CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Nature Precedings

Molecular evolution of RRM-containing proteins and glycine-rich RNA-binding proteins in plants

Judith Lucia Gomez-Porras, Molecular Cell Physiology, Bielefeld University, Universitätsstr. 25, 33615 Bielefeld, Germany. Present Address: Biophysics and Molecular Plant Biology, University of Potsdam, Karl-Liebknecht-Str, 24-25, Haus 20, 14476 Potsdam-Golm, Germany, email: jgomez@uni-potsdam.de

Martin Lewinski, Molecular Cell Physiology, Bielefeld University, Universitätsstr. 25, 33615 Bielefeld, Germany, email: martin.lewinski@uni-bielefeld.de

Diego Mauricio Riaño-Pachón, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany. Current Address: Computational and Evolutionary Biology Group, Biological Sciences Department, Universidad de los Andes, Cra 1A No. 18A-12, Bogotá D.C., Colombia, email: dm.riano122@uniandes.edu.co

Dorothee Staiger, Molecular Cell Physiology, Bielefeld University, Universitätsstr. 25, 33615 Bielefeld, Germany, email: dorothee.staiger@uni-bielefeld.de

Corresponding authors: Dorothee Staiger and Judith Lucia Gomez-Porras

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Molecular evolution of RRM-containing proteins and glycine-rich RNA-binding proteins in plants

Abstract

Background:

In angiosperms, RNA-binding proteins with an RNA recognition motif (RRM)-type RNA interaction domain play an important role in developmental and environmental responses. Despite their pivotal role, a comprehensive analysis of their number and diversity has only been performed in *Arabidopsis* so far.

Results:

Here we present a detailed phylogenetic analysis of RRM-containing proteins in plants, the red algae *Cyanidioschyzon merolae* and cyanobacteria. We identified two major events during the diversification of the RRM in plants, one at the emergence of green plants, and the other at the water-to-land transition. We focused on proteins that combine a single RRM with a glycine-rich stretch, known as glycine-rich RNA-binding proteins (GRPs). We found that GRPs are present in cyanobacteria, however plant and cyanobacterial GRPs are not of monophyletic origin. We provide evidence that plant GRPs form a polyphyletic group.

Conclusion:

Our work provides insights into the origin of GRPs in plants. We determined that the RRM from plants and cyanobacteria do not have a common origin. We could also determine that the acquisition of the glycine-rich stretch has happened at least on three separate occasions during the evolution of GRPs. One event led to the emergence of cyanobacterial GRPs, while later acquisition events led to the emergence of GRPs in the green lineage. No GRPs were found in red or marine green algae. We found a subgroup of GRPs exclusive to land plants, and its appearance may be linked to challenges related to the water-to-land transition.

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Background

RNA-binding proteins (RBPs) play a crucial role in all aspects of RNA processing including pre-mRNA splicing, polyadenylation, mRNA transport, mRNA stability and translation [1, 2]. RBPs are characterised by the presence of one or more RNA-binding domains often combined with other domains involved in protein–protein interaction, protein targeting, or with zinc fingers that provide the basis for additional mechanisms of interaction with nucleic acids. The so-called RRM is the most abundant RNA binding domain. It is found in about one percent of human genes [3]. The domain is around 80 amino acids long and folds into four β-strands and two α-helices. The surface of the β-sheet is engaged in the RNA interaction. The β_1 strand harbours the conserved hexapeptide RNP2 (ribonucleoprotein consensus sequence 2) and the β_3 strand harbours the highly conserved octapeptide RNP1. Aromatic and basic amino acid side chains within the RNPs are exposed to the surface and are in direct contact with the RNA [1].

In plants, a prevalent class of RRM-containing RBPs is the family of GRPs that combine an N-terminal RRM with a glycine-rich domain of variable length at the C-terminus. The glycine-rich domain has been described as a set of glycine repeats believed to be involved in protein-protein interactions [4]. Mangueon *et al.* (2010) proposed to classify GRP into four subclasses. Glycine-rich proteins with an additional RRM are designated as class IVa GRPs by these authors [5].

GRPs have been found in a wide range of plant species including maize [6], tobacco [7], barley [8], sorghum [9], white mustard [10], Arabidopsis [11], rice [12] and moss [13]. They have also been identified in several cyanobacteria. However, only few have been characterised experimentally [14-16]. In general, the glycine-rich stretch is considerably shorter in cyanobacteria than in plant GRPs [15].

Given the involvement of the RRM in RNA metabolism, GRPs may play important roles in plant physiology and development. Several GRPs respond to a suite of environmental stimuli including cold and wounding [17-19], are implicated in abscisic acid (ABA) signalling [6, 20],

and play a regulatory role in circadian timekeeping [21, 22], flower induction [23] and pathogen defence [24].

Evidence for RNA-binding activity of plant GRPs by and large is based on *in vitro* binding to ribohomopolymers. For *At*GRP7 (*Arabidopsis thaliana* glycine-rich RNA-binding protein 7) and *At*GRP8 RNA substrates have been identified. They bind to their pre-mRNAs [25, 26]. This leads to negative autoregulation via the generation of an alternative splice form that is subjected to Nonsense-mediated decay [27, 28].

Despite the availability of a large number of fully sequenced plant genomes and the prevalence of GRPs in many plant species, a complete genome survey for RRM proteins, including GRPs, has been only performed for the model plant *Arabidopsis thaliana*, based on the first draft of the genome [29, 30]. Recently, a genome survey of RBPs with three RRMs has been compiled for *Arabidopsis*, rice and Poplar [31]. In this study we identified all proteins with one of more RRMs in the fully sequenced genomes of 11 land plant genomes, 7 green algae, 1 red alga and 36 cyanobacteria (Additional file 1). Among the proteins with only one RRM we identified the set of non-redundant RRMs per proteome and species.

We performed a phylogenetic analysis of proteins with a single RRM and studied the distribution of GRPs among the different species. GRPs were found widely in cyanobacteria and plants, but were absent in red and marine green algae. Furthermore, our phylogenetic analysis allows us to conclude that neither the cyanobacterial nor the plant GRPs are monophyletic.

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Results and Discussion

RRM proteins in cyanobacteria, red and green algae and land plants

To identify proteins containing at least one RRM, the program HMMER was used to search for protein sequences with a RRM as defined by the Hidden Markov Model (HMM) of the Pfam domain PF00076. The full conceptual proteomes of 11 plants, 1 red alga, 7 green algae and 36 cyanobacterial genomes were screened (see Additional file 1 for sources). For

most of these species this is the first report of the complete set of RBPs with one or multiple RRMs. Previous reports focused on the set of RRM proteins in Arabidopsis [30], the number of RBPs with three RRMs in poplar, rice and Arabdiopsis [31] and the characterization of a restricted number of RRM proteins in the cyanobacteria *Synechocystis* PCC 6803, *Anabaena variabilis* and *Nostoc* PCC 7120 [14-16, 32].

In cyanobacteria the RRM proteins identified harbour only one RRM domain per protein. Previously, a correlation between the number of RRM genes and the genome size had been pointed out for a limited number of cyanobacteria [15] and attributed to whole genome duplications. According to our data, in cyanobacteria the number of RRM proteins correlates only weakly with the genome size (correlation coefficient, 0.5228) (Figure 1). Most of the species show slightly more RRMs than expected according to their genome size, pointing towards a gain of RRMs by means different to whole genome duplications. In Table 1 we have colour-coded the strains according to the gain/loss of RRMs judged by genome size (column 1). Strains with less RRMs than expected are marked in blue, while strains with more RRMs are marked with red. Strains that show a good correlation between number of RRMs and the genome size are left white. We found that *Acaryochloris marina, Microcystis aeruginosa*, *Nostoc punctiforme*, and *Trichodesmium erythraeum* strain IMS101 have two, three or even four RRM proteins less than expected based on genome size (see Table 1). All these species have very large genomes and show a higher morphological complexity than cyanobacteria with smaller genomes. To fully understand the reason for an apparent loss of RRMs the genome structure must be analysed in more detail, to detect if this loss has come at the cost of expansions in other families.

