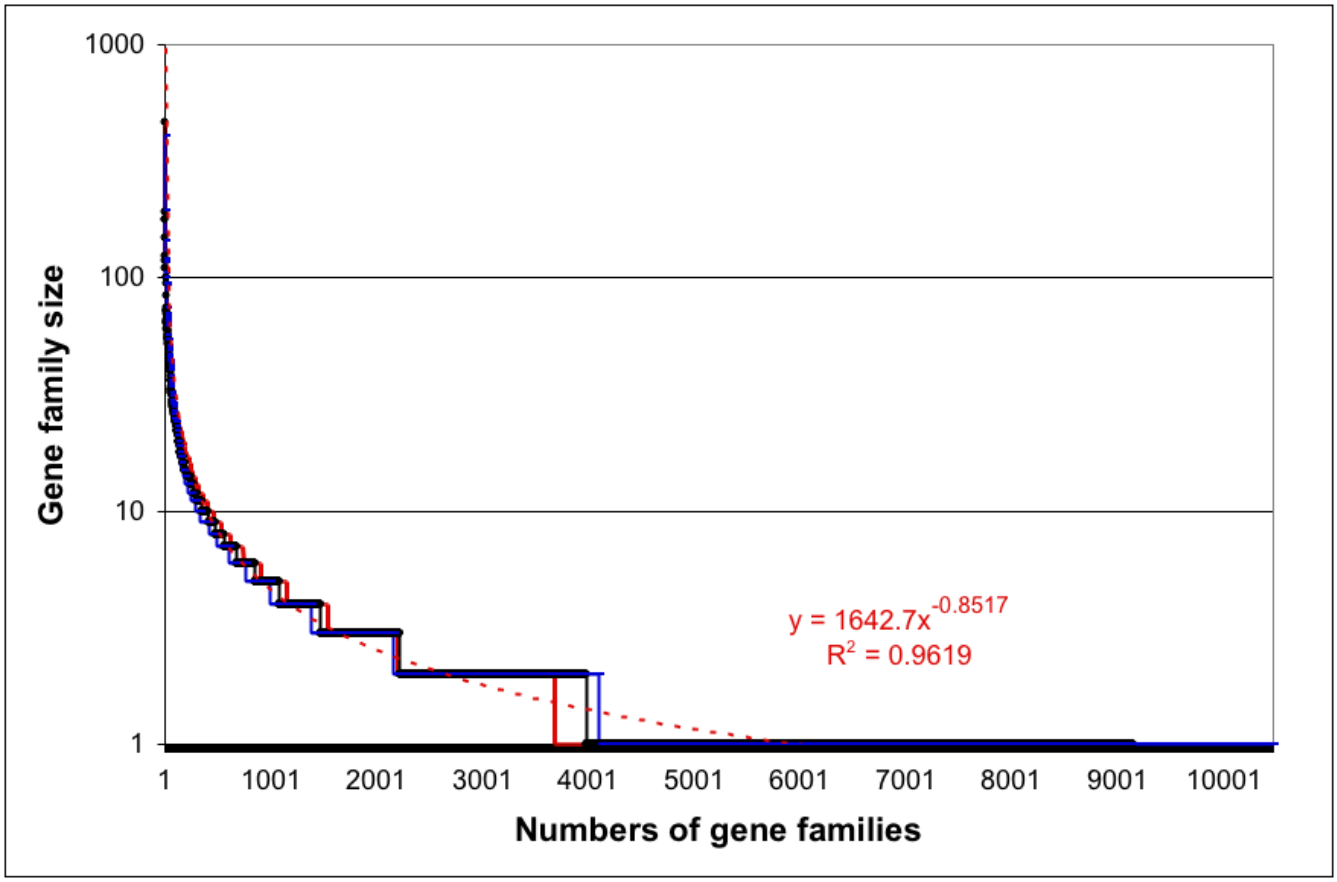

Figure 1
Sizes of gene families in *A. thaliana* Approximate gene family sizes were calculated using single-linkage clustering of BLASTP similarities below E-value thresholds of 10^{-10} (red), 10^{-20} (black), and 10^{-30} (blue). At the resolution of this graph, these lines follow nearly the same path. The curves follow a power law distribution. The best-fit power law equation for the 10^{-10} curve is indicated on the graph.

8) Annotate the gene phylogenies with information on duplication block membership. Infer nodes that likely originated through segmental duplications, and annotate the phylogeny with this information.

9) Use gene position information to infer which closely related genes (defined in terms of position in gene phylogenies) are located physically "close" to one another (defined in terms of the physical distance between genes, as described below). Infer nodes that likely originated through tandem duplications, and annotate the phylogeny with this information.

10) Add translated EST consensus sequences from other species to help provide additional context. This involved using each *A. thaliana* sequence in each gene family to query TIGR unigene sets for soybean, *M. truncatula*, *Lotus japonicus*, tomato, potato, and corn, then choosing the longest translations, aligning these to the HMM, and recalculating the phylogenies using the same procedures as for *A. thaliana* (step 5). Though generally not integral to this project, this information was helpful in determining evolutionary patterns for some families – and particularly for families consisting of small, highly-expressed proteins. Because of space constraints, figures 5, 6, 7, 8 include only *A. thaliana*, *Medicago*, and tomato sequences, though phylogenies for all sequences are included at [24].

Study set selected from all large A. thaliana gene families

A high-throughput phylogenetic analysis of many gene families is complicated at the start by questions of what constitutes a gene family [25-28]. Conceptually, gene families have a common ancestor, arise by gene duplication, and may share similar functions. The diversity of sequence and function in gene families often makes delimiting gene families difficult. Operationally, gene

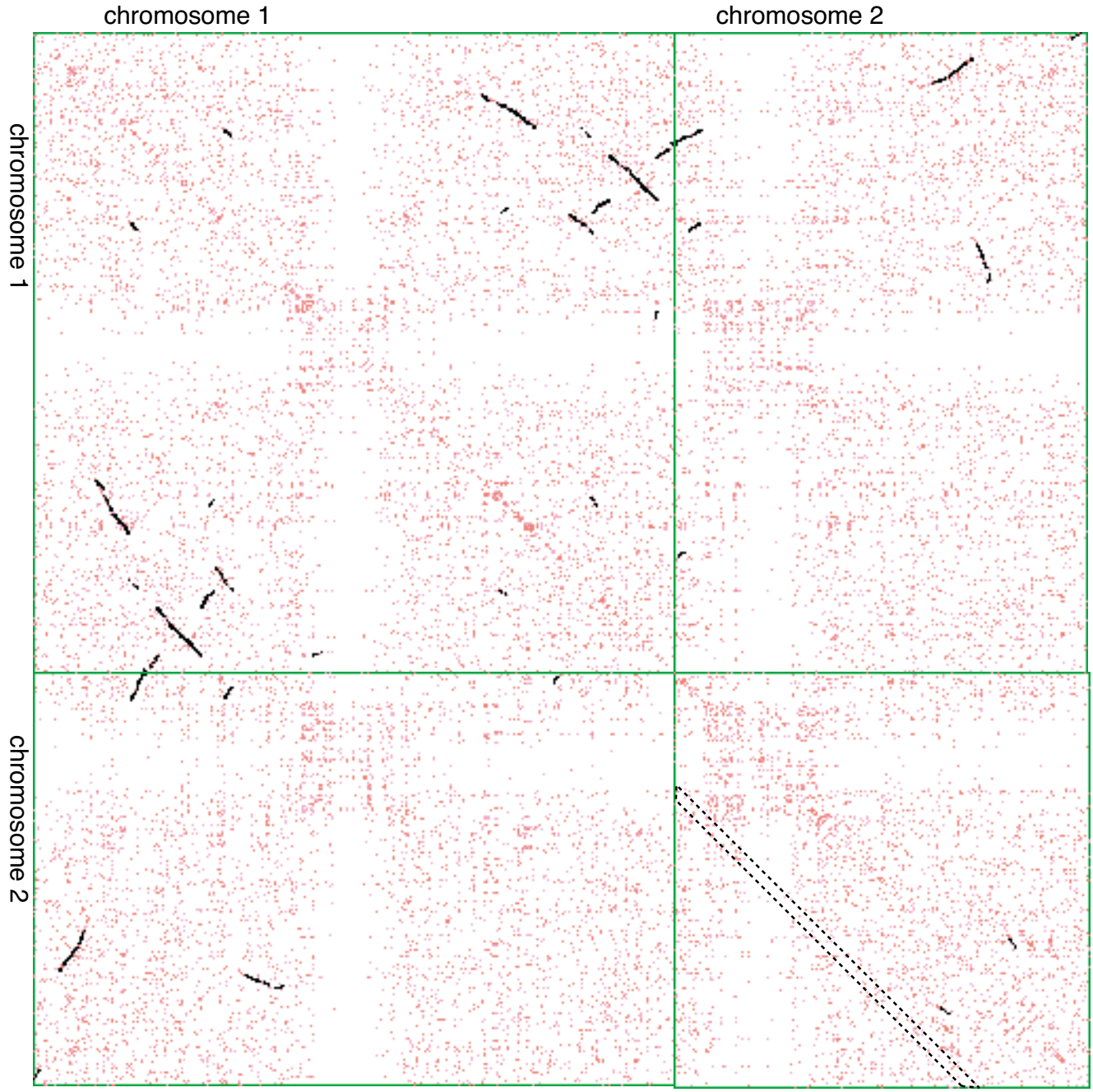

Figure 2
Dot plots of similarities in *A. thaliana* chromosomes 1 and 2 Chromosome 1 is shown to the top and left, chromosome 2 on the bottom and right. Dots represent BLASTP similarities at bit score thresholds of 500. Synteny blocks identified by DiagHunter [20,21] are shown in black (larger images are available at [24]). Hits of proteins to themselves have been suppressed. A large excess of local duplications is apparent in higher densities near the main diagonal. The average density at any given distance between genes can be calculated from diagonal strips through the dot plot. One such strip is highlighted in chromosome 2 × 2.

