Running title:

Evolutionary dynamics of the LRR-RLK gene family

Corresponding Authors:

Iris Fischer (irisfischer402@gmail.com) and Nathalie Chantret (chantret@supagro.inra.fr)
INRA SupAgro, UMR AGAP, 2 Place Pierre Viala, Bât. 21, 34060 Montpellier, France
Telephone: +33 (0)4 99 61 60 83

Research Area: Genes, Development and Evolution
Evolutionary dynamics of the Leucine-Rich Repeats Receptor-Like Kinase (LRR-RLK) subfamily in angiosperms.

Iris Fischer¹,*, Anne Diévart², Gaetan Droc², Jean-François Dufayard² and Nathalie Chantret¹,*

¹ INRA, UMR AGAP, F-34060 Montpellier, France
² CIRAD, UMR AGAP, F-34398 Montpellier, France

*Corresponding Authors

Summary: Phylogenetic analysis of leucine-rich repeat containing receptor-like kinases demonstrates the dynamic nature of gene duplication, loss and selection in this family.
I.F., A.D., and N.C. designed the study; G.D. performed the LRR-RLK extraction; J-F.D. performed the phylogenetic clustering; I.F., A.D., and N.C. performed the data analysis and statistics; I.F. drafted the manuscript with the help of A.D. and N.C.

Financial sources: IF was funded by the German Research Foundation (DFG): FI 1984/1-1 and the ARCAD project (Agropolis Resource Center for Crop Conservation, Adaptation and Diversity) of the Agropolis Foundation. This work was partially funded by grant #ANR-08-GENM-021 from Agence Nationale de la recherche (ANR, France).

Corresponding Authors:
Iris Fischer (irisfischer402@gmail.com) and Nathalie Chantret (chantret@supagro.inra.fr)
ABSTRACT

Gene duplications are an important factor in plant evolution and lineage specific expanded (LSE) genes are of particular interest. Receptor-like kinases (RLK) expanded massively in land plants and Leucine-Rich Repeat (LRR)-RLKs constitute the largest RLK family. Based on the phylogeny of 7,554 LRR-RLK genes from 31 fully sequenced flowering plant genomes, the complex evolutionary dynamics of this family was characterized in depth. We studied the involvement of selection during the expansion of this family among angiosperms. LRR-RLK subgroups harbor extremely contrasted rates of duplication, retention or loss and LSE copies are predominantly found in subgroups involved in environmental interactions. Expansion rates also differ significantly depending on the time when rounds of expansion or loss occurred on the angiosperm phylogenetic tree. Finally, using a $d_{\infty}/d_{S}$-based test in a phylogenetic framework, we searched for selection footprints on LSE and single-copy LRR-RLK genes. Selective constraint appeared to be globally relaxed at LSE genes and codons under positive selection were detected in 50% of them. Moreover, the LRR domains – and specifically four amino acids in them – were found to be the main targets of positive selection. Here, we provide an extensive overview of the expansion and evolution of this very large gene family.
INTRODUCTION

Receptor-like kinases (RLK) comprise one of the largest gene families in plants and expanded massively in land plants (Embryophyta) (Lehti-Shiu et al., 2009; Lehti-Shiu et al., 2012). For plant RLK gene families, the functions of most members are often not known (especially in recently expanded families) but some described functions include innate immunity (Albert et al., 2010), pathogen response (Dodds and Rathjen, 2010), abiotic stress (Yang et al., 2010), development (De Smet et al., 2009), and sometimes multiple functions (Lehti-Shiu et al., 2012). The RLKs usually consist of three domains: an amino-terminal extracellular domain (ECD), a transmembrane (TM) domain, and a carboxy-terminal kinase domain (KD). In plants, the KD usually has a serine/threonine specificity (Shiu and Bleecker, 2001) but tyrosine specific RLKs were also described (e.g. BRASSINOSTEROID INSENSITIVE1 (Oh et al., 2009)). Interestingly, it was estimated that ~20% of RLKs contain a catalytically inactive KD (e.g. STRUBBELIG, CORYNE (Chevalier et al., 2005; Castells and Casacuberta, 2007; Gish and Clark, 2011)). In Arabidopsis, 44 RLK subgroups (SG) were defined by inferring the phylogenetic relationships between the KDS (Shiu and Bleecker, 2001). Interestingly, different SGs show different duplication/retention rates (Lehti-Shiu et al., 2009). Specifically, RLKs involved in stress response show a high number of tandemly duplicated genes whereas those involved in development do not (Shiu et al., 2004) which suggests that some RLK genes are important for responses of land plants to a changing environment (Lehti-Shiu et al., 2012). There seem to be relatively few RLK pseudogenes compared to other large gene families and copy retention was argued to be driven by both, drift and selection (Zou et al., 2009; Lehti-Shiu et al., 2012). As most SGs are relatively old and RLK subfamilies expanded independently in several plant lineages, duplicate retention cannot be explained by drift alone and natural selection is expected to be an important driving factor in RLK gene family retention (Lehti-Shiu et al., 2009).

Leucine-Rich Repeat (LRR)-RLKs, which contain up to 30 LRRs in their extracellular domain, constitute the largest RLK family (Shiu and Bleecker, 2001). Based on the kinase domain, 15 LRR-RLK SGs have been established in Arabidopsis (Shiu et al., 2004; Lehti-Shiu et al., 2009). So far, two major functions have been attributed to them: defense against pathogens and development (Tang et al., 2010b). LRR-RLKs involved in defense are predominantly found in lineage specific expanded (LSE) gene clusters whereas LRR-RLKs involved in development are mostly found in non-expanded groups (Tang et al., 2010b). It was also discovered that the LRR domains are significantly less conserved than the remaining domains of the LRR-RLK genes (Tang et al., 2010b). In addition, a study on four plant
genomes (*A. thaliana*, grape, poplar, rice) showed that LRR-RLK genes from LSE gene clusters show significantly more indication of positive selection or relaxed constraint than LRR-RLKs from non-expanded groups (Tang et al., 2010b).

The genomes of flowering plants (angiosperms) have been shown to be highly dynamic compared to most other groups of land plants (Leitch and Leitch, 2012). This dynamic is mostly caused by the frequent multiplication of genetic material, followed by a complex pattern of differential losses (*i.e.* the fragmentation process) and chromosomal rearrangements (Langham et al., 2004; Leitch and Leitch, 2012). Most angiosperm genomes sequenced so far show evidence for at least one whole genome multiplication event during their evolution (see e.g. Jaillon et al. (2007); D'Hont et al. (2012); The Tomato Genome Consortium (2012)). At a smaller scale, tandem and segmental duplications are also very common in angiosperms (The Arabidopsis Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Rizzon et al., 2006). Although the most common fate of duplicated genes is to be progressively lost, in some cases they can be retained in the genome and adaptive as well as non-adaptive scenarios have been discussed to play a role in this preservation process (for review see Moore and Purugganan (2005); Hahn (2009); Innan (2009); Innan and Kondrashov (2010)). Whole genome sequences also revealed that the same gene may undergo several rounds of duplication and retention. These lineage specific expanded genes were shown to evolve under positive selection more frequently than single-copy genes in angiosperms (Fischer et al., 2014). The study by Fischer et al. analyzed general trends over whole genomes. Here, we ask if, and to what extent, this trend is observable at LRR-RLK genes. As this gene family is very dynamic and large – and in accordance with the results of Tang et al. (2010b) – we expect the effect of positive selection to be even more pronounced than in the whole genome average.