Notably, nineteen species have one or two RRM proteins more than expected, based on their genome size. Interestingly, most of the species with additional RRM proteins have small genomes (13 out of 19 have a genome size equal or smaller than 2.5 Mb). In the case of *Prochlorococcu*s sp and *Synechococcus* sp. analyses of the genome evolution leads to the assumption that the common ancestor of all *Prochlorococcu*s species and maybe the last common ancestor of *Prochlorococcu*s and *Synechococcus* had a genome size of around 2.4

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Mb [33]. Consequently, we propose that at least in the case of *Prochlorococcu*s species instead of a gene gain we observe a reduction of genome size and that the number of RRM proteins is kept constant around the number present in the last common ancestor.

The other species that show a gain of RRMs are *Synechococcus* sp., *Cyanothece* sp. and *Anabaena* sp. All are diazotrophic cyanobacteria and have larger genomes than non-fixing species. Based on what is known about genome evolution of cyanobacteria [33, 34] we suggest that the additional RRM proteins found in these species are the result of horizontal gene transfer (HGT). Unfortunately, unlike enzymes such as nitrogenases whose acquisition has been clearly determined to be via HGT [34, 35], little is known about the frequency of such events for RRM proteins. We suggest that the acquisition of RRMs in these species may offer some selective evolutionary advantage [36] and that this particular protein domain, although of ancient origin, does not necessarily belong to the core genome, where gene transfer events are very rare.

In red and green algae, we note the appearance of proteins with multiple RRMs (Figure 2). In general, the number of RRMs per protein increases with increasing genome size. In the red alga *C. merolae* and marine green algae 24 to 44% of the RRM proteins have multiple RRMs. We found that the percentage of proteins with multiple RRMs in the symbiotic green alga *Chlorella sp.* NC64A is closer to marine algae (33%) than to the closest relative (*Coccomyxa sp*. C-169). The Genome Project of *Coccomyxa sp*. C-169 was firstly annotated as the Genome from *Chlorella vulgaris*, nevertheless our results support the observations pointing that the genomes from *Coccomyxa sp*. C-169 and *Chlorella sp.* NC64A are evolutionary distant. Remarkably, the freshwater alga *Coccomyxa sp*. C-169 is the species with the largest percentage of proteins with multiple RRMs in the green plant lineage (44%) (Figure 2).

Interestingly, land plants and freshwater green algae (*Volvox carteri*, *Chlamydomonas reinhardtii* and *Coccomyxa sp*. C-169) have a similarly high percentage of proteins with multiple RRMs, ranging from 36 to 44% (Figure 2). Considering that the most recent common ancestor (MRCA) in the chlorophyte lineage is remarkably closer than the MRCA between

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

chlorophyte and embryophyte (land plants) [37] this observation may be associated to an evolutionary convergence related to the habitat.

In mosses and monocotyledonous plants similar numbers of proteins with multiple RRMs were found (below 40%). Sorghum was highest with 41% of RRM proteins having multiple RRMs as well as the protein with the largest number of RRMs, namely seven (see Figure 2). Dicotyledonous plants have more proteins with multiple RRMs than monocotyledonous plants, ranging from 40 to 43 percent. The maximum number of RRMs per protein is five. With only 39% *C. papaya* has slightly less proteins with multiple RRMs than the other dicots.

In the model plant *A. thaliana* we identified 334 proteins corresponding to 227 loci that contained one or more RRMs (Figure 2). This exceeds by 31 the number reported previously by Lorkovic and Barta, who found 196 RRM-containing proteins [30]. It is worth to mention that our screen is based on a more recent annotation of the *Arabidopsis* genome.

Changes in the number of RRMs in plants

In order to account for the fact that cyanobacteria only have proteins with a single RRM domain (sRRM), we restricted further analyses to sRRM domains in plants as well.

We identified a total of 2453 proteins with a sRRM domain in eukaryotes and 136 in cyanobacteria. In the green lineage compare to the red algae *C. merolae* the number of sRRM domains increases dramatically and apparently uncorrelated to the genome size. For instance, *C. merolae* has a genome size of 16 Mb and 16 sRRM domains while green algae with a similar genome size such as the marine algae *O. tauri* or *M. pusilla* (genome sizes 11.5 Mb and 15 Mb respectively) show at least twice the number of sRRM proteins (Figure 2, lower table). This indicates a gain of RRMs in the green lineage.

To base further analysis on non-redundant sequences we identified the number of non-redundant sRRM domains. Identical domains were identified by means of pairwise alignments. The number of sequences was reduced from 2453 sRRM domains to 1898 non-redundant sRRMs. Results are summarized in Table 2.

We observed that in algae most sRRM domains are unique; the only exception is *Chlorella sp.* NC64A with 53 unique sRRM domains and one non-unique domain. Unlike algae, land

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

plants show many identical sRRM domains (see Table 2). Particularly two species show a high redundancy of sRRM domains. In *S. moellendorffii*, 35% of the total sRRM domains are redundant and in *Z. mays* even 62% of the total sRRM domains are redundant. This drastic reduction of non-redundant sequences may reflect specific events of gene duplication within these species.

In *O. sativa* ssp. *indica* most identified sRRM domains are non-redundant domains (142 out of 144) while in *O. sativa* ssp. *japonica* only one-third (82 out of 256) are non-redundant sRRM domains (Table 2). Interestingly, although *O. sativa* ssp. *japonica* has an smaller genome than the ssp. *indica* (389 Mb vs 466 Mb) it shows a larger amount of sRRM domains along with a larger redundancy in sequence.

We have included two types of dicotyledonous plants, herbaceous (*A. thaliana* and *A. lyrata*) and woody plants (grapevine, poplar and papaya). We found that the grapevine proteome drafts used in our study are partially redundant, thus one-fourth of the proteins found in one draft are already described in the other under a different name, but correspond essentially to the same protein. However, if we consider the other two woody plants and compare to the herbaceous plants (*A. thaliana* and *A. lyrata*), we see that while in woody plants almost all sRRM are non-redundant, in both *Arabidopsis* species one-third of the sRRM proteins are redundant (Table 2). Probably, the redundant sRRM domains in both species correspond to closely related genes.

In cyanobacteria we identified 13 identical sRRM domains (see Additional file 2). The largest set of identical sRRM domains in cyanobacteria corresponds to the eight sRRM identified in *Anabaena variabilis* and *Nostoc* PCC 7120 (data not shown). Nonetheless, both are different species. *Anabaena variabilis* has a total of 6914 proteins while *Nostoc* PCC 7120 has 7987 proteins.

To assess the rate of expansion or contraction in the number of sRRMs along the phylogeny, the number of sRRMs in ancestral species was estimated using the software CAFÉ [38]. For the calculations the species tree presented in the Additional file 3 and the number of sRRMs

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

in extant species was used as input. The probability of both birth and death per unit of time (λ) was estimated as 0.0097 by expectation maximization analysis.

In Figure 3 we show the number of non-redundant sRRM domains in extant species and the the calculated number for the MRCAs along the phylogeny. Due to lack of information regarding the divergence times of *Chlorella sp* NC64A (C64A) and *Coccomyxa sp*. C-169 (C169) this two species could not be included in the analysis, but are depicted in the figure. The number of sRRM domains in the MRCA of green algae and land plants was estimated to be 32. Along the different branches of green algae we observed a significant expansion in the number of sRRM domains in almost all organisms. The exceptions are the extant species: *C. reinhardtii, M. pusilla* and *O. tauri*, as well as the MRCA of both *Ostreococcus* species included in this study. Based on these results we propose a first expansion of the number of proteins with a sRRM domains dated at the point of green plant emergence.

Regarding gain/losses in land plants, we see a substantial gain of sRRM domains in the MRCA of embryophytes. While the MRCA of green algae and land plants may have had 32 proteins with a sRRM domain, the MRCA of embryophytes was estimated to have 91 sRRM domains (almost three times the number in the ancestor). These changes can hardly be attributed to correlated changes in the genome size. Although generally the genomes of land plants are larger than algal genomes, within the studied species we have some land plants with genomes as large as those of green algae and twice as many sRRM domains. Such is the case for instance of *A. thaliana* and *V. carterii*, both with a genome size of 120 Mb, *A. thaliana* has 106 sRRM domains while *V. carterii* has 51.