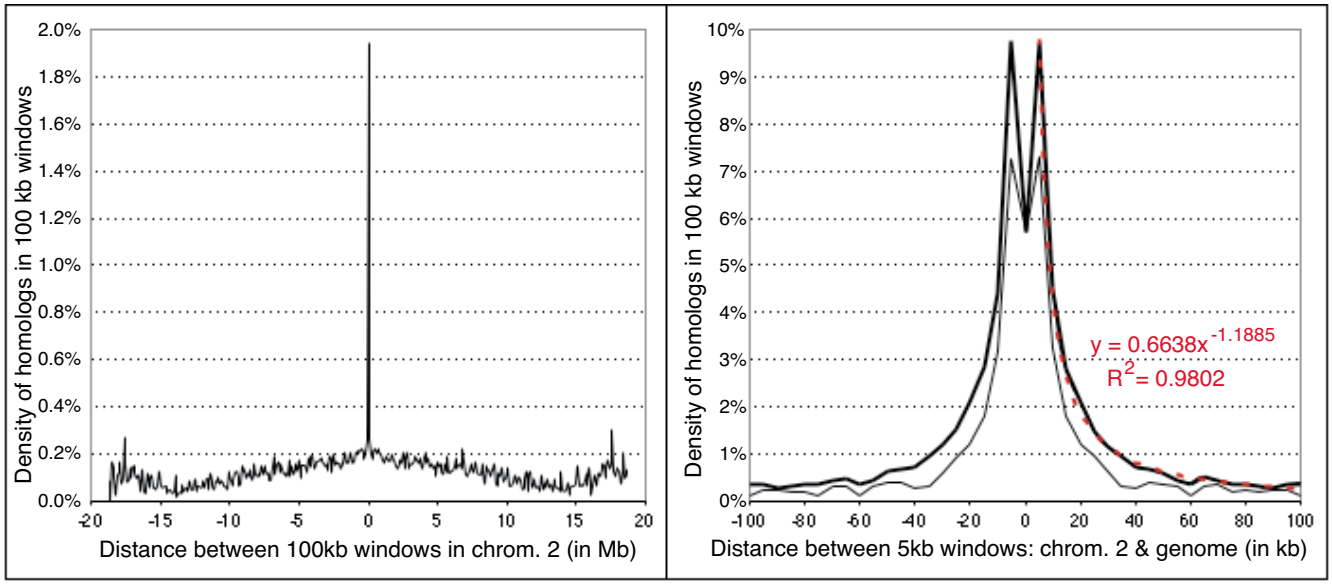

Figure 3

Densities of homologs by genomic distance in *A. thaliana* chromosome 2 and genome-wide The graph on the left (3A) shows average densities in 100 kb diagonal strips through the chromosome 2 × 2 dot plot of similarities. The value at any position in the graph represents the number of homologs between 100 kb windows around a query location and a target location. The graph on the right (3B) shows similar density measurements, but within 5 kb windows and spanning up to 200 kb between genes. The x-axis measures the difference between the query and target locations. The thin line shows the density-by-distance plot for chromosome 2 × 2. The bold line shows the comparable plot for the whole genome, with scores averaged across all five *A. thaliana* chromosome comparisons. The red dotted line shows the best-fit exponential equation to the whole-genome curve, fitted from 5 kb to 100 kb.

families can be defined in terms of levels of sequence similarity and domain composition, but a simple similarity threshold may be misleading if the threshold inappropriately splits a divergent superfamily, or inappropriately groups together separate gene families that share a common domain [25,26].

To limit the scope of this study and to avoid some of the complexities presented by superfamilies with diverse domain arrangements, we arbitrarily chose 50 gene families with at least 20 members, functional domains in common, and consistent family membership. Consistent family membership was judged by distributions of expectation scores in HMM searches (using hmmer [29]) of the *A. thaliana* proteome. Preference was given to families in which there is a clean drop-off in HMM E-value, with members having scores of at worst 10^{-10} and nonmembers generally having much poorer scores. Apart from the minimum family size, we chose better-studied families, though some have no members with known functions or Pfam domains [18]. Lastly, we chose families with a range of family sizes, from families of 20 members up to the approximately 225-member cytochrome P450 superfamily (though the total number of P450 genes in *A. thaliana*, including members of all diverse subfamilies, is estimated to be 275 – 285 genes [1,4,5]). The families used in this study are shown in Table 1.

To get a sense of the distribution of gene family sizes, we also conducted a simple whole-proteome homology search and single-linkage clustering at two BLASTP [30] thresholds. In this context, single linkage clustering transitively merges sets of genes in which any gene is sufficiently similar to some other gene in the set. These results are shown in Figure 1. The distribution closely follows a power-law (Figure 1; $y = 1642.7x^{-0.8517}$, $R^2 = 0.96$). In such a distribution, there are few families with large numbers of members and many families with few members. A power-law distribution is worth noting in part because it calls for a mechanism for the evolution and maintenance of family sizes. Any proposed mechanism will need to be consistent with the mechanisms of individual gene duplications and losses in various families.

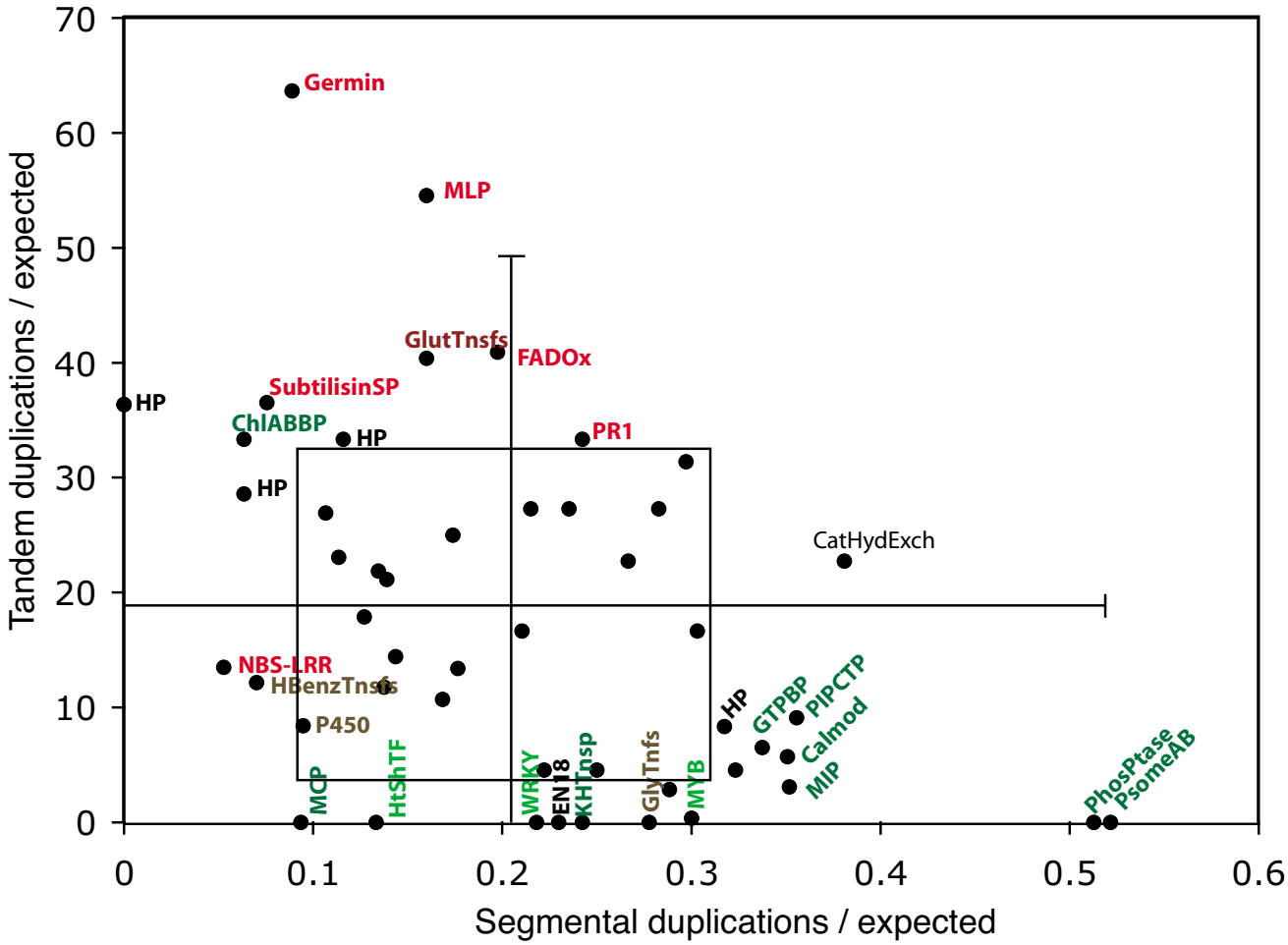

Figure 4

Comparison of observed/expected tandem and segmental duplications for 50 large *A. thaliana* gene families

Ratios of observed to expected tandem duplications in the 50 gene families in the study are shown on the vertical axis, and ratios of observed to expected segmental duplications on the horizontal axis. For purposes of discussion, one and two standard deviations around the means on each axis are shown with a box plot. Among families outside of one standard deviation, families with members that play roles in pathogen defense are indicated in red. Transcription factor families are shown in light green. Several housekeeping genes are shown in dark green. Several broad-function enzyme families are shown in brown. Notice the relative scarcity of gene families that are high in both categories, and eight families that have no apparent tandem duplications.

Using a BLASTP E-value threshold of 10^{-10} (Figure 1) followed by single-linkage clustering, produces 181 potential gene families with at least 20 members, 46 with at least 50 members, and 13 with at least 100 members. At least 85.7% of *A. thaliana* genes have one or more homologs at this threshold. Using a BLASTP threshold of 10^{-20}, single-linkage clustering generates 140 potential gene families with at least 20 members, 40 with at least 50 members, and 10 with at least 100 members. At least 80.6% of *A. thaliana* genes have one or more homolog at this threshold. These BLAST and clustering results provide approximate descriptions of gene family size distributions. The 50 gene families chosen for further analysis were further refined as described in Methods.

Tandem duplications quantified

Our goal is to distinguish between gene duplication resulting from segmental duplication of chromosomal regions, and tandem duplication generating nearby gene copies. This required operational definitions of both gene similarity and genomic proximity. Similarity should be determined in the context of the gene phylogeny because

Table 1: Fifty *A. thaliana* gene families Gene family names or typical gene annotations are given in the first column. The second column contains abbreviated names or mnemonics for the families. Unnamed gene families are given PIR family numbers (e.g. HypProt131). The third column indicates the number of predicted gene sequences included in final *A. thaliana* gene family phylogenies.

Annotation	Short Name	Seqs
calcineurin-like phosphoesterase	CalcinPEst	19
calmodulin	Calmod	79
cation/hydrogen exchanger	CatHydExch	28
chlorophyll a/b-binding	ChlABBP	21
cysteine proteinase	CystProt	31
cytochrome P450	CytP450	225
Enod16	Enod16	32
Enod18/ER6 protein	Enod18_ER6	29
exocyst subunit EXO70	ExocystEX070	23
expansin	Expansin	34
FAD-linked oxidoreductase	FADOxidor	27
flavin-containing monooxygenase	FlavMonoOx	28
germin-like	Germin	30
glutathione transferase; dehydroascorbate reductase	GlutTnsfs	50
glycosyl hydrolase family 1	GlycosHdls	47
glycosyltransferase family 8	GlyTnsf8	24
glycosyl hydrolase family 9	GlyTnsf9	25
auxin-independent growth promoter	GrthRegul	33
GDSL-motif lipase/hydrolase	GSDLLipase	97
GTP-binding; Ras-related GTP-binding	GTPBP	72
acyltransferase	HBenzTnsfs	57
heat shock transcription factor	HtShkTncFct	20
hypothetical	HypProt131	28
hypothetical; esterase-like	HypProt2752	23
hypothetical	HypProt317	42
hypothetical	HypProt536	25
hypothetical	HypProt688	21
lysine/histidine transporter; amino acid permease	KHTnsptr	22
major intrinsic protein (MIP) family	MajIntrinsProt	38
MATE	MATE	50
mitochondrial carrier	MCP	57
MFS	MFS	68
major latex protein (MLP)-related; Bet v I allergen	MLP	25
MYB transcription factor	MYB	120
NBS-LRR disease resistance	NBSLRR	152
oxidoreductase; 20G-Fe(II) oxygenase	Oxidored	95
pathogenesis-related protein 1	PathRelPr1	22
phosphoprotein phosphatase; ser/thr phosphatase	PhosphPtase	26
phototropic response protein	PhotResp	30
phosphatidylinositol/phophatidylcholine transfer prot.	PIPCTP	30
plastocyanin-like domain; blue copper p.; Enod 20	PlastocEn20	37
polygalacturonase	Polygalns	65
oligopeptide transport	POT	48
proteasome alpha and beta subunits	PsomeAB	23
short-chain dehydrogenase/reductase	SCDehydRed	84
subtilisin-like serine protease	SubtilisinSP	53
thaumatin-like	Thaumatin	22
UDP-glycosyltransferase	UDPGlycTnsf	109
WRKY transcription factor	WRKY	55
xyloglucan endotransglycosylase	XyloTGlyc	33

different genes in different families evolve at different rates. We limited the search for tandem duplications to sequences with ≤75% of the average evolutionary distance from terminal nodes to an approximate midpoint root in the phylogeny, the maximum search depth. This is somewhat arbitrary cutoff, but avoids very early duplications in the phylogenies, for which mechanisms are difficult to infer. To determine whether two genes are physically close enough to conclude that they probably arose through tandem duplications, we measured the average genomic distance at which there is an excess number of duplications above the genome average.

