New Phytologist

Research

The evolution and function of protein tandem repeats in plants

Elke Schaper[1,2,3] and Maria Anisimova[4]

[1]Department of Computer Science, ETH Zürich, Zürich 8092, Switzerland; [2]Institute of Integrative Biology, ETH Zürich, Zürich 8092, Switzerland; [3]Vital-IT Competency Center, Swiss Institute for Bioinformatics (SIB), Lausanne 1015, Switzerland; [4]Institute of Applied Simulation (IAS), School of Life Sciences and Facility Management, Zürich University of Applied Sciences (ZHAW), Wädenswil 8820, Switzerland

Authors for correspondence:
Elke Schaper
Tel: +41 44 63 28 26 0
Email: elke.schaper@isb-sib.ch

Maria Anisimova
Tel: +41 58 934 58 82
Email: maria.anisimova@zhaw.ch

Received: *22 August 2014*
Accepted: *18 October 2014*

New Phytologist (2015) **206:** 397–410
doi: 10.1111/nph.13184

Key words: conservation, leucine-rich repeat (LRR), minisatellites, phylogenetic analysis, plant genomics, protein evolution, *R* genes, tandem repeats (TRs).

Summary

- Sequence tandem repeats (TRs) are abundant in proteomes across all domains of life. For plants, little is known about their distribution or contribution to protein function. We exhaustively annotated TRs and studied the evolution of TR unit variations for all Ensembl plants.
- Using phylogenetic patterns of TR units, we detected conserved TRs with unit number and order preserved during evolution, and those TRs that have diverged via recent TR unit gains/losses. We correlated the mode of evolution of TRs to protein function.
- TR number was strongly correlated with proteome size, with about one-half of all TRs recognized as common protein domains. The majority of TRs have been highly conserved over long evolutionary distances, some since the separation of red algae and green plants *c*. 1.6 billion yr ago. Conversely, recurrent recent TR unit mutations were rare.
- Our results suggest that the first TRs by far predate the first plants, and that TR appearance is an ongoing process with similar rates across the plant kingdom. Interestingly, the few detected highly mutable TRs might provide a source of variation for rapid adaptation. In particular, such TRs are enriched in leucine-rich repeats (LRRs) commonly found in *R* genes, where TR unit gain/loss may facilitate resistance to emerging pathogens.

Introduction

Tandem repeats (TRs) are consecutive perfect or imperfect repetitions of a sequence motif (TR unit), stemming from duplications and losses of an ancestral unit. TRs in proteins are transcribed and translated from TRs in coding nucleic sequences, which may, however, be interspersed by introns. TRs represent an abundant feature of proteomes across all domains of life (Marcotte *et al.*, 1999; Hanada *et al.*, 2008). They differ strongly in their unit length (l), varying from repetitions of single amino acids (homorepeats) to whole-domain repetitions.

In plants, several important protein families feature long TRs. These include the pentatricopeptide repeats (PPRs; $l = 35$ amino acids (aa)) which are found in several hundreds of proteins in most angiosperms. PPRs play a major role, amongst others, in different RNA processing activities (Marcotte *et al.*, 1999; Fujii & Small, 2011). The leucine-rich repeats (LRRs; $l = c.$ 20–30 aa) constitute a similarly prolific motif in plant proteomes. LRR-containing proteins comprise the majority of disease resistance proteins in plants (McHale *et al.*, 2006; Fujii & Small, 2011), with the LRR domain being thought to promote protein–protein interactions (Kobe & Kajava, 2001; McHale *et al.*, 2006) and act in molecule recognition. Most of the other common plant TRs are thought to contribute to the promotion of protein–protein interactions: these include the tetratricopeptide repeat (TPR; $l = c.$ 34 aa), the Kelch repeat ($l = c.$ 47 aa), the WD40 repeat ($l = c.$ 39 aa) and Ankyrin repeats (ANKs; $l = c.$ 33 aa) (Groves & Barford, 1999; Adams *et al.*, 2000; Kobe & Kajava, 2001; Stirnimann *et al.*, 2010; Xu & Min, 2011).

Other TRs exist in plant proteomes at lower frequencies, and rare TR domains are not likely to be annotated in sequence motif databases, such as PFAM (Groves & Barford, 1999; Adams *et al.*, 2000; Stirnimann *et al.*, 2010; Punta *et al.*, 2011; Xu & Min, 2011). For example, in humans, > 12% of all validated protein TRs with $l \geq 15$ aa were not found in PFAM A, but nevertheless were detected by *de novo* algorithms (Punta *et al.*, 2011; Schaper *et al.*, 2014). However, for shorter TRs, the ratio of nonannotated TRs is expected to be much higher. To obtain a preferably exhaustive dataset of TRs in plants, TR annotations need to be derived from both sequence motif databases and specifically devised TR *de novo* detection algorithms. Here, we infer and analyze TR annotations from both sources, focusing on plant proteomes.

Several mutational mechanisms act on TR regions. Substitutions and indels may alter the TR units, so that the original TR units may ultimately diverge beyond recognition. The shorter the TR unit, the fewer substitutions/indels are necessary to be unrecognizable, so that, in general, longer annotated TRs span wider

divergence ranges. Second, as a result of mechanisms such as DNA slippage, TRs may be subject to TR unit mismatch mutations, leading to an expansion or a contraction of the TR region through TR unit gains/losses. This mechanism is comparable with the amplification of microsatellites on short length scales (Shi *et al.*, 2013; Schaper *et al.*, 2014) and, on larger length scales, to (tandem) gene duplications that create gene clusters. Both types of process have high rates in plants (see, for example, Symonds & Lloyd, 2003; Marriage *et al.*, 2009; Shi *et al.*, 2013), with tandem gene duplications thought to be a considerable source of plant evolutionary innovation (e.g. Symonds & Lloyd, 2003; Cannon *et al.*, 2004; Marriage *et al.*, 2009 for microsatellites), playing a role in adaptation to rapidly changing environments (Cannon *et al.*, 2004; Hanada *et al.*, 2008). Interestingly, many of the studies on gene duplications involve proteins containing TRs (Marcotte *et al.*, 1999; Leister, 2004; McHale *et al.*, 2006). Finally, TR unit conversion (analogous to gene conversion) could be another mutational mechanism contributing to the evolution of TR regions.

In contrast with the available results on plant microsatellite evolution and whole-gene duplications in plants, our current knowledge of how protein TRs evolve in terms of TR unit gain/loss is limited in the plant kingdom. Microsatellites are known to evolve rapidly, but it is not known whether protein TRs may be subject to similar rates of evolution. It has been proposed that fast population-scale unit gain/loss rates in TRs generate genetic diversity, allowing for adaptation, for example, in an evolutionary arms race with a pathogen (Marcotte *et al.*, 1999; Levdansky *et al.*, 2007; Richard *et al.*, 2008; Chevanne *et al.*, 2010; Fujii & Small, 2011; Riegler *et al.*, 2012). However, different mispair mechanisms with highly different mutation rates may dominate the TR unit gains/losses for different TR unit length scales, and selective pressure on protein sequences, in particular, may act to support sequence conservation, rather than favoring expansion/contraction of TR regions (McHale *et al.*, 2006; Schaper *et al.*, 2014). A prolonged conservation of a TR might indicate that the TR region size and sequence are important to maintain protein structure/function, whereby TR unit gains/losses lead to decreased protein fitness and therefore are selected against. For example, structurally, many of the conserved human TRs act as scaffolds to support protein–protein interaction. As the same holds for the largest groups of protein TRs in plants (LRR, TPR, Kelch, WD40, ANK), a high level of TR unit conservation may also be expected for many plant TRs.

Here, we present a proteome-wide evolutionary analysis of plant TRs, providing insight into the functional relevance of this large group of protein sequence motifs. Recently, we proposed a phylogenetic method to dissect TR unit gains/losses in samples from several species: reconstructed TR unit phylogenies, including all TR units from two orthologous proteins, provide sufficient signal to trace the evolutionary history of the TR units since the speciation event (Kobe & Kajava, 2001; Schaper *et al.*, 2014). When no TR unit gains/losses have occurred since the speciation, the order and number of TR units is preserved in the respective TR unit phylogeny, and any *i*th TR unit from the first species clusters with (i.e. is most similar to) the *i*th TR unit of the second species. We call such a pair of orthologous TRs perfectly conserved (Schaper *et al.*, 2014). For example, in Fig. 1(d), we show the phylogeny of a TR region with eight units in Arabidopsis and in its orthologous TR in the red alga *Cyanidioschyzon merolae* , which has been conserved in both lineages at least since their speciation. By contrast, with recurrent TR unit gains/losses, the units in the TR region of the same species will homogenize, so that the TR units from different species form monophyletic clusters in the TR unit phylogeny. We refer to such pairs of TRs as completely separated (see, for example, Fig. 2).

Therefore, the reconstructed TR unit phylogeny allows us to state with certainty whether a pair of TRs has been conserved at least since the speciation of both species or, alternatively, whether a TR region has been shaped by unit gains/losses in at least one of the two lineages, so that TR units have separated between the lineages. Combining the results from multiple pairwise comparisons allows us to deduce whether and when TR unit gains/losses occurred with respect to speciation events.