Similarly to estimations made for green algae, in land plants there have been successive gains in the number of sRRM proteins along the phylogeny albeit less pronounced than for the MRCA of embryophytes. Based on this observation we propose a second expansion of the domain at the point of water-to-land transition.

Phylogeny of proteins with single RRMs

We inferred a maximum likelihood tree (ML) in order to understand the phylogenetic relationships between cyanobacterial, algae and plant sRRM domains. We conducted our

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

analysis with 1834 sequences using only the RRM domain in our alignments. Alignments were checked manually and sequences that lack the conserved RNP1 and/or RNP2 motifs of the RRM, or that have large insertions or deletions affecting the alignment were not considered in further analyses. As an example, we show in Additional file 4 ten correctly aligned sequences (upper part of the alignment) and ten sequences with insertion or deletions that disturbed the alignment (lower part of the alignment). We decided to discard 64 sequences, including the ten shown in this figure. The resulting alignment is available upon request. In Table 2 the last column refers to the number of sequences from each species that were kept in the alignment. It becomes evident that *Chlorella* sp. NC64A, *Sorghum bicolor*, *Oryza sativa* ssp. *indica, Vitis vinifera*, and *Populus trichocarpa* are the species with most sequences either lacking any of the conserved motifs or with long insertions or deletions, respectively.

Phylogenetic reconstructions were performed using the software FastTree [39, 40], and bootstraps were performed to assess the statistical significance of the groups. Due to the large number of sequences considered and the small size of the RRM domain the bootstrap values for many branches were low. To identify reliable clades we computed ML trees for different combinations of organisms. We found 81 clades that group the same sequences in independent tree reconstructions, suggesting that the common structure must come from a common evolutionary history. The resulting phylogenetic tree (Figure 4) has been color-coded, the outer ring indicates whether the sRRM domain is from cyanobacteria (blue), red algae (red), green algae (light green), mosses (lime-green), monocots (yellow) or dicots (dark green). The clades are colored and sequentially numbered. Almost all sequences were assigned to a clade (1618 from 1834). Remarkably, all cyanobacteria sequences are grouped together in clades 58 and 59, additionally clade 59 with 90 sequences form the biggest clade. This clear separation between cyanobacterial sRRMs and plants/red alga sRRMs leads to the assumption that that they do not share a common ancestor. In fact, this result further support the results published by Anantharaman *et al* [29] that proposed that the RRM is an eukaryote-specific domain and evolved from an ancient nucleic acid-binding

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

domain. According to the authors, the few RRM domains found in only some bacterial species originated from another kind of nucleic acid-binding protein than the one that gave rise to RRM domains in eukaryotes. These authors also reported the expansion of the RRM domain along with other RNA Binding Domains (RBDs) almost exclusive to eukaryotes. The expansion of RRMs in plants and vertebrates is linked to the advent of alternative splicing as a source of transcriptional diversity.

Regarding the sRRMs observed in algae and plants, we found that in the 79 remaining clades, only a few of all possible combinations of organisms (red alga, green algae, mosses, monocots and dictos) are observed. In Table 3 the different phylogenetic groups observed in each clade are shown. As a further confirmation of an expansion of sRRMs in the green lineage, the most common grouping of organisms involves sequences from **G**reen Algae, **M**osses, **Mo**nocots and **D**icots (GMMoD). Thirty one clades grouping 837 sequences belong to this group. The largest clade other than clade 59 (clade 34) groups 88 sequences and is a group of the MMoD kind. In total only six clades (2, 15, 19, 38, 62 and 69) group sequences from *C. merolae* together with sequences from the green lineage. The groups represented are RGMMoD and RG, R stands for red alga. Few clades feature sequences of only one kind of organism, either green algae, mosses, monocots or dicots.

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Phylogeny of glycine-rich proteins

We focused our attention on a specific subclass of sRRM proteins, the GRPs. Cyanobacterial , plant and metazoan GRPs are mostly studied for their response to diverse stimuli from the environment, especially low temperatures [15, 16, 32, 41]. The analysis of plant GRPs is hampered by the ambiguous nomenclature used in the literature.

In an attempt to incorporate current knowledge in our analysis, we have created a Table with names given by different authors to GRPs from *Arabidopsis*, *Physcomitrella*, rice and cyanobacteria (Table 4).

We located the known GRPs in our phylogeny. We could not find in our phylogeny the sequence GR-RBP1 or At2g16260[30]. We could establish that GR-RBP1 has been reannotated as a pseudogene after the genome version 7 of Arabidopsis (TAIR7).

As expected, all cyanobacterial GRPs belonged to clades 58 or 59 (Table 4). Strikingly, known plant GRPs do not belong to a single clade. Most of the described GRPs belonged to clades 7 or 10. However, two genes from rice were grouped in clades 14 and among the sequences that do not form a reliable clade, between clades 33 and 34, respectively. The sRRM sequences in clades 7, 10, 58, 59 and the corresponding sequences for known GRPs found in other clades where tested for the presence of a glycine-rich stretch at the C-terminus after the sRRM domain (see Materials and Methods). We found that some sequences described as GRPs in the literature actually lack the characteristic glycine-rich stretch. This is the case for instance for *Os*GRP2, *Os*GRP4, *Os*GRP5, *Os*GRP6, *Pp*GRP3 [13, 42] and the cyanobacteria GRPs ORF 339, *RbpB, RbpD* and *RbpG* gene from *A. variabilis* [16] (see Table 4, GR-pattern column).

The presence of know GRPs in clades 7 and 10 leads us to the assumption that known plant GRPs are not of monophyletic origin. This statement is further confirmed by the fact that clade 7 groups sequences from green algae, mosses, monocots and dicots and has been labeled as GMMoD, while clade 10 groups only land plant sequences and is label as MMoD (see Table 3). Taking the organisms represented in clades 7 and 10, one may speculate that sequences in clade 7 diverged first, around the emergence of green plants. For the sequences grouped in clade 10 we speculate they diverged around the emergence of land plants.

For the model *Arabidopsis* sRRMs with a GR-stretch belong to clade 7 (*At*GRPs 2 to 6) (see Figure 7A). The structure of clade 7 is a subtree with an upper branch were the know GRPs *At*GRP3, 5 and 6 and *Os*GRP6 are grouped. Interestingly, for all the sequences in the upper branch of the tree that show a GR-stretch, the stretch is just a part of a longer sequence rich in asparagine. The subtree in the lower branch harbours 11 sequences with a GR-strech

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

(including five characterized GRPs). Contrary to the sequences in the upper branch, the C-terminus of these sequences is short.

Clade 7 groups 4 algae sequences from *O. lucimarinus*, *C reinhardtii*, *M. pusilla* and *V. carteri*. Both algae sequences from freshwater algae are the only that harbour a GR-stretch. The GR-stretch found in *C. reinhardtii* (Cr_184151) and in *V. carteri* (Vc_103546) closest resembles the GR-stretches found in *At*GRP2 and *At*GRP4.

Clade 10 groups only land plant sequences. Likewise clade 7, not all sequences present in clade 10 harbour a GR-stretch. All characterized GRPs are grouped in the lower branch of the subtree (see Figure 5B). Surprisingly, the orthologs of *At*GRP7 and 8 in *A. lyrata* lack the GR-stretch. Furthermore, the known GRPs from monocots are more closely related to sequences from woody dicots (*V. vinifera* and *C. papaya*) than to *A. thaliana* GRPs (see Figure 5B).

For the sequences of clade 10 we check for the presence of other Pfam domains at the C-terminus of the RRM-domain. None of the GRPs show further Pfam domains. Noteworthy, some sequences in the clade that are closely related to GRPs present the fusion of the sRRM domain and the RNA binding domain *z*f-CCHC. We suggest that the sequences grouped in clade 10 have diverged recently. The presence of two almost exclusive eukaryotic domains in this clade and the presence of only land plant sequences leads us to conclude that this clade groups genes whose function has emerged late in eukaryotic evolution, such as alternative splicing.