Following the approach of Vision et al. [9], we use a dot plot to map the occurrence of two similar sequences located in different genome regions. Locations along the linear sequence of genomic regions are graphed as the X- and Y-axis with each dot at an XY coordinate marking a similarity "hit" (Figure 2). Dot plots of chromosomes compared to themselves, e.g. chrom. 1 by chrom. 1, map local tandem duplication as well as segmental events within the chromosome. Segmental events are represented as dense linear arrays of dots in the same or opposite orientation as the main diagonal but located off the main diagonal.

The density of dots in any portion of the dot plot represents the density of matches between the genome regions being compared. If a large amount of tandem duplication has occurred in a chromosomal region, this will be visible as a densely dotted region near the main diagonal of a chromosome plotted against itself. The dot plots we present do not include the main diagonal itself (showing the similarity of each protein to itself). The average density at any given distance between genes can be calculated from diagonal strips through the dot plot. One strip is presented as an example in Figure 2.

Figure 3 shows the average densities of diagonal strips through the chromosome 2 dot plot. The whole chromosome 2 comparison is shown in the left panel, and duplications within 100 kb upstream or downstream are shown in the right panel. Chromosome 2 has no detectable internal segmental duplications by our analysis, and therefore has the cleanest duplication plot (Figure 3, left panel), with the peak centered on 0 representing tandem duplications.

Breaking down the density by distance observations into closer intervals, we next plotted density in 5 kb windows. The graph in Figure 3b shows similar density measurements, but within 5 kb windows, from 100 kb downstream from a query to 100 kb upstream in chromosome 2. The plot obtained is very similar for all five chromosomes and Figure 3b also shows a genome average. In each, we find a dramatic excess of local duplications falling within 50 kb physical distance. The units are homologs per area, where a meaningful unit area might be $(5\ kb)^2$. The rationale for using $(5\ kb)^2$ in the denominator is that in *A. thaliana*, the average gene density is approximately one gene per 5 kb, so if all genes were homologous, the number of homologs between any two 5 kb regions would be 1. The value of one homolog per $(5\ kb)^2$ in *A. thaliana* might therefore be described as one density unit (d.u.; a term novel to this paper). As would be expected, the highest densities of local duplications are seen at 5 kb (Figure 3b). In windows extending from 5 kb to 10 kb from any gene, the density of apparently duplicated genes (BLASTP threshold of 10^{-10}) is 0.098 d.u. genome-wide. This means that on average, there are ~0.1 homologs within any two 5 kb windows that are separated by 5 kb, or that one duplicated gene in ten is likely to have a homolog very close by. In the 100 kb window centered on any gene, the corresponding density of duplicated genes is approximately 0.020 – 0.035 d.u., depending on the chromosome. In all chromosomes, a clear local duplication effect is not seen beyond 50 kb. Thus, we define tandem duplications as those closely related genes falling within 50 kb of one another.

The distribution of densities of locally duplicated genes by distance follows an exponential distribution (R^2 value of 0.98, Figure 3). Integrating under this curve, 90% of the area under the curve in the interval between 5 kb and 100 kb is found within the smaller interval of 5 kb to 50 kb and represents densities above an average background density of 0.002 d.u. Again, this supports the use of 50 kb as a reasonable threshold for identifying local duplications in *A. thaliana*.

Expected values for tandem and segmental duplications

To compare the relative contributions of tandem and segmental duplications when gene families differ substantially in size, we generated expected values for tandem and segmental duplication events for gene families of each size class, calculated a ratio of observed to expected values for these two mechanisms, and compared the ratios for each family.

We simulated distributions of expected numbers of tandem duplications that would occur by chance for a gene family of a given size in a genome of a given size. The simulation procedure is to randomly place N genes in a 100,000 kb genome (the approximate extent of euchromatic DNA in *A. thaliana*), and to count the number of genes that are within 50 kb of one another. A total of 1000 simulation runs generates a distribution for each gene family size. For small gene families, the probability that two genes will fall near one another follows a Poisson distribution. For example, a mean and variance of 0.12

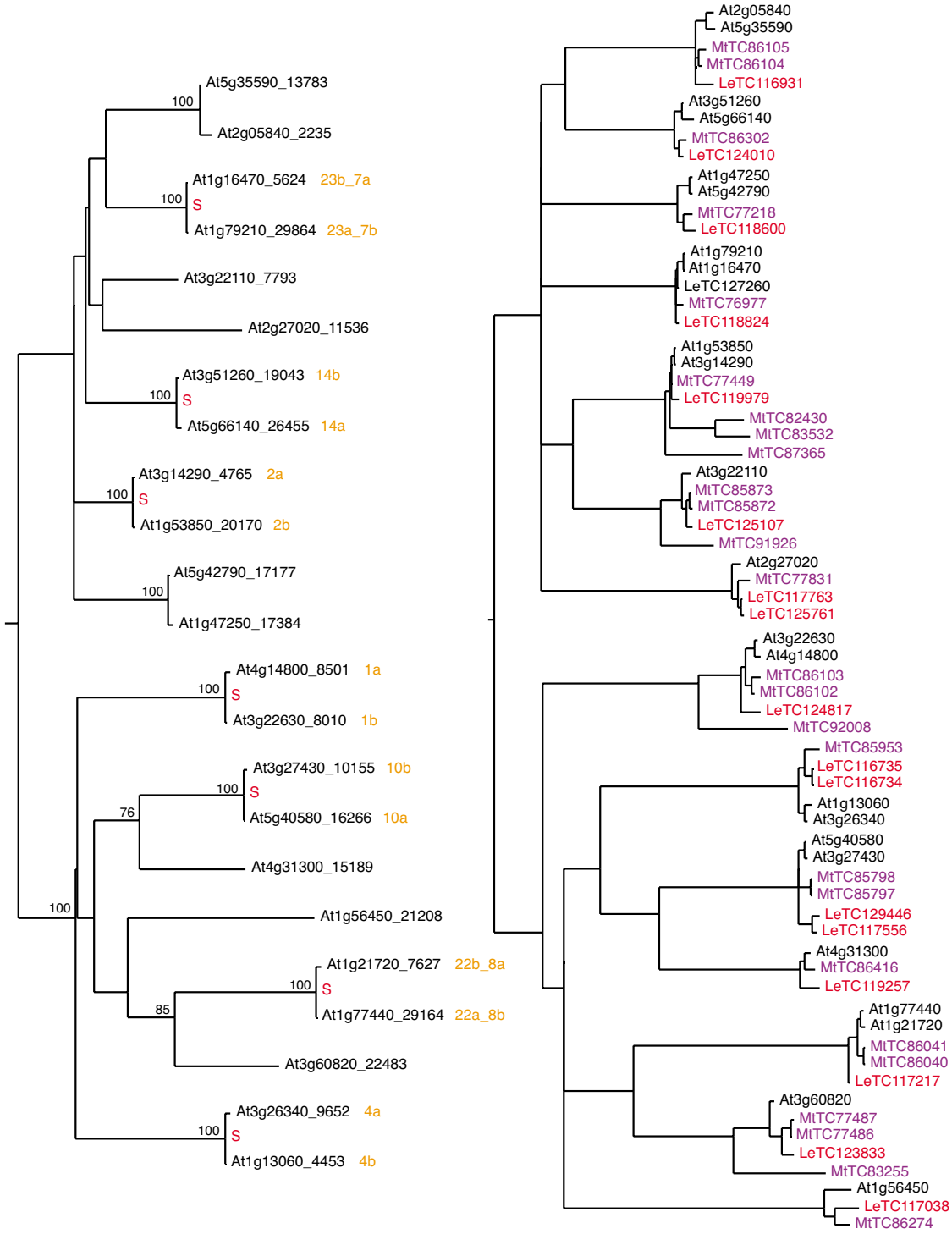

Figure 5

Proteasome 20S subunit family: low tandem, high segmental The phylogeny on the left shows segmental duplications in the *A. thaliana* proteasome 20S subunit family, which lacks tandem duplications. The phylogeny on the right represents the same *A. thaliana* sequences but with *M. truncatula* and tomato EST sequences added to evaluate the degree to which these homologs are conserved. Relationships of clades represented in both phylogenies are in general agreement, with some differences due to instabilities of some deep nodes.

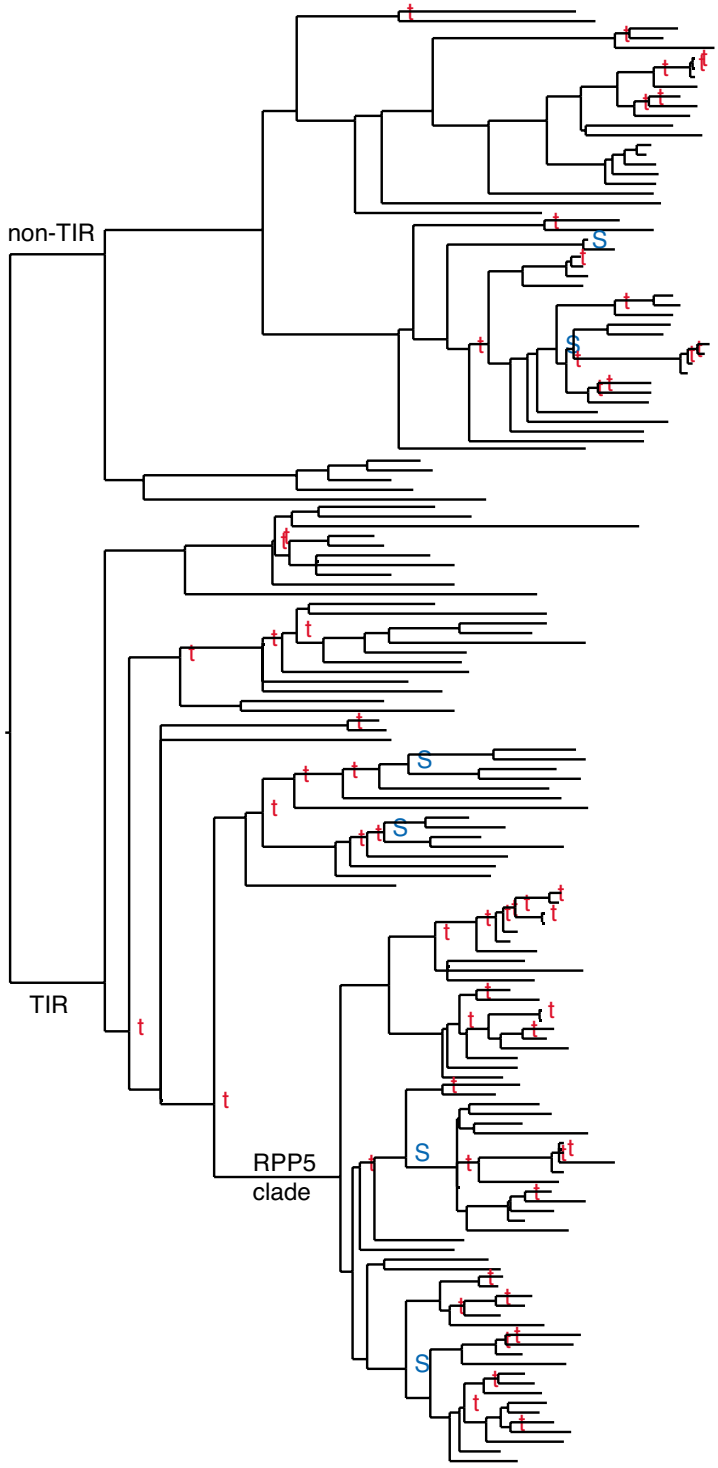

Figure 6
NBS-LRR disease resistance family: moderate tandem, low segmental duplications The NBS-LRR disease resistance family is divided into two subfamilies: the non-TIR subfamily (top third of the phylogeny) and the TIR subfamily (the bottom two-thirds). Tandem duplications are indicated with "t" and segmental with "S." Other duplications are not classified by our methods. For clarity in the large tree, gene names and positions have been removed. The complete phylogeny, including bootstrap values, is available at [24].