We analyzed 33 Embryophyta genomes to investigate the evolutionary history of the LRR-RLK gene family in a phylogenetic framework. Twenty LRR-RLK subgroups were identified and from this dataset we deciphered the evolutionary dynamics of this family within angiosperms. The expansion/reduction rates were contrasted between SGs and species as well as in ancestral branches of the angiosperm phylogeny. We then focused on genes which number increased dramatically in a subgroup- and/or species-specific manner (*i.e.* LSE genes). Those genes are likely to be involved in species-specific cellular processes or adaptive interactions and were used as a template to infer the potential occurrence of positive selection. This led to the identification of sites on which positive selection likely acted. We discuss our results in the light of angiosperm genome evolution and current knowledge of LRR-RLK
functions. Positive selection footprints identified in LSE genes highlight the importance of combining evolutionary analysis and functional knowledge to guide further investigations.
RESULTS

We extracted genes containing both LRRs and a KD from 33 published Embryophyte genomes. Here, we mostly describe the findings for the 31 angiosperm (eight monocot and 23 dicot) genomes we analyzed. The 7,554 LRR-RLK genes were classified in 20 subgroups (SGs). This classification was inferred using distance related methods because the high number of sequences to be analyzed would imply excessive computation time for methods relying on maximum likelihood. Since we decided to study the evolutionary dynamic of LRR-RLK gene family using the SG classification as a starting point, we first wanted to verify that each SG was monophyletic. Ten subsets of about 750 sequences were created by picking one sequence out of ten to infer a PHYML tree (data not shown). Analysis of the trees shows that most SGs (14) are monophyletic with strong branch support. On the other hand, for six SGs (SG_I, SG_III, SG_VI, SG_Xb, SG_XI and SG_XV), the topology differs slightly between trees: in at least five trees out of ten, either the SG appears to be paraphyletic or few sequences are placed outside the main monophyletic clade with low branch support. As we could not confirm that these SG are monophyletic, they were tagged with a "*" throughout the manuscript.

Next, we determined the number of ancestral genes present in the last common ancestor of angiosperms (LCAA) using a tree reconciliation approach (see Materials and Methods). In short, tree reconciliation compares each SG-specific LRR-RLK gene tree to the species tree to infer gene duplications and losses. Note that since only LRR-RLKs with at least one complete LRR were considered, some of the inferred gene losses might correspond to RLKs without, or with degenerated LRRs. Using this method, we predicted the number of LRR-RLK genes in the LCAA to be 150. All SGs were present in the LCAA but the number of genes between SGs was highly variable (Table I). SG_III* and SG_XI* show the highest number of ancestral genes, with 32 and 29 genes, respectively. The lowest numbers of ancestral genes are recorded for SG_VIIb, SG_Xa, SG_XIIIa, and SG_XIIIb which only possessed two genes and SG_XIV which only contained one. These results show that already in the LCAA that lived ~150 million years ago (Supplemental Table SI) some SGs were more prone to retain copies than others. We wanted to determine if this ancestral pattern was preserved during the course of angiosperm evolution and if different SGs expanded or contracted compared to the LCAA.

Expansion rates of LRR-RLK genes differ between subgroups and species

To gain a more comprehensive understanding of LRR-RLK evolution, we first looked at SG-specific expansion rates in two complementary ways. First, we calculated the global SG
expansion rate (= ratio of contemporary LRR-RLK genes per species in one SG divided by the ancestral number) for each SG (Fig. 1). Second, we inferred the branch-specific expansion rate of each SG on the phylogenetic tree of the 31 angiosperm species. We did this by automatically computing the ratio of descendant LRR-RLKs divided by the ancestral number of LRR-RLKs at every node (see Materials and Methods) (Fig. 2). Looking at the global SG expansion, we found that SG_Xa, SG_XIIa, SG_XIIb, and SG_XIV expanded more than 2-fold on average, and SG_I* and SG_IX around 2-fold (Fig. 1, Supplemental Table SII). Interestingly, SG_XIIa already had a moderate-high ancestral gene number (nine) and therefore seems to be generally prone to high retention rates. Indeed, SG_XIIa was subject to
repeated rounds of major expansion events (i.e. expansion > 2-fold) during its evolutionary history, e.g. in Poaceae, the Solanum ancestor, Malvaceae, and the Arabidopsis ancestor; but also species-specific expansions, e.g. in THECC, GOSRA, ARALY, SCHPA, MALDO, LOTJA, POPTR, and JATCU (Fig. 2, see Table II for five-digit species code). On the other hand, SG_I* and SG_XIIb had a medium number of copies in the ancestral genome (seven and four, respectively) but the pattern of expansion is quite different when analyzed in detail (Fig. 2). For SG_I*, the expansion rate is mostly due to ancestral expansion events rather than species-specific ones. For example, the high number of copies in ARATH and EUTSA (Fig. 1) is not due to expansions specific to these species but rather an expansion in Brassicaceae.
Subsequently, copies were lost in the other species of this family analyzed here (ARALY, SCHPA, BRARA) but retained in ARATH and EUTSA (Fig. 2). Species-specific expansions can also be observed in SG_I*, mostly in PRUPE and POPTER. For SG_XIIb, on the other hand, the high expansion rate is mostly due to recent species-specific expansions in PHODC, MUSAC, VITVI, GOSRA, MALDO, POPTER, and JATCU. But one major ancestral expansion can be observed in Rosids.

SG IX, SG Xa, and SG XIV had only few copies in the LCAA (three, two, and one, respectively) and all show a relatively high global expansion rate (Fig. 1). For these subgroups also, a contrasted branch-specific expansion patterns can be observed (Fig. 2). SG Xa went through relatively few major expansions: one can be detected in the dicots ancestor and a species-specific one in POPTER. Likewise, SG IX shows only one ancestral expansion in Malvaceae but more species-specific expansions in PHODC, MUSAC, MALDO, and GLYMA. Finally, SG XIV went through several rounds of ancestral (monocots, dicots, Malvaceae, and Brassicaceae) as well as species-specific expansion (PHODC, MALDO, and POPTER). The other SGs show a moderate expansion rate (1.3-1.75) or no expansion at all (Fig. 1, Supplemental Table SII). SG XV* is the only SG for which the number of copies was decreasing on average compared to the LCAA genome (0.77). It is important to note that the LCAA ancestral gene number could have been slightly over-estimated for those SGs without a confirmed monophyletic origin (denoted by “*”), resulting in an under-estimation of global expansion rate. However, we re-calculated the global expansion rates for each of those SGs using the largest subset of sequence that always include a stable monophyletic clade. The obtained global expansion rate differed only slightly from the ones presented here (data not shown) and the conclusions drawn remain unchanged.

Because some species underwent whole genome duplication (WGD) or whole genome triplication (WGT) relatively recently compared to others (Table III), we determined species-specific patterns of LRR-RLK expansions and looked if those patterns are consistent with the recent history of the species. Therefore, we computed the global species expansion rate (= ratio of LRR-RLK genes per SG in one species divided by the ancestral number) for each of the 31 angiosperm species. As expected, the global expansion rate differs significantly between species (Fig. 3, Supplemental Table SII). Compared to the LCAA (150 genes), the number of LRR-RLK genes did not decrease for most species except for LOTJA (114) and CARPA (127). This indicates that, on average, LRR-RLK genes are more prone to retention than loss. Some species, however, did not significantly expand their average number of LRR-RLK genes compared to the common ancestor: PHODC (158), CUCME (149), CUCSA
(180), SCHPA (194), BRARA (185), MEDTR (183), and RICCO (182). LRR-RLK genes expanded more than 2-fold in GLYMA (477), MALDO (441), POPTR (400), and GOSRA (372), and around 2-fold in MUSAC (280), MAIZE (241), SETIT (301), ORYSJ (317), ORYSI (301), SOLTU (254), PRUPE (260), MANES (238), and EUTSA (240). The remaining species show a moderate expansion rate (1.4-1.75): CACJA (222), THECC (238), JATCU (208), ARATH (222), SOLLCC (232), SORBI (225), BRADI (225), VITVI (193), and
ARALY (195). As expected, the four species with the highest global expansion rate (GLYMA, MALDO, POPTR, GOSRA) are recent polyploids in which most SGs have expanded (Fig. 2). However, some SGs expanded more than 2-fold, indicating that small scale duplication events have occurred in addition to polyploidy. In POPTR, for instance, the global expansion rates of SG_Xa and SG_XIIb are more than 8-fold (Fig. 3) and a strong branch-specific expansion rate is detected on the terminal POPTR branch (3.25 for SG_Xa and 5.4
for SG_XIIb) (Fig. 2). Surprisingly, SG_VIIa and SG_VIIb show a high branch-specific expansion rate in POPTR (4.0 and 3.0, respectively), which is not reflected in the global expansion rate in this species (Fig. 3). This is due to the fact that SG_VIIa and SG_VIIb went through strong reduction in Malpighiales (0.33) and Fabids (0.5), respectively. Thus, the cumulative effect of successive reductions and expansions is not evident in the global expansion rate. These contrasted evolutionary dynamics can also be observed in MALDO. A
global expansion of SG_IX was not detected because of the strong reduction in
Amygdaloideae. To summarize, this data can be integrated into the species phylogeny to draw
an image of the complex evolutionary dynamics of the LRR-RLK gene family through time
(Fig. 4).