Conclusion

Our screening for RRM-type RNA-binding proteins in plants and cyanobacteria showed that an expansion of the domain has occurred in the green lineage. A second expansion took place at the point of land plant emergence. We show that the family of proteins called GRPs are not of monophyletic origin. The results shown in our study and by Anantharaman *et al* [29] suggest that the sRRM domain in cyanobacteria has either evolved from a different RNA-binding domain or been acquired not only in cyanobacteria, but also in few other bacteria and archeas probably by horizontal transfer. We found that plant GRPs belonged to

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

two distinct clades. On the one hand in clade 7 we have GRPs from freshwater algae and land plants. On the other hand, in clade 10 we have another subgroup of GRPs grouped together with other proteins that in addition to the sRRM domain harbour other prevalently eukaryotic domain, the zf-CCHC. This domain combination seems to be linked to the advent of alternative splicing. Since GRPs are not of monophyletic origin we propose that the acquision of the GR stretch is an event that occurred after the divergence of the sRRM.

Materials and Methods:

Protein sequences

The conceptual proteomes of 8 algae, 36 cyanobacteria and 11 plants were downloaded from the sources listed in Additional data file 1. A HMM for the Pfam domain RNA-recognition motif (PF00076.15) was used to search against the protein sequences using HMMER 2.3.2. The program 'hmmpfam' was called with the option '-cut_ga' in order to retrieve hits with scores higher than the specified gathering cut-off for the HMM in the PFAM library v23.0 [43]. Among the RRM-containing proteins retrieved, glycine-rich stretches were identified using a regular expression (G[3,5]x[0,6]G[3,5]). The regular expression used search for at least three consecutive glycine residues, at least twice in tandem, separated by no more than six non-glycine residues. Furthermore, the glycine-rich stretch must not overlap the RRM.

Multiple sequence alignment

Multiple sequence alignments were performed using only the RRM domain unless otherwise specified. The domain was extracted from the retrieved sequences taking the start and end as reported by HMMER. Redundant sequences were identified and removed from the final data set via pairwise alignments using the program stretcher from EMBOSS [44].

Sequences were aligned using the program MAFFT v6.6 (http://align.bmr.kyushu-u.ac.jp/mafft/online/server/) [45, 46]. For large datasets of sequences the fast and moderately accurate FFT-NS-2 algorithm was used. For small sets of sequences the slower

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

but more accurate algorithms L-INS-I for domains or G-INS-I for full-length sequences were used. In all cases, the following parameters were used: scoring matrix BLOSUM62, Gap opening penalty 1.53 and Gap extension penalty 0.1. Alignments were checked and optimised manually using Jalview [47]. Poorly aligned sequences, including those lacking the conserved RNP1 and RNP2 motifs and/or including deletion/insertions within were omitted from the final alignments (see example of poorly aligned sequences in Additional data file 4).

We used the program ProtTest (http://darwin.uvigo.es/software/prottest.html) [48] to estimate the empirical model of amino acid substitutions that best describes the evolutionary processes that produced our alignment. The WAG model [49] with parameters +G +F was determined as best fitting according to the Akaike Information Criterion (AIC) [48].

Changes in the number of RRMs in extant species to determine expansions and contractions trough the evolution of RRMs were analysed using the software CAFÉ [38]. The parameter for the birth and death of gene families (λ) was optimized using the expectation.maximization algorithm (EM) and estimated from the data as 0.097 for all analyses. *P*-values were computed using 1000 bootstrap resamplings. Identification of the branch that was the most likely cause of deviations from a random model was determined by Viterbi and Likelihood ratio test procedures [38]. We considered P-values ≤0.01 to be significant.

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Phylogenetic tree reconstruction

Phylogenetic reconstruction was conducted using Maximum Likelihood (ML). Large sets of sequences were analysed using FastTree [39, 40] while small data sets were analysed using the program TREE-PUZZLE v5.2 [50]. The reliability of branches was assessed with 1000 bootstrap resamplings. All sequences and alignments used in this study are available upon request.

Trees were displayed using the programs TreeGraph 2 [51] and FigTree [52].

Authors' contribution

JLGP and ML performed the phylogenetical analyses. DMRP screened the genomes using HMMER and performed some of the phylogenies. JLGP and DS wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Alexander Goesmann for the computational support of the CeBiTec. We are grateful to Jorge E. Mayer for the critical reading of the manuscript and helpful discussions. This work was supported by the DFG (STA 653/2).

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

References

1. Glisovic T, Bachorik JL, Yong J, Dreyfuss G: **RNA-binding proteins and post-transcriptional gene regulation**. *FEBS Lett* 2008, **582:**1977-1986.

2. Lorkovic ZJ: **Role of plant RNA-binding proteins in development, stress response and genome organization**. *Trends Plant Sci* 2009, **14:**229-236.

3. Maris C, Dominguez C, Allain FH: **The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression**. *FEBS J* 2005, **272:**2118-2131.

4. Fusaro AF, Sachetto-Martins G: **Blooming time for plant glycine-rich proteins**. *Plant Signal Behav* 2007, **2:**386-387.

5. Mangeon A, Junqueira RM, Sachetto-Martins G: **Functional diversity of the plant glycine-rich proteins superfamily**. *Plant Signal Behav* 2010, **5:**99-104.

6. Gomez J, Sanchez-Martinez D, Stiefel V, Rigau J, Puigdomenech P, Pages M: **A gene induced by the plant hormone abscisic acid in response to water stress encodes a glycine-rich protein**. *Nature* 1988, **334:**262-264.

7. Hirose T, Sugita M, Sugiura M: **cDNA structure, expression and nucleic acid-binding properties of three RNA-binding proteins in tobacco: occurrence of tissue-specific alternative splicing**. *Nucleic Acids Res* 1993, **21:**3981-3987.

8. Dunn MA, Brown K, Lightowlers R, Hughes MA: **A low-temperature-responsive gene from barley encodes a protein with single-stranded nucleic acid-binding activity which is phosphorylated in vitro**. *Plant Mol Biol* 1996, **30:**947-959.

9. Cretin C, Puigdomenech P: **Glycine-rich RNA-binding proteins from Sorghum vulgare**. *Plant Mol Biol* 1990, **15:**783-785.

10. Heintzen C, Melzer S, Fischer R, Kappeler S, Apel K, Staiger D: **A light- and temperature-entrained circadian clock controls expression of transcripts encoding nuclear proteins with homology to RNA-binding proteins in meristematic tissue**. *Plant J* 1994, **5:**799-813.

11. Macknight R, Love K, Dean C: **Identification of an Arabidopsis cDNA encoding a novel glycine rich RNA-binding protein (Accession no. AJ002892) and mapping of the gene family onto the arabidopsis physical map (98-152)**. *Plant Physiol* 1998, **117:**1525.

12. Macknight R, Lister C, Dean C: **Rice cDNA clones OsGRP1 and OsGRP2 (Accession Nos. AJ002893 and AJ002894, respectively) define two classes of glycine-rich RNA-Binding Proteins (98-153)**. *Plant Physiol* 1998, **117:**1525.

13. Nomata T, Kabeya Y, Sato N: **Cloning and characterization of glycine-rich RNA-binding protein cDNAs in the moss Physcomitrella patens**. *Plant Cell Physiol* 2004, **45:**48-56.

14. Hamano T, Murakami S, Takayama K, Ehira S, Maruyama K, Kawakami H, Morita EH, Hayashi H, Sato N: **Characterization of RNA-binding properties of three types of RNA-binding proteins in Anabaena sp. PPC 7120**. *Cell Mol Biol (Noisy - le-grand)* 2004, **50:**613-624.

15. Maruyama K, Sato N, Ohta N: **Conservation of structure and cold-regulation of RNA-binding proteins in cyanobacteria: probable convergent evolution with eukaryotic glycine-rich RNA-binding proteins**. *Nucleic Acids Res* 1999, **27:**2029-2036.

16. Mulligan ME, Jackman DM, Murphy ST: **Heterocyst-forming filamentous cyanobacteria encode proteins that resemble eukaryotic RNA-binding proteins of the RNP family**. *J Mol Biol* 1994, **235:**1162-1170.

17. Carpenter CD, Kreps JA, Simon AE: **Genes encoding glycine-rich Arabidopsis thaliana proteins with RNA-binding motifs are influenced by cold treatment and an endogenous circadian rhythm**. *Plant Physiol* 1994, **104:**1015-1025.