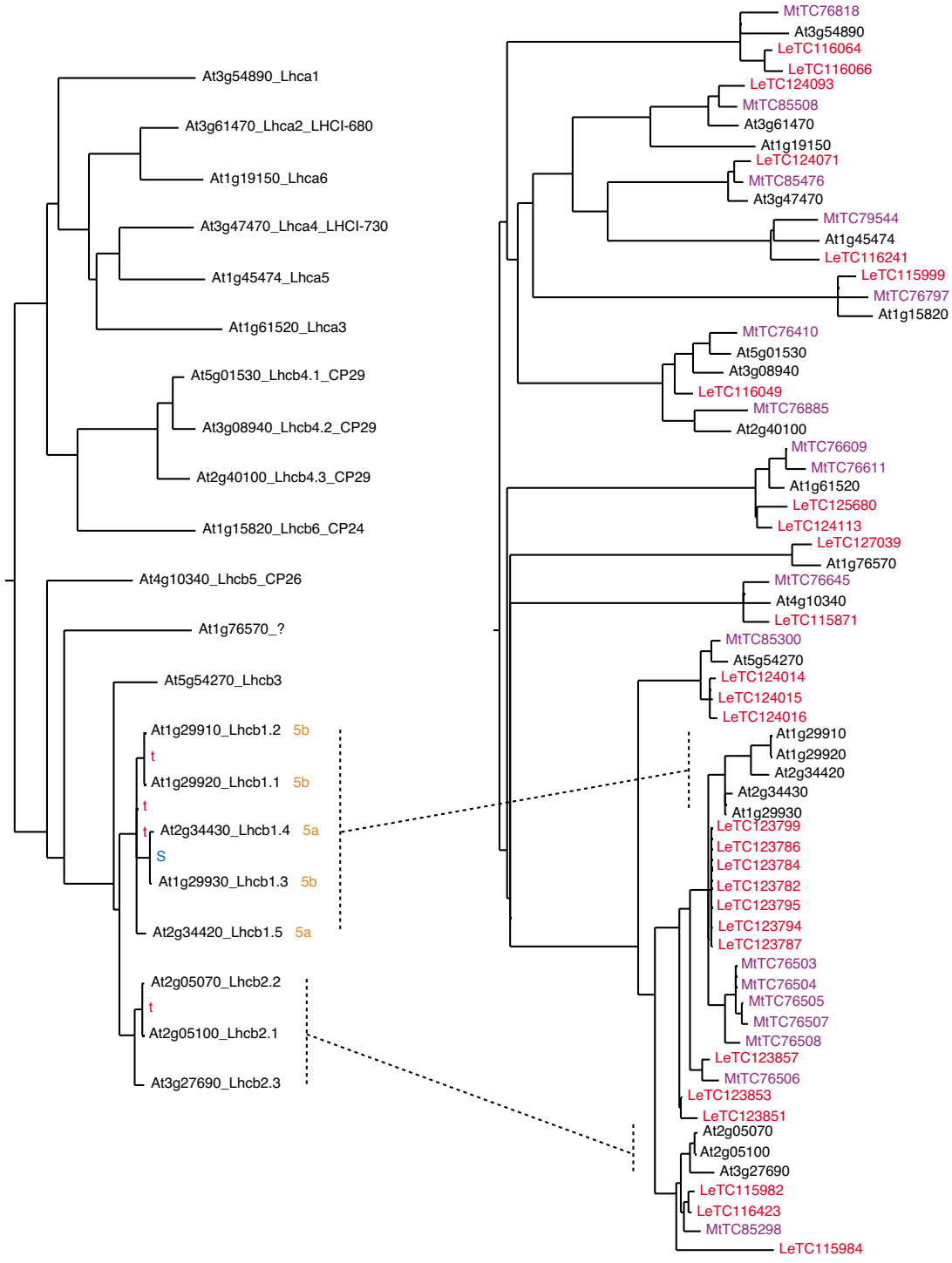

Figure 7

Chlorophyll a/b binding protein family: high tandem, low segmental duplications The phylogeny on the left shows segmental and tandem duplications in the *A. thaliana* chlorophyll a/b binding protein family. Gene names used in the photosynthesis literature are included in this tree. The phylogeny on the right shows the same *A. thaliana* sequences, with *M. truncatula* and tomato EST sequences added to provide an indication of degree of conservation of these sequences and lineages. Notice the tandem duplications in the *A. thaliana lhc1-3* clade, and the corresponding duplications in *Medicago* and tomato, many of which appear to have occurred after separation of these plant families.

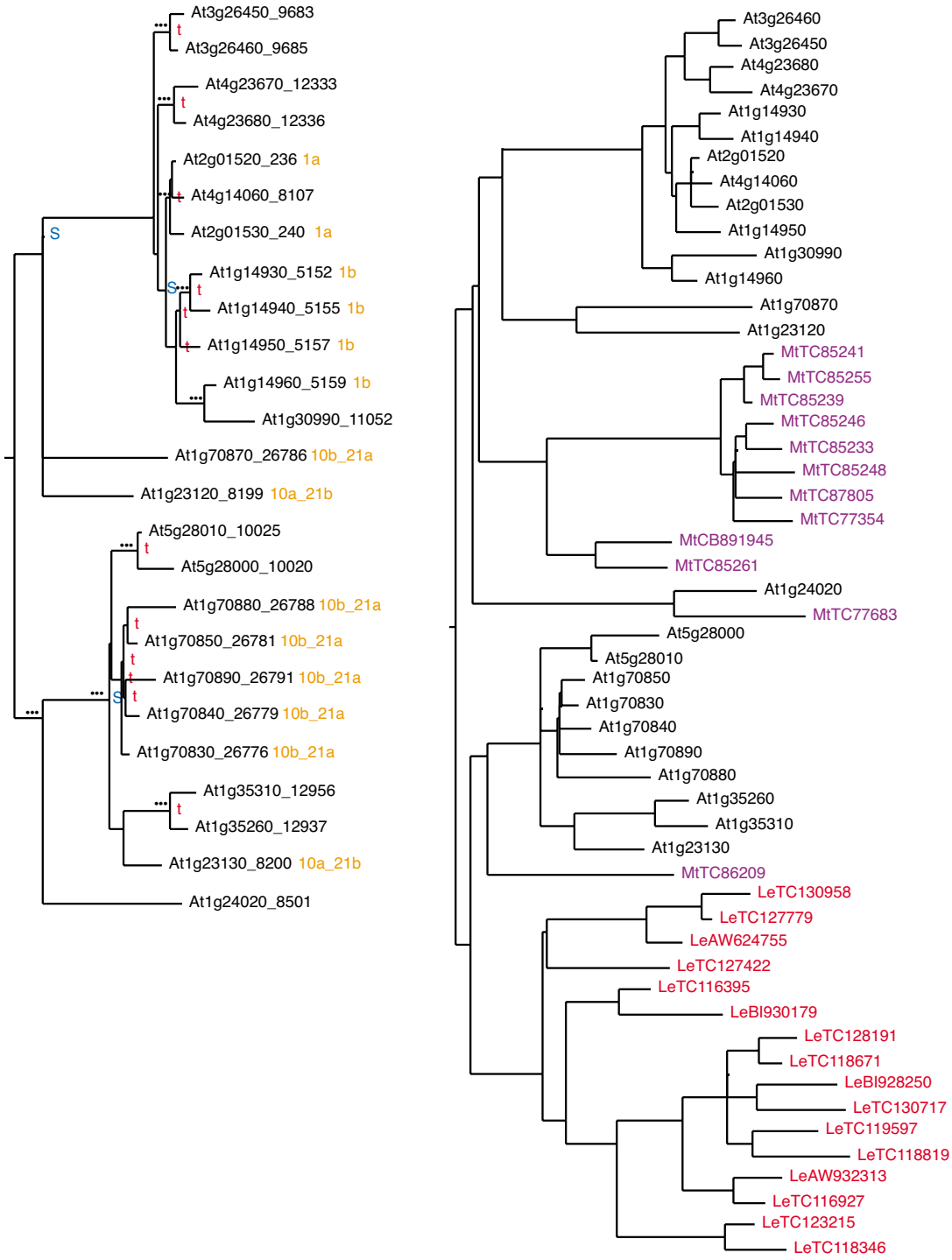

Figure 8

Major latex protein family: high tandem, low segmental duplications The phylogeny on the left shows segmental and tandem duplications in the *A. thaliana* major latex protein family. The phylogeny on the right shows the same *A. thaliana* sequences with *M. truncatula* and tomato EST sequences added to provide an indication of degree of conservation of these sequences and lineages. Clades are generally represented in comparable relationships, with some differences due to instabilities of some deep nodes. Bootstrap values are indicated as follows: *** >90%; ** >=80%; * >=70%. Note the expansion of several clades in each species following separation of these taxa.

