Immune receptors with exogenous domain fusions form evolutionary hotspots in grass genomes

Paul C. Bailey[1], Gulay Dagdas[2], Erin Baggs[1], Wilfried Haerty[1] and Ksenia V. Krasileva[1,2,*]

[1] Earlham Institute, Norwich Research Park, Norwich, NR4 7UG, UK

[2] The Sainsbury Laboratory, Norwich Research Park, Norwich, NR4 7UH, UK

* Corresponding author: ksenia.krasileva@earlham.ac.uk

Understanding evolution of plant immunity is necessary to inform rational approaches for genetic control of plant diseases. The plant immune system is innate, encoded in the germline, yet plants are capable of recognizing diverse rapidly evolving pathogens. Plant immune receptors (NLRs) can gain pathogen recognition through point mutation, recombination of recognition domains with other receptors, and through acquisition of novel 'integrated' protein domains. The exact molecular pathways that shape immune repertoire including new domain integration remain unknown. Here, we describe a non-uniform distribution of integrated domains among NLR subfamilies in grasses and identify genomic hotspots that demonstrate rapid expansion of NLR gene fusions. We show that just one clade in the *Poaceae* is responsible for the majority of unique integration events. Based on these observations we propose a model for the expansion of integrated domain repertoires that involves a flexible NLR 'acceptor' that is capable of fusion to diverse domains derived across the genome. The identification of a subclass of NLRs that is naturally adapted to new domain integration can inform biotechnological approaches for generating synthetic receptors with novel pathogen 'traps'.

INTRODUCTION

Plants have powerful defence mechanisms, which rely on an arsenal of plant immune receptors (Jones, Vance and Dangl, 2016; Dodds and Rathjen, 2010). The Nucleotide Binding Leucine Rich Repeat (NLR) proteins represent one of the major classes of plant immune receptors. Plant NLRs are modular proteins characterized by a common NB-ARC domain similar to the NACHT domain in mammalian immune receptor proteins (Jones, Vance and Dangl, 2016). On the population level, NLRs provide plants with enough diversity to keep up with rapidly evolving pathogens (Hall et al., 2009; Joshi et al., 2013). With over 50 fully sequenced plant genomes today, it is

timely to apply comparative genomics approaches to investigate common trends in NLR evolution across the plant kingdom, including key crop species.

In contrast to the highly conserved NB-ARC domains, the Leucine Rich Repeats (LRRs) of NLRs show high variability (Noel et al., 1999; Jacob, Vernaldi and Maekawa, 2013). The functional consequence of high LRR variation is thought to be the generation of novel recognition specificities (Bakker et al., 2006; Sukarta, Slootweg and Goverse, 2016). In addition, recent findings show that novel pathogen recognition specificities can also be acquired through the fusion of non-canonical domains to NLRs (Le Roux et al., 2015; Kroj et al., 2016). These exogenous domains can serve as 'baits' mimicking host targets of pathogen-derived effector molecules and therefore act in concert with LRR variation to broaden the spectra of recognised pathogen-derived effectors (Cesari, Bernet al., 2014a; Cesari et al., 2014b; Le Roux et al., 2015).

NLRs plant immune receptors were discovered over 20 years ago through cloning of plant disease resistance genes in Arabidopsis (Mindrinos et al., 1994; Bent et al., 1994). Sequencing of the Arabidopsis genome allowed annotation of the NLR repertoire based on a genome-wide scan for the conserved NB-ARC domain that subsequently revealed common and non-canonical NLR architectures. Application of this method to newly sequenced plant genomes has revealed common principles in NLR composition. Additionally, genome scans have contributed to our understanding of the genome-wide architecture of NLRs, including a tendency for NLRs to form major resistance clusters (Christopoulou et al., 2015; Christie et al., 2016). The relatively poor quality of assembled genome sequence in repetitive regions has hampered accurate identification and annotation of NLR genes, which are present at high copy number in the genome and also encode repetitive LRR domains. To overcome this problem, a method called resistance gene enrichment sequencing was developed (Jupe et al., 2013; Witek et al., 2016; Andolfo et al., 2014); it involves enrichment of NLRs from genomic or transcribed DNA and enables their accurate assembly. The identification of NLRs across plant genomes using uniform computational methods, such as scanning genomes with Hidden Markov Models (HMMs) for the NB-ARC domain, has allowed the NLR repertoire to be compared across species (Sarris et al., 2016; Kroj et al., 2016; Yue et al., 2016). This has led to identification of plant families with a significantly expanded or reduced number of NLRs (Sarris et al., 2016; Kroj et al., 2016; Zhang et al., 2016) and the identification of co-evolutionary links between NLR diversification and their regulation by miRNAs (Zhang et al., 2016). Comparative genomics analyses also revealed that formation of NLRs with non-canonical architectures is common across flowering plants (Sarris et al., 2016; Kroj et al., 2016).

The NLR copy number variation identified in genomic and RenSeq scans of different plant genomes has been attributed to the birth and death process of gene evolution (Michelmore, Meyers and Young, 1998). The mechanisms by which new NLR genes are created and upon which selection can act remains elusive. The prevailing consensus holds that NLR diversity is likely to be generated through a variety of mechanisms including duplication, unequal crossing over, non-homologous (ectopic) recombination, gene conversion and transposable elements (Jacob, Vernaldi and Maekawa, 2013). Identification of the selection pressures acting on

the NLR gene family has also proved to be a challenging question to answer with often only subtle or divergent selection pressure signatures identified for individual NLRs. This has led to the conclusion that NLRs are generally under purifying and neutral selection (Bakker et al., 2006).

The most recent paradigm in NLR diversification involves fusion to exogenous protein domains, also called integrated domains (NLR-ID), a mechanism deployed across flowering plants (Sarris et al 2016, Kroj et al 2016). The availability of sequenced genomes now allows the evolution and diversification of NLRs with integrated domains to be addressed, including the following questions. First, are NLR-IDs distributed uniformly across different subclasses of NLRs or are there specialized clades that are more prone to exogenous domain integration? Previous blast analyses of known NLR genes, such as RGA5 and Sr33 hinted at diversity of integrated domains fused to their homologs, however, no evolutionary links between these genes have been established (Cesari et al., 2014a; Periyannan et al., 2013). Second, is NLR-ID diversification associated with particular genomic locations and if so are these locations syntenic across species? The answer to this question might shed a light to the mechanisms of how NLR-IDs are formed. Third, previous functional analyses of two NLR-ID genes demonstrated that they require activation partners that are co-located in same genomic locus and even share the same regulatory sequence being expressed from the opposite strand (Okuyama et al., 2011; Le Roux et al., 2015; Zhai et al., 2014; Sarris et al., 2016). It is still not clear whether such requirement for a paired NLR is a rule for all NLR-IDs (Sarris et al., 2016) and if so, how diversification of NLR-IDs would be coupled to the diversity of the pair. While the first question about adaptability of NLRs to new gene fusions can be addressed by studying evolutionary history of NLR-IDs themselves, definitive answers to the second question might require near-complete genomes with high continuity as well as population genetics analyses.

The grasses (*Poaceae*) represents a highly successful family of flowering plants that originated as early as 120 million years ago (Prasad et al., 2005; Prasad et al., 2011) and rapidly colonized diverse environments, becoming the most abundant plant family on Earth. Among grasses are three major cereals that form the basis of modern day agriculture and human diet: maize (*Zea mays*), rice (*Oryza sativa*) and wheat (*Triticum* species). It has been suggested that high genomic plasticity of grasses contributed to their adaptability and success in agriculture (Dubcovsky and Dvorak, 2007). The genomes of grasses range in size, ploidy and chromosome number from 270 Mb genome of *Brachypodium distachyon* and 400 Mb genome of rice (*O. sativa*) to 17 Gb hexaploid genome of bread wheat (*T. aestivum*). With genome expansion, transposable elements proliferated from comprising 21% of the Brachypodium genome to over 80% of the wheat genome and play a major role in genome evolution (Vogel et al., 2010; Choulet et al., 2010; Wicker et al., 2016). Genomes of the *Poaceae* have also undergone major re-arrangements through chromosome fusions, duplications and translocations, leading to divergent chromosome numbers, yet maintaining long syntenic blocks which allow the identification of common and divergent genome regions (Vogel et al., 2010; Salse et al., 2008). Through both global and local rearrangements, the genomes of grasses acquired diverse variation in gene copy numbers, including high

copy number of NLRs (Sarris et al., 2016; Zhang et al., 2016), which makes *Poaceae* an attractive system to study NLR evolution.

In this paper, we have examined the evolutionary dynamics of NLR-IDs in the genomes of nine grass species, their distribution within the NLR phylogeny and the diversity of their integrated domains within and across species. We identified several "hotspot" clades and were able to define one ancient monophyletic clade of NLRs that is highly amenable to new domain integrations, in which most diversity was attained in the *Triticeae* species. This clade is present at syntenic locations in grasses following the evolutionary history of genomes as well as local species-specific chromosomal translocation events. We also observed that it is absent in maize, likely as a consequence of overall contraction of NLRs in this species. The identification of this NLR-ID hotspot can form the basis for new biotechnological approaches for designing NLR receptors with synthetic fusions to new pathogen traps.