18. Kim JS, Jung HJ, Lee HJ, Kim KA, Goh CH, Woo Y, Oh SH, Han YS, Kang H: **Glycine-rich RNA-binding protein 7 affects abiotic stress responses by regulating stomata opening and closing in Arabidopsis thaliana**. *Plant J* 2008, **55:**455-466.

19. Sturm A: **A Wound-Inducible Glycine-Rich Protein from Daucus carota with Homology to Single-Stranded Nucleic Acid-Binding Proteins**. *Plant Physiol* 1992, **99:**1689-1692.

20. Cao S, Jiang L, Song S, Jing R, Xu G: **AtGRP7 is involved in the regulation of abscisic acid and stress responses in Arabidopsis**. *Cell Mol Biol Lett* 2006, **11:**526-535.

21. Staiger D: **RNA-binding proteins and circadian rhythms in Arabidopsis thaliana**. *Philos Trans R Soc Lond B Biol Sci* 2001, **356:**1755-1759.

22. Staiger D, Koster T: **Spotlight on post-transcriptional control in the circadian system**. *Cell Mol Life Sci* 2011, **68:**71-83.

23. Streitner C, Danisman S, Wehrle F, Schoning JC, Alfano JR, Staiger D: **The small glycine-rich RNA binding protein AtGRP7 promotes floral transition in Arabidopsis thaliana**. *Plant J* 2008, **56:**239-250.

24. Fu ZQ, Guo M, Jeong BR, Tian F, Elthon TE, Cerny RL, Staiger D, Alfano JR: **A type III effector ADP-ribosylates RNA-binding proteins and quells plant immunity**. *Nature* 2007, **447:**284-288.

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

25. Fuhrmann A, Schoening JC, Anselmetti D, Staiger D, Ros R: **Quantitative analysis of single-molecule RNA-protein interaction**. *Biophys J* 2009, **96:**5030-5039.

26. Schuttpelz M, Schoning JC, Doose S, Neuweiler H, Peters E, Staiger D, Sauer M: **Changes in conformational dynamics of mRNA upon AtGRP7 binding studied by fluorescence correlation spectroscopy**. *J Am Chem Soc* 2008, **130:**9507-9513.

27. Schoning JC, Streitner C, Meyer IM, Gao Y, Staiger D: **Reciprocal regulation of glycine-rich RNA-binding proteins via an interlocked feedback loop coupling alternative splicing to nonsense-mediated decay in Arabidopsis**. *Nucleic Acids Res* 2008, **36:**6977-6987.

28. Staiger D, Zecca L, Wieczorek Kirk DA, Apel K, Eckstein L: **The circadian clock regulated RNA-binding protein AtGRP7 autoregulates its expression by influencing alternative splicing of its own pre-mRNA**. *Plant J* 2003, **33:**361-371.

29. Anantharaman V, Koonin EV, Aravind L: **Comparative genomics and evolution of proteins involved in RNA metabolism**. *Nucleic Acids Res* 2002, **30:**1427-1464.

30. Lorkovic ZJ, Barta A: **Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA-binding proteins from the flowering plant Arabidopsis thaliana**. *Nucleic Acids Res* 2002, **30:**623-635.

31. Peal L, Jambunathan N, Mahalingam R: **Phylogenetic and expression analysis of RNA-binding proteins with triple RNA recognition motifs in plants**. *Mol Cells* 2011, **31:**55-64.

32. Sato N: **A family of cold-regulated RNA-binding protein genes in the cyanobacterium Anabaena variabilis M3**. *Nucleic Acids Res* 1995, **23:**2161-2167.

33. Dufresne A, Garczarek L, Partensky F: **Accelerated evolution associated with genome reduction in a free-living prokaryote**. *Genome Biol* 2005, **6:**R14.

34. Shi T, Falkowski PG: **Genome evolution in cyanobacteria: the stable core and the variable shell**. *Proc Natl Acad Sci U S A* 2008, **105:**2510-2515.

35. Swingley WD, Blankenship RE, Raymond J: **Integrating Markov clustering and molecular phylogenetics to reconstruct the cyanobacterial species tree from conserved protein families**. *Mol Biol Evol* 2008, **25:**643-654.

36. Clery A, Blatter M, Allain FH: **RNA recognition motifs: boring? Not quite**. *Curr Opin Struct Biol* 2008, **18:**290-298.

37. Lewis LA, McCourt RM: **Green algae and the origin of land plants**. *American Journal of Botany* 2004, **91:**1535-1556.

38. De BT, Cristianini N, Demuth JP, Hahn MW: **CAFE: a computational tool for the study of gene family evolution**. *Bioinformatics* 2006, **22:**1269-1271.

39. Price MN, Dehal PS, Arkin AP: **FastTree: computing large minimum evolution trees with profiles instead of a distance matrix**. *Mol Biol Evol* 2009, **26:**1641-1650.

40. Price MN, Dehal PS, Arkin AP: **FastTree 2--approximately maximum-likelihood trees for large alignments**. *PLoS One* 2010, **5:**e9490.

41. De LF, Zhang T, Wauquier C, Huez G, Kruys V, Gueydan C: **The cold-inducible RNA-binding protein migrates from the nucleus to cytoplasmic stress granules**

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

by a methylation-dependent mechanism and acts as a translational repressor. *Exp Cell Res* 2007, **313:**4130-4144.

42. Kim JY, Kim WY, Kwak KJ, Oh SH, Han YS, Kang H: **Glycine-rich RNA-binding proteins are functionally conserved in Arabidopsis thaliana and Oryza sativa during cold adaptation process**. *J Exp Bot* 2010, **61:**2317-2325.

43. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A: **The Pfam protein families database**. *Nucleic Acids Res* 2010, **38:**D211-D222.

44. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16:**276-277.

45. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic Acids Res* 2002, **30:**3059-3066.

46. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program**. *Brief Bioinform* 2008, **9:**286-298.

47. Clamp M, Cuff J, Searle SM, Barton GJ: **The Jalview Java alignment editor**. *Bioinformatics* 2004, **20:**426-427.

48. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution**. *Bioinformatics* 2005, **21:**2104-2105.

49. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach**. *Mol Biol Evol* 2001, **18:**691-699.

50. 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.

51. Stover BC, Muller KF: **TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses**. *BMC Bioinformatics* 2010, **11:**7.

52. **http://tree.bio.ed.ac.uk/software/figtree/.**; 2011.

53. Kim JY, Park SJ, Jang B, Jung CH, Ahn SJ, Goh CH, Cho K, Han O, Kang H: **Functional characterization of a glycine-rich RNA-binding protein 2 in Arabidopsis thaliana under abiotic stress conditions**. *Plant J* 2007, **50:**439-451.

54. Vermel M, Guermann B, Delage L, Grienenberger JM, Marechal-Drouard L, Gualberto JM: **A family of RRM-type RNA-binding proteins specific to plant mitochondria**. *Proc Natl Acad Sci U S A* 2002, **99:**5866-5871.

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

FIGURE LEGENDS

Figure 1. **Correlation between number of RRM proteins and genome size for 36 cyanobacterial species**. The number of RRM proteins exhibits only a weak correlation with genome size (R^2=0.523).

Figure 2. **Number of RRM-containing proteins in plant genomes**. Tabular and graphical representation of the number of RRM-containing proteins in all eukaryotes analysed in our study. In the species tree branch width relates to the number of RRM domains per organism. Number of RRM domains is depicted as percentage right from the organisms. Color codes indicate the number of sRRMs per protein.

Figure 3. **Changes in the number of sRRM domains in green algae and land plants.** Numbers represent sRRMs in extant species and estimates in ancestral species (bold). Significant (*p*-values ≤ 0.01) expansions (+) or contractions (-) in the number of sRRMs are represented in each branch. Data was computed using the software CAFE [38].

Figure 4. **Unrooted ML phylogenetic tree based on the RMM domain of sRRM-containing proteins**. Sequences were aligned using MAFFT [45, 46] tree. The tree was inferred with FastTree [35]. The colored clades are reliable clades 1 to 81. The colored ring corresponds to organisms color code displayed in the lower bar. R: *C. merolae*, G: Green algae, M: mosses, Mo: Mocotos and D: Dicots.