Table 2: Tandem and segmental duplications in fifty *A. thaliana* gene families For full gene family names, see Table 1. Gene families described in the text are underlined. Families are organized by high, medium, and low tandem or segmental groups, which are defined by standard deviations above or below the median values for the observed/expected tandem or segmental duplications. The categories in the normalized tandem and segmental columns are indicated as: plain italic = 1 standard deviation below; bold = 1 standard deviation above; bold italic = two standard deviations above.

short name	seqs	tan	seg	exp tan	exp seg	**tan obs/ exp**	**seg obs/ exp**	tan obs/ exp	seg obs/ exp
MCP	57	0	4	0.7	43	*0*	0.09	-1sd	med
HtShkTncFct	20	0	2	0.1	15	*0*	0.13	-1sd	med
WRKY	55	0	9	0.7	41	*0*	0.22	-1sd	med
Enod18_ER6	29	0	5	0.2	22	*0*	0.23	-1sd	med
KHTnsptr	22	0	4	0.1	17	*0*	0.24	-1sd	med
GlyTnsf8	24	0	5	0.1	18	*0*	0.28	-1sd	med
PsomeAB	23	0	9	0.1	17	*0*	***0.52***	-1sd	+1sd
PhosphPtase	26	0	10	0.2	20	*0*	***0.51***	-1sd	+2sd
HBenzTnsfs	57	9	3	0.7	43	12.2	*0.07*	med	-1sd
NbsLrr	152	54	6	4	114	13.5	*0.05*	med	-1sd
HypProt317	42	10	2	0.4	32	28.6	*0.06*	med	-1sd
MYB	120	1	27	2.6	90	0.38	0.3	med	med
PlastocEn20	37	1	8	0.4	28	2.86	0.29	med	med
PhotResp	30	1	5	0.2	23	4.55	0.22	med	med
Enod16	32	1	6	0.2	24	4.55	0.25	med	med
HypProt688	21	1	5	0.1	16	8.33	0.32	med	med
CytP450	225	76	16	9	169	8.43	0.09	med	med
Oxidored	95	20	12	1.9	71	10.7	0.17	med	med
GSDLLipase	97	22	10	1.9	73	11.8	0.14	med	med
MFS	68	13	9	1	51	13.4	0.18	med	med
Polygalns	65	14	7	1	49	14.4	0.14	med	med
CalcinPEst	19	2	3	0.1	14	16.7	0.21	med	med
Thaumatin	22	2	5	0.1	17	16.7	0.3	med	med
SCDehydRed	84	22	8	1.2	63	17.9	0.13	med	med
POT	48	11	5	0.5	36	21.2	0.14	med	med
UDPGlycTnsf	109	49	11	2.2	82	21.9	0.13	med	med
GlyTnsf9	25	5	5	0.2	19	22.7	0.27	med	med
GlycosHdls	47	12	4	0.5	35	23.1	0.11	med	med
ExocystEX070	23	3	3	0.1	17	25	0.17	med	med
MATE	50	14	4	0.5	38	26.9	0.11	med	med
CystProt	31	6	5	0.2	23	27.3	0.22	med	med
Expansin	34	6	6	0.2	26	27.3	0.24	med	med
XyloTGlyc	33	6	7	0.2	25	27.3	0.28	med	med
GTPBP	72	3	19	1	54	3.09	**0.35**	med	+1sd
GrthRegul	33	1	8	0.2	25	4.55	**0.32**	med	+1sd
MajIntrinsProt	38	2	10	0.4	29	5.71	**0.35**	med	+1sd
Calmod	79	8	20	1.2	59	6.5	**0.34**	med	+1sd
PIPCTP	30	2	8	0.2	23	9.09	**0.36**	med	+1sd
CatHydExch	28	5	8	0.2	21	22.7	**0.38**	med	+1sd
ChlABBP	21	4	1	0.1	16	**33.3**	*0.06*	+1sd	-1sd
HypProt131	28	8	0	0.2	21	**36.4**	*0*	+1sd	-1sd
HypProt536	25	8	0	0.2	19	**36.4**	*0*	+1sd	-1sd
SubtilisinSP	53	19	3	0.5	40	**36.5**	*0.08*	+1sd	-1sd
FlavMonoOx	28	7	6	0.2	21	**31.8**	0.29	+1sd	med
HypProt2752	23	4	2	0.1	17	**33.3**	0.12	+1sd	med
PathRelPr1	22	4	4	0.1	17	**33.3**	0.24	+1sd	med
GlutTnsfs	50	21	6	0.5	38	**40.4**	0.16	+1sd	med
FADOxidor	27	9	4	0.2	20	**40.9**	0.2	+1sd	med
Germin	30	14	2	0.2	23	***63.6***	*0.09*	+2sd	-1sd
MLP	25	12	3	0.2	19	***54.5***	0.16	+2sd	med

neighboring genes is observed for a 20-member gene family. For large gene families, the probability approaches a normal distribution. For example, a mean of 4.0 neighboring genes, a variance of 15.4, and a standard deviation of 3.92 is observed for a 100-member gene family. The simulations provide a means of accounting for tandem duplications expected by chance alone, against which we compare observed values (Table 2). As we show below, the expected values are far lower than the observed values for most gene families because tandem duplication processes have not randomly distributed copies across the genome.

Our goal for calculating expected numbers of segmental duplications was to establish an easily interpretable normalizing constant (a different objective than establishing values for a standard null hypothesis). Our assumption was that the majority of genes resulting from segmental duplications have been lost, and we wanted a way to compare extent of loss between families beyond the level expected due only to the loss of large duplicate regions. By our method of identifying segmental duplication blocks [20,21], approximately 75% of the euchromatic portion of the *Arabidopsis* genome exists in at least one duplication block. If all genes within those duplicated regions had been retained, then (all other factors being equal) the fraction of gene copies expected in segmentally duplicated regions would also be 75%. In fact, the proportion of retained gene copies is much lower than this, but 75% provides a baseline and normalizing constant for comparing observed counts of gene copies due to segmental duplication in gene families of different sizes (Table 2).

Counts of tandem and segmental duplications

Table 2 shows counts of tandem and segmental duplications in each family, together with ratios of these counts to the expected genome average for tandem or segmental counts for each family size. Other types of events, including transpositions or remnants of segmental duplications, were not classified. The ratio of observed/expected tandem duplication counts demonstrate an enormous range from 0 to 63; some families are the apparent result of no tandem duplication while one, the Germin family, demonstrates 63 times as many gene copies in tandem arrays as would be expected by chance. The ratio of observed/expected segmental duplication events range from 0 to 0.52; some families have lost all segmental duplicates predicted by the model, while some have lost only about half the duplicates predicted by the model. Table 2 presents gene families grouped by low, medium, and high ratios of observed/expected tandem duplication events, and then grouped by low, medium, and high ratios of observed/expected segmental duplication events. To generate these classes, cutoffs are set at one standard deviation above and below the median. Several families in the tandem and segmental categories also fall above two standard deviations, and these are also indicated in Table 2.

Gene families represented in Table 2 tend either to fall into high-tandem/low-segmental duplication classes or vice versa as is evident in the moderate negative correlation found in a plot of expected/observed segmental and tandem duplications (correlation coefficient = -0.47; R^2 = 0.22; p = 0.00057 for ANOVA F-statistic; Figure 4). Among the eight low-tandem duplication families, none are in the low-segmental duplication category, and two of the eight have segmental duplication counts that place them approximately two standard deviations above the segmental-duplication median. Among the eight low-segmental duplication families, none are in the low-tandem duplication category, and five fall more than one standard deviation above the median ratio of observed/expected tandem-duplication events. There are gene families such as PR1 and CatHydExch with high numbers of segmental or tandem duplications compared to that expected, but not high numbers of both. Neither the ratios of observed/expected tandem events nor the ratios of observed/expected segmental duplications are correlated with gene family size (the R^2 values are 0.044 and 0.066, with p-values 0.14 and 0.08, respectively).

In our set of 50 gene families, the low-tandem duplication class appears to be represented by highly conserved, housekeeping or key regulatory gene families, while the medium- and high-tandem duplication classes are represented by families involving pathogen defense or diverse enzymatic functions. Families involved in pathogen defense all fall in the medium- or high-tandem duplication classes; the NBS-LRR [31,32], Thaumatin [33], Germin [34,35], PR1 [36], and Major Latex Protein/PR10 families [37]. The low-tandem duplication class includes two of the three transcription factor families (heat shock and WRKY) and some housekeeping gene families (mitochondrial carrier proteins [38,39], proteasome 20S subunit family [40,41]).

Gene phylogenies from multiple species

Some phylogenies that include only *A. thaliana* sequences appear to have long internal branches – potentially indicating rapid evolution. Addition of homologous sequences from other species provides a means of testing whether genes in these families have evolved rapidly, or whether long internal branches indicate ancient differences between highly conserved protein sequences. This approach is shown in Figure 5, a phylogeny of the 20S proteasome subunit family [41,42]. The right-hand phylogeny includes representatives from three species: *A. thaliana*, tomato, and *M. truncatula*. The tight clustering of sequences at the end of long internal branches indicates

that this family consists of highly conserved amino acid sequences that have been retained in these genomes for extended times – though it should be said that taxa represented here are fairly closely related dicotyledons, and sequences from basal angiosperms or gymnosperms would likely be placed much more deeply in the phylogeny. Similarly, sequences from multiple species were also used for comparisons shown in Figures 7 and 8.

The multi-species approach taken here generally provides qualitative rather than quantitative measures of evolutionary patterns. For tomato and *Medicago*, we used translated EST tentative consensus (TC) sequences, which are error-prone. Nevertheless, for gene families with highly-expressed, relatively short transcripts, the information gives estimates of minimum evolutionary distances between *A. thaliana* and other dicot gene homologs.

Discussion

This paper describes differences in the relative importance of tandem and segmental duplication to the size and evolution among large gene families in the *A. thaliana* genome. Tandem duplications are clearly an important engine generating new gene copies in genomic clusters, where unequal crossovers generate new diversity. Segmental duplication events have a different effect as they may widely disperse gene copies throughout the genome where they experience few recombinational exchanges with parental copies [43]. To study the joint effects of these genome processes on multigene family evolution, we placed gene families into low, medium, and high tandem duplication classes and low, medium, and high segmental duplication classes, and investigated attributes of some better-studied families within each duplication class.

Distribution of gene family sizes

The frequency distribution of A. thaliana gene family sizes closely follows a power-law relationship (Figure 1). A plausible explanation for this distribution in other genomes was proposed by Huynen and van Nimwegen [44]. In their model, gene families are founded by a single ancestor, and through duplications and deletions, the family size fluctuates over time, with the possibility of the family going extinct from the genome. The requirements of the model are that all members in a family have the same probability of duplication or loss at any given time, different gene families may have different probabilities at any given time, and the average of all duplication probabilities is less than one (preventing gene families from growing to infinity). Under these general conditions, the model generates a power-law distribution of the sizes of surviving gene families [44-46] and thus, selection need not be invoked to explain gene copy number distribution *per se*.

Still, our observations show that since the time of segmental events (the most recent of which is estimated to have occurred 20 – 40 Mya [7,9,12], varying numbers of gene copies have been maintained in segmentally duplicated regions across different gene families. For example, gene copies generated by segmental duplication are more often retained following polyploidy in the more slowly-evolving MYB gene family (Table 2, [24]) whereas, in the large, rapidly-evolving NBS-LRR disease resistance family, duplication in local genomic clusters is common with surprisingly low retention of segmental duplications. Below, we consider the possible biological significance of gene duplication patterns for each class of gene families.

Low tandem, low segmental duplication

Eight gene families were classified as low-tandem duplication and of these, most fell in the moderate or high segmental duplication classes. A few families demonstrating low tandem duplication levels also demonstrate relatively low segmental duplication levels compared to a genome-wide average. The mitochondrial carrier protein family (MC or MCP [38,39]) and heat shock transcription factor family [47] each have retained few segmental duplications and yet demonstrate no apparent tandem duplications or a clustered organization. The MC proteins serve as antiporters, preferentially exchanging one solute for another [38,39,48]. Structurally characterized members of the MC family are dimers. Conceivably, additional gene copies might disrupt the stoichiometry of protein dimers in these transmembrane complexes, particularly once duplicated genes were lost following polyploidy. It remains to be tested whether gene duplication and loss patterns in members of protein complexes generally differ from patterns for monomeric proteins.