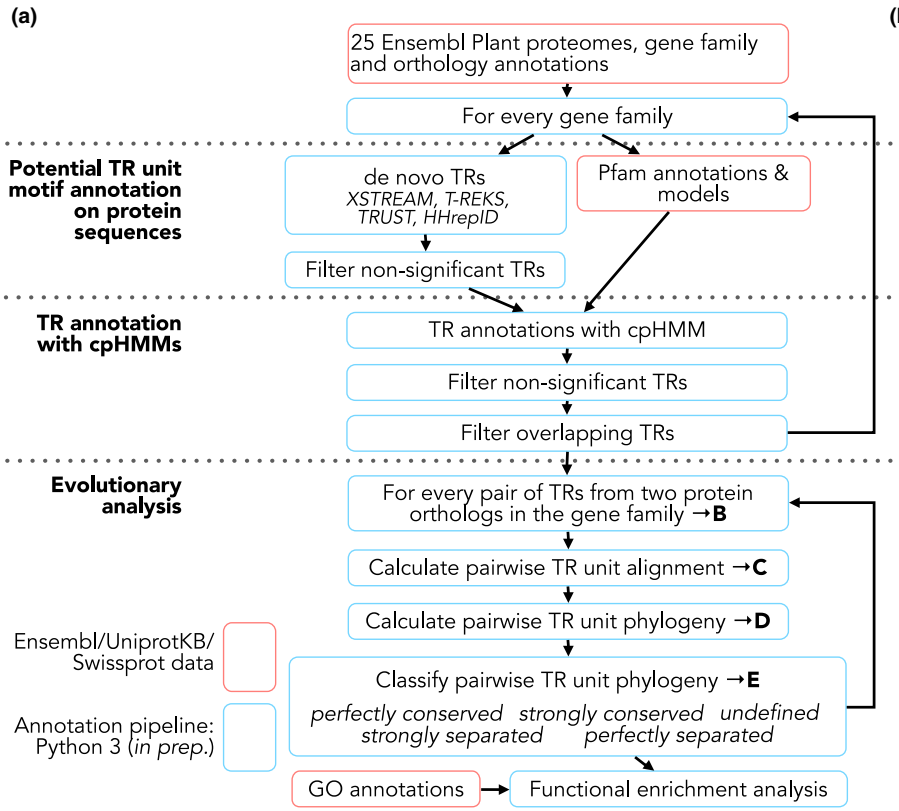
In this study, we provide an exhaustive annotation of protein TRs and their in-depth phylogenetic analysis across 25 diverse plant species for which full genomes were available in Ensembl (Groves & Barford, 1999; Adams *et al.*, 2000; Stirnimann *et al.*, 2010; Xu & Min, 2011; Flicek *et al.*, 2012). Unlike in our previous study of human TRs (Punta *et al.*, 2011; Schaper *et al.*, 2014), here we do not focus on a single species, but provide TR unit distribution and evolution data for all 25 species, based on patterns observed in TR unit phylogenies for all possible species pairs. With these data, we propose a set of candidate proteins that play a role in the adaptation of plant species, that is, those with frequent TR unit gains/losses for closely related plant species. In addition, we analyze the attributes of TRs that have been conserved deep into the tree of plants, and show how these conserved TRs differ from separated TRs in terms of the unit configuration (i.e. number and order of TR units), exon structure and functional annotation of the TR-containing protein.

Materials and Methods

Fig. 1 provides a schematic overview of the applied methods, with details in this text.

Proteome-wide annotation of TRs in protein orthologs

The entire proteomes, including gene trees, orthology annotation and alignments of orthologous sequences, of 25 plant species were downloaded from Ensembl Compara Plants v.20 (Vilella *et al.*, 2009; Flicek *et al.*, 2012; Schaper *et al.*, 2014). TRs were annotated in all sequences from two sources: PFAM domain annotations (Punta *et al.*, 2011; see, for example, Shi *et al.*, 2013) provided by Ensembl and *de novo* TR detections with HHRepID (Biegert & Söding, 2008), T-REKS (Jorda & Kajava, 2009), TRUST (Szklarczyk & Heringa, 2004) and XSTREAM (Newman & Cooper, 2007). To refine the TR annotation, we constructed circular sequence profile hidden Markov models (cpHMMs; Schaper *et al.*, 2014) – directly from the PFAM model for PFAM annotations, or indirectly from the predicted



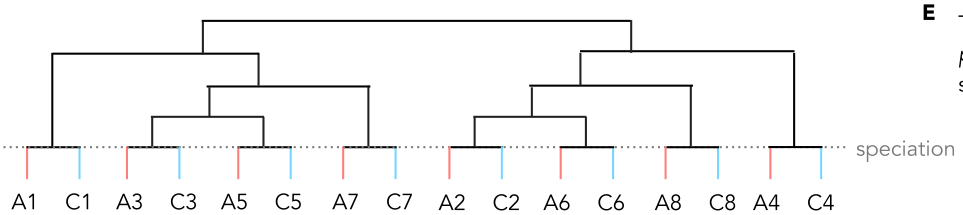
(b)

> A: Q9LJX4
[…]VGGMMSHFGRRSETRYHQQGPSRNTGIYPGG
WQGNRGGASSIVDDLKRHSFLDELKSPNARKLEL
SDIAGRVVEFSVDQHGSRFIQQKLEHCSDEEKAS
VFSEVLPQASKLMTDVFGNYVIQKFIEHGTPAQR
EELVKQLAGQMVSLSLQMYGCRVIQKALEVIDVD
QKTELIRELDGNVLKCVRDQNGNHVIQKCIESMP
AGRIGFVIAAFRGQVATLSTHPYGCRVIQRILEH
CSDDEETHCIIDEILESAFALAHDQYGNYVTQHV
LERGKPDERRQIIEKLTGNVVQMSQHKYASNVVE
KCLEHADSTEREFLIEEIMGKSEEDNHLLAMMKD
QFANYVVQKVLEISKDQQREILVQRMKIHLQSLR
KYTYGKHIVARFEQLFGEESEVSEEGTEG

> C: CMR410CT
[…]ANPPNSGVIASVGRSALLERFRTALLASDEA
FTSWKLEDIRGQIVEFSTDQHGSRFIQTKLETAT
PDQVGWVLQEVLAEMNRLVTDVFGNYVVQKLLEH
GTARDLQAIAMKLKNRILALSLHMYGCRAVQKAL
EVLPASTQAELVIELDGHVLKCIRDQNGNHVIQK
CIERVPGQHVQFIVDAVRGQAVSLAEHSYGCRVI
QRILEYSPEEQKVPIMREIMQACRTLIRDQYGNY
VIQHVVEHGKEEERAHILRMVRDQCISMSQHKYA
SNVVERCLQHGSPADRKALIDILLGRSDVAGSGS
GGGGSSGATALPRNSVPLIDLVQDQFGNYVVQRV
LDVAGDEQRQQAAELLRANLNVIKRFSYGKHILA
RLESTGAAAMAPTFANREGAPGGAHPCVQNLQNT
GKRAPAQRLQLEPSSNCSSASNSSSSSATTTSSA
SSTWNLRTASIASPSKRRGRATGCFSSEVLELEA
SMKKRSEGVTAGGFPQASPASATATAATAERNPR
AAGRARRQTHTR

(c)

(d)



E The TRs have been *perfectly conserved* since speciation.

Fig. 1 Data assembly scheme and example of conserved tandem repeat (TR) unit phylogeny. (a) Overview of the steps from unannotated Ensembl Plant gene families to bi-species TR phylogenies (in prep.: M. Anisimova, J. Pečerska, S. Zoller, E. Schaper). cpHMM, circular sequence profile hidden Markov model; GO, gene ontology. (b) PUF repeats (Pumillo-family RNA binding, PF00806) in the *Arabidopsis thaliana* RNA-binding translation regulator Pumillo homolog 5 (A; Q9LJX4) and its *Cyanidioschyzon merolae* ortholog (C; CMR410CT). (c) Alignment of all PUF repeat units enumerated according to their order in the protein sequence. (d, e) The bi-species TR unit phylogeny of the PUF repeats gives an example of the strongly conserved mode of TR evolution: all duplications leading to the currently observed TR regions in the mouse-cress and in the red algae occurred before their divergence *c*. 1.6 billion yr ago (Herron *et al.*, 2009).

TR units for *de novo* annotated TRs. These cpHMMs were then used to refine the cpHMM on the same sequence, but also to consistently annotate TRs in all orthologous sequences. A model-based statistical significance test was conducted for all candidate TRs in order to diminish the number of false-positive annotations ($\alpha = 0.1$ (0.01) for PFAM (*de novo*) annotations;

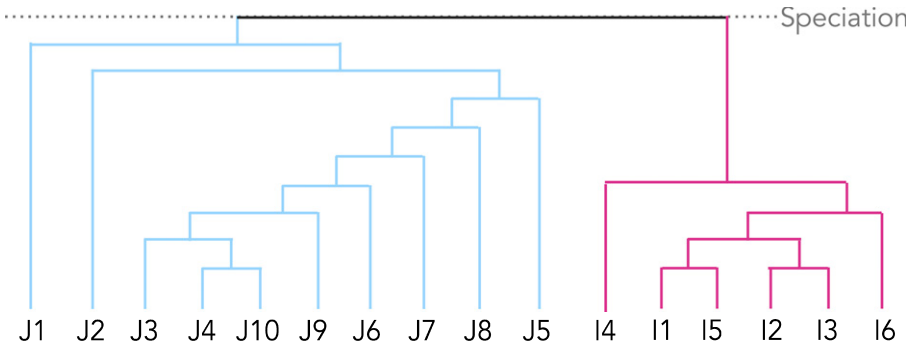


Fig. 2 An example of the bi-species tandem repeat (TR) unit phylogeny representing the strongly separated mode of TR evolution inferred for a separated leucine-rich repeat (LRR) (PF12799) found in the *Oryza sativa* Japonica putative blight resistance protein (J; Q5JMK0) and its *O. sativa* Indica ortholog (I; BGIOSGA000152). Since divergence of these species *c.* 0.4 million yr ago (Vaughan *et al.*, 2008), a number of TR unit gains/losses in at least one of the lineages have led to complete separation of the TR units of the two species on the TR unit phylogeny. TR unit gains/losses completely mask the ancestral duplication history of this TR region before speciation.

details on the assumed model of TR evolution are given in Schaper *et al.*, 2014; an introduction to the statistical test used is given in Schaper *et al.*, 2012). As a byproduct of the statistical significance test, we obtained a maximum likelihood estimate of the between-unit TR unit divergence ($\hat{d}_{\text{TR units}}$), which measured the average expected substitution rate between TR units in the same TR region (so, for identical TR units, $\hat{d}_{\text{TR units}} = 0$).

Next, we discarded TRs with number of units $n < 4$ or with unit length $l < 15$. Our analyses of TR evolution were based on phylogenies of TR units. Sizable unit lengths are the prerequisite for trustworthy reconstruction of TR unit phylogenies. The longer the TR unit, the greater is the chance that the accumulated substitutions in the TR region will be informative about the history of TR gains/losses. For this reason, our evolutionary analyses focused on the set of TRs with $l \geq 15$, following Schaper *et al.* (2014). Results for $10 \leq l < 15$ are available online, but TR unit phylogenies may be more error prone in this range (see, for example, Yang, 1998). Similarly, we discarded TRs with number of units $n < 4$ to ensure statistical significance of the TR unit phylogenies.

To avoid redundant annotations, we scanned for overlapping TR annotations. In the case of an overlap of PFAM and *de novo* annotations in the alignment of orthologous sequences, only the PFAM annotations were retained. In the case of further overlap, for example, between *de novo* annotations, only the TR with the highest statistical significance was kept. The applied procedure is described in greater detail in Schaper *et al.* (2014). Nonoverlapping TR annotations with no restriction on l are available online: ftp://ftp.vital-it.ch/papers/vital-it/Phytologist-Schaper/index.html.