RESULTS AND DISCUSSION

NLR-IDs are distributed non-uniformly across NLR Protein subgroups

We examined the evolution of NLRs across nine grass species with available genomes - *Setaria italica*, *Sorghum bicolor*, *Zea mays* (maize), *Brachypodium distachyon, Oryza sativa* (rice), *ordeum vulgare* (barley), *Aegilops tauschii, Triticum urartu* and *Triticum aestivum* (hexaploid bread wheat). The phylogeny of 4,130 NLRs from these species, based on the common NB-ARC domain, showed that proteins within a few clades are highly prone to domain integrations compared to other clades, although NLR-ID formation is not exclusive to these clades (Figure 1A). One hotspot clade was particularly enriched in NLR-ID proteins (59 % are NLR-IDs) compared to 8% of proteins with NLR-IDs across all clades (Figure 1A, hotspot 1, highlighted in red). This clade was found to be nested within an outer clade (Figure 1A, highlighted in blue) with only 0 to 14 % of proteins containing NLR-IDs. These two clades include proteins representative of all the studied grass species with the exception of *Z. mays* (Figure 1E). Therefore, we predict that this hotspot clade originated before the split of Panicodae, Ehrhartoidae and Pooidae (BEP and PACCMAD clades) from the rest of the Poaceae 60 MYA (Vogel et al., 2010). Supporting our hypothesis, an outer ancestral clade was apparent (Figure 1A, highlighted in cyan) that contained proteins from all the grass species present in the tree, including *Z. mays*, although the bootstrap support value leading to this clade was only 85 %. Separate NLR phylogenies for each of the grass species showed that the pattern of integrated domain hotspots was strongest in *Brachypodium* and in the *Triticeae* species (Figure 2; Supplemental Figures 1 – 9). It is also clear that NLR(-ID) protein duplication has proliferated most strongly in these species for this hotspot clade (Figure 1E). However, the relative ratio of NLRs with and without extra domains in this clade has remained relatively constant at around 59% suggesting that the rate of domain recycling has been constant across these species (Figure 1B; Supplemental Table1).

Two other major NLR-ID hotspots were investigated (Figure 1, hotspot 2 and 3). Hotspot 2 contains 81 proteins from the grass species present in the tree, except for *O. sativa* and *H. vulgare*. It is located in an inner clade - 38% are NLR-IDs - nested within an outer clade but no ancestral clade was apparent. For hotspot

3, 42 % of proteins (46 out of 109) are NLR-IDs and there is no outer clade or ancestral clade detectable. In each of these hotspots, one integrated domain dominates, a DDE superfamily endonuclease domain (hotspot 2) and a BED-type zinc finger domain (hotspot 3) in contrast to hotspot 1 which contains 34 different domains.

Expansion of the NLR-ID hotspot 1 clade is linked to diversification through new gene fusions

The NLR-ID proteins from evolutionary hotspot 1 were examined further to test the hypothesis that the increase in the number of NLRs with integrated domains was due to the creation of novel gene fusions rather than the duplication of existing ID fusions. A clear expansion of the ID domain repertoire was found for this group of proteins, particularly for the *Triticeae* species (Table 1). It is possible that differences in the observed repertoires can be explained partly by incomplete annotation of genomes or fragmented assembly of NLRs; such proteins were omitted from the phylogenetic analysis if they were < 70 % complete across the NB-ARC domain. However, the overall trend across the genomes strongly suggests that differences cannot be explained solely by differences in genome assemblies. Moreover, genomes such as *B. distachyon*, *Z. mays* and *O. sativa* are assembled to much higher quality than those of the *Triticeae* species, yet they contain fewer NLR-IDs and have lower ID diversity. This fact suggests that the trend we observed is not only biologically relevant, but could in reality be even more pronounced when complete *Triticeae* genomes become available.

To further understand the evolution of ID fusions, the section of the tree in Figure 1 for hotspot 1 and the associated outer and ancestral clades were re-aligned and analyzed by maximum likelihood phylogeny (Figure 3A; Supplemental Dataset 4). Each gene was annotated with a cartoon showing both canonical and non-canonical domains. We observed examples in the *Triticeae* in which neighbouring proteins in the tree - clustering with high bootstrap support - share the same domain at the same position indicating common ancestry and suggesting selection to maintain a functional fusion. We were only able to find such evidence of conservation between proteins from the *Triticeae* and *Brachypodium*.

In contrast to these observations of gene architecture conservation across species, throughout the tree, we also observed orthologs and closely related paralogs with distinct domain fusions. This observation suggests that this subfamily of NLR protein has a high ability to form independent fusions with a wide variety of domains, unlike other types of NLR protein in the rest of the NB-ARC protein family.

The clades highlighted in figure 3A differ most strikingly for the position of the integrated domain within the protein. The majority of proteins in the outer clade have integrated domains at the N-terminal end. These integrated domains belong to the same Pfam family which suggests a single integration event followed by gene duplication and secondary losses (Figure 3B) such as observed for hotspots 2 and 3. In contrast, the proteins from the inner clade have IDs integrated primarily at their C-terminal end and are much more diverse (Figure 3C). Most of the NLR-ID diversity was observed in *T. aestivum* with 68 proteins in this clade including at least one domain that was representative of 21 non-redundant IDs. The number of different ID domains for proteins in the individual *T. aestivum*

subgenomes is higher and more diverse than for two of the diploid progenitors, *T. urartu* (A genome) and A. tauschii (D genome), indicating that new domains have continued to be integrated *de novo* into *T. aestivum* proteins following the divergence of *T. aestivum* from these progenitors (Supplemental Dataset 1 and 2).

Genomic locations involved in proliferation and diversification of NLR-IDs

We observed that NLRs from the hotspot clade were found on different chromosomes across and within species. For five species analyzed in this study, the chromosomal location of NLR-IDs was available from the genome annotation. We looked to see whether there was any enrichment of NLR-IDs from the hotspot clade on any particular chromosome and investigated whether these inter-species differences could be explained by whole-genome rearrangement during evolution (Table 2). Indeed, in most species NLR-IDs from hotspot clade were concentrated only on 1-2 chromosomes, such as chromosomes 2 and 5 in *S. bicolor*, chromosome 11 in O. sativa, chromosome 4 in *B. distachyon* which form known syntenic blocks (Vogel et al, 2010) indicating an ancient origin of the locus that was present in the common ancestor of grasses. With the proliferation of NLR-ID hotspot 1 in wheat, we also observed more divergent locations of NLR-IDs, mostly concentrated on chromosomes 7AS, 7DS, 4AL, but also on chromosomes 1, 3 and 6 (Table 2, Supplemental Table 3). Such proliferation can be explained by more recent large scale genomic rearrangements, such as translocation of chromosomal region from 7BS to 4AL and other known chromosomal translocation and duplications (Salse et al., 2008; Clavijo et al., 2016). This indicates that proliferation of NLR-IDs in *Triticeae* might be linked to greater plasticity of its genome. Since some of the larger translocations in wheat occurred after the formation of NLR-ID hotspot 1, it is also possible that the interaction across members of NLR-ID locus contribute to larger genomic rearrangement events.

When we examined orthologous NLRs located on different wheat sub-genomes, we identified rapid local proliferation of domain fusions (Figure 4). In some instances, orthologous copies were subjected to simple domain loss, such as the sub-clade with the Kelch domain, while others exhibited domain swap, such as the sub-clade harbouring NPR1/AP2/Myb_DNA_Binding domains (Figure 4). This indicates that active and very rapid gene rearrangement continues to take place in the local genomic context of NLR-ID hotspot 1.

Possible mechanisms driving NLR-ID diversification

Any mechanism that creates gene fusions requires a move or a copy and paste event of an exogenous gene from one location to another. Since NLRs from the hotspot clade are mostly found at syntenic locations, yet harbour diverse fusions, it is most likely that these NLRs act as hotspot ‘acceptors’ for exogenous genes to create NLR-IDs rather than move themselves. We observed that the overall number of NLRs in the hotspot increases proportionally to the total increase of NLRs in the genome. Therefore, we hypothesize that duplication of NLRs at hotspots create more ‘acceptor’ sites which results in greater NLR-ID diversity. What makes these particular NLRs more amenable to new gene fusions compared to other NLRs remains unclear.

How can exogenous domains become fused to NLRs? Wicker, Buchmann and Keller (2010) formulated three

main models for gene movement/duplication into non-homologous locations and observed all of them taking place in cereal genomes. The first model involves transposable elements (TEs) acting alone that either excise and transpose genes from one location to another (DNA transposons) or copy and paste them via RNA intermediate (retrotransposons). In the latter case, the gene at the new site does not contain introns. The second model relies on the endogenous host machinery alone that repairs double-stranded DNA breaks using non-homologous exogenous DNA template. The third and most prevalent process in grasses combines the activity of TEs and DNA repair. In this model, TEs insert in new locations themselves or act on common repeats to induce double stranded breaks without bringing in any exogenous genes, then these double stranded breaks are repaired with endogenous machinery using non-homologous DNA fragments. In all cases, gene movement has potential to create new gene fusions, such as NLR-IDs.