It is likely that a relatively large portion of the variance in the segmental losses across the 50 gene families is a result of the stochastic process of genomic loss following polyploidy and thus, will appear a more course-grained process than tandem duplication and loss. In *Arabidopsis*, many megabase duplication blocks have been retained, while other very large regions have been lost. In any case, the more extreme cases of very high or very low apparent segmental duplication warrant further description below; the high-segmental duplications for the proteasome 20S subunit family and the low-segmental duplications for the NBS-LRR family. These have observed/expected segmental ratios of 0.41 and 0.05, respectively.

Low tandem, high segmental duplications

A large proportion of the families in or near the low-tandem duplication class also fall in the high-segmental duplication class (Table 2). These include proteins involved in a variety of roles: transcription factors (MYB), signalling (GTP binding proteins, calmodulin,

phosphoprotein phosphatase), various enzymatic functions (glycosyl transferase, plastocyanin), membrane transport (major intrinsic protein), and cellular housekeeping roles (Proteasome 20S subunits).

The proteasome 20S subunit family provides an interesting case study with which to consider possible constraints on duplication processes and gene copy numbers [22,23]. In eukaryotes, the proteasome recycles proteins by degradation of ubiquitin-tagged proteins [41,42]. It is a large protein complex, consisting of a 28-subunit catalytic cylindrical structure, called the 20S proteasome, and an ATP-dependent 19S regulatory particle consisting of an additional set of approximately 18 subunits [49]. The 20S proteasome is made up of four stacked rings. The two middle rings are each composed of seven 20S beta polypeptides, and these rings are sandwiched between two alpha rings, each composed of a ring of seven polypeptides, giving an 7α 7β 7β 7α structure [42]. In most eukaryotes described to date, each of the seven alpha and seven beta subunits is somewhat different from one another, requiring 14 types of proteasome subunits to make up the 20S proteasome [50].

In the *A. thaliana* 20S proteasome there are 23 genes encoding 20S proteasome subunits [40,41,51,52] – rather than 14. The phylogeny in Figure 5 suggests the origin of the additional subunits. There are two large clades of 20S proteasome sequences, each representing the alpha or beta subunits and each alpha or beta clade is composed of seven clades or lineages representing different alpha (or beta) sequences [40,42] (Figure 5). Interestingly, it appears that there are two, nearly complete sets of alpha and beta subunits because rather than the expected 14 sequences, we find 23 sequences comprised of nine pairs (18 total) plus five as singletons. We identified seven instances of segmental duplication on the tree, and the short branch lengths for two additional pairs suggest that these also may represent remnants of the same, recent, polyploidy event (Figure 5).

The scenario suggested by the combination of biological, phylogenetic, and genome contextual information is that following a round of genome doubling, roughly 20–40 Mya [7,10-12], most members of the duplicated proteasome subunits have been maintained but that five copies have been lost. There are no tandem duplications in this gene family, thus these two, nearly complete sets of 20S subunits were apparently generated by segmental duplication alone. Maintenance of seven alpha and beta lineages suggests maintenance of the stoichiometry of the 20S components, while tolerance of duplicated subunits might provide greater regulatory or catalytic flexibility [49]. Moore and Purugganan [53] describe precedence for positive selection driving the fixation and preservation of at least some duplicate genes in *Arabidopsis*. Whatever the cause of higher-than-expected retention of segmental duplicates in the Proteasome subunit family, this pattern contrasts with the following example in NBS-LRR resistance genes, which shows rapid turnover of gene family members and loss of major gene lineages in some plant families.

Moderate tandem, low segmental duplication

Three families in the moderate-tandem duplication class are also in the low-segmental duplication class: a protein phosphatase family, an acyltransferase family, and the NBS-LRR disease resistance family. The NBS-LRR disease resistance gene family contains 152 members in *A. thaliana* by our HMM-search criteria. Members of the family have been shown to confer resistance to a wide variety of pathogens [31,32,54-56] and are of tremendous economic importance so, we focus on these as an example here.

We found 54 tandem duplication events and six segmental duplication events on the tree. For one of the largest clades, the 60-member RPP5-containing clade in the TIR subfamily, we identified 25 tandem duplications within this single RPP5 clade, 24 of which occur after segmental duplications. The segmental duplications map to duplication blocks dated to the recent round of polyploidy and estimated to have occurred after the separation of Brassicaceae from other dicot families, roughly 20 – 40 Mya [10,57]. Given that no close homologs outside of Brassicaceae species [31] have been found to date for the RPP5 family, a likely scenario is that a RPP5 ancestral sequence underwent tandem and perhaps another transposing duplication, was subsequently amplified via polyploidy, and then experienced multiple rounds of tandem duplication.

In the NBS-LRR resistance gene family, Baumgarten et al. [43] find that segmental duplications largely explain the genome-wide distribution of NBS-LRR homologs. Here, we extend these findings to show that tandem duplications and losses play the dominant role in affecting *copy number*, as in the expansion of the RPP5 homologs. In fact, net gene loss following polyploidy, coupled with dynamic expansion and loss, could lead to quite variable numbers of sequences in any given clade. Across the very large NBS-LRR gene family, we observe several clades for which we could demonstrate few or no tandem duplications [31]. Perhaps the most dramatic example of sequence loss is the complete lack of TIR NBS-LRR sequences in the Poaceae, despite abundant sequence for both the grasses and the NBS-LRR gene family [32,58]. The absence of TIR sequences in the grasses has to be inferred as a loss of this sequence type from the grasses because TIR homologs have been found in pine [31] and moss [59]. Just as

certain clades have expanded rapidly, such as the RPP5 clade, other lineages such as the entire TIR subfamily in grasses appear to have been lost.

High tandem, low segmental duplication

Among the 11 high- or very-high tandem duplication families, five are in the low-segmental duplication class, and four other families demonstrate a lower than median level of segmental duplications. The high tandem, low to moderate segmental duplication families fall into several broad functional categories. Two families in which the level of tandem duplication is more than two standard deviations above the genome median are the germin and major latex protein (MLP) families. Both families are involved in pathogen defense as well as other functions. Germins have been found to play a variety of roles, including cell wall formation during germination, stress-related signaling, production of active oxygen species, and degradation of oxalate presented as a fungal toxin [60,61]. Other families with members shown to have roles in defense against pathogens are the subtilisin-like serine proteases and the pathogen-related PR1 family [36]. In contrast to the apparent advantage of diverse defense sequences, why might the chlorophyll a-b binding (CAB) family have retained high numbers of tandem duplications? *A priori*, the CAB family might be expected to evolve much more conservatively than the very highly duplicated defense-related families because these proteins form the large multi-protein complexes of photosystems I and II [62]. We will describe the unusual CAB family first and then the MLP family.

High tandem, low segmental duplication, example 1: CAB

The CAB proteins are components of the complex multi-subunit photosystems I and II (PS I and PS II) [63,64]. Both photosystems consist of a chlorophyll-binding core, and a peripheral antenna or light harvesting complex (LHC). In addition to the light-harvesting function, the antenna is able to dissipate excess energy through a process called feedback de-excitation [65]. There are at least 10 distinct types of proteins in the LHCs for PS I and PS II [62,64], encoded by the nuclear *lhc* genes. The basic structure of photosystems have been conserved since the evolution of early land plants [62,64]. Four *lhc* genes associated with PS I are denoted *lhca1-4*. Genes associated with PS II are denoted *lhcb1-6*, and *lhcb1* and *lhcb2* can also associate with PS I [62-64,66]. Although the basic structure of photosystems have been conserved since the evolution of early land plants [62,64], our results show surprisingly dynamic copy numbers, especially for *lhcb1*, *lhcb2*, and *lhcb3*.

Figure 7 shows a phylogeny for the *A. thaliana* CAB family (left side) and for comparison, a phylogeny that also includes consensus ESTs from tomato and *M. truncatula* (right side). In the *A. thaliana* gene phylogeny, four tandem duplications and the one segmental duplication were detected in the clade that contains *lhcb1*, *lhcb2*, and *lhcb3*. In the three-species gene phylogeny, there are multiple *lhcb1*, *lhcb2*, and *lhcb3* homologs in both tomato and *Medicago*. These paralogs show recent independent expansions in each species lineage to generate sets of paralogs more similar to one another than to homologs from another species. In the *A. thaliana lhcb1* clade, for example, there are five paralogs that have arisen through three tandem duplications following divergence of Brassicaceae, and one segmental duplication prior to divergence of Brassicaceae. In the corresponding clade of the three species gene tree, there are at least 11 tomato sequences, with six arising before the tomato/*A. thaliana* split, and at least six *Medicago* sequences, with one arising before the *Medicago*/*A. thaliana* split. Similar phylogenetic patterns are apparent in EST data for corn, soybean, and potato (not shown). In contrast, none of the *lhca* genes nor the *lhcb4-6* genes show recent amplification of gene copy number.

There appear to be different evolutionary modes at work in different parts of the CAB family. The phylogeny suggests high rates of turnover in the *lhcb1-3* genes and low turnover with possible loss of segmental duplicates in the other *lhc* genes. Structural and functional studies show that in PS II, a dimer of the core complexes is flanked by two proteins each encoded by *lhcb4*, *lhcb5*, and *lhcb6*. These, in turn, are flanked by a total of four trimers of *lhcb1*, *lhcb2*, and *lhcb3* [64,67,68]. *Lhcb4* is essential to formation of a functioning PS II, but functional photosystems are still formed if expression of *lhcb1* and *lhcb2* is inhibited [67,68] and transcription of *lhcb5* is strongly upregulated. However, *Lhcb1* and *Lhcb2* proteins do play important roles in low light conditions [64] and in establishing the proper formation of grana stacks, and *Lhcb5* can not entirely compensate for these functions [64]. Thus, the evolutionary flexibility of *lhcb1-lhcb3* genes may provide a mechanism to tune the light harvesting complex for different light conditions [65], while in contrast, the genomically dispersed and evolutionarily more stable, *lhcb4*, *lhcb5*, and *lhcb6* genes maintain the photosynthetic core of PS II.

The Major Latex Protein (MLP) family encodes proteins that were originally isolated from the latex of opium poppy [69,70] but also found in a wide range of plants and tissues [71]. Functions of MLP are not known, but they do show significant similarity to a pathogenesis-related proteins (IPR or PR10 proteins [37]) which show increased expression with pathogen or stress challenge [37,72-74]. Members of the two gene families (MLP and IPR-PR10) show only about 25% identity to each other, but sequence and structural analyses indicate that they are

similar enough to be considered to be part of a single superfamily [37]. Interestingly, there are no *A. thaliana* homologs that group with the IPR-PR10 subfamily [37], but we located 11 tandem and three segmental duplications for the MLP family (Figure 8), resulting in a tandem duplication observed/expected ratio of 54.5 and segmental duplication observed/expected ratio of 0.16.

Evolutionary distances among sequences resulting from the predicted segmental duplications are greater in the MLP than among segmentally duplicated sequences in the proteasome family. In the MLP family, pairwise distances among segmentally duplicated sequences range from about 15 to 60 PAM units [75], but in the proteasome 20S family, range from 0 to about 4 PAM. Nevertheless, the MLP duplications do appear to come from the same polyploidy event as the proteasome duplications (Blanc et al. [6,10]). Clearly, the MLP members have been evolving much more rapidly following polyploidy than have the proteasome 20S subunits or most members of the CAB family. As above, we used *Medicago* and tomato sequences to mark divergence times and the results support duplication due to a recent polyploidy event in *Arabidopsis*.