Phylogenetic analysis of TRs in protein orthologs

To study the mode of TR evolution, we used pairs of orthologous proteins and first built multiple sequence alignments of all TR units from both proteins using Mafft (v7.017b; default parameters; Katoh & Toh, 2008; for an example, see Fig. 1c). Next, for each such alignment, we reconstructed the TR unit phylogenies using PhyML 3.0 (Guindon *et al.*, 2010) (see, for example, Figs 1d, 2). Previously, we have derived the exact probabilities of obtaining perfectly separated or perfectly conserved TR unit phylogenies on a random tree (Schaper *et al.*, 2014). Already, for $n = 4$, both cases are rare (2.9×10^{-4} for perfect conservation and 2.16×10^{-2} for perfect separation). However, any error in the phylogenetic reconstruction will obscure the true evolutionary history. To account for these cases, we introduced two additional and slightly weaker measures, strong conservation and strong separation. A detailed description of these measures and their significance can be found in Schaper *et al.* (2014).We reconstructed TR unit phylogenies for all pairs of TRs in orthologous proteins (including 1 : 1, 1 : many and many : many orthologs) and classified them as conserved, separated or unknown (Fig. 1e).

To establish a lower boundary for the duration of conservation of a TR in one species, we searched for the most distantly related second species, where both TRs are still strongly conserved. This provides evidence that the TR has been conserved at least since the split of both species. By contrast, to establish an upper boundary for the time to separation of TR units, we searched for the most closely related second species, where both TRs are already strongly separated. This provides evidence that the TR has undergone TR unit gains/losses on at least one of the two lineages since their time of separation.

As a result of errors in gene and orthology annotation, the numbers of conserved TRs are generally underestimated, whereas the numbers of separated TRs might be both under- and overestimated. For tests on the robustness of our results, see Supporting Information Notes S1 and Table S1.

Results

The distribution of TRs across plant proteomes

We searched for TRs with at least $n \geq 4$ units across all orthologous proteins of Ensembl plant genomes. The total number of detected TRs per proteome varied between 986 in the red alga *C. merolae* and 17 788 in soybean *Glycine max*, corresponding to 18.6% and 14.5%, respectively, of TR-containing proteins in the proteome. Table 1 summarizes the most prominent TR types

Table 1 Summary of annotated tandem repeats (TRs) across plant proteomes

	All TRs	PPR	LRR	TPR	EF hand	Kelch	WD40	ANK	*De novo* $l < 15$	*De novo* $l \geq 15$
(a) TR characteristics ($n \geq 4$) for *A. thaliana*										
$\bar{n}$	22.4	40.7	32.4	43.1	34.0	49.6	42.0	43.6	3.9	26.2
$\bar{l}$	7.0	10.7	9.4	6.9	4.0	5.1	6.1	5.5	5.6	5.9
SD (n)	4.8	4.6	5.2	2.8	0.4	0.7	1.6	1.6	4.3	4.0
$\hat{d}_{TR\,\text{units}}$	0.88	1.29	1.05	1.45	1.07	1.28	1.26	1.14	0.61	0.57
(b) Tandem repeat count ($n \geq 4$) per species										
Eudicots → Rosids → Brassicaceae										
Arabdidopsis thaliana	9222	17.8	14.1	2.9	2.4	2.0	1.9	1.5	49.2	2.1
Arabdidopsis lyrata	9349	17.5	15.1	2.9	2.4	2.1	1.9	0.2	50.6	2.1
Brassica rapa	12 438	13.2	14.7	2.9	2.6	1.9	2.0	1.2	53.5	2.1
Eudicots → Rosids → Fabids										
Glycine max	17 788	16.3	18.0	3.2	2.2	1.7	1.9	1.7	48.4	1.3
Medicago truncatula	9351	16.1	23.5	2.0	1.6	0.9	1.3	1.4	46.3	2.6
Populus trichocarpa	12 758	15.0	23.5	2.7	1.7	1.6	1.7	2.0	46.0	1.5
Eudicots → Rosids → Vitales										
Vitis vinifera	9177	20.7	21.9	2.7	1.8	1.5	1.9	2.5	41.2	1.1
Eudicots → Asterids → Solanales										
Solanum tuberosum	9929	17.0	18.2	2.3	1.9	1.0	1.5	1.5	49.6	3.3
Solanum lycopersicum	9487	16.2	15.1	2.6	2.2	1.4	1.9	1.5	52.0	2.5
Monocots → *Poaceae* → *Oryza*										
Oryza sativa Japonica	11 582	13.1	15.0	2.0	1.5	1.1	1.2	1.6	59.8	1.6
Oryza sativa Indica	13 943	11.4	16.9	1.9	1.3	1.0	1.2	1.9	59.7	1.6
Oryza glaberrima	11 665	12.3	14.0	1.9	1.4	1.2	1.2	1.7	61.4	1.5
Oryza brachyantha	9756	16.2	16.0	2.5	1.8	1.4	1.4	1.6	53.7	1.7
Monocots → Poaceae → Panicoideae										
Triticum urartu	9117	15.9	21.2	1.8	1.5	1.1	1.3	2.2	49.2	2.4
Hordeum vulgare	8028	18.8	17.6	2.1	1.6	1.4	1.3	1.4	50.3	1.3
Aegilops tauschii	10 229	16.2	24.8	1.9	1.4	1.0	1.2	2.2	46.1	1.7
Brachypodium distachyon	10 541	15.5	13.9	2.4	1.7	1.3	1.5	1.5	56.7	1.4
Monocots → Poaceae → Pooideae										
Zea mays	13 129	13.1	10.9	2.0	1.8	1.4	1.3	1.2	62.8	1.9
Sorghum bicolor	12 831	12.7	13.7	1.8	1.3	1.2	1.0	1.7	61.6	1.9
Setaria italica	11 867	14.1	17.2	2.1	1.4	1.2	1.3	1.7	56.2	1.3
Monocots → Zingiberales										
Musa acuminata	10 517	15.6	17.4	2.5	2.7	1.4	2.1	1.3	50.6	0.8
(Nonangiosperms)										
Selaginella moellendorffii	12 049	31.0	11.6	3.8	1.4	2.7	2.1	0.3	40.7	2.1
Physcomitrella patens	7098	4.5	16.2	3.6	2.1	2.2	2.7	0.9	59.6	2.5
Chlamydomonas reinhardtii	7762	0.3	2.0	1.4	0.5	0.6	1.1	1.0	89.7	1.5
Cyanidioschyzon merolae	986	1.2	2.0	8.7	0.3	5.8	4.2	1.2	69.8	1.6

(a) Characteristics for all analyzed *Arabidopsis thaliana* TRs, averaged over the seven most frequent TR types and *de novo* annotated TRs: l, TR unit length; n, number of TR units per TR; $\hat{d}_{\text{TR units}}$, average within-unit TR divergence (see the 'Materials and Methods' section). (b) The first column shows the total number of TRs for each species. Other columns show the percentage of TRs belonging to different TR types: for example, 17.8% of all *A. thaliana* TRs were pentatricopeptide repeats (PPRs). LRR, leucine-rich repeat; TPR, tetratricopeptide repeat; ANK, Ankyrin repeat.

and the TR count for all species. TR annotations were built from two sources: PFAM A annotations and *de novo* detections. For example, for *Arabidopsis thaliana*, 4731 of 9222 TRs (51.3%) were *de novo* detections. Every PFAM A annotation and *de novo* detection was converted to a cpHMM (details in the 'Materials and Methods' section; Schaper *et al.*, 2014). We refer to all TRs that were detected with the same cpHMM to be of the same TR type. Further, for some common TRs annotated in PFAM (e.g. LRR), several pHMMs are available (e.g. LRR1, LRR2), and we refer to all of them under one name.

We observed a strong correlation between the number of proteins encoded in a genome and the number of predicted TRs (Fig. 3; $R^2 = 0.86$ for $l \geq 15$; $R^2 = 0.37$ for $l < 15$). On average, TRs were detected in 37% of all plant genes, but ranged widely from *c.* 28% in the barrel clover *Medicago truncatula* to *c.* 82% in the green alga *Chlamydomonas reinhardtii*. Interestingly, *C. reinhardtii* also exhibited comparably high densities of short nucleic TRs in coding and noncoding sequences (Zhao *et al.*, 2013). Only *c.* 3.4% of TRs were unique to a single species. However, the ratio of unique TRs was significantly higher among *de novo* detections (*c.* 13.3%) compared with TRs detected based on cpHMM matches to known PFAM domains (*c.* 2.2%). Within each species, the distribution of TRs was dominated by just a few TR types. For example, 42 577 (16.3%; varying from 10.9% in *Zea mays* to 24.8% in *Aegilops tauschii*) of all TRs in the entire dataset spanning all 25 plant species were LRRs. Further, 39 742 (15.1%; varying from 11.4% in *Oryza sativa* Indica to 20.7% in *Vitis vinifera*) were