**Different patterns of lineage specific expansion in LRR-RLK subgroups**

Given the differences of LRR-RLK expansion rates between species, we wanted to identify
cases of LSE, *i.e.* cases where a high duplication/retention rate is specific to one species.
Using a tree reconciliation approach (see Materials and Methods), we built a dataset
consisting of ultraparalog clusters (UP – only related by duplication) which represents the
LSE events and a superortholog reference gene set (SO – only related by speciation). We only
considered clusters containing five or more sequences. After cleaning, our final dataset
comprised 75 UP and 189 SO clusters containing 796 and 1,970 sequences, respectively
(Table IV). The median number of sequences in the UP clusters is not significantly different
from the median number in SO clusters (8 in both cases) (Supplemental Fig. S1). For UP
clusters, however, the alignments are significantly longer (Mann-Whitney test: \( p < 0.001 \)) with
a median of 3,237 base pairs (bp) compared to 2,841 for SO clusters. One possible
explanation for this could be that UP clusters are more dynamic and might contain more
LRRs. PRANK, the alignment algorithm we used, introduces gaps instead of aligning
ambiguous sites and therefore produces longer alignments when sequences are divergent.
However, this phenomenon does not influence the outcome of further test for positive
selection using codeml (Yang, 2007).

We then wanted to determine which SGs are represented in the SO and UP dataset.
Unsurprisingly, all SGs were present in SO clusters (Fig. 5). This could be expected as all
SGs were already present in the LCAA and remained stable or expanded (except SG_XV*).
In general, the frequency of SO clusters (and sequences) for each SG reflects the number of
copies in the LCAA (Table I, Fig. 5). On the other hand, only eleven of the 20 subgroups
were represented in UP clusters (SG_I*, SG_III*, SG_VI*, SG_VIII-2, SG_IX, SG_Xa,
SG_Xb*, SG_XI*, SG_XIIa, SG_XIIb, SG_XIIIa) and these SGs harbor a total of 837
sequences. SG_I*, SG_VIII-2, SG_XIIa, and SG_XIIb are clearly over-represented which is
in accordance with their expansion pattern. Other expanded SGs, however, have only a low
number of UP clusters or – in the case of SG_IV – no UP clusters at all. Therefore, it seems
that recently duplicated genes are more prone to be retained in some SGs.
Differences of selective constraint between subgroups, domains, and amino acids

To provide further insight into the LRR-RLK gene family evolution, we wanted to determine under which kind of selective pressures the LRR-RLK genes evolved. We focused on the dataset described above, *i.e.* LSE and orthologous genes. We inferred the $d_N/d_S$ -ratio (or $\omega$, *i.e.* the ratio of non-synonymous vs. synonymous substitutions rates) at codons of the alignments and branches of the phylogeny of the UP and SO clusters. An $\omega=1$ indicates neutral evolution/relaxed constraint, an $\omega<1$ indicates purifying selection, and an $\omega>1$ can indicate positive selection. We used mapNH ([Dutheil et al., 2012; Romiguier et al., 2012](#)) to compute the $\omega$ for each branch. mapNH ran for 71 UP and 176 SO clusters containing 1,246 and 2,960 branches, respectively (Table IV). We first wanted to test for relaxation of selective constraint in UP and SO clusters and looked for branches with $\omega>1$. We found 6.04% of UP branches but only 0.49% of SO branches to have an $\omega>1$. The mean $\omega$ for branches with $\omega>1$ is significantly larger in UP clusters (1.45) compared to SO clusters (1.13, $p=0.004$). The same is true for branches with $\omega<1$ where $\omega$ is significantly larger in UP clusters (0.48) compared to SO clusters (0.24, $p<0.001$). Overall, the mean $\omega$ is significantly larger for branches from UP clusters (0.54) than for SO clusters (0.24, $p<0.001$) (Table IV, Supplemental Fig. S2).

We found 38 out of 75 UP clusters (= 50.67%) containing codons under positive selection (see Supplemental Table SIII for more details) after manual curation but only six out of 186 SO clusters (= 3.23%). Additionally, codons under positive selection found in UP
clusters are not distributed evenly over domains (Fig. 6). To account for differences in domain size, a hit frequency, *i.e.*, the number of sites under positive selection we found relative to all sites possible for each domain, was calculated (see Materials and Methods). The domain showing the highest hit frequency is the LRR domain, followed by the cysteine-pairs and their
flanking regions (Fig. 6A). Hits in both domains are distributed over all SGs and species tested. The KD and its surrounding domains contain very few codons under positive selection. Domains classified as “other” combine domains important for the function of the LRR-RLK genes but vary between SGs. For example, SG_I* (Fig. 6B) contains a malectin domain. All hits classified as “other” here fall in the malectin-like domain (MLD) of a POPTR SG_I* cluster.

Finally, we wanted to investigate whether some AAs in the LRR are more frequently targeted by positive selection. The LRR typically contains 24 AAs and sometimes islands between them (Fig. 6C). Four AAs were predominantly subject to positive selection: 6, 8, 10, and 11 which all lie in the LRR-characteristic LXXLXLXX β-sheet/β-turn structure.
DISCUSSION

We studied the SG- and species-specific expansion dynamics in LRR-RLK genes from 31 angiosperm genomes in a phylogenetic framework. We also analyzed the lineage-specifically expanded genes in this family to determine to which extent positive selection occurred on them using a $d_{NS}/d_S$-based test. We found differences in expansion patterns depending on SGs and species but only a few SGs that were subjected to LSE. A significantly higher proportion of LSE LRR-RLK genes was affected by positive selection compared to single-copy genes and the LRR domain (specifically four AAs within this domain) were targeted by positive selection. In the following, we will discuss our findings in more detail.

Subgroup- and species-specific expansions

We observed significant variations in the global expansion rates between LRR-RLK SGs. These are due to a complex history of expansion-retention-loss cycles that are specific to each SG. The phylogenetic approach allowed us to determine the relative importance of ancestral versus recent species-specific expansions for each SG and to characterize precisely the loss/retention dynamics along evolutionary history of the studied species (summarized in Fig. 4). For example, SG_III* and SG_XI* had a high copy number of LRR-RLK in the LCAA and kept a stable copy number over the last 150 million years. On the other hand, SG_I*, SG_XIIa, and SGXIIb, which had a moderate copy number in the LCAA, keep expanding. Some functions have been described for genes of these SGs, mainly in *A.thaliana* (Supplemental Table SIV). For SG_III* and SG_XI*, mostly genes involved in development are described. The high numbers of ancestral genes in these two SGs combined with their size stability during angiosperm evolution may be interpreted as an early high level of diversification/specialization of these genes which are needed to orchestrate common developmental features. This hypothesis can be reinforced by the high number of superorthologous genes in these subgroups. For SG_I* and SG_XIIa, on the other hand, mostly genes involved in response to biotic stress are described up to now. These observations confirm that different expansion/retention patterns appear to be related to gene function although one has to keep in mind that functions have only been assigned to few LRR-RLK genes. Three SGs (SG_IX, SG_Xa, and SG_XIV) expanded compared to their very low ancestral number (1-3) leading to a high total expansion rate. As it has been postulated that duplications are the raw material for adaptation (Nei and Rooney, 2005; Fischer et al., 2014), the evolution of those SGs was likely driven by adaptation – to varying degrees in different angiosperm species, depending on the environment they evolved in. The known functions are
both related to response to biotic or abiotic stress and development. Because so far our knowledge of LRR-RLK functions is limited and mostly restricted to *A. thaliana*, further studies are needed to make more reliable statements on the link between function and expansion/retention dynamics in different SGs.