Patterns of gene duplication

We observed a moderate negative correlation between levels of predicted tandem and segmental duplications in gene families. If either sequence variation or gene copy number must be maintained within some bounds, one possible source of selection against tandem duplication is that unequal cross-over and gene loss will generate variation and high turn-over of gene copies [15,76]. Conversely, segmental duplicates may be more often retained due to subfunctionalization, without increasing the likelihood of gene rearrangment [77,78]. In families that demonstrate moderately high levels of segmental and tandem duplication, gene family members have been retained within segmental duplication blocks, while gene copy number in some clades has been expanded by tandem duplication. An example is found in flavin-containing monooxygenase family, which has six segmental duplications at various nodes in the tree, and seven tandem duplications accounting for members of one clade. Another example includes the chlorophyll a/b binding protein family. On the other end of the spectrum, we found few families in our study set with both low segmental and low tandem duplication. No families fall below one standard deviation below the median in both the segmental and tandem categories, but several are close: the mitochondrial carrier proteins and the heat shock transcription factor families have 9% and 13% of the expected segmental duplications, respectively. Conceivably, protein copy stoichiometry is critical in some families representing multi-subunit protein complexes [77,78].

Conclusions

The relative contributions of tandem and segmental duplication to the generation and maintenance of 50 large *A. thaliana* gene families was characterized using ratios of observed/expected tandem duplication and observed/expected segmental duplication. Counts of tandem and segmental duplications were negatively correlated; no families exhibited both high levels of tandem and segmental duplication. Although the distribution of gene family sizes across the genome can be accounted for by a stochastic model, by comparing the relative levels of tandem and segmental duplication in large gene families, we can speculate that gene function might feedback on copy number and genome organization, and thus result in the widely varying patterns of observed tandem and segmental duplication.

Methods

Gene family selection, alignment, and phylogeny construction

Initial candidate gene families were identified using 2001 *A. thaliana* PIR superfamilies [79], based on the 2001 MIPs *A. thaliana* predicted proteins [80]. Though helpful, these were found to be somewhat inconsistent, splitting some families unnecessarily and producing some with overlapping membership. Families initially considered contained at least 20 genes and, where Pfam [18] domains were identifiable, had at least one Pfam domain (E-value < 0.01) in common across all gene family members and the organization of domains was consistent throughout the family. Initial selection of gene family members was conducted using the TIGR 2001 predicted *A. thaliana* proteins. All alignments, phylogenies, and analyses were recalculated based on newer gene predictions (the 2003 TIGR *A. thaliana* release 4.0).

Predicted proteins for all gene families were aligned using T-Coffee [19] and a maximum of 30 randomly selected proteins sequences from each family. These initial alignments were used to create HMMs, which were in turn used to re-align the full protein sets. HMM parameters in hmmer [29] were: "hmmbuild --archpri .7 --fast -- gapmax .3". The HMMs were calibrated using hmmcalibrate, and then were used to search the full set of predicted *A. thaliana* protein sequences, using the hmmsearch program in hmmer [29]. Sequences scoring at least 10^{-10} were generally retained as gene family members and genes scoring worse than this threshold were excluded, although scores were evaluated in the context of all scores in the putative gene families. Families with gradually declining scores in the range of 0.1 – 10^{-15} were generally excluded from the study because the difficulty in unambiguously assigning family membership.

Alignments were prepared for use in phylogenetic reconstructions as follows. To remove highly variable or indel regions, sequence positions falling outside of HMM match states were removed. Genes matching fewer than 75% of the remaining positions were removed entirely. Alignments were also manually inspected, and other particularly poorly aligning regions were removed. Both full-length and trimmed sequences for all gene families are available at [24].

Parsimony and bootstrapped neighbor joining trees were calculated for each gene family. Parsimony trees were calculated using protpars in the Phylip suite [81]. The input sequence order was jumbled five times, and a topology calculated based on each data order. One most-parsimonious tree was chosen at random to serve as the basis for branch length calculations. Maximum likelihood branch lengths were calculated on the parsimony topologies using TreePuzzle [82]. The model of substitution was of Adachi and Hasegawa [83], amino acid frequencies were calculated from the input trees, and rate heterogeneity was allowed with four Gamma rate categories. Neighbor joining trees were calculated using Clustalw, without the Kimura distance correction, and with 1000 bootstrap replicates. All trees are available at [24].

Prediction of **A. thaliana** *duplication blocks and gene family segmental and tandem duplications*

Internal genomic duplications were predicted using DiagHunter [20,21]. All duplication block predictions (genes, genomic coordinates, and dot plot images of genomic similarities and predicted duplications) are available at [24]. Predictions of segmental or tandem duplications in gene families were made using the OrthoParaMap suite [22,23].

The approach for identifying segmental duplicates consists of identifying pairs of sufficiently similar genes in a phylogeny that fall sufficiently close to their respective corresponding regions in a synteny block. Pairs of gene family members falling within a synteny block are annotated as such in the phylogeny, using the extended New Hampshire (NHX) format [84]. For all but the 13 largest gene families, the threshold for "sufficiently similar" was set at 10^{-25}, and the threshold for "sufficiently close" was set at 50 kb. Because the number of potential false positive hits to a synteny block rises approximately proportionally to the square of the number of genes in a gene family, more stringent thresholds ("similar" = 10^{-30} and "close" = 30 kb) were used for the following families (see Table 1 for full names): CytP450, MATE, MFS, Myb, NBS-LRR, WRKY, GSDLLipase, GTPBP, MajIntrins, Prot, Oxidored, Polygalns, SCDehydRed, UDPGlycTnsf.

Nodes giving rise to tandem gene duplications were inferred using the ParaMap program in the OrthoParaMap suite [22,23]. This recursively walks through the tree, identifying internal nodes that give rise to genes or other nodes that are physically near one another (<50 kb) on the chromosome.

Calculation of gene family size distributions

All predicted proteins were used in a BLASTP [30] search against one another, using three different E-value thresholds (10^{-10}, 10^{-20}, and 10^{-30}). BLAST results were parsed using a BioPerl [85] – based script. Approximate gene families were constructed using a single linkage clustering approach implemented in Perl. The 2003 TIGR *A. thaliana* 4.0 assembly and protein predictions were used for these procedures.

Calculation of gene duplication densities by distance

Predicted protein sequences in the 2003 TIGR *A. thaliana* 4.0 assembly were assigned genomic positions based on the nucleotide position halfway between the predicted 5' and 3' positions. Protein sequences from each chromosome were used in BLASTP [30]searches against all other sequences in that chromosome, to give lists of BLAST hits and query/target midpoint positions for each chromosome. Hits to self were excluded.

Calculation of expected tandem and segmental duplications

Tandem duplications expected to occur by chance in a gene family of a given size were simulated under the assumption of a 100,000 kb genome (approximately the size of the *A. thaliana* euchromatic genome). Gene families of sizes ranging from 20 to 230 genes were simulated, using size classes in increments of 10. Approximate distributions were calculated using 1000 simulation runs for each gene family size class.

Normalizing constants for segmental duplications in gene families of given sizes were calculated under the assumption that the maximum proportion of segmental duplications retained in an average gene family should be the same as the percentage of the genome that exists "In duplicate" (in synteny blocks). Arithmetic for the proportion of expected segmental duplicates (in the absence of local gene losses or duplications following polyploidy) is shown in the Results section.

Additional data

Alignments, gene phylogenies, annotations, analyses of genomic position, relationships to internal genomic duplications, and comparisons with homologous ESTs from various species are available at http://www.tc.umn.edu/~cann0010/genefamilyevolution/

Authors' contributions

SBC developed software used in the analysis, carried out the analyses on all gene families, and drafted the manuscript. AM conducted background literature reviews of families at [24]. SBC and AB developed the method for quantifying tandem duplications in *Arabidopsis*. NDY and GM advised throughout the project, and helped in manuscript preparation and revision. All authors read and approved the final manuscript.

Acknowledgements

Thanks to Bridgette Barry for discussions about the CAB family, to Martina Stromvik for discussions about the MLP family, to Jeff Doyle for suggestions on the manuscript, and to the Minnesota Supercomputing Institute (MSI) for access to computing resources. This work was supported in part by NSF award DBI-9975866 to GM and NSF award DBI-0110206 to NDY and by a USDA National Needs fellowship and a University of Minnesota Plant Molecular Genetics Institute fellowship to SBC.