In order to distinguish among these mechanisms, we extracted coding DNA sequence of integrated domains for 40 *T. aestivum* genes from the hotspot clade and aligned them back to the genome (blastn, e-value 1e-3). Similar to the NLR portion of the genes, most of the integrated domain contained introns, which indicates that they were not acquired by retrotransposition. Similarly, integrated domains that were acquired recently in cereals and have been validated previously (Sarris et al., 2016), such as NPR1 and Exo70, contained an unintegrated single copy (one in sub-genome) paralogs elsewhere in the genome, providing evidence against gene movement through DNA transposition. Therefore, we hypothesize that the integration of exogenous domains would follow ‘copy-and-paste’ mechanisms observed previously in cereals (Wicker, Buchmann and Keller, 2010). Such mechanisms involve double-stranded DNA breaks and repair with non-homologous template. Whether this process is driven by endogenous plant machinery alone or triggered by the movement of transposable elements remains unclear. Expansion of the hotspot clade in *Triticeae*, its proliferation to multiple genomic locations as well as increased diversity of integrated domains might be linked to the overall increased fraction of TEs in these genomes compared to smaller genomes of other grasses.

Evolutionary model of NLR-ID hotspot formation and proliferation

The processes underpinning genome evolution include domain duplication, fission and fusion (Moore et al., 2008) which have recently been implicated in NLR evolution (Zhong and Cheng, 2016; Kroj et al., 2016; Sarris et al., 2016). Our model (Figure 5) summarizes how these processes could have driven the expansion and diversification during evolution of the proteins in NLR-ID hotspot 1. The model in Figure 5 can be used to illustrate how a subset of NLR-ID diversity (Figure 4) may have been generated. Subsequent to the hotspot clades ancestral genes’ acquisition of high ability to form fusions, diversification occurs resulting in the large variety of exogenous integrated domains. The NLR-IDs in the upper clade of figure 4 contains only fusions to the Kelch domain and could therefore be represented by the models ID1 domain, as it appears to be a conserved fusion despite several duplications. The absence of the Kelch domain in NLRs (TRIAE AA2059910.1) nested within clades where all other members have the domain, suggests the absence of the domain is likely the result of fission events. Conversely, a diversity of exogenous

integrated domains can be seen in the lower clade of figure 4 - this suggests an NLR with evolutionary history similar to the models ID2. Thus, the B3 NLR fusion could be analogous to ID2 with the NLR-ID having undergone successive duplications resulting in both domain preservation and domain swap where ID3 could hypothetically be a Myb_DNA_Binding domain.

CONCLUSIONS

The source of plants' ability to rapidly acquire new pathogen recognition specificities remains a key question in plant immunity. Its answer is closely linked to the evolution of plant immune receptors and their diversification. Recently NLR diversification has been shown to involve acquisition of new exogenous domains, likely through gene fusion events. These integrated domains can be baits for the pathogens and resemble their original targets in the host, thus rapidly expanding recognition potential of plant NLRs. Here we found that formation of NLR-IDs is not random in NLR evolution and we observed a clear hotspot of NLR-ID proliferation and diversification in grasses, particularly in the *Triticeae*. This hotspot is of ancient origin and is present in all surveyed grasses except for maize. Such proliferation of NLR-IDs involves new diverse domain integrations. Genomic locations of NLR-IDs from the hotspot clade indicate that it evolved very early near the origin of grasses (or before) and expanded alongside whole genome evolution of these species. In the *Triticeae,* it shows more rapid movement across the genome as well as rapid local rearrangements. Although the exact mechanisms of NLR-ID formation remain to be uncovered, we predict that it involves double-stranded DNA breaks and could be driven either by endogenous machinery such as non-homologous (ectopic) recombination which has previously been shown to drive evolution of Mla locus in barley (Leister et al., 1998), or alternatively by local activity of transposable elements and endogenous DNA repair machinery as has been previously documented for other types of gene duplications in cereals (Wicker, Buchmann and Keller, 2010).

In the future, the availability of higher quality genome assemblies as well as multiple genomes for each species will allow more detailed analyses of syntenic gene clusters and will identify precise location of DNA breakpoints that lead to NLR-ID formation. Combining long molecule sequencing RenSeq (Giolai et al., 2016) with population genetic analyses will allow us to estimate how rapidly new gene fusions are formed within populations and how fast selection of advantageous combinations occur in nature.

Modern plant breeding practices have dramatically reduced the genetic variability of crops. Isolation and utilization of major race-specific resistance genes has been one of the major genetic methods of disease management. However, these approaches are hampered by the appearance of pathogen races that are overcoming resistance. Reliance on synthetic fungicides for pathogen control has been successful but it imposes heavy economic costs and has suffered from same consequences as over reliance on antibiotics in medicine, leading to selection of highly virulent drug-resistant pathogens. Moreover, fungicides adversely affect human health and the environment, which has resulted in stringent rules on pesticide use and banning of chemicals deemed harmful to human health (Ilbery et

al., 2013). Furthermore, future bans are planned to reduce environmental damage. Consequently, it is predicted there will be a significant yield reduction due to pathogen pandemics. There is an urgent need for new genetic sources of resistance for future sustainable crop production (Dangl, Horvath and Staskawicz, 2013; Ellis et al., 2014). Our identification of NLRs that are highly amenable to integration of exogenous domains can be efficiently exploited for advancing understanding of how new immune receptor specificities are formed and provide new avenues to generate novel synthetic fusions.

METHODS

Identification of NLRs and NLR-IDs in plant genomes

NLR plant immune receptors were identified in nine monocot species by the presence of common NB-ARC domain (Pfam PF00931) as described previously (Sarris et al., 2016). *T. aestivum* (TGAC v1) and *A. tauschii* genomes (ASM34733v1) were downloaded from EnsemblPlants and analyzed using the same pipeline as before (Sarris et al., 2016). All up to date scripts are available from https://github.com/krasileva-group/plant_rgenes.

Phylogenetic Analysis

An HMM model - based on the pFAM model, PF00931 - was built to include the ARC2 subdomain which is also present in plant NB-ARC proteins (Supplemental File 1). To build the model of NB-ARC1-ARC2, eight proteins (Swissport identifiers: APAF_HUMAN, LOV1A_ARATH, K4BY49_SOLLC, RPM1_ARATH, R13L4_ARATH, RPS2_ARATH, DRL24_ARATH, DRL15_ARATH) were aligned using the PRANK program (Loytynoja and Goldman, 2008) and the HMM profile was built from this alignment with the HMMER3 HMMBUILD program (Mistry et al 2013), using default parameters for both programs. Amino acid sequences encoding the NB-ARC proteins were aligned to this hmm model using the HMMER3 HMMALIGN program (version 3.1b2) (Mistry et al 2013). The resulting alignment of the NB-ARC1-ARC2 domain was converted to fasta format using the HMMER ESL-REFORMAT program. Any amino acids with non-match states in the hmm model were removed from the alignment. Sequences with less than 70 % coverage across the alignment were removed from the data set. The longest sequence for each gene out of the available set of splice versions was used for phylogenetic analysis. In addition, 35 proteins encoding genes with characterized and known functions in pathogen defence from the literature were also included; the list of genes was based on a curated R-gene dataset by Sanseverino et al, 2012 (http://prgdb.crg.eu). Phylogenetic analysis was carried out using the MPI version of the RAxML (v8.2.9) program (Stamatakis, 2014) with the following method parameters set: -f a, -x 12345, -p 12345, -# 100, -m PROTCATJTT. The tree contained 4,130 sequences, 338 columns, took 67 hours to generate and required 17 GB RAM memory. Separate trees for each species were also prepared for Figure 2 using the same methods. Overall species phylogeny was constructed using NCBI taxon identification numbers at phyloT (phylot.biobyte.de). The trees were mid-point rooted and visualized using the Interactive Tree of Life (iToL) tool (Letunic and Bork, 2016) and are publicly available at http://itol.embl.de under 'Sharing data' and 'KrasilevaGroup' and in Newick format in Supplemental Dataset 3. Annotation files were prepared for displaying the presence of ID domains in the proteins, identifying species gene identifiers by colour and visualising the location of individual domains within the protein backbone. An ID domain was defined as being any

domain, except for LRR, AAA, TIR and RPW8 which are often associated with NB-ARC-containing proteins.

Supplemental Data

Supplemental Table 1. Number of NLRs and NLR-IDs in hotspot 1, hotspot 2 and hotspot 3 in nine grass species.

Supplemental Table 2. Integrated domains found in nine grass species at relaxed e-value cutoff (<0.05).

Supplemental Table 3: Genomic locations of all NLRs and NLR-IDs present in the tree in Figure 1A, anchored to the genetic map of *T. aestivum* CS42.

Supplemental Figure 1 Maximum likelihood phylogeny based on NB-ARC domain of all NLRs and NLR-IDs of *S. italic*.

Supplemental Figure 2. Maximum likelihood phylogeny based on NB-ARC domain of all NLRs and NLR-IDs of *S. bicolor*.