Next, we wanted to ascertain species-specific expansions of LRR-RLK genes and how they are related to the recent history of the species in our study. Whole genome multiplication has been argued to be a major force in diversification of angiosperms (Soltis et al., 2009; Soltis and Burleigh, 2009; Renny-Byfield and Wendel, 2014). All angiosperms share two ancient WGDs (Jiao et al., 2011). Likewise, all monocots share a WGD ~130 Mya (Tang et al., 2010a) and most dicots (Eudicots) share a WGT around the same time (Jaillon et al., 2007; Wang et al., 2012), but more recent WGDs and WGTs occurred in many angiosperm species (Fig. 4, Table III). The link between WGD/Ts and the number of LRR-RLK genes is not straightforward. We found that in *Glycine max*, *Gossypium raimondii*, and *Malus x domestica*, which were subject to relatively recent WGDs (15-13, 17-13 and 45-30 Mya, respectively) (Pfeil et al., 2005; Velasco et al., 2010; Wang et al., 2012), the number of LRR-RLK genes expanded more than 2-fold compared to the LCAA. These results are in accordance with what was already described for these species. Indeed, it was found that *G. max* contains a very large number of retained genes from this WGD (Cannon et al., 2014). Additionally, recent studies on large gene families in *G. raimondii* indicate that their copy number is either driven by retention after the last WGD (*e.g.* NAC transcription factors) (Shang et al., 2013) or a combination of segmental (SD) and tandem duplications (TD) (*e.g.* WRKY transcription factors) (Dou et al., 2014). For *M. domestica* (most recent WDG after the divergence for peach according to Verde et al. (2013)), a recent study on nucleotide binding site (NBS) LRR genes showed that they also stem mostly from SDs and TDs (Arya et al., 2014).

More contrasted results are observed in Brassicaceae where two WGDs occurred (Barker et al., 2009; Fawcett et al., 2009). Most SGs expand their number of genes on this ancestral branch, but the species belonging to this clade mostly retain or loose genes on average (Fig. 2 and 4). The only exception concerns *Eutrema salsugineum* (an *A. thaliana* relative) which is the only species with a more than 2-fold average expansion rate. The global expansion rate in *E. salsugineum* is mostly due to two SGs (SG_I* and SG_XIIIa). In the original genome paper (Wu et al., 2012), the authors found that genes from the category “response to stimulus” (response to salt stress, osmotic stress, water deprivation, ABA stimulus, and hypoxia) are significantly over-represented in *E. salsugineum* compared to *A. thaliana*. This over-
representation is described as mostly caused by SDs and TDs (Wu et al., 2012) in accordance with what we observed in SG_XIIIa. This could be of functional importance to this halophile plant.

Finally, of all species analyzed here, Zea mays and Brassica rapa (and maybe Manihot esculenta) show the most recent cases of WGD/T (12-5 and 9-5 Mya, respectively) (Schnable et al., 2011; Wang et al., 2011) yet their expansion rates are moderate. This is further evidence for the dynamic nature of angiosperm genomes that has been discussed before (Leitch and Leitch, 2012; Fischer et al., 2014). After a WGD event, genomes tend to return to the diploid (or previous) state by losing redundant duplicated genes (fractionation process) – although the gene loss is biased (Bowers et al., 2003; Schnable et al., 2009). Which genes are lost or retained strongly depends on their function (De Smet et al., 2013). However, it has been shown that genes involved in stress response are mostly created by TDs rather than WGD (Hanada et al., 2008). Indeed, it was hypothesized before that RLK genes involved in stress response mostly duplicate by TD (Shiu et al., 2004). Here, we provide a detailed representation of expansion-retention-loss dynamics of the whole LRR-RLK gene family in 31 angiosperm species (Fig.4). Each new genome sequenced will improve the accuracy of the expansion-retention-loss event predictions and will help identifying new elements that can be useful for future functional analysis and/or linked to adaptive traits.

**Studying selection pressures in a large and dynamic gene family**

As described above, the composition of LRR-RLKs in each of the 31 studied angiosperm species results from a complex dynamic of species- and SG-specific expansion/loss events. To further investigate the potential role of this family in plant adaptation we analyzed to which selective pressures the LRR-RLKs were submitted. Such an analysis cannot be considered for the phylogeny of the entire gene family because of the high number of sequences and high sequence divergence (the phylogeny on which we divided the SGs was inferred on the conserved kinase domain only). We then chose to focus on two specific cases: (i) LSE as a specific case of duplication/retention, and (ii) a subset of strictly orthologous genes. Indeed, LSE has been shown to fuel adaptation in angiosperms (Fischer et al., 2014) and we wanted to test the prevalence of this mode of duplication in our large dataset. Therefore, we evaluated to which extant LRR-RLK genes were subject to LSE and how positive selection acted on those genes. As a reference, we chose the strictly orthologous subset. This approach allows the interpretation of LSE evolution compared to the general LRR-RLK selective background (Fischer et al., 2014).
The power of this phylogenetic approach relies on the number of species analyzed and we profit from an ever increasing number of sequenced plant genomes. Another important requirement for this approach is the quality of sequencing and annotation – especially for a large gene family – as sequencing errors and mis-annotations can lead to false positives when testing for positive selection (Han et al., 2013). We profit from a recently developed pipeline designed to automatically perform different steps of the analysis (Fischer et al., 2014). This allowed us to quickly incorporate sequenced genomes of choice and future studies can easily expand this analysis as new reliable data becomes available. Finally, we set great value on manually verifying the data throughout the process – from the identification of the LRR-RLKs to the inference of positive selection. Although this is tedious work for such a large dataset, it is nevertheless important. As we recently showed, ~50% automatically reported instances of positive selection turned out to be false positives after manual curation (Fischer et al., 2014).

We found that all SGs are represented in the single-copy reference set with an over-representation of SG_III* and SG_XI*. This is in accordance with the fact that these two SGs had the highest number of copies in the genome of the LCAA and did not significantly expand since (see above). In general, the frequency of clusters from the single-copy gene set (and sequences) for each SG reflects the number of copies in the LCAA (Table I, Fig. 5). On the other hand, only eleven of the 20 SGs were represented in the LSE dataset. This is mainly because the majority of expansions are rather old in these SGs, whereas they happened relatively recently in SG_I*, SG_VIII-2, SG_XIIa, and SG_XIIb (see above). Fourteen species (or clades) are represented in the LSE dataset: MUSAC (2 ultraparalog clusters), SETIT (1), ORYZA (10), VITVI (3), SOLAN (6), MEDTR (3), GLYMA (2), PRUPE (6), MALDO (11), POPTR (8), BRASS (11), GOSRA (5), THECC (2), and PHYPA (5). Again, not every species is affected to the same extent, but this does not necessarily reflect recent WGD/T. Additionally, LSE can also arise from SD and TD which frequency of occurrence is not uniform within or between genomes. Our results indicate that different species are more prone to retain recently duplicated genes than others. This in turn might reflect on their recent evolution or domestication which should be examined in more detail in future studies.