References

1. AGI: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature* 2000, **408:**796-815.
2. Shiu SH, Bleecker AB: **Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases.** *Proc Natl Acad Sci U S A* 2001, **98:**10763-10768.
3. Tichtinsky G, Vanoosthuyse V, Cock JM, Gaude T: **Making inroads into plant receptor kinase signalling pathways.** *Trends Plant Sci* in press.
4. Feldman KA: **Cytochrome P450s as genes for crop improvement.** *Plant Biotech* 2001, **4:**162-167.
5. Nelson DR: **Arabidopsis P450 statistics.** [http://drnelson.utmem.edu/Arabfam.html].
6. Blanc G, Barakat A, Guyot R, Cooke R, Delseny M: **Extensive duplication and reshuffling in the Arabidopsis genome.** *Plant Cell* 2000, **12:**1093-1101.
7. Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y: **The hidden duplication past of Arabidopsis thaliana.** *Proc Natl Acad Sci U S A* 2002, **99:**13627-13632.
8. Vandepoele K, Saeys Y, Simillion C, Raes J, Van De Peer Y: **The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice.** *Genome Res* 2002, **12:**1792-1801.
9. Vision TJ, Brown DG, Tanksley SD: **The origins of genomic duplications in Arabidopsis.** *Science* 2000, **290:**2114-2117.
10. Blanc G, Hokamp K, Wolfe KH: **A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome.** *Genome Res* 2003, **13:**137-144.
11. Ermolaeva MD, Wu MM, Eisen JA, Salzberg SL: **The age of the Arabidopsis thaliana genome duplication.** *Plant Mol Biol* 2003, **51:**859-866.
12. Bowers JE, Chapman BA, Rong J, Paterson AH: **Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events.** *Nature* 2003, **422:**433-438.
13. Zhang L, Vision TJ, Gaut BS: **Patterns of nucleotide substitution among simultaneously duplicated gene pairs in Arabidopsis thaliana.** *Mol Biol Evol* 2002, **19:**1464-1473.
14. Ziolkowski PA, Blanc G, Sadowski J: **Structural divergence of chromosomal segments that arose from successive duplication events in the Arabidopsis genome.** *Nucleic Acids Res* 2003, **31:**1339-1350.
15. Achaz G, Coissac E, Viari A, Netter P: **Analysis of intrachromosomal repeats in yeast Saccharomyces cerevisiae: a possible model for their origin.** *Mol Biol Evol* 2000, **17:**1268-1275.
16. Hughes AL, Friedman R, Ekollu V, Rose JR: **Non-random association of transposable elements with duplicated genomic blocks in Arabidopsis thaliana.** *Mol Phylogenet Evol* 2003, **29:**410-416.
17. MIPS: **PIR superfamilies in Arabidopsis.** [http://mips.gsf.de/proj/thal/db/tables/tables_func_frame.html].
18. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30:**276-280.
19. Notredame C, Holm L, Higgins DG: **T-COFFEE: an objective function for multiple sequence alignments.** *Bioinformatics* 1998, **14:**407-422.
20. Cannon SB: **DiagHunter web site.** 2003 [http://www.tc.umn.edu/~cann0010/Software.html].
21. Cannon SB, Kozik A, Chan B, Michelmore R, Young ND: **DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization.** *Genome Biol* 2003, **4:**R68.
22. Cannon SB: **OrthoParaMap web site.** 2003 [http://www.tc.umn.edu/~cann0010/Software.html].
23. Cannon SB, Young ND: **OrthoParaMap: Distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies.** *BMC Bioinformatics* 2003, **4:**35.
24. Cannon SB: **genefamilyevolution web site.** 2003 [http://www.tc.umn.edu/~cann0010/genefamilyevolution/].
25. Gogarten JP, Olendzenski L: **Orthologs, paralogs and genome comparisons.** *Curr Opin Genet Dev* 1999, **9:**630-636.
26. Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L: **Gene families: the taxonomy of protein paralogs and chimeras.** *Science* 1997, **278:**609-614.
27. Ohta T: **Multigene families and the evolution of complexity.** *J Mol Evol* 1991, **33:**34-41.
28. Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, Srinivasarao GY, Xiao C, Yeh LS, Ledley RS, Janda JF, Pfeiffer F, Mewes HW, Tsugita A, Wu C: **The protein information resource (PIR).** *Nucleic Acids Res* 2000, **28:**41-44.
29. Eddy SR: **HMMER: Profile hidden Markov models for biological sequence analysis: The HMMER User's Guide.** 2001 [http://hmmer.wustl.edu/].
30. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
31. Cannon SB, Zhu H, Baumgarten AM, Spangler R, May G, Cook DR, Young ND: **Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies.** *J Mol Evol* 2002, **54:**548-562.
32. Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW, Young ND: **Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily.** *Plant J* 1999, **20:**317-332.
33. Batalia MA, Monzingo AF, Ernst S, Roberts W, Robertus JD: **The crystal structure of the antifungal protein zeamatin, a member of the thaumatin-like, PR-5 protein family.** *Nat Struct Biol* 1996, **3:**19-23.
34. Carter C, Graham RA, Thornburg RW: **Arabidopsis thaliana contains a large family of germin-like proteins: characterization of cDNA and genomic sequences encoding 12 unique family members.** *Plant Mol Biol* 1998, **38:**929-943.
35. Membre N, Bernier F, Staiger D, Berna A: **Arabidopsis thaliana germin-like proteins: common and specific features point to a variety of functions.** *Planta* 2000, **211:**345-354.
36. Santamaria M, Thomson CJ, Read ND, Loake GJ: **The promoter of a basic PR1-like gene, AtPRB1, from Arabidopsis establishes an organ-specific expression pattern and responsiveness to ethylene and methyl jasmonate.** *Plant Mol Biol* 2001, **47:**641-652.
37. Osmark P, Boyle B, Brisson N: **Sequential and structural homology between intracellular pathogenesis-related proteins and a group of latex proteins.** *Plant Mol Biol* 1998, **38:**1243-1246.
38. Kuan J, Saier M. H., Jr.: **The mitochondrial carrier family of transport proteins: structural, functional, and evolutionary relationships.** *Crit Rev Biochem Mol Biol* 1993, **28:**209-233.
39. Borecky J, Maia IG, Arruda P: **Mitochondrial uncoupling proteins in mammals and plants.** *Biosci Rep* 2001, **21:**201-212.
40. Parmentier Y, Bouchez D, Fleck J, Genschik P: **The 20S proteasome gene family in Arabidopsis thaliana.** *FEBS Lett* 1997, **416:**281-285.
41. Vierstra RD: **The ubiquitin/26S proteasome pathway, the complex last chapter in the life of many plant proteins.** *Trends Plant Sci* 2003, **8:**135-142.

42. Fu H, Doelling JH, Arendt CS, Hochstrasser M, Vierstra RD: **Molecular organization of the 20S proteasome gene family from Arabidopsis thaliana.** *Genetics* 1998, **149:**677-692.
43. Baumgarten A, Cannon S, Spangler R, May G: **Genome-Level Evolution of Resistance Genes in Arabidopsis thaliana.** *Genetics* 2003, **165:**309-319.
44. Huynen MA, van Nimwegen E: **The frequency distribution of gene family sizes in complete genomes.** *Mol Biol Evol* 1998, **15:**583-589.
45. Sornette D, Cont R: **Convergent multiplicative processes repelled from zero: power laws and truncated power laws.** *J Physique I* 1997, **7:**431-444.
46. Kesten H: **Random difference equations, and renewal theory for products of random matrices.** *Acta Math* 1973, **131:**207-248.
47. Kim BH, Schoffl F: **Interaction between Arabidopsis heat shock transcription factor 1 and 70 kDa heat shock proteins.** *J Exp Bot* 2002, **53:**371-375.
48. Aquila H, Link TA, Klingenberg M: **Solute carriers involved in energy transfer of mitochondria form a homologous protein family.** *FEBS Lett* 1987, **212:**1-9.
49. Fu H, Doelling JH, Rubin DM, Vierstra RD: **Structural and functional analysis of the six regulatory particle triple-A ATPase subunits from the Arabidopsis 26S proteasome.** *Plant J* 1999, **18:**529-539.
50. Hochstrasser M, Johnson PR, Arendt CS, Amerik AYu, Swaminathan S, Swanson R, Li SJ, Laney J, Pals-Rylaarsdam R, Nowak J, Connerly PL: **The Saccharomyces cerevisiae ubiquitin-proteasome system.** *Philos Trans R Soc Lond B Biol Sci* 1999, **354:**1513-1522.
51. von Arnim AG: **A hitchhiker's guide to the proteasome.** *Sci STKE* 2001, **2001:**PE2.
52. Gray WM, Estelle I: **Function of the ubiquitin-proteasome pathway in auxin response.** *Trends Biochem Sci* 2000, **25:**133-138.
53. Moore RC, Purugganan MD: **The early stages of duplicate gene evolution.** *Proc Natl Acad Sci U S A* 2003, **100:**15682-15687.
54. Michelmore R, Meyers BC: **Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process.** *Genome Res* 1998, **8:**1113-1130.
55. Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW: **Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis.** *Plant Cell* 2003, **15:**809-834.
56. Jones JD: **Putting knowledge of plant disease resistance genes to work.** *Curr Opin Plant Biol* 2001, **4:**281-287.
57. Blanc G, Wolfe K: **Paralagons in Arabidopsis thaliana.** 2002 [http://wolfe.gen.tcd.ie/athal/].
58. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Li J, Liu Z, Qi Q, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Zhao W, Li P, Chen W, Zhang Y, Hu J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Tao M, Zhu L, Yuan L, Yang H: **A draft sequence of the rice genome (Oryza sativa L. ssp. indica).** *Science* 2002, **296:**79-92.
59. Akita M, Valkonen JP: **A novel gene family in moss (Physcomitrella patens) shows sequence homology and a phylogenetic relationship with the TIR-NBS class of plant disease resistance genes.** *J Mol Evol* 2002, **55:**595-605.
60. Patnaik D, Khurana P: **Germins and germin like proteins: an overview.** *Indian J Exp Biol* 2001, **39:**191-200.
61. Schweizer P, Christoffel A, Dudler R: **Transient expression of members of the germin-like gene family in epidermal cells of wheat confers disease resistance.** *Plant J* 1999, **20:**541-552.
62. Blankenship RE: **Origin and early evolution of photosynthesis.** *Photosynth Res* 1992, **33:**91-111.
63. Oullette AJA, Barry BA: **Tandem mass spectrometric identification of spinach Photosystem II light-harvesting components.** *Photosynth Res* 2002, **72:**159-173.
64. Andersson J: **Dissecting the photosystem II light-harvesting antenna.** *[dissertation]* 2003.
65. Bailey S, Walters RG, Jansson S, Horton P: **Acclimation of Arabidopsis thaliana to the light environment: the existence of separate low light and high light responses.** *Planta* 2001, **213:**794-801.
66. Yakushevsha AE, Keegstra W, Boekema EJ, Dekker JP, Andersson J, Jansson S, Ruban AV, Horton P: **The structure of photosystem II in Arabidopsis: localization of the CP26 and CP29 antenna complexes.** *Biochemistry* 2003, **42:**.
67. Ruban AV, Wentworth M, Yakushevsha AE, Andersson J, Lee MM, Keegstra W, Dekker JP, Boekema EJ, Jansson S, Horton P: **Plants lacking the main light harvesting complex retain photosystem II macro-organization.** *Nature* 2003, **in press:**.
68. Andersson J, Wentworth M, Walters RG, Howard CA, Ruban AV, Horton P, Jansson S: **Absense of the main light-harvesting complex of photosystem II affects photosynthetic function.** *Plant J* 2003.
69. Nessler CL, Burnett RJ: **Organization of the major latex protein gene family in opium poppy.** *Plant Mol Biol* 1992, **20:**749-752.
70. Nessler CL: **Sequence analysis of two new members of the major latex protein gene family supports the triploid-hybrid origin of the opium poppy.** *Gene* 1994, **139:**207-209.
71. Stromvik MV, Sundararaman VP, Vodkin LO: **A novel promoter from soybean that is active in a complex developmental pattern with and without its proximal 650 base pairs.** *Plant Mol Biol* 1999, **41:**217-231.
72. Bufe A, Spangfort MD, Kahlert H, Schlaak M, Becker WM: **The major birch pollen allergen, Bet v 1, shows ribonuclease activity.** *Planta* 1996, **199:**413-415.
73. Flores T, Alape-Giron A, Flores-Diaz M, Flores HE: **Ocatin. A novel tuber storage protein from the andean tuber crop oca with antibacterial and antifungal activities.** *Plant Physiol* 2002, **128:**1291-1302.
74. Moiseyev GP, Fedoreyeva LI, Zhuravlev YN, Yasnetskaya E, Jekel PA, Beintema JJ: **Primary structures of two ribonucleases from ginseng calluses. New members of the PR-10 family of intracellular pathogenesis-related plant proteins.** *FEBS Lett* 1997, **407:**207-210.
75. Dayhoff MO: **Atlas of Protein Sequences and Structure. Volume 5, Supplement 3, pp. 353-358.** Washington, DC, USA, National Biomedical Research Foundation; 1979.
76. Boore JL: **The duplication/random loss model for gene rearrangement exemplified by mitochondrial genomes of deuterostome animals.** *Comparative Genomics* Edited by: Sankoff D and Nadeau J. Dordrecht, NL, Kluwer Academic Press; 2000:133-147.
77. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290:**1151-1155.
78. Lynch M, Conery JS: **The evolutionary demography of duplicate genes.** *J Struct Funct Genomics* 2003, **3:**35-44.
79. Barker WC, Garavelli JS, Hou Z, Huang H, Ledley RS, McGarvey PB, Mewes HW, Orcutt BC, Pfeiffer F, Tsugita A, Vinayaka CR, Xiao C, Yeh LS, Wu C: **Protein Information Resource: a community resource for expert annotation of protein data.** *Nucleic Acids Res* 2001, **29:**29-32.
80. Schoof H, Zaccaria P, Gundlach H, Lemcke K, Rudd S, Kolesov G, Arnold R, Mewes HW, Mayer KF: **MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome.** *Nucleic Acids Res* 2002, **30:**91-93.
81. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genetics, University of Washington, Seattle.** 2000.
82. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18:**502-504.
83. Adachi J, Hasegawa M: **Model of amino acid substitution in proteins encoded by mitochondrial DNA.** *J Mol Evol* 1996, **42:**459-468.
84. Zmasek CM, Eddy SR: **ATV: display and manipulation of annotated phylogenetic trees.** *Bioinformatics* 2001, **17:**383-384.
85. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12:**1611-1618.