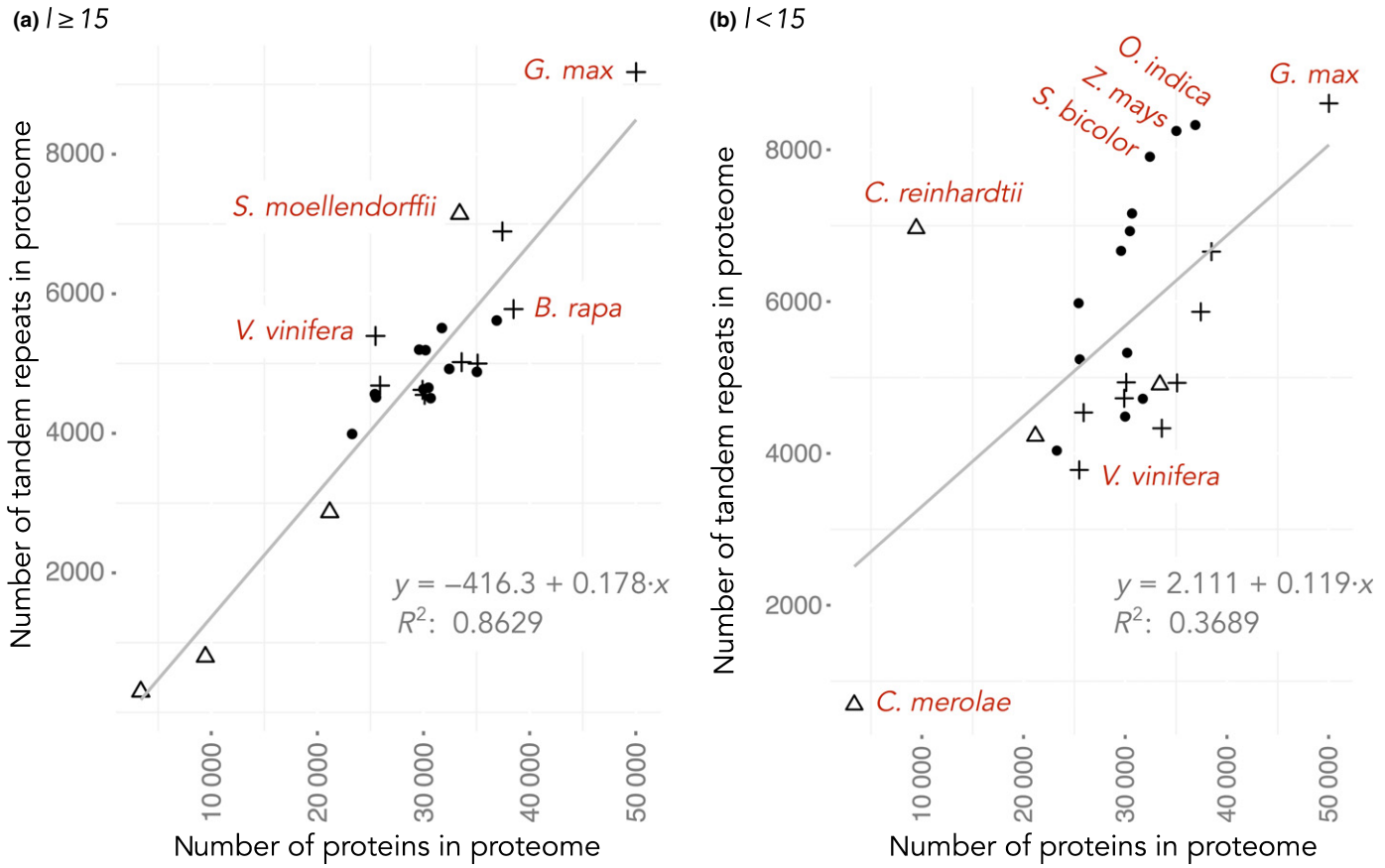


Fig. 3 Correlation between the number of tandem repeats (TR) and the number of proteins for proteomes of Ensembl Compara Plants (v. 20). The number of proteins per proteome is strongly correlated with the number of whole-genome duplications that have occurred during species evolution. (a) TR with unit length $l \geq 15$. (b) TR with unit length $l < 15$. *G. max*, *Glycine max*; *S. moellendorffii*, *Selaginella moellendorffii*; *V. vinifera*, *Vitis vinifera*; *B. rapa*, *Brassica rapa*; *C. reinhardtii*, *Chlamydomonas reinhardtii*; *O. indica*, *Oryza sativa* Indica; *Z. mays*, *Zea mays*; *S. bicolor*, *Sorghum bicolor*; *C. merolae*, *Cyanidioschyzon merolae*.Crosses, eudicots; circles, monocots; triangles, nonangiosperms.

PPRs. In total, roughly every third *A. thaliana* TR belonged to one of these two TR types. This includes all PFAM TRs detected in our dataset, as well as a small fraction of all *de novo* annotations (see Table 1).

The majority of plant TRs show complete long-term conservation

Inference of conserved TRs in plants was based on reconstructed TR unit phylogenies for all pairs of TRs in orthologous proteins (or 'orthologous TRs'). If a pair of orthologous TRs was classified as strongly conserved, this provided evidence that no TR unit gains/losses had occurred since the divergence of the two lineages. Therefore, for any given TR, we were able to estimate the minimum duration of the TR conservation by tracing the most distant pair of strongly conserved orthologous TRs. For example, if a given TR from *A. thaliana* was found to be strongly conserved with respect to the corresponding TR in a protein ortholog in *Brassica rapa*, we concluded that this TR had been conserved in both species since the root of the Brassicaceae. However, if the TR was not strongly conserved with respect to any orthologous TR from a more distant species, we could not draw any conclusions beyond that point.

Fig. 4 provides an overview of the conservation patterns for different TRs in our dataset at different evolutionary distances across the kingdom of plants. The data are summarized for all TRs of the same type from within-species paralogs (e.g. the 4684 *A. thaliana* TRs are summarized to 1378 nonparalogous TRs). The majority of TRs were found to be conserved, particularly within single genera. For example, 1219 (88%) of the TRs were conserved between *A. thaliana* and *A. lyrata*, which split *c.* 13 million yr ago (Beilstein *et al.*, 2010). Most remaining TRs could not be clearly assigned to a mode of evolution (could not be classified as either separated or conserved).

Further, 1226 TRs (70%) were conserved between monocots and eudicots, thus providing evidence for TR conservation at least to the splitting of these groups *c.* 150 million yr ago (time estimate from Chaw *et al.*, 2004). Surprisingly, 162 orthologous TRs were detected in the proteomes of both red algae and Viridiplantae, and 68 of these (42%) were conserved at least since the ancient split *c.* 1.6 billion yr ago (Herron *et al.*, 2009). We conclude that strong conservation over long evolutionary times must be the predominant mode of evolution of domain-like TRs across the examined plant species. Apart from some differences in the frequency of TR types and, consequently, the total number of TRs, the relative proportion

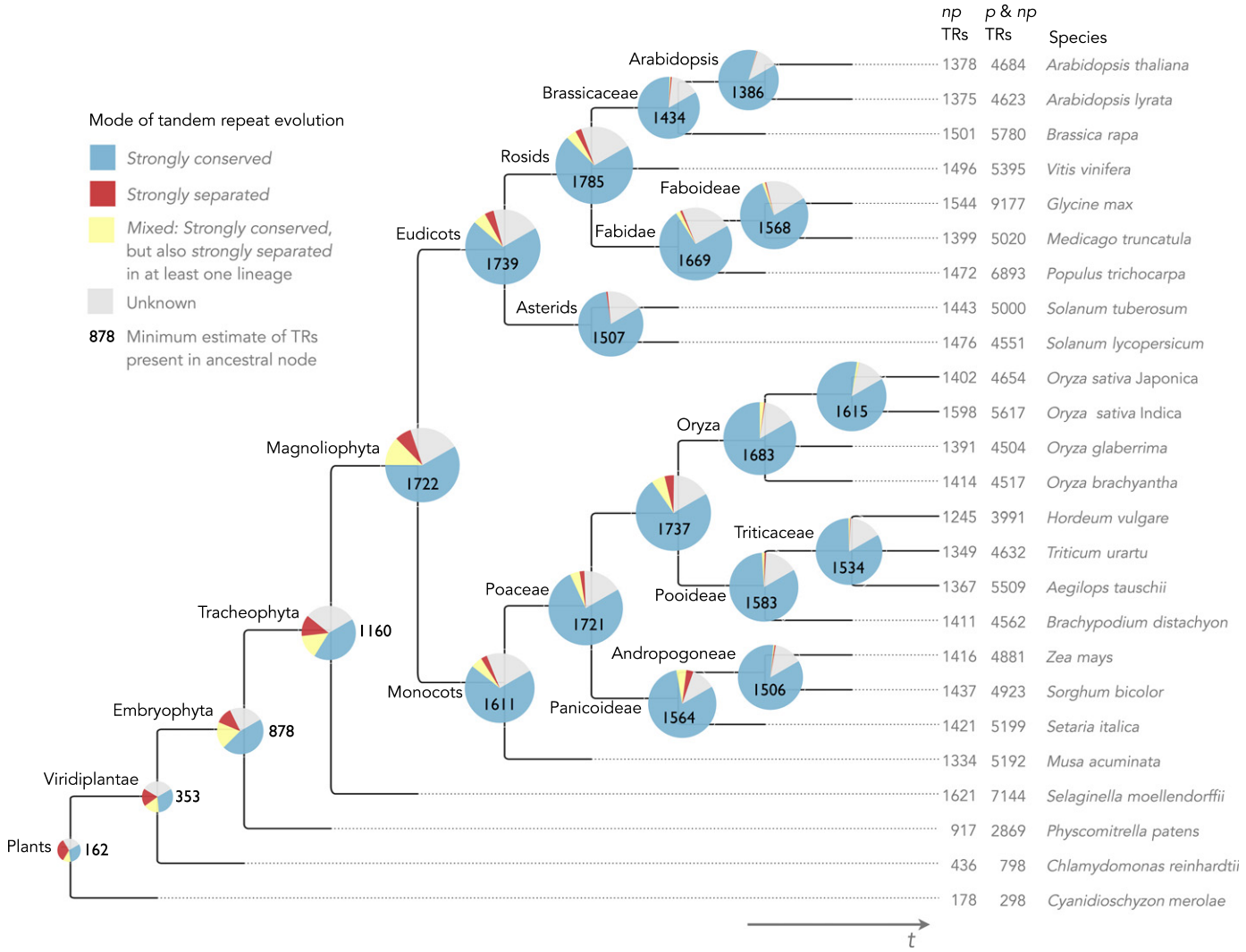


Fig. 4 Conserved and separated tandem repeats (TRs) in plants. The rooted cladogram is shown for all plant species in our analysis. The paralogs and nonparalogs (*p* & *np*) column next to the species name denotes the count of all TRs with unit length $l \geq 15$ for each species. Next, nonparalogs (*np*) are all groups of within-species paralogs; the count of this group is marked in the leftmost column. All results presented on the cladogram are based on *np*. For every tree node, we calculated the number of TRs that were present in at least one ortholog of the two lineages descendant from the node. For example, 1785 unique TRs were found in at least one Brassicaceae ortholog, as well as in at least one Fabidae ortholog, providing evidence that the most recent common ancestor of both lineages (denoted by the ancestral node) also contained these TRs. This number was depicted on the pie chart for each node, and is further illustrated by the size of each pie chart. Further, each pie chart shows the frequency of different modes of TR evolution. We checked whether each TR has been strongly conserved (blue) or strongly separated (red) in a pair of orthologs from both lineages. In addition, a TR might have been strongly conserved since the split of the lineages in two species, but still have undergone TR unit gains/losses in a third species. In this case, we denote the TR as both strongly conserved and strongly separated (yellow). If no pairwise TR unit phylogeny provides evidence for either conservation or separation, the evolutionary mode of the TR is denoted as unknown (grey). The cladogram was taken from the National Center for Biotechnology Information (NCBI) (November 2013): http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi. FigTree was used to visualize the plant phylogeny: http://tree.bio.ed.ac.uk/software/figtree.

of conserved TRs was remarkably similar across different plant families.