When focusing on the study of selective pressures, we first looked at $\omega$ at the branches of the LSE and single-copy gene clusters and found that selective constraint was relaxed in the LSE dataset. This outcome was expected as it was previously shown that LSE genes evolve more relaxed constraint than single-copy genes in angiosperms (Fischer et al., 2014). This study, however, looked at whole angiosperm genomes but a similar pattern has already been
demonstrated in other large gene families (e.g. Johnson and Thomas, 2007; Xue et al., 2012; Yang et al., 2013a; Yang et al., 2013b) and in LRR-RLK genes in particular (Tang et al., 2010b). Previous studies on that subject only had a limited dataset (four angiosperm species; Tang et al. 2010b). Here, we demonstrate that this is still true when a larger and more representative sample of angiosperms is considered.

Next, we wanted to identify codons which evolved under positive selection in the LSE and the single-copy dataset. A recent study on gene families in the whole genomes of ten angiosperms found that 5.4% of LSE genes contained codons showing positive selection footprints (Fischer et al., 2014). Here, we ask if and to what extend this is also true for the large and dynamic LRR-RLK gene family. We discovered that for LSE LRR-RLK genes, the rate of codons under selection is almost 10-fold higher (50.67%) than the genome average. In addition, we found >3% of single copy genes containing codons under selection whereas Fischer et al. (2014) described no case of positive selection at the single-copy gene clusters in their study. Together with the high rate of branches with \( \omega > 1 \) in LSE gene clusters (6.04%, compared to 0.49% for single-copy genes) this indicates that LRR-RLK genes are more prone to evolve under positive selection than the average of angiosperm gene families. As it might be expected, all ultraparalog clusters with codons under positive selection come from the four over-represented SGs: SG_I* (1 ultraparalog cluster), SG_VIII-2 (3), SG_XIIa (24), and SG_XIIb (10). The single-copy gene clusters with codons under selection come from six SGs: SG_III, SG_VIIa, SG_Xa, SG_Xb, SG_XIIa, and SG_XIIb. Therefore, recent expansion and retention only affect a few SGs but in those SGs positive selection plays an important role. For SG_XIIa, positive selection has been already previously inferred for genes involved in environmental interactions: \( \text{Xa21} \), which confers resistance to the bacterial blight disease, was found to have evolved under positive selection in rice (Wang et al., 1998; Tan et al., 2011); and FLS2, involved also in response to biotic stress, shows a signature of rapid fixation of an adaptive allele in Arabidopsis (Vetter et al., 2012). Future studies on smaller subsets of SGs will surely cast further light on selection patterns in LRR-RLK genes. Only eleven species (or clades) are represented in the LSE dataset with codons under positive selection: SETIT (1 ultraparalog cluster), ORYZA (2), SOLAN (4), MEDTR (2), GLYMA (2), PRUPE (3), MALDO (8), POPTRO (7), BRASS (2), GOSRA (5), and THECC (2). Not every species is affected the same level by positive selection and again future studies might bring more details concerning the evolutionary history of specific species and SGs to light.

In addition, we found that not every domain of the LRR-RLK genes was similarly affected by positive selection. Most codons under selection fall in the LRR domain. This
outcome might be expected as LRRs are very dynamic and plasticity in this region provides plants with a broad toolset to face environmental challenges and therefore undergoes positive selection frequently (Zhang et al., 2006; Tang et al., 2010b). Only very few codons under positive selection were found in the kinase domain and its surrounding regions. This result is consistent with the fact that the KD is very conserved among species and SGs and evolved mostly under purifying selection (Shiu et al., 2004; Tang et al., 2010b). A more surprising result was the identification of a significant number of positively selected sites in the malectin-like domain of a *Populus trichocarpa* SG_I* cluster. So far, the function of extracellular malectin-like domains of RLKs is not well understood (Lindner et al., 2012). However, a malectin-like domain-containing SG_I* LRR-RLK has been described to confer susceptibility to a downy mildew pathogen in *A. thaliana* and to have similarities to Symbiosis RLKs (SYMRKs) which are important for the regulation of bacterial symbiont accommodation (Markmann et al., 2008; Hok et al., 2011). Therefore, our results suggest that it could be interesting to further investigate the function and evolutionary history of this SG_I* domain, particularly in *P. trichocarpa*. Another unexpected finding was the frequent occurrence of positive selection at the cys-pairs and flanking regions which are involved in folding and/or the binding to other proteins. To what extent the function of LRR-RLKs is affected by mutations in the cys-pair regions depends on the function of the gene (Song et al., 2010; Sun et al., 2012) and it would be interesting to study this in more detail in the future.

Finally, we took a closer look at the AAs in the LRR primarily affected by positive selection. Only four, out of the 24 AAs a LRR typically contains, were predominantly and strongly subject to positive selection. These variable AAs lie in the un-conserved part of the LRR-characteristic LXXLXLXX β-sheet/β-turn structure which is involved in protein-protein interactions (Jones and Jones, 1997; Enkhbayar et al., 2004). Specifically, solvent-exposed residues were targeted by positive selection (Parniske et al., 1997; Wang et al., 1998). Further investigation on the functional consequences of these nucleotide variations need to be done to confirm their adaptive potential but our findings align very well to the current understanding of LRR ligand binding. Taken together, our results could be very useful for further functional investigations of LRR-RLK genes in different species.

**CONCLUSIONS**

We studied LRR-RLK genes from 33 land plant species to investigate SG- and species-specific expansion of these genes, to which extent they were subject to LSE, and to determine
the role positive selection played in the evolution of this large gene family. We described that some SGs are more prone to expansion/retention than others and that the expansions occurred at different times in the evolution of LRR-RLK genes. This fine-scale analysis of the dynamic allowed us to identify branches and species for which a higher than average retention rate could indicate a potential adaptive event for some SGs. We also described that only a few SGs show patterns of recent LSE and that at those genes selective constraint is relaxed. More than 50% of the LSE genes contain codons which show evidence for positive selection which is almost 10-fold the frequency previously described at gene families in angiosperms (Fischer et al., 2014). Finally, we found that over the LRR-RLK genes the LRR domain and specifically four AAs responsible for ligand interaction are most frequently subject to selection.

MATERIALS AND METHODS

Studied Genomes

We analyzed 31 angiosperm genomes (eight monocot (sub)species and 23 dicot species) (see Table II): Phoenix dactylifera (Al-Dous et al., 2011), Musa acuminata (D'Hont et al., 2012), Oryza sativa subsp. japonica (International Rice Genome Sequencing Project, 2005), Oryza sativa subsp. indica (Yu et al., 2002), Brachypodium distachyon (The International Brachypodium Initiative, 2010), Zea mays (Schnable et al., 2009), Sorghum bicolor (Paterson et al., 2009), Setaria italica (Zhang et al., 2012), Solanum tuberosum (Potato Genome Sequencing Consortium, 2011), Solanum lycopersicum (The Tomato Genome Consortium, 2012), Vitis vinifera (Jaillon et al., 2007), Lotus japonicus (Sato et al., 2008), Cajanus cajan (Varshney et al., 2012), Arabidopsis thaliana (The Arabidopsis Genome Initiative, 2000), Arabidopsis lyrata (Hu et al., 2011), Schrenkiella parvula (a synonym is Eutrema parvula, we used the nomenclature from Oh et al. (2014)) (Dassanayake et al., 2011), Eutrema salsugineum (a synonym is Thellungiella halophila, we chose the nomenclature according to phytosome: http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Esalsugineum) (Wu et al., 2012), Brassica rapa (Wang et al., 2011), Populus trichocarpa (Tuskan et al., 2006), Glycine max (Schmutz et al., 2010), Medicago truncatula (Young et al., 2011), Prunus persica (Ahmad et al., 2011), Malus x domestica (Velasco et al., 2010), Ricinus communis (Chan et al., 2010), Jatropha curcas (Sato et al., 2011), Manihot esculenta (Prochnik et al., 2012), Cucumis sativus (Huang et al., 2009), Cucumis melo (Garcia-Mas et al., 2012), Carica papaya (Ming et al., 2008), Gossypium raimondii (Wang et al., 2012), and Theobroma cacao (Argout et al., 2011). We also extracted LRR-RLKs from the moss Physcomitrella patens.
(Rensing et al., 2008) and the spikemoss Selaginella moellendorfii (Banks et al., 2011). Throughout the article we refer to the species using five-digit identifiers which can be found in Table II. Altogether, we analyzed 33 genomes from 39 proteomes (we used several annotation versions of the A. thaliana and O. sativa genomes). A table containing the details on which genome versions we used can be found in Supplemental Table SV. The phylogeny of those species is provided in Fig. 4.