Supplemental Figure 3. Maximum likelihood phylogeny based on NB-ARC domain of all NLRs and NLR-IDs of *Z. mays.*

Supplemental Figure 4. Maximum likelihood phylogeny based on NB-ARC domain of all NLRs and NLR-IDs of *O. sativa*.

Supplemental Figure 5. Maximum likelihood phylogeny based on NB-ARC domain of all NLRs and NLR-IDs of *B. distachyon*.

Supplemental Figure 6. Maximum likelihood phylogeny based on NB-ARC domain of all NLRs and NLR-IDs of *H. vulgare*.

Supplemental Figure 7. Maximum likelihood phylogeny based on NB-ARC domain of all NLRs and NLR-IDs of *A. tauschii*.

Supplemental Figure 8. Maximum likelihood phylogeny based on NB-ARC domain of all NLRs and NLR-IDs of *T. aestivum*

Supplemental Figure 9. Maximum likelihood phylogeny based on NB-ARC domain of all NLRs and NLR-IDs of *T. urartu*.

Supplemental Dataset 1. Domains in NLRs and NLR-IDs of *T. aestivum CS42 TGAC v1*.

Supplemental Dataset 2. Domains in NLRs and NLR-IDs of *A. tauschii*.

Supplemental Dataset 3. All phylogenies in newick format.

Supplemental Dataset 4. Gene identification numbers for NLRs from the red hotspot clade, blue neighbouring clade and cyan ancestral clade.

Supplemental File 1. The hmm model used to align NB LRR proteins for phylogenetic analysis.

AUTHOR CONTRIBUTIONS

PB, KVK and WH designed the study. PB, GD, EB and KVK analyzed the data. All authors contributed to writing of the manuscript.

ACKNOWLEDGEMENTS

Authors are grateful to all members of Krasileva group and many colleagues, especially Sophien Kamoun, for thoughtful discussions of the presented material. We thank Daniil Prigozhin for suggestions on data analyses and the manuscript. We are grateful to Matthew Moscou and William Jackson as well to the 2016 European Research Council interview panel members for providing additional motivation to write this manuscript. KVK is strategically supported by the Biotechnology and Biological Science Research Council (BBSRC) and the

Gatsby Charitable Foundation. This project was also supported by BBSRC and Institute Strategic Programme Grant at The Earlham Institute (BB/J004669/1) and BBSRC National Capability in Genomics at The Earlham Institute (BB/J010375/1). The high-performance computing resources and services used in this work were supported by the EI Scientific Computing group alongside the NBIP Computing infrastructure for Science (CiS) group.

REFERENCES

Andolfo, G., Jupe, F., Witek, K., Etherington, G.J., Ercolano, M.R., and Jones, J.D.G. (2014). Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. BMC Plant Biol. **14**: 120.

Bakker, E.G., Toomajian, C., Kreitman, M., and Bergelson, J. (2006). A genome-wide survey of R gene polymorphisms in Arabidopsis. Plant Cell **18**: 1803–1818.

Bent, A.F. , Kunkel, B.N., Dahibeck D., Brown, K.L., Schmidt, R., Giraudat J., Leung, J., Staskawicz, B.J. (1994). RPS2 of *Arabidopsis thaliana* a leucine-rich repeat class of plant disease resistance genes. Science (80-.). **265**: 1856–1860.

Cesari, S., Bernoux, M., Moncuquet, P., Kroj, T., and Dodds, P.N. (2014). A novel conserved mechanism for plant NLR protein pairs: the “integrated decoy” hypothesis. Front. Plant Sci. **5**: 606.

Césari, S., Kanzaki, H., Fujiwara, T., Bernoux, M., Chalvon, V., Kawano, Y., Shimamoto, K., Dodds, P., Terauchi, R., and Kroj, T. (2014). The NB-LRR proteins RGA4 and RGA5 interact functionally and physically to confer disease resistance. EMBO J. **33**: 1941 LP – 1959.

Choulet, F. et al. (2010). Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell **22**: 1686–701.

Christie, N., Tobias, P. a, Naidoo, S., and Külheim, C. (2016). The *Eucalyptus grandis* NBS-LRR Gene Family: Physical Clustering and Expression Hotspots. Front. Plant Sci. **6**: 1238.

Christopoulou, M., McHale, L.K., Kozik, A., Reyes-Chin Wo, S., Wroblewski, T., and Michelmore, R.W. (2015a). Dissection of Two Complex Clusters of Resistance Genes in Lettuce (*Lactuca sativa*). Mol. Plant-Microbe Interact. **28**: 751–765.

Christopoulou, M., Wo, S.R.-C., Kozik, A., McHale, L.K., Truco, M.-J., Wroblewski, T., and Michelmore, R.W. (2015b). Genome-Wide Architecture of Disease Resistance Genes in Lettuce. G3 (Bethesda). **5**: 2655–69.

Clavijo, B.J. et al. (2016). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. bioRxiv: http://biorxiv.org/content/early/2016/10/13/080796.

Dangl, J.L., Horvath, D.M., and Staskawicz, B.J. (2013). Pivoting the plant immune system from dissection to deployment. Science **341**: 746–51.

Dodds, P.N. and Rathjen, J.P. (2010). Plant immunity: towards an integrated view of plant–pathogen interactions. Nat. Rev. Genet. **11**: 539–548.

Dubcovsky, J. and Dvorak, J. (2007). Genome Plasticity a Key Factor. Science **316**: 1862–1866.

Ellis, J.G., Lagudah, E.S., Spielmeyer, W., and Dodds, P.N. (2014). The past, present and future of breeding rust resistant wheat. Front Plant Sci **5**: 641.

Giolai, M., Paajanen, P., Verweij, W., Percival-Alwyn, L., Baker, D., Witek, K., Jupe, F., Bryan, G., Hein, I., G Jones, J.D., Clark, D., and Clark, M.D. Targeted capture and sequencing of gene sized DNA molecules.: 1–8.

Hall, S. a., Allen, R.L., Baumber, R.E., Baxter, L. a., Fisher, K., Bittner-Eddy, P.D., Rose, L.E., Holub, E.B., and Beynon, J.L. (2009). Maintenance of genetic variation in plants and pathogens involves complex networks of gene-for-gene interactions. Mol. Plant Pathol. **10**: 449–457.

Ilbery, B., Maye, D., Ingram, J., and Little, R. (2013). Risk perception, crop protection and plant disease in the UK wheat sector. Geoforum **50**: 129–137.

Jacob, F., Vernaldi, S., and Maekawa, T. (2013). Evolution and conservation of plant NLR functions. Front. Immunol. **4**: 1–16.

Jones, J.D.G., Vance, R.E., and Dangl, J.L. (2016). Intracellular innate immune surveillance devices in plants and animals. Science (80-.). **354**: aaf6395–aaf6395.

Joshi, R.K. and Nayak, S. (2013). Perspectives of genomic diversification and molecular recombination

towards R-gene evolution in plants. Physiol. Mol. Biol. Plants **19**: 1–9.

Jupe, F. et al. (2013). Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. Plant J. **76**: 530–544.

Kroj, T., Chanclud, E., Michel-Romiti, C., Grand, X., and Morel, J.B. (2016). Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. New Phytol. **210**: 618–626.

Le Roux, C. et al. (2015). A Receptor Pair with an Integrated Decoy Converts Pathogen Disabling of Transcription Factors to Immunity. Cell **161**: 1074–1088.

Leister, D., Kurth, J., Laurie, D. a, Yano, M., Sasaki, T., Devos, K., Graner, a, and Schulze-Lefert, P. (1998). Rapid reorganization of resistance gene homologues in cereal genomes. Proc Natl Acad Sci U S A **95**: 370–375.

Letunic, I. and Bork, P. (2007). Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. Bioinformatics **23**: 127–128.

Löytynoja, A. and Goldman, N. (2008). A model of evolution and structure for multiple sequence alignment. Philos Trans R Soc Lond, B, Biol Sci **363**: 3913–3919.

Michelmore, R.W., Meyers, B.C., and Young, N.D. (1998). Cluster of resistance genes in plants evolveby divergent selection and a birth-and-death process. Curr. Opin. Plant Biol. **3**: 285–90.

Mindrinos, M., Katagiri, F., Yu, G.-L., and Ausubel, F.M. (1994). The *A. thaliana* disease resistance gene RPS2 encodes a protein containing a nucleotide-binding site and leucine-rich repeats. Cell **78**: 1089–1099.

Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. **41**.

Moore, A.D., Björklund, Å.K., Ekman, D., Bornberg-Bauer, E., and Elofsson, A. (2008). Arrangements in the modular evolution of proteins. Trends Biochem. Sci. **33**: 444–451.

Noel, L., Moores, T., van der Biezen, E., Parniske, M., Daniels, M., Parker, J., and Jones, J. (1999). Pronounced intraspecific haplotype divergence at the RPP8 complex disease resistance locus of Arabidopsis. Plant Cell **11**: 2099–2111.

Okuyama, Y. et al. (2011). A multifaceted genomics approach allows the isolation of the rice Pia-blast resistance gene consisting of two adjacent NBS-LRR protein genes. Plant J. **66**: 467–479.