Figure 5. **Details of clades 7 and 10 from the ML phylogenetic tree**. Known GRPs are highlighted in bold and sequences with a GRP stretch are underlined. **A**. Clade 7. **B**. Clade 10. The subtree where known GRPs are grouped is highlighted in green. Proteins with the zf-CCHC domain are marked with a red diamond. Bootstrap values are the result of 1000 BSs.

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Table 1. Genome size and number of RRM proteins in cyanobacteria.

The blue and red squares in the first column denote strains that show less or more RRM proteins than expected according to the genome size.

	Strain	RRM proteins	Genome Size (Mb)	Expected RRM proteins
	Acaryochloris marina	5	6.5	7
	Anabaena variabilis	8	6.4	6
	Cyanothece ATTC 51142	6	4.9	5
	Cyanothece PCC 7424	6	5.9	6
	Cyanothece PCC 7425	6	5.4	5
	Cyanothece PCC 8801	4	4.7	5
	Gloeobacter violaceus	4	4.7	5
	Microcystis aeruginosa	3	5.8	6
	Nostoc PCC 7120	8	6.4	6
	Nostoc punctiforme	5	8.2	8
	Prochlorococcus marinus CCMP1986	3	1.7	2
	Prochlorococcus marinus AS9601	2	1.7	2
	Prochlorococcus marinus CCMP1375	3	1.8	2
	Prochlorococcus marinus MIT 9211	3	1.7	2
	Prochlorococcus marinus MIT 9215	2	1.7	2
	Prochlorococcus marinus MIT 9301	2	1.6	2
	Prochlorococcus marinus MIT 9303	3	2.7	3
	Prochlorococcus marinus MIT 9312	3	1.7	2
	Prochlorococcus marinus MIT 9313	3	2.4	2
	Prochlorococcus marinus MIT 9515	3	1.7	2
	Prochlorococcus marinus NATL1A	3	1.9	2
	Prochlorococcus marinus NATL2A	3	1.8	2
	Synechococcus CC9311	3	2.6	3
	Synechococcus CC9605	4	2.5	3
	Synechococcus CC9902	4	2.2	2
	Synechococcus elongatus PCC 6301	3	2.7	3
	Synechococcus elongatus PCC 7942	3	2.7	3
	Synechococcus JA-2-3Ba NC 007776	4	3.0	3
	Synechococcus JA-3 NC 007775	4	2.9	3
	Synechococcus PCC 7002	3	3.0	3
	Synechococcus RCC307	3	2.2	2
	Synechococcus sp WH8102	4	2.4	2
	Synechococcus WH 7803	3	2.4	2
	Synechocystis PCC 6803	3	3.6	4
	Thermosynechococcus elongatus	3	2.6	3
	Trichodesmium erythraeum IMS101	4	7.8	8

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Table 2. Total number of proteins with an sRRM domain, non-redundant sRRM domains and sRRM domains included in the multiple sequence alignments

Abb.	Species	sRRM domains	Non-redundant sRRM domains	Alignment
Cm	*Cyanidioschyzon merolae*	16	16	15
C64A	*Chlorella* sp. NC64A	54	53	47
C169	*Coccomyxa* sp.C-169	41	41	39
Vc	*Volvox carteri*	53	53	51
Cr	*Chlamydomonas reinhardtii*	48	48	45
Mp299	*Micromonas* sp. RCC299	50	50	47
Mp	*Micromonas pupilla* CCMP 1545	41	41	37
Ot	*Ostreococcus tauri*	33	33	31
Ol	*Ostreococcus lucimarinus*	39	39	39
Pp	*Physcomitrella patens*	110	104	102
Sm	*Selaginella moellendorffii*	147	96	94
Zm	*Zea mays*	477	182	178
Sb	*Sorghum bicolor*	128	125	120
Osi	*Oryza sativa* spp *indica*	144	142	136
Osj	*Oryza sativa* spp *japonica*	256	82	80
Vv	*Vitis vinifera*	213	161	152
Pt	*Populus trichocarpa*	179	171	166
Cp	*Carica papaya*	90	88	86
At	*Arabidopsis thaliana*	193	142	140
Al	*Arabidopsis lyrata*	141	108	106

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Table 3 Phylogenetic groups observed in the clades shown in figure 4. D: Dicots, G: Green Algae, M: Mosses, Mo: Monocots and R: red alage.

Group	Clades
G	44, 64
M	26
Mo	9, 45
D	8, 22, 30
MD	49, 68
MoD	6, 14, 18, 23, 32, 33, 53, 63, 67, 73
RG	19, 38
GMD	65
GMoD	20, 41, 43, 46, 50, 57, 60, 70
MMoD	3, 10, 11, 27, 29, 31, 34, 35, 36, 39, 48, 52, 74
GMMoD	1, 4, 5, 7, 12, 13, 16, 17, 21, 24, 25, 28, 37, 40, 42, 47, 51, 54, 55, 56, 61, 66, 71, 72, 75, 76, 77, 78, 79, 80, 81
RGMMoD	2, 15, 62, 69

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Table 4 Gene numbers and old nomenclature for plant and cyanobacterial GRPs.

Organism	Gene number	Other names	Clade	GR-Pattern	References
At		AtGR-RBP1	Pseudogene	NA	[30]
At	AT4G13850	GR-RBP 2, At-mRBP1a	7	Yes	[30, 53, 54]
At	AT5G61030	GR-RBP3, At-mRBP2a	7	Yes	[30, 54]
At	AT3G23830	GR-RBP 4, At-mRBP1b	7	Yes	[30, 54]
At	AT1G74230	GR-RBP5, At-mRBP2b	7	Yes	[30, 54]
At	AT1G18630	*At*GR-RBP6	7	Yes	[30]
At	AT2G21660	AtGRP7, CCR2, GR-RBP7	10	Yes	[17, 30]
At	AT4G39260	AtGRP8, CCR1	10	Yes	[17, 30]
Os	Os12g43600	OsGRP1	10	Yes	[12]
Os	Os01g68790	OsGRP1	7	Yes	[12, 42]
Os	Os03g56020	OsGRP2	33/34	No	[42]
Os	Os03g46770	OsGRP3	10	Yes	[42]
Os	Os04g33810	OsGRP4	14	No	[42]
Os	Os05g13620	OsGRP5	10	No	[42]
Os	Os12g31800	OsGRP6	7	Yes	[42]
Pp	Phypa1_1_73609	PpGRP1	10	Yes	[13]
Pp	Phypa1_1_16354	PpGRP2	10	Yes	[13]
Pp	Phypa1_1_ 208328	PpGRP3	7	No	[13]
Sb	SbGR-RNP	AF310215	10	Yes	[13]
Zm	GRMZM2G080603	ZmCHEM2	10	Yes	[49]
Zm	GRMZM2G120995	ZmMA16	10	Yes	[6]
Av	YP_320548.1	ORF291; DNA topoisomerase 1; ORF339; ORF97	59	No	[16]
Av	YP_320649.1	RbpF	59	Yes	[16]
Av	YP_321493.1	RbpB	59	No	[32]
Av	YP_322196.1	RbpD	59	No	[32]
Av	YP_322501.1	RbpC	59	Yes	[32]
Av	YP_323803.1	RbpG	58	No	[32]

Additional data files

The following additional data files are available with the online version of this paper:

Additional file 1:

Table S1 – Cyanobacteria, red algae, green algae and plant genomes analysed in Gomez-Porras *et al*. Sources

Additional file 2:

Table S2 - List of RRM-containing proteins in cyanobacteria. Total number of proteins with an sRRM domain and redundant sRRM domains.