**LRR-RLK extraction, clustering, phylogeny, and identification of gain/loss events**

We used the hmmsearch program (Eddy, 2009) to extract peptide sequences containing both, intact (i.e. non-degenerated) LRR(s) and a KD from the proteomes as previously described by (Dievart et al., 2011). We classified SGs using the KD by a global phylogenetic analysis (the tree can be found at http://phylogeny.southgreen.fr/kinase/index.php - Global Analysis). First, sequences were aligned using MAFFT (Katoh et al., 2005) with a progressive strategy. Second, the alignments were cleaned using TrimAl (Capella-Gutiérrez et al., 2009) with settings to remove every site with more than 20% of gaps or with a similarity score lower than 0.001. Third, a similarity matrix was computed by ProtDist (Felsenstein, 1993) using a JTT model. Fourth, a global distance phylogeny was inferred using FastME (Desper and Gascuel, 2006) with default settings and SPR movements to optimize the tree topology. Fifth, SGs were defined manually in the global phylogeny using the Arabidopsis genes as reference which lead us to 20 SGs in contrast to the 15 previously described (Shiu et al., 2004; Lehti-Shiu et al., 2009).

More accurate phylogenies were then inferred for each of the 20 SGs. The KD of the sequences attributed to each SG were re-aligned using MAFFT with an iterative strategy (maximum of 100 iterations). Alignments were cleaned using TrimAl with settings to only remove sites with more than 80% of gaps. Then, maximum likelihood phylogenies were inferred by PhyML 3.0 (Guindon et al., 2010) using a LG+gamma model and the best of NNI and SPR topology optimization. Statistical branch support was computed using the aLRT/SH-like strategy (Guindon et al. 2010). This left us with 20 phylogenies, one for each SG (all phylogenies are available at http://phylogeny.southgreen.fr/kinase/index.php - SG_I - SG_XV).

Each of the 20 phylogenetic trees has been reconciled with the species tree using RAP-Green (Dufayard et al., 2005; https://github.com/SouthGreenPlatform/rap-green). By comparing the gene tree to the species tree, this analysis allows to root phylogenetic trees and to infer duplication and loss events (Dufayard et al., 2005). We tested this approach of rooting...
(by minimizing the number of inferred duplications and losses) and compared it to rooting with outgroups (data not shown). The two methods provided very close root locations that did not change the overall conclusions. Using this RAP-Green tree reconciliation approach (parameters: Maximum support for reduction 0.95), we inferred the number of duplications and losses at each node of the species tree. Briefly, each duplication and loss respectively increases and decreases by one the number of copies in the common ancestor of the taxonomic group analyzed.

We determined the global SG- and species-specific expansion rate by computing the number of LRR-RLK genes in one SG divided by the ancestral number and number of LRR-RLK genes in one species divided by the ancestral number, respectively. An ANOVA analysis showed that the expansion rate differed significantly between the SGs/species (p<2e-16 in both cases). We used the TukeyHSD test of the agricolae package (http://cran.r-project.org/web/packages/agricolae/index.html) in R (R Development Core Team, 2012) to further explore which groups of SGs/species differ from each other. This test compares the range of sample means and defines an Honest Significance Difference (HSD) value which is the minimum distance between groups to be considered statistically significant. In short, TukeyHSD is a post-hoc test which groups subsets by significance levels after ANOVA showed significant differences between subsets.

**LSE dataset and testing for positive selection**

Testing for adaptation can be done by comparing positive (Darwinian) selection footprints in lineages with recently and specifically duplicated genes to reference lineages containing only single-copy genes. One way to infer positive selection is by analyzing nucleotide substitution data at the codon level in a phylogenetic framework. As nucleotide substitutions can either be nonsynonymous (i.e. protein changing, thereby potentially impacting the fitness) or synonymous (i.e. not protein changing, thereby theoretically without consequences for the fitness, but see (Lawrie et al., 2013)), the nonsynonymous/synonymous substitution rate ratio, denoted as \( d_{NS}/d_S \) or \( \omega \), can be used to infer the direction and strength of natural selection. An \( \omega<1 \) indicates purifying selection and the closer \( \omega \) is to 0, the stronger purifying selection is acting. Under neutral evolution, \( \omega=1 \). An \( \omega>1 \) indicates that positive selection is acting.

We identified ultraparalog clusters (UP – only related by duplication) using a tree reconciliation approach (Dufayard et al., 2005). Those represent our LSE gene set. As a single-copy gene reference, we chose a superortholog gene set (SO – only related by speciation). We chose clusters with a minimum of five sequences. To address the question of
whether or not positive selection is more frequent after LSE events, we compared the results obtained on UPs with those obtained on SO gene sets. Species which diverged <15 Mya were merged for the LSE detection (see Fig. 4) in order not to overly reduce the ultraparalog (UP) dataset and to not induce bias due to very recent speciation events: ANDRO (= ZEAMA and SORBI), ORYZA (= ORYSJ and ORYSI), SOLAN (= SOLLC and SOLTU), CUCUM (= CUCSA and CUCME), BRASS (= ARATH, ARALY, BRARA, SCHPA and EUTSA). We then applied the pipeline developed by Fischer et al. (2014) to the extracted UP and SO clusters. In short, the pipeline consists of following steps: (i) The clusters were aligned using PRANK+F with codon option (Löytynoja and Goldman, 2005). The alignments were cleaned by GUIDANCE (Penn et al., 2010) with the default sequence quality cut-off and a column cut-off of 0.97 to remove problematic sequences and unreliable sites from the alignments. We used PRANK and GUIDANCE here as previous benchmarks (Fletcher and Yang, 2010; Jordan and Goldman, 2012) showed that these programs lead to a minimum of false positives when inferring positive selection using codeml. The cleaned alignments can be retrieved here: http://phylogeny.southgreen.fr/kinase/alignments.php – Alignments: Manually curated alignments for positive selection analysis. (ii) We relied on the egglib package (De Mita and Siol, 2012) to infer the maximum likelihood phylogeny at the nucleotide level for every alignment using PhyML 3.0 (Guindon et al., 2010) under the GTR substitution model. (iii) We ran the codeml site model implemented in the PAML4 software (Yang, 2007) to infer positive selection on codons under several substitution models. In clusters identified to have evolved under positive selection, Bayes empirical Bayes was used to calculate the posterior probabilities at each codon and detect those under positive selection (i.e. those with a posterior probability of $\omega>1$ strictly above 95%). All alignments detected to be under positive selection at the codon level were curated manually for potential alignment errors. A table containing the details on all codons showing a signal of positive selection using codeml can be found in Supplemental Table SIII. (iv) We used mapNH (Dutheil et al., 2012; Romiguier et al., 2012) to infer $\omega$ at the branch level.