A large variety of TR types showed TR conservation. For example, Fig. 5(a) (top bar) shows the distribution of TR types (including all paralogs) found in *A. thaliana* that have been conserved at least since the split of monocots and eudicots. At the same time, the conserved TRs were dominated by only a few TR types that occurred in very high frequencies, including PPRs, LRRs, TPRs and Kelch repeats.

Although the conserved TRs originally expanded to their current size by TR unit gains/losses, the TR region has evolved into a stable supra-domain, and TR unit gains/losses did not contribute to the current function of the TR-containing protein. For example, plant proteins with conserved TRs were enriched in binding functions (e.g. LRR, WD40, TPR, EF hand, RCC1, see Fig. 5), which has also been observed in metazoans (Schaper *et al.*, 2014). Frequently, the TR region plays the role of a structural scaffold that promotes the formation of molecular

New Phytologist

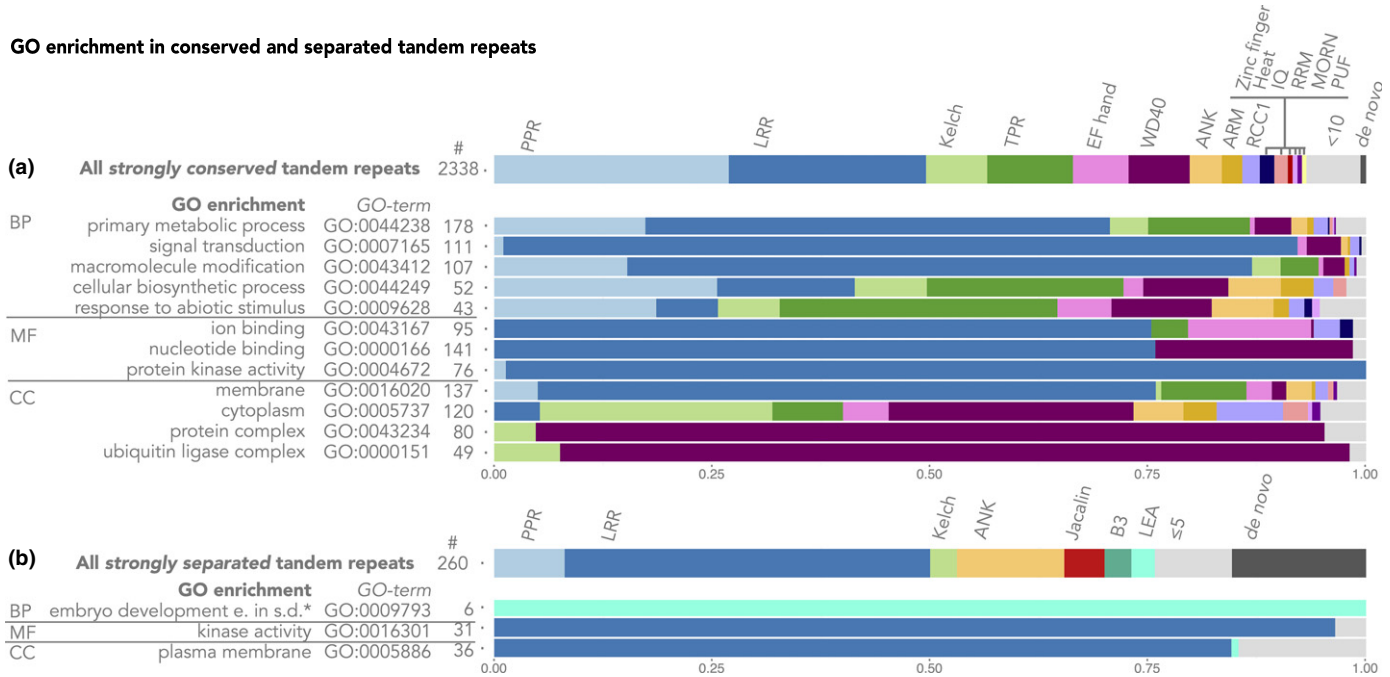


Fig. 5 The distribution of tandem repeat (TR) types and enriched gene ontology (GO) terms for *Arabidopsis thaliana* proteins with conserved and separated TRs. GOrilla (Eden *et al.*, 2009) was used to perform the enrichment analysis assuming a hypergeometrical model with all TR-containing *A. thaliana* proteins ($l \geq 15$) as background distribution. (a) TRs ($l \geq 15$) that have been strongly conserved at least since the split of the monocots and eudicots. The first summary bar shows the frequency of the different TR types: there are 2338 conserved TRs, with pentatricopeptide repeats (PPRs) being the most frequent. All TR types based on *de novo* TR detections were binned into one category (dark grey), although they may describe very diverse motifs. Likewise, TR types based on PFAM annotations with low frequencies (< 10 TRs) were binned together (light grey). The thinner bars below the summary bars show representative enriched GO terms ordered by their frequency. Each bar corresponding to a GO term depicts the distribution of different TR types in proteins annotated with this GO term. GO terms are grouped by their respective ontology: biological process (BP), molecular function (MF) or cellular component (CC). (b) Corresponding plot for the 260 *A. thaliana* TRs ($l \geq 15$) that have been strongly separated in at least one magnoliophyte (monocot or eudicot in our dataset). Here, TR types based on PFAM annotations with low frequencies (≤ 5 TRs) were binned together. The GO enrichment data comprising directed acyclic graphs of enriched GO terms are available online within the full dataset. embryo development e. in s.d.*, embryo development ending in seed dormancy; ANK, Ankyrin repeat; ARM, Armadillo/beta-catenin like; IQ, IQ calmodulin-binding motif; LEA, late embryogenesis abundant; LRR, leucine-rich repeat; MORN, Membrane Occupation and Recognition Nexus; PUF, Pumillo-family RNA binding; RCC, Regulator of chromosome condensation; RRM, RNA recognition motif; TPR, tetratricopeptide repeat.

complexes and interactions. The high conservation of TRs with binding functions suggests that any change in the TR constellation may have detrimental effects on fitness, such as disturbance of protein functions, and therefore would be selected against.

Proteins containing conserved plant TRs were enriched in a diversity of biological processes (Fig. 5a). These included mostly processes in plant primary metabolism, such as signal transduction (e.g. LRR, WD40, ARM), macromolecule modification (e.g. LRR, PPR) and cellular biosynthetic processes (e.g. PPR, TPR, WD40), but also processes in plant secondary metabolism, such as response to abiotic stimuli (e.g. TPR, PPR, WD40, ANK).

Few protein TRs evolve by unit gains and losses in plants

Using bi-species TR unit phylogenies, we also searched for strong separation of TRs. If a pair of TRs in two orthologous proteins show strongly separated TR units, this indicates that a series of TR unit gains/losses have occurred in at least one of the lineages since speciation.

Fig. 4 illustrates the numbers of separated nonparalogous TRs (red or yellow) for different clades across plants. For closely related species, very few or no TRs have separated. However, the relative proportion of TRs that have evolved by repeated TR unit gains/losses in at least one species increased with the depth of the clade. For example, 345 TRs (20%; counting every group of within-species paralogs as one) were separated between two magnoliophytes, but a much larger proportion of TRs (68 TRs representing 42%) were separated between at least any two species in our dataset. Similarly, the number of TRs that have been conserved in some lineages and separated in others also increased with the depth of the clade. This is to be expected: analogous with the evolution for gene families, TR units in different lineages may be subjected to different selective pressures as a result of environmental changes over time. In addition, it is possible that, within one species, one paralog has been conserved, whereas another paralog has undergone TR unit gains/losses or other mutations concealing the TR structure. This is probably common, as plant genomes are marked by a large number of gene paralogs because of frequent sequence duplication events.

Fig. 5(b) shows the distribution of TR types among all protein TRs in *A. thaliana* which have been strongly separated in comparison with at least one other magnoliophyte. The total number of these separated TRs was 260 of all 4684 detected TRs,

representing 5.6%. The difference in the percentage of separated TRs should be noted: although this was 5.6% when only one lineage was considered, it was 20% when all lineages on the phylogeny were considered (see above).

The majority of separated TRs were LRRs (109; 42%), followed by ANKs (32; 12%) and PPRs (21; 8%). Another large fraction of separated TRs were *de novo* detections (40; 15%). In addition, several rare domains were especially enriched in separated TRs, including Jacalin ($l = c.$ 130) and late embryogenesis abundant protein TR (LEA) ($l = c.$ 73). Jacalin-containing proteins have recently been shown to be a proponent player in plant adaptation to environmental stresses in wheat (Song *et al.*, 2013). Similarly, LEA-containing proteins are thought to be involved in the abiotic stress response (e.g. Hundertmark & Hincha, 2008). In line with our data, it is possible that fast TR unit gains/losses of Jacalin- or LEA TRs is a means of fast adaption to environmental change. A gene ontology (GO) term analysis of all proteins with separated TRs yielded an enrichment in kinase activity, plasma membrane proteins and, interestingly, 'embryo development ending in seed dormancy' (Fig. 5); however, because of the small sample size, it may be inept to draw generalized conclusions for the function of these TRs. Presumably, the molecular function of separated TRs should be addressed in case-wise studies.

To focus on particular TRs that might be subject to population-scale TR unit gains/losses, we searched closely related species (e.g. *A. thaliana* and *A. lyrata*; *Solanum tuberosum* and *Solanum lycopersicum*) for pairs of orthologous TRs with strongly separated TR units (data in Table 2). In general, there were very few such cases, for example, 7/4684 (0.1%) of all *A. thaliana* TRs. Many separated TRs were detected *de novo* (3/7 in *A. thaliana*; 5/11 in *S. tuberosum*). These had mostly short TR units with low sequence divergence, which usually facilitates TR unit gain/loss mutations. The role of these TRs in protein function is currently unknown.

Most of the separated TRs were found within a single exon, suggesting that unit gains/losses may have occurred in tandem at the nucleic level. Possible exceptions to this rule in our data were one separated PPR (AT5G55000.2) and one separated LRR8 (AT5G019501) in *A. thaliana*. When the number of exons matches the number of TR units, it is most likely that the TR region did not arise through slippage, but rather by an exon shuffling-like process (e.g. Björklund *et al.*, 2006).

Interestingly, most of the strongly separated TRs belong to the LRR family of domains (3/7 in *A. thaliana*; 6/12 in *S. tuberosum*; 8/8 in *Triticum urartu*), which are often found in plant genes associated with resistance properties. We discuss the significance of this finding below.