Periyannan, S.K. et al. (2013). The Gene Sr33, an Ortholog of Barley Mla Genes, Encodes Resistance to Wheat Stem Rust Race Ug99. Science **341**: 786–788.

Prasad, V., Strömberg, C. a. E., Leaché, a. D., Samant, B., Patnaik, R., Tang, L., Mohabey, D.M., Ge, S., and Sahni, a. (2011). Late Cretaceous origin of the rice tribe provides evidence for early diversification in Poaceae. Nat. Commun. **2**: 480.

Prasad, V., Stromberg, C., Alimohammadian, H., and Sahni, A. (2005). Dinosaur Coprolites and the Early Evolution of Grasses and Grazers. Science **310**: 1177–1180.

Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegu, B., Quraishi, U.M., Calcagno, T., Cooke, R., Delseny, M., and Feuillet, C. (2008). Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. Plant Cell **20**: 11–24.

Sanseverino, W., Hermoso, A., D'Alessandro, R., Vlasova, A., Andolfo, G., Frusciante, L., Lowy, E., Roma, G., and Ercolano, M.R. (2013). PRGdb 2.0: Towards a community-based database model for the analysis of R-genes in plants. Nucleic Acids Res. **41**: 1167–1171.

Sarris, P.F. et al. (2015). A plant immune receptor detects pathogen effectors that target WRKY transcription factors. Cell **161**: 1089–1100.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**: 1312–1313.

Sukarta, O.C. a, Slootweg, E.J., and Goverse, A. (2016). Structure-informed insights for NLR functioning in plant immunity. Semin. Cell Dev. Biol. **56**: 134–149.

Vogel, J.P. et al. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature **463**: 763–768.

Wicker, T., Buchmann, J.P., and Keller, B. (2010). Patching gaps in plant genomes results in gene movement and erosion of colinearity. Genome Res. **20**: 1229–1237.

Wicker, T., Yu, Y., Haberer, G., Mayer, K.F.X., Marri, P.R., Rounsley, S., Chen, M., Zuccolo, A., Panaud, O., Wing, R. a., and Roffler, S. (2016). DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses. Nat. Commun. **7**: 12790.

Witek, K., Jupe, F., Witek, A.I., Baker, D., Clark, M.D., and Jones, J.D.G. (2016). Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing. Nat. Biotechnol. **34**: 656–660.

Yue, J.X., Meyers, B.C., Chen, J.Q., Tian, D., and Yang, S. (2012). Tracing the origin and evolutionary history of plant nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes. New Phytol. **193**: 1049–1063.

Zhai, C., Zhang, Y., Yao, N., Lin, F., Liu, Z., Dong, Z., Wang, L., and Pan, Q. (2014). Function and interaction of the coupled genes responsible for Pik-h encoded rice blast resistance. PLoS One **9**.

Zhang, Y., Xia, R., Kuang, H., and Meyers, B.C. (2016). The Diversification of Plant NBS-LRR Defense Genes Directs the Evolution of MicroRNAs That Target Them. Mol. Biol. Evol. **33**: 2692–2705.

Zhong, Y. and Cheng, Z.-M. (Max) (2016). A unique RPW8-encoding class of genes that originated in early land plants and evolved through domain fission, fusion, and duplication. Sci. Rep. **6**: 32923.

Table 1. Summary of unique Pfam domains found in NLR-ID hotspot 1 and neighbouring clade. Only domains with e-value <1e-3 are shown. For the full list of domains with lower stringency (e-value <0.05), see Supplemental Table 2.

Species	Total NLR-ID	Total Unique ID	Outgroup 1 Unique ID	Hotspot 1 Unique ID	Hotspot 1 IDs
S. italica	9	7	-	2	NAM, WRKY
S. bicolor	22	13	-	4	WRKY, HLH, NAM, Glutaredoxin
Z. mays	7	8	-	-	-
O. sativa	18	16	-	3	AvrRpt-cleavage, Thioredoxin, DUF761,
B. distachyon	16	9	-	7	AP2, Jacalin Myb_DNA-binding, Pkinase, Pkinase_Tyr, WRKY
H. vulgare	19	15	-	11	AvrRpt-cleavage, B3, CG-1, DUF581, Exo70, Kelch_1, PP2C, Pkinase, Pkinase_Tyr, WRKY, zf-LSD1
***A. tauschii* (D)**	67	32	2	8	AvrRpt-cleavage, B3, Kelch_1, Pkinase, Pkinase_Tyr, RVT_2, WRKY, p450
***T. aestivum* subgenomes:**	133	46	4	21	AP2, Ank_2, Ank_5, B3, BTB, CG-1, DUF3420, DUF793, Exo70, GRAS, Kelch_1, Myb_DNA-binding, NPR_1-like, PGG, PP2C, Pkinase, Pkinase_Tyr, RIP, TIG, WRKY, zf-RING_2
A				13	
B				8	
D				13	
U				4	
***T. urartu* (A)**	32	25	4	9	B3, CG-1, EF_hand_5, Exo70. Kelch_1, PP2C, Pkinase, Pkinase_Tyr, RVT_3
Average	35	19	1	7	-

Table 2. Genomic locations of NLRs overall and from the hotspot 1 clade

Species	Chromosome	Total No. NLRs	No. of NLRs in hotspot 1 ID clade	No. of NLRs in outer clade of hotspot 1
S. bicolor	Chr 1	13	-	-
	Chr 2	32	2	1
	Chr 3	12	-	-
	Chr 4	6	-	-
	Chr 5	86	2	3
	Chr 6	13	-	-
	Chr 7	20	-	3
	Chr 8	37	1	3
	Chr 9	21	-	-
	Chr 10	20	-	-
	Chr U	2	1	-
Z. mays	Chr 1	8	-	-
	Chr 2	10	-	-
	Chr 3	6	-	-
	Chr 4	12	-	-
	Chr 5	7	-	-
	Chr 6	8	-	-
	Chr 7	6	-	-
	Chr 8	4	-	-
	Chr 9	1	-	-
	Chr 10	22	-	-
O. sativa	Chr 1	30	-	-
	Chr 2	20	-	-
	Chr 3	18	-	-
	Chr 4	23	-	-
	Chr 5	16	1	-
	Chr 6	27	-	-
	Chr 7	17	-	1
	Chr 8	40	-	4
	Chr 9	18	-	-
	Chr 10	15	1	-
	Chr 11	93	4	-
	Chr 12	43	1	2
	Chr U	2	-	-
B. distachyon	Chr 1	34	2	-
	Chr 2	52	1	-
	Chr 3	40	1	1
	Chr 4	106	8	9
	Chr 5	21	-	-
	ChrU	2	-	-
T. aestivum	Chr 1 A, B, D	51, 122, 72	7, 10, 7	7,8,4
	Chr 2 A, B, D	82, 112, 83	5, 4, 4	1,2,1
	Chr 3 A, B, D	41, 80, 57	4, 12, 7	3,7,5
	Chr 4 A, B, D	129, 21, 16	19, 2, 1	10,1,0
	Chr 5 A, B, D	36, 88, 79	5, 8, 4	3,1,1
	Chr 6 A, B, D	57, 86, 66	9, 11, 7	8,8,7
	Chr 7 A, B, D	117, 110, 124	12, 9, 15	1,6,5
	Chr U	103	2	2

Figures

Figure 1. Maximum likelihood phylogeny of NLRs in grasses identifies evolutionary hotspots of NLRs with integrated domains (A) Maximum likelihood tree of the NB-ARC family in grasses (4,130 proteins, 9 species) shows a non-uniform distribution of proteins with an integrated domain (ID) across the phylogeny (red branches). Based on high bootstrap support (85-99%), a clade containing a particularly high number of NLR-ID proteins (an average of 59% NLR-IDs, compared to 8% background in *T. aestivum*) was identified (highlighted by the red sector). This clade is nested inside an outer subgroup with high bootstrap support (96-98%, highlighted by the blue sector) next to an ancestral clade (bootstrap support, 85%, highlighted by the cyan sector). For each highlighted clade, the percent of NLRs with ID domains is indicated in red. (B) close-up of the hotspot 1 region showing the key bootstrap support values. (C) close-up of the hotspot 2 region showing the key bootstrap support values. (D) close-up of the hotspot 2 region showing the key bootstrap support values. (E) Phylogenetic relationship of the species included in the phylogeny and a summary of the number of NLRs(-ID) proteins that are present in the tree from each species, including the number of NLR(-ID) proteins from hotspot 1 and the associated outer clade. E-value cut-off for presence of an ID domain, 0.001.