In order to analyze the distribution of positively selected sites among domains, we calculated a hit frequency that computes the number of sites under positive selection found in each domain relative to all sites possible. All possible sites for each domain were calculated as follows: First, we extracted the size of each domain of every SG. If SGs were subdivided further we took the average size of each domain. Second, we multiplied the size of each domain by the number of UP clusters we found for each SG. For example: the LRR of SG_I* contains an average of 77 sites. We found 8 UP clusters for SG_I*. Therefore, the total
number of possible LRR sites for SG_1* is 77*8=616 sites. Third, we added the sites for each
domain up for all SGs.
SUPPLEMENTAL DATA

Supplemental Tables SI – SV can be obtained from:

http://phylogeny.southgreen.fr/kinase/index.php - Tables

**Supplemental Table SI:** Estimated divergence times and corresponding references for Fig. 4.

**Supplemental Table SII:** Results of the TukeyHSD test.

**Supplemental Table SIII:** Details on all codons showing a signal of positive selection using codeml.

**Supplemental Table SIV:** Arabidopsis LRR-RLK gene classification according to TAIR (https://www.arabidopsis.org/)

**Supplemental Table SV:** List of genomes used here, with name, link, and version of the genome fasta file.

**Supplemental Figure S1:** Summary of ultraparalog and superortholog cluster size and length.

Each dot represents (A) an ultraparalog or (B) a superortholog alignment. The histogram above the scatter plot represents the count of alignments for each cluster size (= number of sequences in alignment); the histogram right to the scatter plot represents the frequency of alignments for each alignment length.

**Supplemental Figure S2:** $\omega$ distribution of branches of UP and SO clusters.

The distribution of $\omega$ of branches in UP (purple) and SO clusters (yellow).

ACKNOWLEDGEMENTS

The authors thank the reviewers for their helpful comments.
**Table I:** Total number of LRR-RLK in our angiosperm dataset, number of ancestral genes in LCAA, and median global expansion rate for each SG among the 31 species.

<table>
<thead>
<tr>
<th>SG</th>
<th>Total number of genes</th>
<th>Number of ancestral genes</th>
<th>Median global expansion rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>I*</td>
<td>482</td>
<td>7</td>
<td>2.00</td>
</tr>
<tr>
<td>II</td>
<td>349</td>
<td>9</td>
<td>1.22</td>
</tr>
<tr>
<td>III*</td>
<td>1,400</td>
<td>32</td>
<td>1.22</td>
</tr>
<tr>
<td>IV</td>
<td>131</td>
<td>3</td>
<td>1.33</td>
</tr>
<tr>
<td>V</td>
<td>263</td>
<td>5</td>
<td>1.80</td>
</tr>
<tr>
<td>VI*</td>
<td>324</td>
<td>10</td>
<td>1.00</td>
</tr>
<tr>
<td>VIIa</td>
<td>157</td>
<td>3</td>
<td>1.67</td>
</tr>
<tr>
<td>VIIb</td>
<td>84</td>
<td>2</td>
<td>1.50</td>
</tr>
<tr>
<td>VIII-1</td>
<td>216</td>
<td>5</td>
<td>1.40</td>
</tr>
<tr>
<td>VIII-2</td>
<td>355</td>
<td>8</td>
<td>1.25</td>
</tr>
<tr>
<td>IX</td>
<td>193</td>
<td>3</td>
<td>1.67</td>
</tr>
<tr>
<td>Xa</td>
<td>143</td>
<td>2</td>
<td>2.00</td>
</tr>
<tr>
<td>Xb*</td>
<td>367</td>
<td>9</td>
<td>1.11</td>
</tr>
<tr>
<td>XI*</td>
<td>1,177</td>
<td>29</td>
<td>1.28</td>
</tr>
<tr>
<td>XIIa</td>
<td>1,126</td>
<td>9</td>
<td>3.00</td>
</tr>
<tr>
<td>XIIb</td>
<td>423</td>
<td>4</td>
<td>2.00</td>
</tr>
<tr>
<td>XIIIa</td>
<td>84</td>
<td>2</td>
<td>1.50</td>
</tr>
<tr>
<td>XIIIb</td>
<td>77</td>
<td>2</td>
<td>1.00</td>
</tr>
<tr>
<td>XIV</td>
<td>84</td>
<td>1</td>
<td>3.00</td>
</tr>
<tr>
<td>XV*</td>
<td>119</td>
<td>5</td>
<td>0.80</td>
</tr>
<tr>
<td>Total</td>
<td>7,554</td>
<td>150</td>
<td></td>
</tr>
</tbody>
</table>
### Table II: Five-digit code for each species.

<table>
<thead>
<tr>
<th>Species name</th>
<th>Common name</th>
<th>Five-digit code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phoenix dactylifera</td>
<td>Date palm</td>
<td>PHODC</td>
</tr>
<tr>
<td>Musa acuminata</td>
<td>Banana</td>
<td>MUSAC</td>
</tr>
<tr>
<td>Brachypodium distachyon</td>
<td>Purple false brome</td>
<td>BRADI</td>
</tr>
<tr>
<td>Oryza sativa ssp. japonica</td>
<td>Asian rice</td>
<td>ORYSJ</td>
</tr>
<tr>
<td>Oryza sativa ssp. indica</td>
<td>Indian rice</td>
<td>ORYSI</td>
</tr>
<tr>
<td>Setaria italicata</td>
<td>Foxtail millet</td>
<td>SETIT</td>
</tr>
<tr>
<td>Zea mays</td>
<td>Maize</td>
<td>MAIZE</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>Milo</td>
<td>SORBI</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>Potato</td>
<td>SOLTU</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>Tomato</td>
<td>SOLLIC</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>Common grape vine</td>
<td>VITVI</td>
</tr>
<tr>
<td>Theobroma cacao</td>
<td>Cacao tree</td>
<td>THECC</td>
</tr>
<tr>
<td>Gossypium raimondii</td>
<td>Cotton progenitor</td>
<td>GOSRA</td>
</tr>
<tr>
<td>Carica papaya</td>
<td>Papaya</td>
<td>CARPA</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>Thale cress</td>
<td>ARATH</td>
</tr>
<tr>
<td>Arabidopsis lyrata</td>
<td>Out-crossing ARATH relative</td>
<td>ARALY</td>
</tr>
<tr>
<td>Brassica rapa</td>
<td>Turnip</td>
<td>BRARA</td>
</tr>
<tr>
<td>Schrenkiella parvula</td>
<td>A saltwater cress</td>
<td>SCHPA</td>
</tr>
<tr>
<td>Eutrema salsugineum</td>
<td>A saltwater cress</td>
<td>EUTSA</td>
</tr>
<tr>
<td>Cucumis sativus</td>
<td>Cucumber</td>
<td>CUCSA</td>
</tr>
<tr>
<td>Cucumis melo</td>
<td>Melon</td>
<td>CUCME</td>
</tr>
<tr>
<td>Prunus persica</td>
<td>Peach</td>
<td>PRUPE</td>
</tr>
<tr>
<td>Malus x domestica</td>
<td>Apple</td>
<td>MALDO</td>
</tr>
<tr>
<td>Lotus japonicus</td>
<td></td>
<td>LOTJA</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>Barrel medic</td>
<td>MEDTR</td>
</tr>
<tr>
<td>Glycine max</td>
<td>Soybean</td>
<td>GLYMA</td>
</tr>
<tr>
<td>Cajanus cajan</td>
<td>Pigeon pea</td>
<td>CAJCA</td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>Black cottonwood</td>
<td>POPTR</td>
</tr>
<tr>
<td>Ricinus communis</td>
<td>Castor oil plant</td>
<td>RICCO</td>
</tr>
<tr>
<td>Jatropha curcas</td>
<td>Barbados nut</td>
<td>JATCU</td>
</tr>
<tr>
<td>Manihot esculenta</td>
<td>Cassava</td>
<td>MANES</td>
</tr>
<tr>
<td>Selaginella moellendorffii</td>
<td>A spikemoss</td>
<td>SELML</td>
</tr>
<tr>
<td>Physcomitrella patens</td>
<td>A moss</td>
<td>PHYP A</td>
</tr>
</tbody>
</table>
Table III: Polyploidy events

Estimated times of polyploidy event and corresponding references for Fig. 4.