Correlation of TR features and TR mode of evolution

In order to understand which molecular mechanisms contribute to the evolution of plant TRs, we contrasted the conserved and separated TRs from *A. thaliana* in terms of their molecular characteristics (Fig. 6). We found that the between-unit TR divergence was a strong predictor of the mode of TR evolution (Fig. 6a). Separated TRs that had undergone TR unit gains/losses since the split of monocots and dicots showed, on average, a clearly lower sequence divergence than those TRs that had been conserved at least during the same time. This was expected for two reasons. First, the phylogenetic history of separated TRs is younger than that of conserved TRs because of numerous TR unit gains/losses in at least one lineage. Thus, separated TR units had less time to accumulate sequence substitutions. Second, the probability of mismatch mutations leading to TR unit gains/losses is highest for TR units with highest sequence identity (Albà *et al.*, 1999; Faux *et al.*, 2007). Therefore, TRs with low between-unit divergence have a greater chance to become separated. Consistent with this, the group of *de novo* detected TRs, which tend to have a low sequence divergence, was strongly represented within the set of separated TRs, but was very rare within the conserved TRs (Table 1, Fig. 5). Similar trends have been reported for human TRs (Schaper *et al.*, 2014). Apart from the between-unit divergence, all other observed TR characteristics were less predictive of the mode of TR evolution. Conserved TRs had slightly longer TR units compared with separated TRs (Fig. 6c), which may be explained by lower mispair mutation probabilities of long TR units (Schlötterer, 2000; Leclercq *et al.*, 2010).

The TR unit length is often thought to be a determining factor for the mutation rate of TR unit gains/losses, with shorter TRs being more mutable (e.g. more prone to DNA replication slippage). Thus, in the absence of selective pressure, longer TRs would be expected to be more conserved than shorter TRs. In accordance, in our data, conserved TRs have longer average TR unit lengths than separated TRs (Fig. 6c). For example, 51.5% of all *A. thaliana* TRs with $l > 20$ have been strongly conserved since the split of the monocots and eudicots (compared with 30.9% for TRs with $15 \leq l < 20$ and 11.9% for TRs with $10 \leq l < 15$). However, TRs with longer units can tolerate more mutations before the repeat structure is disrupted. Together with a TR detection bias for TRs of different lengths, this might contribute to the observed differences: the longer the TR, the more strongly its sequence can diverge and still be detectable. At the same time, TRs with higher sequence divergence among the TR units tend to be more conserved in terms of TR unit gains/losses (see above).

In terms of the exon structure, two scenarios are plausible for TRs with fast TR unit gain/loss (Schaper *et al.*, 2014). First, TR units may become lost or duplicated through tandem mismatch mutations. In this case, we would expect the TR units to be physically adjacent on the nucleic level, and not separated by introns. Second, TR units may evolve through an exon shuffling-like mechanism. In this case, we would expect the TR units to be divided into exons. For *A. thaliana* TRs, conserved TRs tended to occupy, on average, slightly more exons, having fewer TR units per exon (Fig. 6d,e). However, from our results, we cannot deduce whether a single mechanism is generally responsible for TR unit gains/losses in plants. The mechanism might vary on a case-by-case basis. For example, all separated PPRs from *A. thaliana* occupied a single exon, suggesting a mutation mechanism based on tandem mismatches (see also O'Toole *et al.*, 2008).

Table 2 Strongly separated tandem repeats (TRs) in pairwise species comparisons

Ensembl protein ID	Description	Detection ID	l	n	$\hat{d}_{TR}$	n_{Exon}	Max (n_{TR})
(a) *Arabidopsis thaliana* (I) compared with *A. lyrata* (II)							
AT5G55000.2	Pentapeptide TR in ubiquitination protein	PF00805	40	4.4	0.87	3	4
AT5G01950.1	LRR8 in transmembrane receptor	PF13855	57	3.8	0.76	3	3
AT5G49750.1	LRR8 in signal transducer	PF13855	48	4.0	1.05	3	2
AT1G33612.1	LRR4 in endomembrane system	PF12799	48	6.4	1.14	1	7
AT4G08395.1		*De novo*	25	7.0	0.50	2	5
AT2G41260.2	TR in late-embryogenesis-abundant gene involved in the acquisition of desiccation tolerance	*De novo*	55	3.8	0.04	2	3
AT4G05250.1	TR in ubiquitin-like superfamily protein	*De novo*	37	5.5	0.21	1	6
(b) *Solanum tuberosum* (I) compared with *Solanum lycopersicum* (II)							
PGSC0003DMT400037401	Ubiquitin TR	PF00240	76	5.0	0.01	1	6
PGSC0003DMT400062706	LRR6	PF13516	24	12.5	0.55	1	13
PGSC0003DMT400036468	LRR7	PF13504	22	7.4	1.46	1	8
PGSC0003DMT400061341	LRR8	PF13855	72	13.5	0.77	3	6
PGSC0003DMT400062706	LRR8	PF13855	59	3.8	0.35	1	4
PGSC0003DMT400062712	LRR4	PF12799	48	6.5	0.63	1	7
PGSC0003DMT400062706	LRR4	PF12799	48	6.1	0.31	1	7
PGSC0003DMT400051907	Unannotated	*De novo*	59	5.5	0.06	1	6
PGSC0003DMT400093707	Unannotated	*De novo*	14	4.0	0.42	1	4
PGSC0003DMT400076370	Unannotated	*De novo*	25	12.2	0.33	1	13
PGSC0003DMT400089185	Unannotated	*De novo*	24	8.4	0.29	2	7
PGSC0003DMT400013432	Unannotated	*De novo*	22	3.7	0.12	1	4
(c) *Oryza sativa* Japonica (I) compared with *O. sativa* Indica (II)							
OS04T0483600-01	Kelch	*PF13415*	53	4.9	1.10	4	2
OS04T0628100-01	Ubiquitin	*PF00240*	76	5.0	0.01	1	6
OS04T0483600-01	Kelch5	*PF13854*	55	4.3	1.14	4	4
OS09T0343200-01	ANK4	*PF13637*	62	4.5	1.23	2	10
OS11T0173900-00	LRR4 (two copies) in kinase	*PF12799*	48	9.7	0.75	1	7
OS05T0250700-00	LRR7 in kinase	*PF13504*	23	6.0	1.31	1	6
OS11T0567600-01	LRR4 (two copies)	*PF12799*	48	5.3	0.79	1	5
OS11T0172300-01	LRR8	*PF13855*	61	3.8	0.88	1	7
OS11T0570000-01	LRR8 in putative receptor kinase	*PF13855*	72	5.6	0.74	1	8
OS06T0111300-00	PPR1	*PF12854*	35	9.9	1.32	2	10
OS11T0435300-00	ANK5	*PF13857*	69	4.3	1.11	2	4
OS01T0937400-02	LRR6 in putative blight/disease resistance protein	*PF13516*	24	20.5	0.80	1	21
OS03T0573500-00	LRR4 (two copies) in putative disease resistance protein	*PF12799*	47	9.7	1.83	1	11
OS02T0227900-00	LRR4 (two copies)	*PF12799*	48	4.7	1.05	1	5
(d) *Triticum urartu* (I) compared with *Aegilops tauschii* and *Hordeum vulgare* (II)							
TRIUR3_01633-P1	LRR4	PF12799	48	3.7	0.67	1	4
TRIUR3_00451-P1	LRR1	PF00560	23	4.9	1.33	1	6
TRIUR3_04398-P1	LRR1 in putative disease resistance protein	PF00560	23	4.1	1.01	1	5
TRIUR3_01004-P1	LRR1/7	PF00560	25	3.6	1.21	2	2
TRIUR3_10923-P1	LRR7 in putative disease resistance protein	PF13504	23	4.4	1.06	1	5
TRIUR3_21043-P1	LRR1	PF00560	23	9.0	1.59	1	10
TRIUR3_11748-P1	LRR7 in putative disease resistance protein	PF13504	17	3.8	1.11	1	4
TRIUR3_16358-P1	LRR1 TR in putative disease resistance protein	PF00560	21	5.9	0.65	1	6

Listed are all detected separated TRs in closely related species: for two monocots (a, *A. thaliana*; b, *S. tuberosum*) and two dicots (c, *O. sativa*; d, *T. urartu*). The TR unit length (l), number of TR units (n), between-unit TR divergence ($\hat{d}_{TR}$), number of exons spanned by TR (n_{Exon}) and maximum number of TR units in any of the exons (Max (n_{TR})) are shown for TR in these four species. LRR, leucine-rich repeat; ANK, Ankyrin repeat; PPR, pentatricopeptide repeat.

Discussion

Our study shows the high frequency and high diversity of TRs in proteomes across the plant kingdom. A multitude of motifs occur repeated in tandem – some of these are very rare, and are only found in one or a few proteins. Others, such as PPRs, LRRs and TPRs, are frequently found in large gene families. The most common TR types persist in all plant species, but there is evidence for particularly strong expansions of certain TR types in some lineages (e.g. PPRs in *Selaginella moellendorffi*).

Evolutionary origin of TRs

We found multiple pieces of evidence indicating that TR regions are shaped through an ancient process of motif gain/loss, which acts to generate new domains and, ultimately, new proteins, as