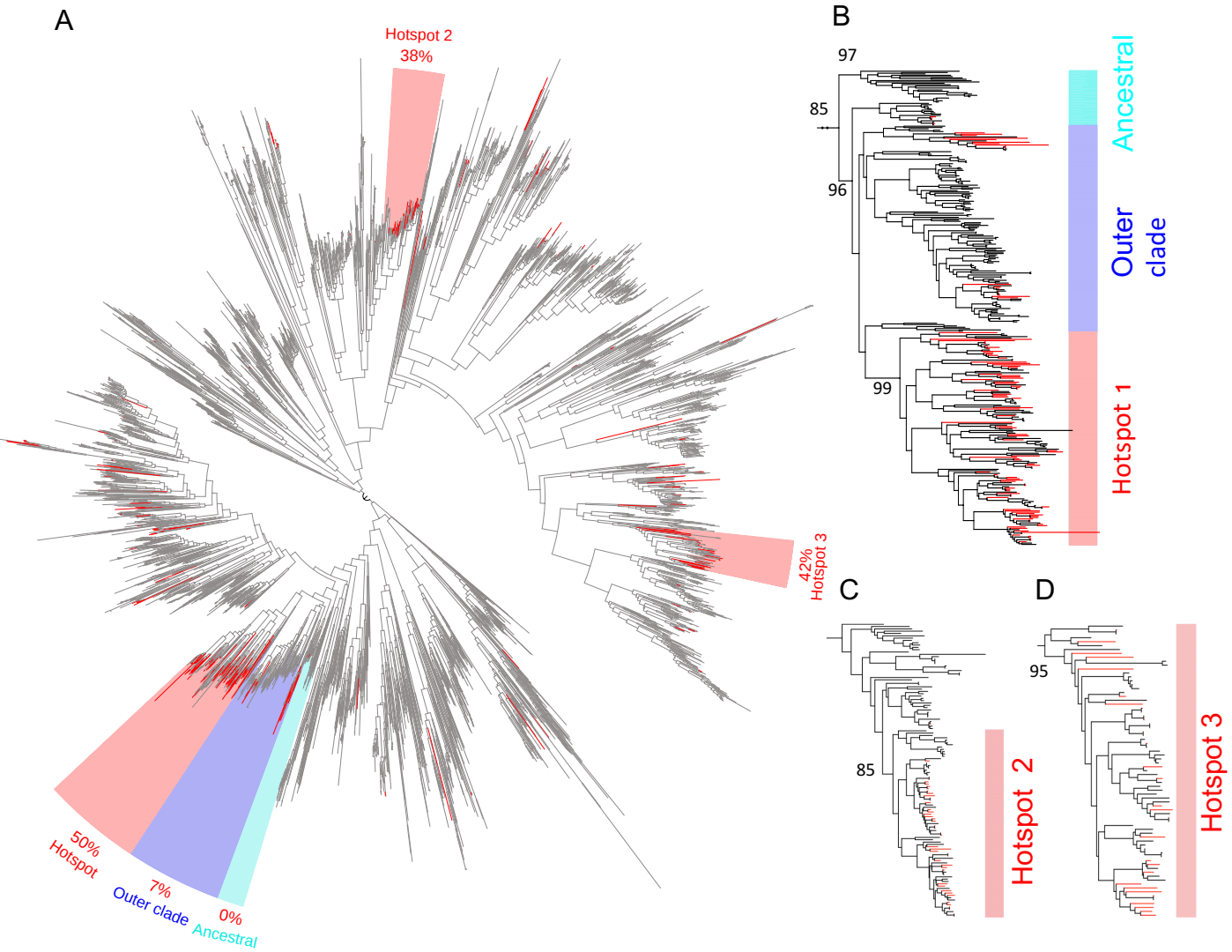
Figure 2. NLR-ID hotspot 1 clade prone to new integrations expands and diversifies in individual grass genomes. (A-I) Maximum likelihood phylogeny based on the NB-ARC domain of all NLRs and NLR-IDs for each species. Clades supported by a bootstrap value >=85% (black circles) were collapsed and indicated with a triangle. Branch colours indicate the clade-type identified in Figure 1: ancestral (cyan), outgroup (blue) and hotspot (red). The phylogenies are annotated with the distribution of fused exogenous domains (red squares) for NLRs within the expanded clades. The domain names given in red refer to Pfam annotation of the exogenous domains found in the hotspot clade for that species. E-value cut-off for presence of an ID domain, 0.001.

Figure 3. Close-up of the NLR-ID hotspot 1 clade displaying a variety of ID domains indicates rapid domain recycling. The branches of the hotspot clade, the outer clade and the ancestral clade are shown in red, blue and cyan respectively. Dots on the branches indicate a bootstrap support value >= 85 %. Alongside the tree are cartoons of each protein, annotated with the domain(s) in the position that they appear in the protein (protein backbone, grey; NB-ARC domain, black; LRR and AAA domains, orange) (B) An example of proteins in the *Triticeae* with the same Pkinase domain present at the N-terminal end of the protein (C) An example of proteins in two related but distinct clades with ID domains, notably transcription factors for one of the clades, present their C-terminal ends. E-value cut-off for presence of an ID domain, 0.001; e-value cut-off for an LRR or AAA domain, 10.0.

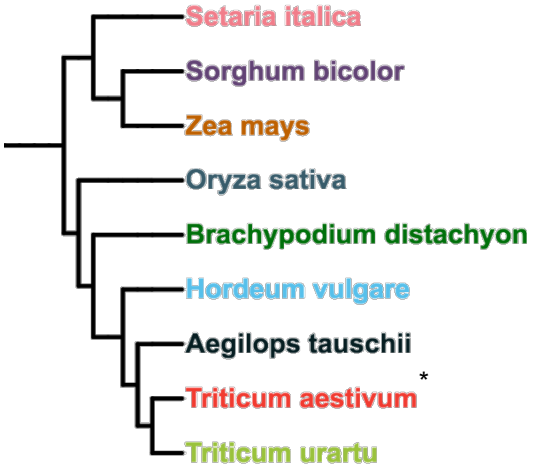
Figure 4. Diversity of integrated domains among syntenic homeologous NLRs in wheat indicates rapid domain recycling. A clade in the phylogeny of wheat NLR-IDs based on NB-ARC1-ARC2 that includes A, B and D genome homeologs from same genetic position.

Figure 5. Evolutionary model of NLR-ID hotspot formation and diversification. Arrows indicate their fate over evolutionary time; orange arrows indicate duplication events. The ancestral protein (cyan) underwent duplication to form the outer clade of proteins (blue). Then proteins in specific clades, especially those proteins in hotspot 1 (red) gained the ability to form fusions with other domains (ID's). Some proteins have maintained the same domain (for example, ID1) whilst other proteins have undergone further diversification through the exchange of the ID domain (for example, ID3 and ID4).

Figure 1



E

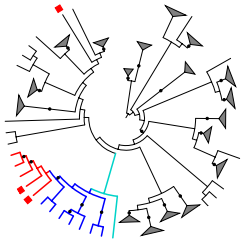


	All NLR, NLR-ID (%)	Outgroup NLR, NLR-ID (%)	Hotspot 1 NLR, NLR-ID (%)
Setaria italica	304, 9 (3%)	5, 0	7, 2 (29%)
Sorghum bicolor	262, 22 (8%)	10, 0	6, 4 (67%)
Zea mays	84, 7 (8%)	0, 0	0, 0
Oryza sativa	362, 18 (5%)	7, 0	7, 3 (43%)
Brachypodium distachyon	255, 16 (6%)	10, 0	12, 7 (58%)
Hordeum vulgare	237, 19 (8%)	3, 0	22, 14 (64%)
Aegilops tauschii	482, 67 (13%)	28, 4 (14%)	16, 12 (75%)
Triticum aestivum*	1732, 133 (8%)	91, 13 (14%)	68, 40 (59%)
Triticum urartu	377, 32 (8%)	17, 6 (35%)	19, 10 (53%)