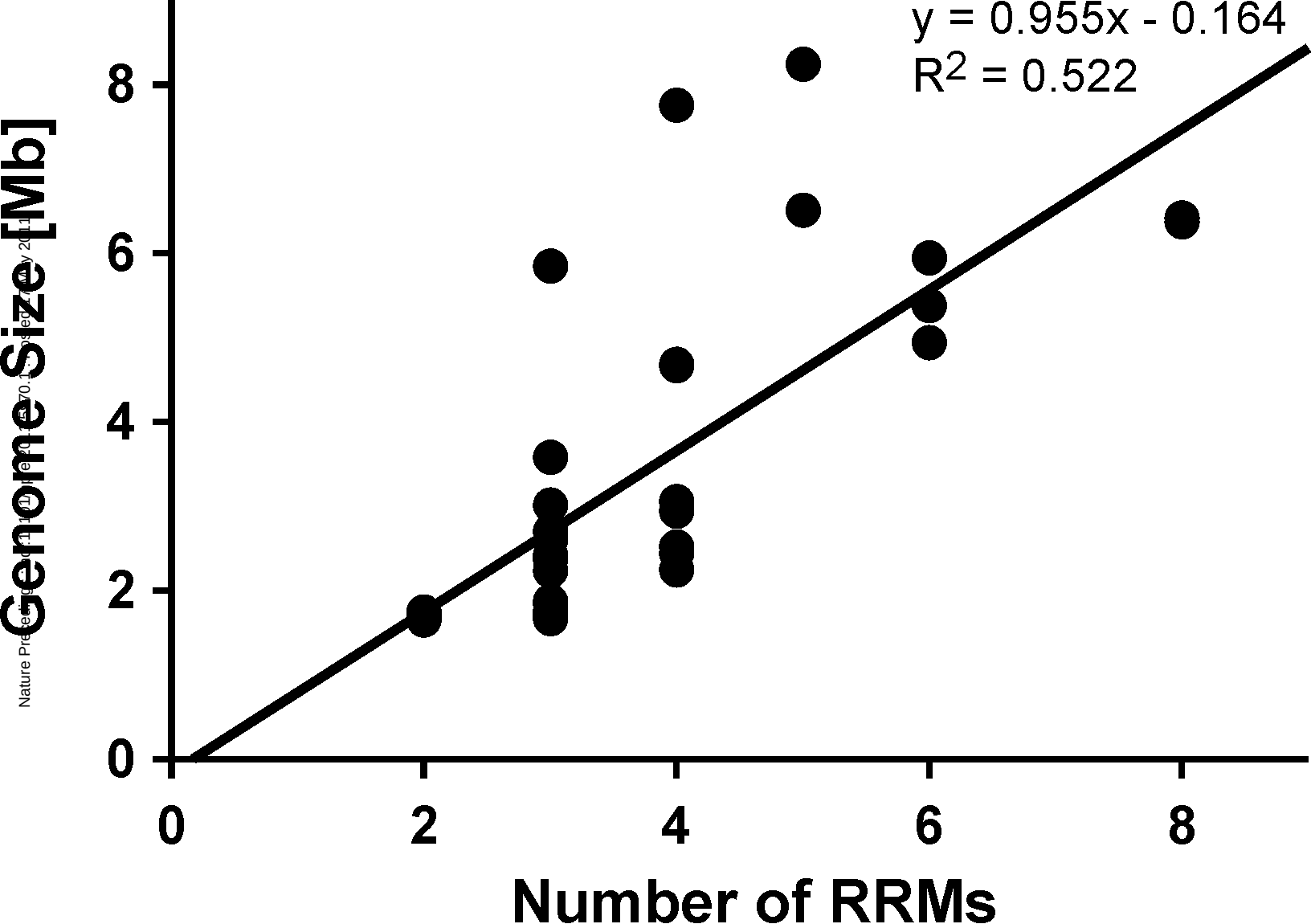
Additional file 3:

Figure S3 – Species tree used for inferences of gain/loss of sRRM domains among the green lineage. Divergence times are shown in million years.

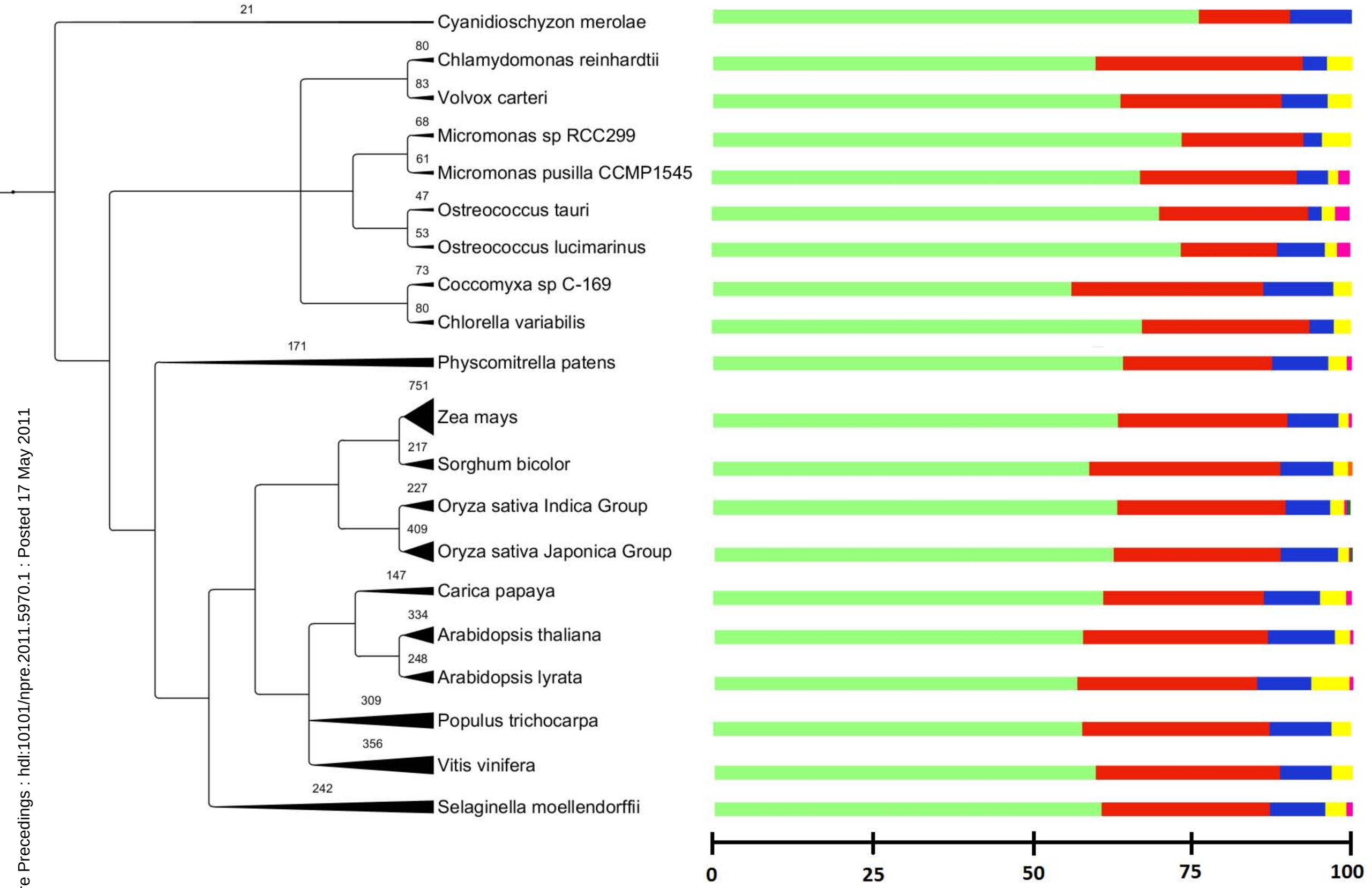
Additional file 4:

Figure S4 – Amino acid sequence alignment of a sub-set of sRRM domains. Top ten sequences show correctly aligned sequences. Lower part of the alignment show some examples of sequences excluded from the phylogenetic analysis due to insertion/deletions in the motifs RNP1 and RNP2.