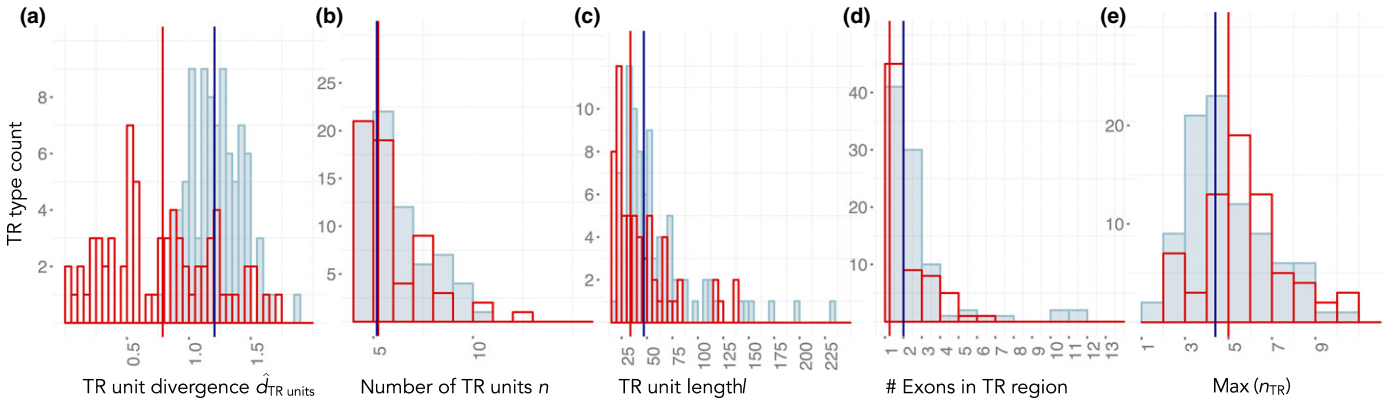


Fig. 6 Characteristics of separated and conserved tandem repeats (TRs). Frequency distributions of TR characteristics for strongly conserved (blue columns) and strongly separated (red columns) *Arabidopsis thaliana* TRs with reference to the magnoliophytes. The mean value was calculated for each TR characteristic and for each of the 122 TR types with strongly conserved TRs and 74 TR types with strongly separated TRs, where a TR type comprises all TRs detected by the same circular sequence profile hidden Markov model (cpHMM). For example, the large family of leucine-rich repeats (LRRs) was condensed into two data points – the first for strongly conserved TRs with a mean length of LRR units of $l = 31.1$, and the second for strongly separated TRs with a mean TR unit length of $l = 42.3$. The TR characteristics shown are: (a) TR unit divergence $\hat{d}_{TR}$, which is the maximum likelihood estimate of the TR between-unit divergence, resulting from the model-based TR significance test (Schaper *et al.*, 2012); $\hat{d}_{TR\ units}$ is measured as the expected number of amino acid substitutions per site since the root of the tandem repeat unit tree; (b) the number of amino acids in the TR multiple sequence alignment (counted only for columns with more amino acids than gaps, which we parsimoniously consider as noninsertion columns) divided by l; (c) TR unit length l, defined as the number of (noninsertion) sites of the TR unit with at least as many observed amino acid characters as gaps in the respective column of the TR multiple sequence alignment; (d) the number of exons (n_{Exon}) that contain at least parts of the TR region; (e) the maximum number of TR units in any of the exons.

has been proposed recently (Bornberg-Bauer & Albà, 2013). Importantly, the majority of plant TRs show long-standing conservation and high stability, and are involved in numerous molecular processes in plants. This suggests that often the protein function requires a stable TR configuration, such that the TR region can be interpreted as one domain, rather than a sequence of single domains. We observed similar results in humans (Schaper *et al.*, 2014), with many frequent TR types shared between metazoans, fungi and plants (including LRRs, Kelch, WD40). This suggests that the origin of these TRs and the molecular mechanisms shaping them predate the split of ophistokonts and plants.

By comparing our results for Arabidopsis TRs and previous results for human TRs (Schaper *et al.*, 2014), we can see that both species share many common TR types, albeit in different proportions. Despite this, there are striking differences among separated TRs in these species. Almost no TR type that shows separation in humans (with respect to mammals; e.g. zinc fingers, neuroblastoma breakpoint family (NBPF) TRs, epidermal growth factor (EGF) TRs, Schaper *et al.*, 2014) also shows TR separation in Arabidopsis (with respect to magnoliophytes; e.g. LRR, PPR, ANKs; Fig. 5), and vice versa. This leads to the conclusion that none of the ancient TR types are specifically prone to TR unit gains/losses. Rather, they are frequently conserved in terms of TR unit number and order, forming a stable domain, which is used in different proteins as an architectural block. However, separated TRs evolved more recently, often in proteins that had been subject to a recent large-scale expansion (such as zinc fingers in mammals and LRRs in magnoliophytes).

Furthermore, despite the diversity of plants in our study, we observed a constant high ratio of TRs per proteome (Fig. 3a). The high ratio shows that TR unit gain/loss is a major source of new domains, whereas the uniformity of the ratio across species suggests that the mechanism of TR generation has persisted over long evolutionary times. In comparison, the ratio of short TRs ($l \leq 10$) per protein is clearly elevated in monocots compared with dicots (Fig. 3b). This indicates that the mechanism generating short TRs must have experienced a shift in at least one of the lineages after their split. The reason behind this distinction between monocots and dicots is an open question. For plants, in particular, whole-genome duplications have contributed strongly to current proteome sizes. Therefore, it is interesting to note that a large number of TRs in plant proteomes often does not signify a large diversity of TRs, but rather a large number of paralogous TRs.

Potential for evolutionary adaptation via diversification of TR unit repertoires

Based on our observations in closely related species, we suggest that evolution by frequent TR unit gains/losses in plant protein TRs is the exception, and not the rule, affecting only a small fraction of protein TRs in plants. The evolution of protein TRs is, in this regard, clearly distinct from micro- and minisatellite amplification. Evidence for the separation of TR units between pairs of closely related plant species was limited to a handful of TR-containing proteins. These TRs were enriched in resistance-related LRRs, which typically act in pathogen effector recognition in the extracellular region, suggesting that they might provide a source of adaptive variation. In our data, LRRs represent the most common TR type in most species, comprising up to one-quarter of all protein TRs (e.g. in the true grass *A. tauschii*, and the fabids

M. truncatula and *Populus trichocarpa*). The diversity of LRRs is large: in our data, the mean expected divergence of LRR units across all *A. thaliana* orthologs was 1.05 substitutions per site (Table 1). Reflecting this, several sequence profile HMMs for LRRs are available in the PFAM database to capture this diversity. In terms of molecular function, LRRs in plants are generally thought to act as receptor domains, for example, for ligand recognition, and are therefore exposed to the extracellular region of transmembrane proteins. Further, many of these receptor proteins feature kinases towards the cytosol, such that ligand recognition can trigger a signaling cascade within the cell. In our data, among proteins with separated tandem LRRs, we found several kinases, transmembrane receptors and signal transducer proteins (Table 2).

A large fraction of LRR-containing proteins are associated with the plant immune system as the main pathogen effector-recognizing agent (for a review, see Jones & Dangl, 2006). These are known as resistance proteins (or R proteins), often contain a nucleotide-binding domain and are therefore commonly referred to as NB-LRR proteins. Diversifying selection of exposed amino acids in the LRR region has been proposed to create the necessary diversity, enabling rapid adaptation to co-evolving parasites (Tameling & Joosten, 2007; Yang *et al.*, 2013). In addition to amino acid substitutions, TR unit gains/losses presumably represent a drastic means to change the ligand recognition properties of LRR-containing proteins. In our dataset, 58% of all LRRs in flowering plants were found to be conserved since the ancestors of eudicots and monocots, compared with 70% for all TRs (including LRRs). At the same time, 41% of all LRRs were separated, compared with 20% of all TRs in the same range (Supporting Information Fig. S1). Thus, although the majority of LRRs were found to be conserved over long evolutionary distances, we found a comparable number of LRRs that were subject to recurrent TR unit gains/losses, which was extremely rare for other TR types. In contrast with other mainly conserved plant TRs, in our data, the LRRs were by far the largest single TR type affected by unit separation.

Genes with separated LRR units might be involved in protein adaptations, leading to improved tolerance and pathogen resistance. Plants have evolved a large repertoire of NB-LRR proteins specific to a wide range of pathogen effectors. This repertoire is largely conserved to enable continuous protection against the pathogen. In our data, the LRRs with conserved units may be representative of these proteins. In addition, the extreme radiation of NB-LRR proteins as a result of gene duplications allows some paralogs to evolve under relaxed selective constraints. Lineage-wise, the resulting diversity may then allow for the adaptation to emergent pathogenic effectors, perhaps in a Red Queen scenario (i.e. in an evolutionary arms race; Jones & Dangl, 2006; Diévart *et al.*, 2011). This would explain the comparatively large fraction of separated LRRs in our data and, in particular, the multiple cases of separated LRRs within closely related species (see above; Table 2).

However, as the repertoire and roles of resistance genes in plant immunity vary widely among the lineages, the more specific reasons behind the cases of separation in LRR regions deserve further investigation. With more specific questions in mind, researchers could generate population data and set up experiments in combination with the computational approach presented here to allow the detection of interesting candidate genes by contrasting the variability of LRR units in *R* genes within populations and between relevant closely related species. This type of analysis would reveal whether the LRR is subject to population-scale TR gains/losses or, alternatively, whether TR separations are rare events that intersperse long periods of TR unit conservation. Knowledge on specific TR unit differences responsible for functional disease resistance-related changes might then be used to generate synthetic proteins that could be introduced to new species and tested for resistance properties.

Likewise, for cases of separation among other TR types, such case-wise studies would be necessary to gain further insight into the role of TRs in function and adaptation. Current large-scale sequencing efforts, such as the *A. thaliana* 1001 Genomes Project (Cao *et al.*, 2011), are interesting candidates to provide the necessary data.

Contrasts in the conservation of noncoding and coding TRs

An interesting question is why we observe a high degree of conservation in protein TRs with $l \geq 15$ aa, whereas noncoding micro- and minisatellites mutate on much smaller timescales. In other words, do the protein TRs in this study underlie very low mutation rates, perhaps as a result of diverged sequences, and/or longer TR units? Or does selection act on the protein TR region to keep the number of TR units conserved, although TR unit mutations do occur over the timescales considered in our study? The existence of examples of TRs of similar length, and similar sequence divergence, but still very different evolutionary timescales, suggests that neutral evolution cannot always accurately describe protein TR evolution (see, for example, Hancock *et al.*, 2001; Verstrepen *et al.*, 2005; Chevanne *et al.*, 2010). In addition, the mentioned structural constraints of, for example, TR domains acting as protein–protein interaction scaffolds would suggest that selection has a stabilizing effect on the number of TR units. To weigh up the effect of low mutation rates and stabilizing selection on protein TRs, estimates of mutation rates of noncoding TRs can be used as a proxy for TR mutation rates under neutral evolution. Then, results could be compared with estimates of substitution rates for protein TRs of similar length and divergence on the nucleic sequence. As of now, mutation rates have only been estimated for micro- and minisatellites with very low divergence, practically limiting this approach to short protein TRs.

The presented study of TR unit evolution relies on relatively lengthy TR units to provide trustworthy TR unit phylogenies. For these TRs, the evolutionary signal is surprisingly clear, and subject to low false-positive rates, as shown recently (Schaper *et al.*, 2014). However, many plant TRs have short units (Table 1), the majority of which have unit lengths comparable with microsatellites. This wealth of data has not been investigated further here and we can only speculate on the role of TRs with

$l < 15$ aa in protein structure or function. Presumably, they do not form large groups of common TR types, such as many of the TRs in this study, but rather constitute a wealth of rare TR unit motifs. Perhaps, some short TRs enable fast adaptation to changing environments or emerging pathogens? Alternatively, short neutrally evolving TRs may provide an evolutionary buffer for protein innovation which again leads to the generation of new configurations with a fitness advantage (e.g. Wagner, 2008). In these cases, there would be a shift in the mode of protein TR evolution from mostly separated to mostly conserved with increased TR unit length. If such a signal was not found, and short TRs were equally conserved as TRs with $l \geq 15$ aa, it would be clear that noncoding and coding TRs were subject to fundamentally different evolutionary modes.