* Hexaploid

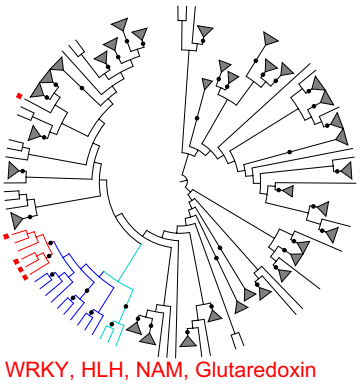
Figure 2

A *S. italica*



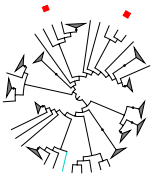
NAM, WRKY

B *S. bicolor*

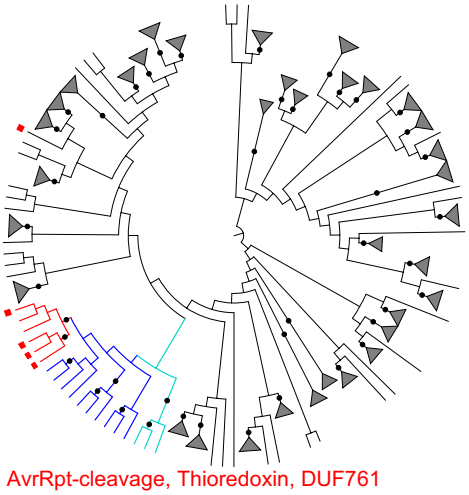


WRKY, HLH, NAM, Glutaredoxin

C *Z. mays*

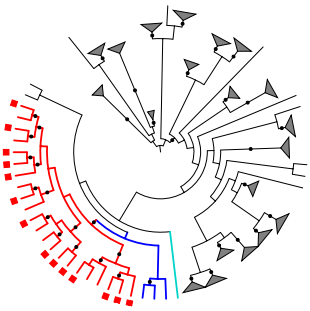


D *O. sativa*



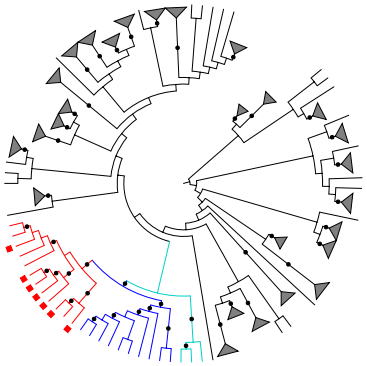
AvrRpt-cleavage, Thioredoxin, DUF761

E *B. distachyon*



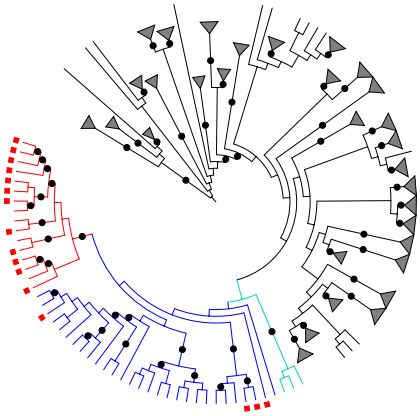
AP2, pKinase, Jacalin, WRKY, pKinase_Tyr, Myb_DNA-binding

F *H. vulgare*



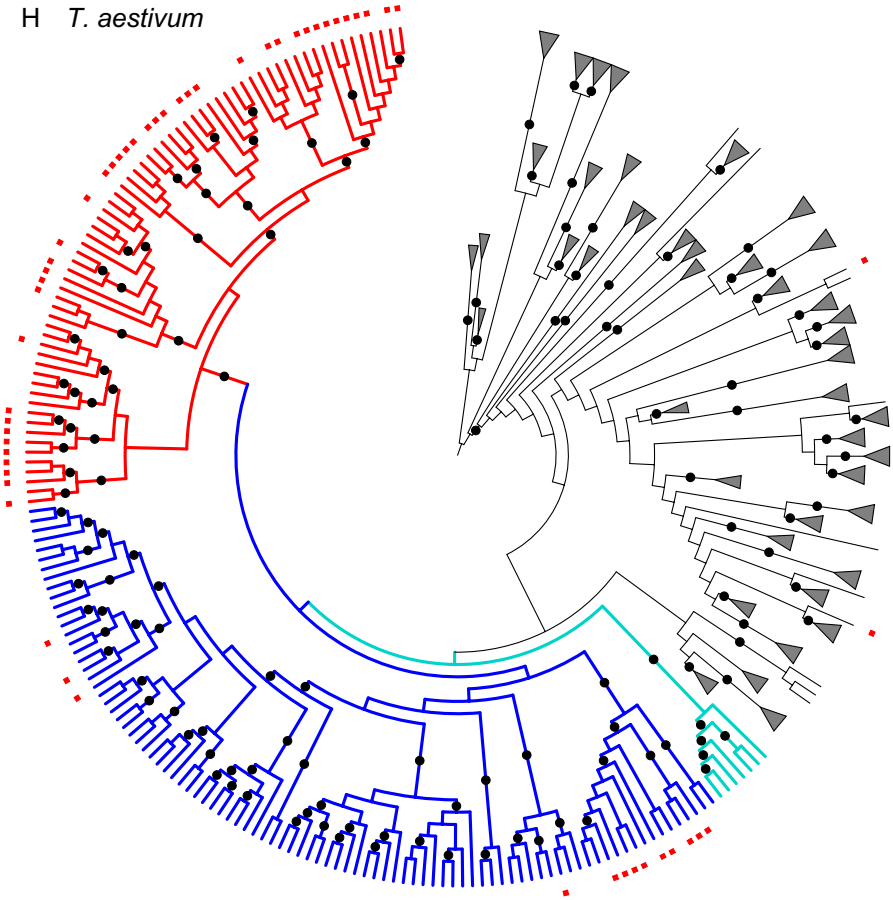
Kelch_1, pKinase, AvrRpt-cleavage, PP2C, Exo70, pKinase_Tyr, B3, DUF581, WRKY, zf-LSD1,CG1

G *A. tauschii*



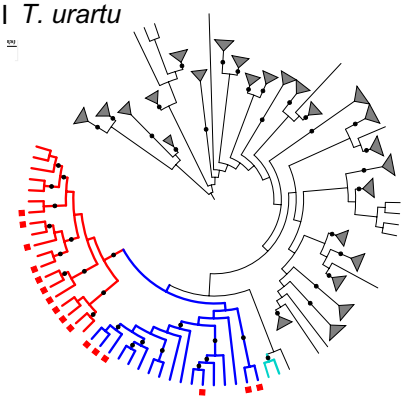
pKinase_Tyr, Kelch_1, RVT_2, pKinase, B3, WRKY, p450, AvrRpt-cleavage

H *T. aestivum*



Ank_2, AP2, Ank_5, B3, BTB, CG-1, DUF3420, DUF793, Exo70, GRAS, Kelch_1, Myb_DNA-binding, NPR1_like_C, PGG, PP2C, Pkinase, Pkinase_Tyr, RIP, TIG, WRKY, zf-RING_2