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011


y = 0.955x - 0.164
R2 = 0.522
Genome Size [Mb]
Number of RRMs
0
2
4
6
8

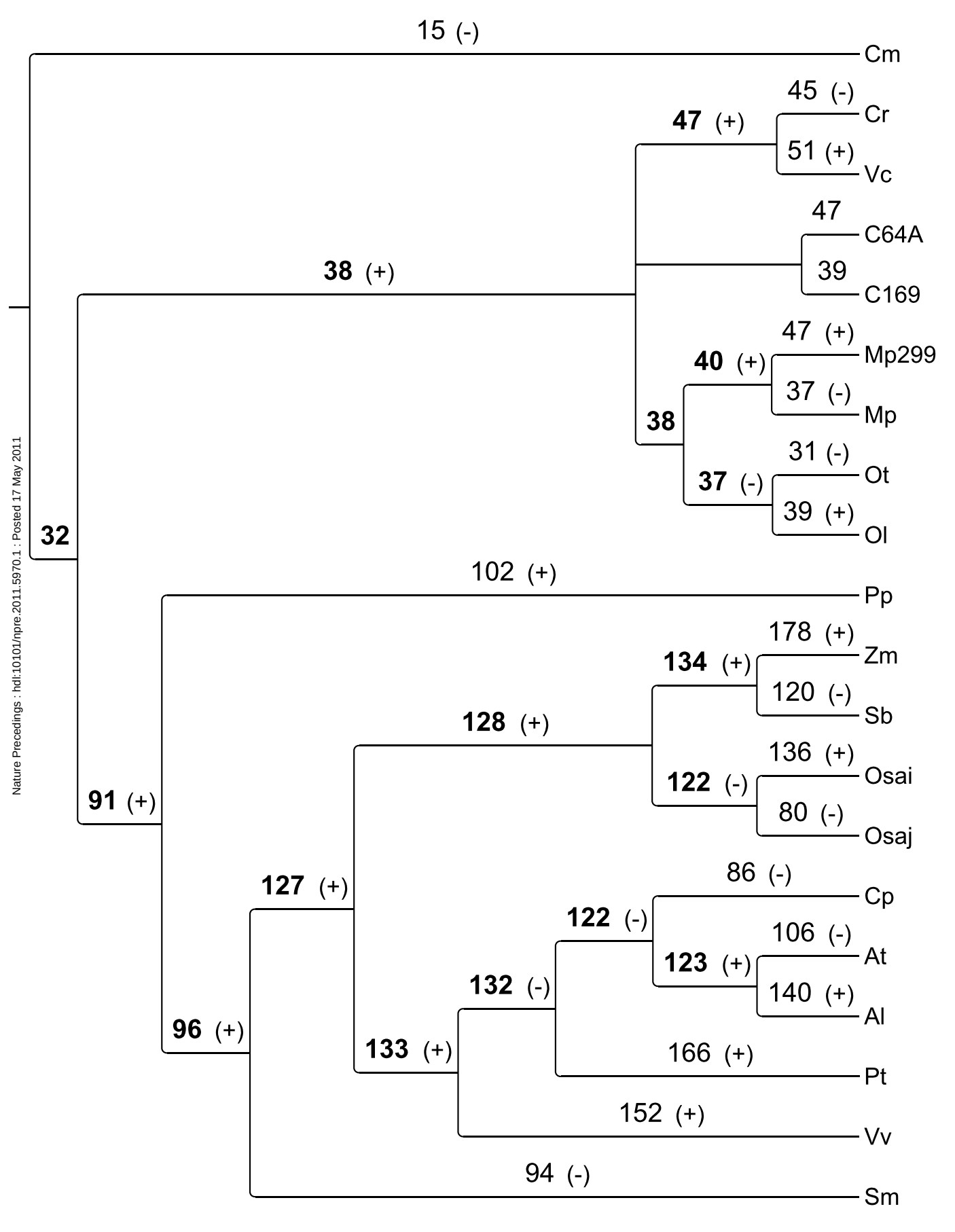
Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011



Number of RRMs per protein

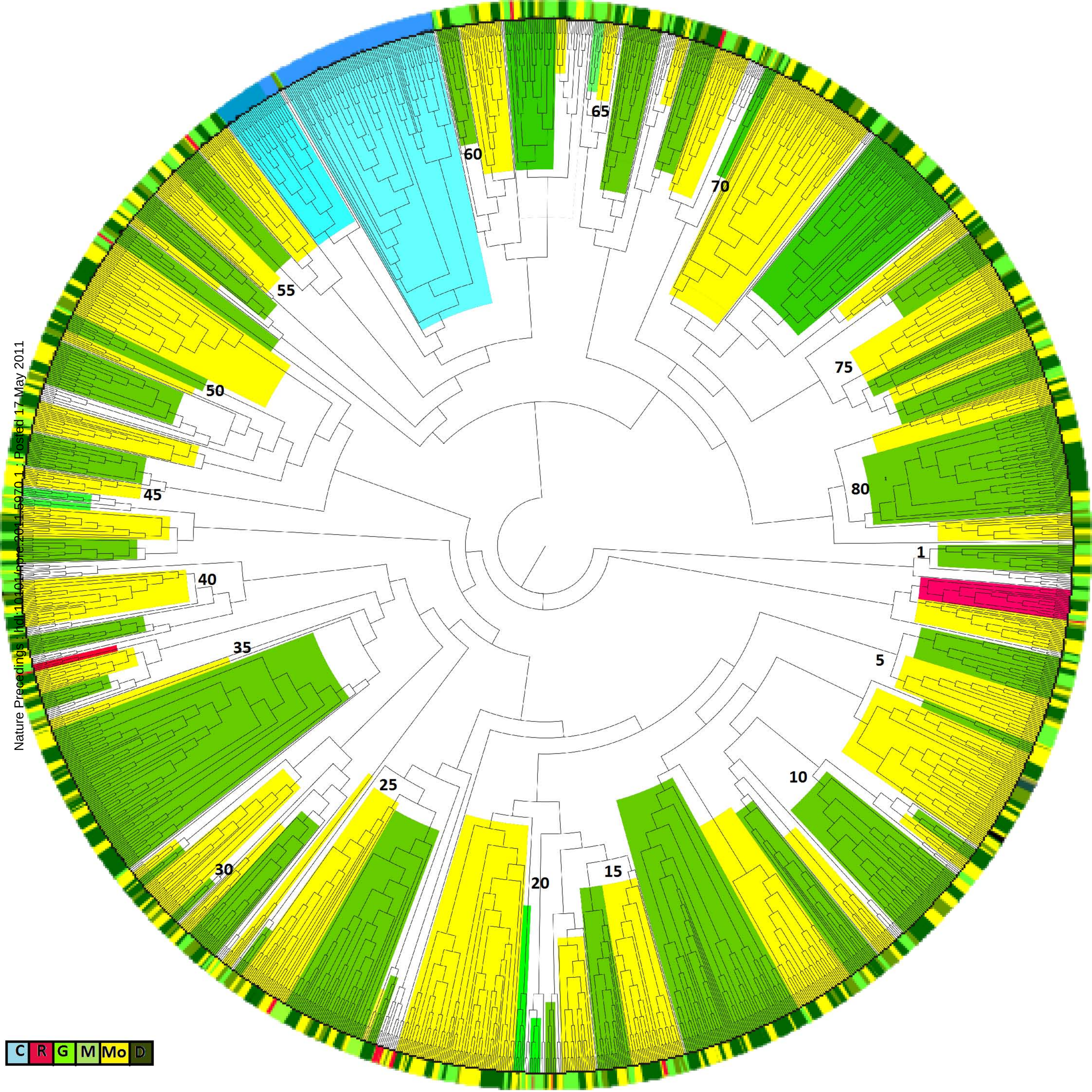
	1	2	3	4	5	6	7
Cm	16	3	2	0	0	0	0
Cr	48	26	3	3	0	0	0
Vc	53	21	6	3	0	0	0
Mp299	50	13	2	3	0	0	0
Mp	41	15	3	1	1	0	0
Ot	33	11	1	1	1	0	0
Ol	39	8	4	1	1	0	0
C169	41	22	8	2	0	0	0
C64A	54	21	3	2	0	0	0
Pp	110	40	15	5	1	0	0

	1	2	3	4	5	6	7
Zm	477	200	60	12	2	0	0
Sb	128	65	18	5	0	0	1
Osi	144	60	16	5	1	1	0
Osj	256	107	37	7	1	1	0
Cp	90	37	13	6	1	0	0
At	193	97	35	8	1	0	0
Al	141	70	21	15	1	0	0
Pt	179	91	30	9	0	0	0
Vv	213	103	29	11	0	0	0
Sm	147	64	21	8	2	0	0



Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011

Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011



Nature Precedings : hdl:10101/npre.2011.5970.1 : Posted 17 May 2011