Practical application of the presented work

The computational approach presented here proposes an efficient means of identification of candidate genes in which TR unit separation or conservation occurs as a result of selective pressures for protein variants with altered properties. Further biological insights may be gathered through specific studies of such genes and TR changes with respect to changes in protein properties (e.g. binding affinity or protein stability) and phenotypic differences (e.g. resistance, tolerance to stress). With progress in protein engineering, our computational predictions of separated and conserved TRs may be used to guide protein design, as has been successfully demonstrated for several types of protein repeat (Jost & Plückthun, 2014). In this context, we believe that our predictions of LRRs, for example, may, in the future, serve to produce synthetically modified species with better pathogen resistance or stress tolerance.

The complete TR annotation data (including very short TRs) from our study are provided online: ftp://ftp.vital-it.ch/papers/vital-it/Phytologist-Schaper/index.html. These data should be used for further work on specific genes or lineages of interest, or to test more general biological hypotheses with respect to evolution of TR, such as the relationships with transposable elements and gene duplications.

Acknowledgements

The authors thank Diana Elena Coman for insightful discussions, and Nives Škunca, Stefan Zoller and three anonymous reviewers for their invaluable feedback on an earlier version of the manuscript. This work was supported by the Swiss National Science Foundation (SNF) grant 31003A_127325/1 to M.A.

References

Adams J, Kelso R, Cooley L. 2000. The kelch repeat superfamily of proteins: propellers of cell function. *Trends in Cell Biology* **10**: 17–24.

Albà MM, Santibáñez-Koref MF, Hancock JM. 1999. Conservation of polyglutamine tract size between mice and humans depends on codon interruption. *Molecular Biology and Evolution* **16**: 1641–1644.

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* **107**: 18724–18728.

Biegert A, Söding J. 2008. *De novo* identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* **24**: 807–814.

Björklund AK, Ekman D, Elofsson A. 2006. Expansion of protein domain repeats. *PLoS Computational Biology* **2**: e114.

Bornberg-Bauer E, Albà MM. 2013. Dynamics and adaptive benefits of modular protein evolution. *Current Opinion in Structural Biology* **23**: 459–466.

Cannon SB, Mitra A, Baumgarten A, Young ND, May G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biology* **4**: 10.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C *et al.* 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics* **43**: 956–963.

Chaw S-M, Chang C-C, Chen H-L, Li W-H. 2004. Dating the monocot–dicot divergence and the origin of core eudicots using whole chloroplast genomes. *Journal of Molecular Evolution* **58**: 424–441.

Chevanne D, Saupe SJ, Clavé C, Paoletti M. 2010. WD-repeat instability and diversification of the *Podospora anserina* hnwd non-self recognition gene family. *BMC Evolutionary Biology* **10**: 134.

Diévart A, Gilbert N, Droc G, Attard A, Gourgues M, Guiderdoni E, Périn C. 2011. Leucine-rich repeat receptor kinases are sporadically distributed in eukaryotic genomes. *BMC Evolutionary Biology* **11**: 367.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**: 48.

Faux NG, Huttley GA, Mahmood K, Webb GI, la Banda de MG, Whisstock JC. 2007. RCPdb: an evolutionary classification and codon usage database for repeat-containing proteins. *Genome Research* **17**: 1118–1127.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S *et al.* 2012. Ensembl 2013. *Nucleic Acids Research* **41**: D48–D55.

Fujii S, Small I. 2011. The evolution of RNA editing and pentatricopeptide repeat genes. *New Phytologist* **191**: 37–47.

Groves MR, Barford D. 1999. Topological characteristics of helical repeat protein. *Current Opinion in Structural Biology* **9**: 383–389.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**: 307–321.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiology* **148**: 993–1003.

Hancock JM, Worthey EA, Santibáñez-Koref MF. 2001. A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. *Molecular Biology and Evolution* **18**: 1014–1023.

Herron MD, Hackett JD, Aylward FO, Michod RE. 2009. Triassic origin and early radiation of multicellular volvocine algae. *Proceedings of the National Academy of Sciences, USA* **106**: 3254–3258.

Hundertmark M, Hincha DK. 2008. LEA (late embryogenesis abundant) proteins and their encoding genes in *Arabidopsis thaliana*. *BMC Genomics* **9**: 118.

Jones JDG, Dangl JL. 2006. The plant immune system. *Nature* **444**: 323–329.

Jorda J, Kajava AV. 2009. T-REKS: identification of tandem repeats in sequences with a K-means based algorithm. *Bioinformatics* **25**: 2632–2638.

Jost C, Plückthun A. 2014. Engineered proteins with desired specificity: DARPins, other alternative scaffolds and bispecific IgGs. *Current Opinion in Structural Biology* **27**: 102–112.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* **9**: 286–298.

Kobe B, Kajava AV. 2001. The leucine-rich repeat as a protein recognition motif. *Current Opinion in Structural Biology* **11**: 725–732.

Leclercq S, Rivals E, Jarne P. 2010. DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. *Genome Biology and Evolution* **2**: 325–335.

Leister D. 2004. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends in Genetics* **20**: 116–122.
Levdansky E, Romano J, Shadkchan Y, Sharon H, Verstrepen KJ, Fink GR, Osherov N. 2007. Coding tandem repeats generate diversity in *Aspergillus fumigatus* genes. *Eukaryotic Cell* **6**: 1380–1391.
Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. 1999. A census of protein repeats. *Journal of Molecular Biology* **293**: 151–160.
Marriage TN, Hudman S, Mort ME, Orive ME, Shaw RG, Kelly JK. 2009. Direct estimation of the mutation rate at dinucleotide microsatellite loci in *Arabidopsis thaliana* (Brassicaceae). *Heredity* **103**: 310–317.
McHale L, Tan X, Koehl P, Michelmore RW. 2006. Plant NBS-LRR proteins: adaptable guards. *Genome Biology* **7**: 212.
Newman AM, Cooper JB. 2007. XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* **8**: 382.
O'Toole N, Hattori M, Andres C, Iida K, Lurin C, Schmitz-Linneweber C, Sugita M, Small I. 2008. On the expansion of the pentatricopeptide repeat gene family in plants. *Molecular Biology and Evolution* **25**: 1120–1128.
Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J *et al.* 2011. The Pfam protein families database. *Nucleic Acids Research* **40**: D290–D301.
Richard G-F, Kerrest A, Dujon B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews: MMBR* **72**: 686–727.
Riegler M, Iturbe-Ormaetxe I, Woolfit M, Miller WJ, O'Neill SL. 2012. Tandem repeat markers as novel diagnostic tools for high resolution fingerprinting of Wolbachia. *BMC Microbiology* **12**: S12.
Schaper E, Gascuel O, Anisimova M. 2014. Deep conservation of human protein tandem repeats within the eukaryotes. *Molecular Biology and Evolution* **31**: 1132–1148.
Schaper E, Kajava AV, Hauser A, Anisimova M. 2012. Repeat or not repeat? – Statistical validation of tandem repeat prediction in genomic sequences. *Nucleic Acids Research* **40**: 10005–10017.
Schlötterer C. 2000. Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**: 365–371.
Shi J, Huang S, Fu D, Yu J, Wang X, Hua W, Liu S, Liu G, Wang H. 2013. Evolutionary dynamics of microsatellite distribution in plants: insight from the comparison of sequenced *Brassica*, *Arabidopsis* and other angiosperm species. *PLoS ONE* **8**: e59988.
Song M, Xu W, Xiang Y, Jia H, Zhang L, Ma Z. 2013. Association of jacalin-related lectins with wheat responses to stresses revealed by transcriptional profiling. *Plant Molecular Biology* **84**: 95–110.
Stirnimann CU, Petsalaki E, Russell RB, Müller CW. 2010. WD40 proteins propel cellular networks. *Trends in Biochemical Sciences* **35**: 565–574.
Symonds VV, Lloyd AM. 2003. An analysis of microsatellite loci in *Arabidopsis thaliana*: mutational dynamics and application. *Genetics* **165**: 1475–1488.
Szklarczyk R, Heringa J. 2004. Tracking repeats using significance and transitivity. *Bioinformatics* **20**: i311–i317.
Tameling WIL, Joosten MHAJ. 2007. The diverse roles of NB-LRR proteins in plants. *Physiological and Molecular Plant Pathology* **71**: 126–134.
Vaughan DA, Ge S, Kaga A, Tomooka N. 2008. Phylogeny and biogeography of the genus *Oryza*. In: Hirano H-Y, Hirai A, Sano Y, Sasaki T, eds. *Biotechnology in agriculture and forestry. Rice biology in the genomics era*. Berlin, Heidelberg, Germany: Springer, 219–234.
Verstrepen KJ, Jansen A, Lewitter F, Fink GR. 2005. Intragenic tandem repeats generate functional variability. *Nature Genetics* **37**: 986–990.
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin RM, Birney E. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research* **19**: 327–335.
Wagner A. 2008. Neutralism and selectionism: a network-based reconciliation. *Nature Reviews Genetics* **9**: 965–974.
Xu C, Min J. 2011. Structure and function of WD40 domain proteins. *Protein & Cell* **2**: 202–214.
Yang S, Li J, Zhang X, Zhang Q, Huang J, Chen J-Q, Hartl DL, Tian D. 2013. Rapidly evolving R genes in diverse grass species confer resistance to rice blast disease. *Proceedings of the National Academy of Sciences, USA* **110**: 18572–18577.
Yang Z. 1998. On the best evolutionary rate for phylogenetic analysis. *Systematic Biology* **47**: 125–133.
Zhao Z, Guo C, Sutharzan S, Li P, Echt CS, Zhang J, Liang C. 2013. Genome-wide analysis of tandem repeats in plants and green algae. *G3 (Bethesda)* **4**: 67–78.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Conserved and separated leucine-rich repeats (LRRs) across the kingdom of plants.

Table S1 Influence of sequence quality on the reconstruction of the mode of evolution of tandem repeats (TRs) for *Arabidopsis thaliana* and *Oryza sativa* Japonica

Notes S1 Testing the robustness of the results to errors in sequence and orthology annotation.

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.