I *T. urartu*



B3, CG-1, EF_hand_5, Exo70, Kelch_1, PP2C, Pkinase, Pkinase_Tyr, RVT_3

Figure 3

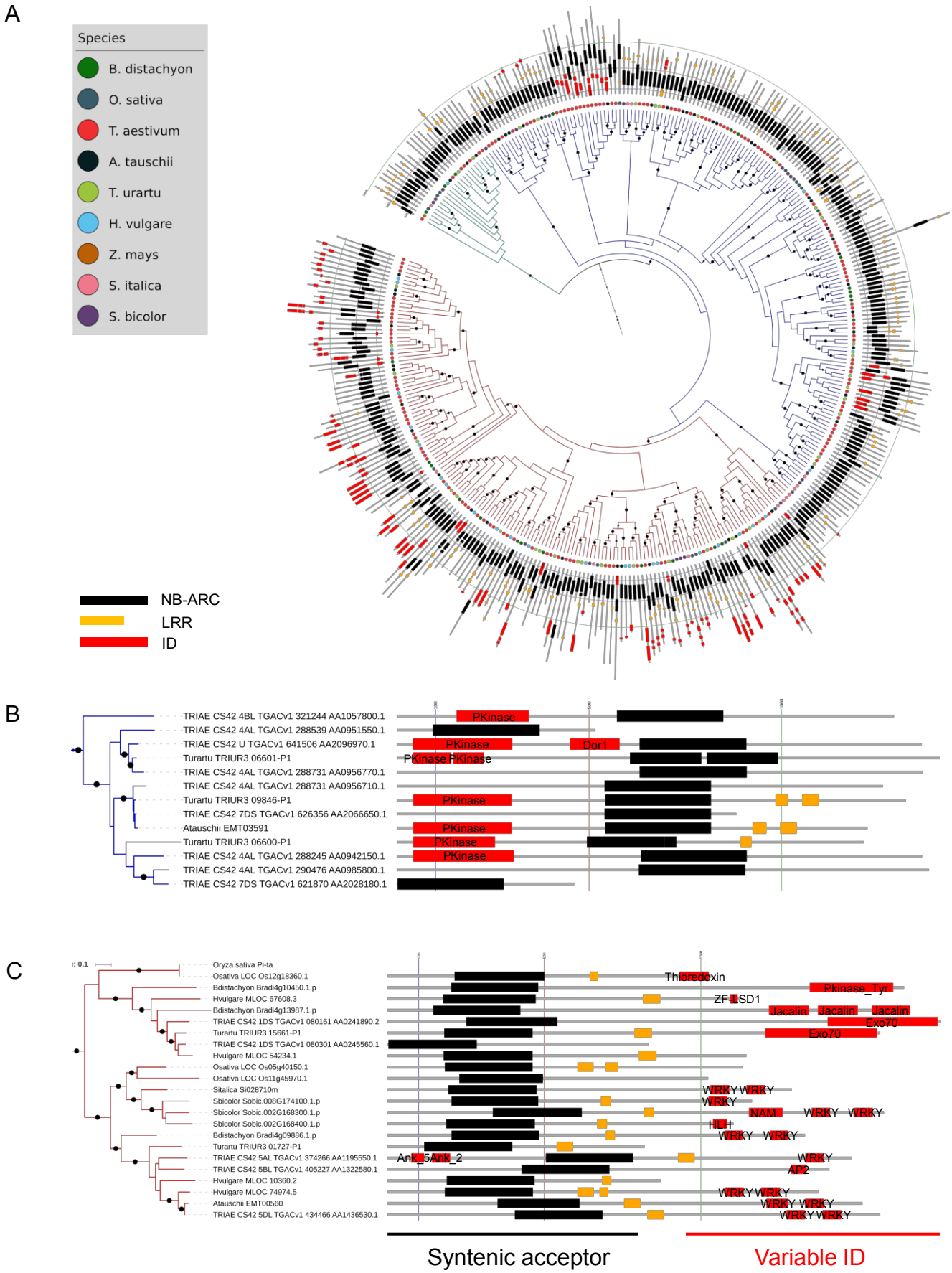

A
Species
B. distachyon
O. sativa
T. aestivum
A. tauschii
T. urartu
H. vulgare
Z. mays
S. italica
S. bicolor
NB-ARC
LRR
ID
B
TRIAE CS42 4BL TGACv1 321244 AA1057800.1
TRIAE CS42 4AL TGACv1 288539 AA0951550.1
TRIAE CS42 U TGACv1 641506 AA2096970.1
Turartu TRIUR3 06601-P1
TRIAE CS42 4AL TGACv1 288731 AA0956770.1
TRIAE CS42 4AL TGACv1 288731 AA0956710.1
Turartu TRIUR3 09846-P1
TRIAE CS42 7DS TGACv1 626356 AA2066650.1
Atauschii EMT03591
Turartu TRIUR3 06600-P1
TRIAE CS42 4AL TGACv1 288245 AA0942150.1
TRIAE CS42 4AL TGACv1 290476 AA0985800.1
TRIAE CS42 7DS TGACv1 621870 AA2028180.1
PKinase
Dor1
C
Oryza sativa Pi-ta
Osativa LOC Os12g18360.1
Bdistachyon Bradi4g10450.1.p
Hvulgare MLOC 67608.3
Bdistachyon Bradi4g13987.1.p
TRIAE CS42 1DS TGACv1 080161 AA0241890.2
Turartu TRIUR3 15661-P1
TRIAE CS42 1DS TGACv1 080301 AA0245560.1
Hvulgare MLOC 54234.1
Osativa LOC Os05g40150.1
Osativa LOC Os11g45970.1
Sitalica Si028710m
Sbicolor Sobic.008G174100.1.p
Sbicolor Sobic.002G168300.1.p
Sbicolor Sobic.002G168400.1.p
Bdistachyon Bradi4g09886.1.p
Turartu TRIUR3 01727-P1
TRIAE CS42 5AL TGACv1 374266 AA1195550.1
TRIAE CS42 5BL TGACv1 405227 AA1322580.1
Hvulgare MLOC 10360.2
Hvulgare MLOC 74974.5
Atauschii EMT00560
TRIAE CS42 5DL TGACv1 434466 AA1436530.1
Thioredoxin
Pkinase_Tyr
ZF-LSD1
Jacalin
Exo70
WRKY
NAM
HLH
Ank_5
Ank_2
AP2
Syntenic acceptor
Variable ID

Figure 4

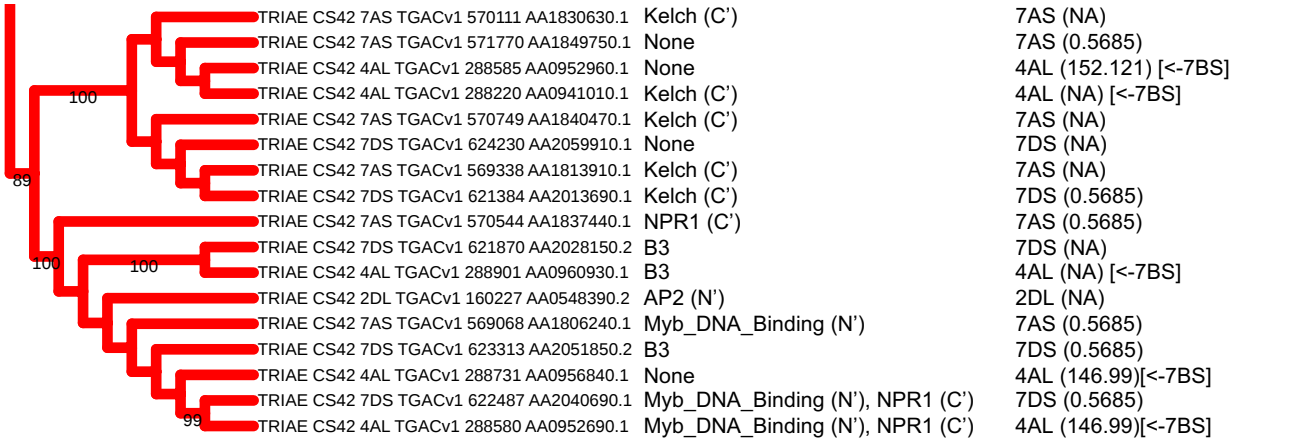
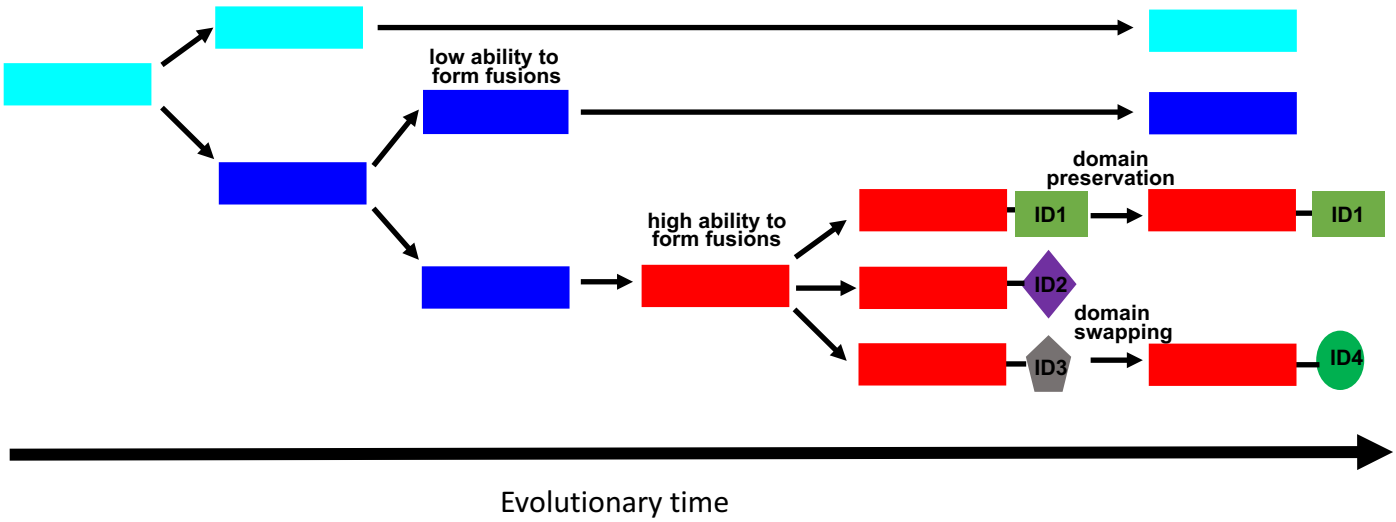

T.aestiivum Protein ID
Integrated Domains
Chromosome (cM)
[genome rearrangement]
TRIAE CS42 7AS TGACv1 570111 AA1830630.1 Kelch (C') 7AS (NA)
TRIAE CS42 7AS TGACv1 571770 AA1849750.1 None 7AS (0.5685)
TRIAE CS42 4AL TGACv1 288585 AA0952960.1 None 4AL (152.121) [<-7BS]
TRIAE CS42 4AL TGACv1 288220 AA0941010.1 Kelch (C') 4AL (NA) [<-7BS]
TRIAE CS42 7AS TGACv1 570749 AA1840470.1 Kelch (C') 7AS (NA)
TRIAE CS42 7DS TGACv1 624230 AA2059910.1 None 7DS (NA)
TRIAE CS42 7AS TGACv1 569338 AA1813910.1 Kelch (C') 7AS (NA)
TRIAE CS42 7DS TGACv1 621384 AA2013690.1 Kelch (C') 7DS (0.5685)
TRIAE CS42 7AS TGACv1 570544 AA1837440.1 NPR1 (C') 7AS (0.5685)
TRIAE CS42 7DS TGACv1 621870 AA2028150.2 B3 7DS (NA)
TRIAE CS42 4AL TGACv1 288901 AA0960930.1 B3 4AL (NA) [<-7BS]
TRIAE CS42 2DL TGACv1 160227 AA0548390.2 AP2 (N') 2DL (NA)
TRIAE CS42 7AS TGACv1 569068 AA1806240.1 Myb_DNA_Binding (N') 7AS (0.5685)
TRIAE CS42 7DS TGACv1 623313 AA2051850.2 B3 7DS (0.5685)
TRIAE CS42 4AL TGACv1 288731 AA0956840.1 None 4AL (146.99)[<-7BS]
TRIAE CS42 7DS TGACv1 622487 AA2040690.1 Myb_DNA_Binding (N'), NPR1 (C') 7DS (0.5685)
TRIAE CS42 4AL TGACv1 288580 AA0952690.1 Myb_DNA_Binding (N'), NPR1 (C') 4AL (146.99)[<-7BS]
100
89
100
100
99

Figure 5


low ability to
form fusions
high ability to
form fusions
domain
preservation
ID1
ID1
ID2
domain
swapping
ID3
ID4
Evolutionary time