<table>
<thead>
<tr>
<th>Event</th>
<th>Name</th>
<th>Reference</th>
<th>Age [million years]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Seed plant tetraploidy</td>
<td>(Jiao et al., 2011)</td>
<td>350-330</td>
</tr>
<tr>
<td>2</td>
<td>Angiosperm tetraploidy</td>
<td>(Jiao et al., 2011)</td>
<td>230-190</td>
</tr>
<tr>
<td>3</td>
<td>Monocot tetraploidy</td>
<td>(Tang et al., 2010a)</td>
<td>130</td>
</tr>
<tr>
<td>4</td>
<td>Date palm WGD</td>
<td>(D’Hont et al., 2012)</td>
<td>75-65 (?)</td>
</tr>
<tr>
<td>5</td>
<td>Banana gamma</td>
<td>(D’Hont et al., 2012)</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>Banana beta</td>
<td>(D’Hont et al., 2012)</td>
<td>65</td>
</tr>
<tr>
<td>7</td>
<td>Banana alpha</td>
<td>(D’Hont et al., 2012)</td>
<td>65</td>
</tr>
<tr>
<td>8</td>
<td>Grass tetraploidy B (sigma)</td>
<td>(D’Hont et al., 2012)</td>
<td>123-109</td>
</tr>
<tr>
<td>9</td>
<td>Grass tetraploidy (rho)</td>
<td>(Paterson et al., 2004)</td>
<td>70</td>
</tr>
<tr>
<td>10</td>
<td>Maize tetraploidy</td>
<td>(Schnable et al., 2011)</td>
<td>12-5</td>
</tr>
<tr>
<td>11</td>
<td>Eudicot hexaploidy (Arabidopsis gamma)</td>
<td>(Jaillon et al., 2007; Cenci et al., 2010; Wang et al., 2012)</td>
<td>150-120</td>
</tr>
<tr>
<td>12</td>
<td><em>Solanum</em> hexaploidy</td>
<td>(The Tomato Genome Consortium, 2012)</td>
<td>91-52</td>
</tr>
<tr>
<td>13</td>
<td>Papilionoid tetraploidy</td>
<td>(Pfeil et al., 2005)</td>
<td>55-54</td>
</tr>
<tr>
<td>14</td>
<td>Soybean tetraploidy</td>
<td>(Pfeil et al., 2005)</td>
<td>15-13</td>
</tr>
<tr>
<td>15</td>
<td>Apple tetraploidy</td>
<td>(Velasco et al., 2010; Verde et al., 2013)</td>
<td>45-30</td>
</tr>
<tr>
<td>16</td>
<td>Poplar tetraploidy</td>
<td>(Tuskan et al., 2006)</td>
<td>65-60</td>
</tr>
<tr>
<td>17</td>
<td>Arabidopsis beta</td>
<td>(Fawcett et al., 2009)</td>
<td>70-40</td>
</tr>
<tr>
<td>18</td>
<td>Arabidopsis alpha</td>
<td>(Barker et al., 2009)</td>
<td>23</td>
</tr>
<tr>
<td>19</td>
<td>Brassica hexaploidy</td>
<td>(Wang et al., 2011)</td>
<td>9-5</td>
</tr>
<tr>
<td>20</td>
<td>Cotton WGD</td>
<td>(Wang et al., 2012)</td>
<td>20-13</td>
</tr>
<tr>
<td>21</td>
<td>Cassava WGD</td>
<td>(Mühlhausen and Kollmar, 2013)</td>
<td>? (after Crotonoideae split)</td>
</tr>
</tbody>
</table>
**Table IV:** Details of the LSE and mapNH analysis for UP and SO clusters.

<table>
<thead>
<tr>
<th></th>
<th>UP</th>
<th>SO</th>
</tr>
</thead>
<tbody>
<tr>
<td>total number of clusters</td>
<td>75</td>
<td>189</td>
</tr>
<tr>
<td>clusters for final mapNH analysis</td>
<td>71</td>
<td>176</td>
</tr>
<tr>
<td>median cluster size (1st; 3rd Qu)</td>
<td>8 (6; 12)</td>
<td>8 (6; 14)</td>
</tr>
<tr>
<td>min; max cluster size</td>
<td>5; 38</td>
<td>5; 25</td>
</tr>
<tr>
<td>median alignment length (1st; 3rd Qu)</td>
<td>3,237</td>
<td>2,841</td>
</tr>
<tr>
<td></td>
<td>(2,952; 3,574)</td>
<td>(2,034; 3,192)</td>
</tr>
<tr>
<td>min; max alignment length</td>
<td>1,749; 8,691</td>
<td>861; 6,216</td>
</tr>
<tr>
<td>branches analyzed/total number of</td>
<td>1,193/1,246</td>
<td>2,860/2,960</td>
</tr>
<tr>
<td>branches</td>
<td></td>
<td></td>
</tr>
<tr>
<td>clusters with branches omega &gt; 1.0 (%)</td>
<td>25 (35.21)</td>
<td>10 (5.68)</td>
</tr>
<tr>
<td>branches with omega &lt; 1.0 (%)</td>
<td>1,121 (93.96)</td>
<td>2,846 (99.51)</td>
</tr>
<tr>
<td>mean omega for &lt; 1.0 branches +/- sd</td>
<td>0.48 +/- 0.17</td>
<td>0.24 +/- 0.12</td>
</tr>
<tr>
<td>branches with omega &gt; 1.0 (%)</td>
<td>72 (6.04)</td>
<td>14 (0.49)</td>
</tr>
<tr>
<td>mean omega for &gt; 1.0 branches +/- sd</td>
<td>1.45 +/- 0.51</td>
<td>1.13 +/- 0.14</td>
</tr>
<tr>
<td>mean omega +/- sd</td>
<td>0.54 +/- 0.31</td>
<td>0.24 +/- 0.13</td>
</tr>
</tbody>
</table>
**FIGURE LEGENDS**

**Figure 1:** Global expansion rate in each subgroup: Total number of genes in each species divided by the ancestral number (see Table I). An ANOVA test showed that the expansion rate differs significantly between SGs ($p<2e-16$). Therefore, we performed a TukeyHSD test to determine which SGs exactly show a significant difference between each other and group those SGs by significance level (a-e). Letters above the boxplot indicate TukeyHSD significance group (see Supplemental Table SII). The significance groups are color coded according to the mean expansion rate: orange: >2.25 fold expansion; red: 1.75-2.25 fold expansion; purple: 1.30-1.75 fold expansion; blue: 0.75-1.30 fold expansion (i.e. no expansion). The outlier species are labeled for each SG. See Table II for species IDs.

**Figure 2:** Branch-specific expansion/diminution of LRR-RLK genes for every SG on every branch in the phylogenetic tree. The tree on the left displays all the nodes and branches, polyploidy events are marked with dots. Every line gives the expansion rate where the current (descendant) node is compared to the previous (ascendant) node. Red boxes indicate expansion, blue boxes indicate diminution, and blank boxes indicate stagnation. For example: SG_I* has the same number of copies in monocots compared to the ascendant node (= angiosperms) indicated by a blank box. In PHODC, a diminution occurred compared to the ascendant node (= monocots) indicated by a blue box. In MUSAC, an expansion occurred compared to the ascendant node (= monocots) indicated by a red box and so on.

**Figure 3:** Global expansion rate in each species: Total number of genes in each species divided by the ancestral number (see Table I). An ANOVA test showed that the expansion rate differs significantly between species ($p<2e-16$). Therefore, we performed a TukeyHSD test to determine which species exactly show a significant difference between each other and group those species by significance level (a-e). Letters above the boxplot indicate TukeyHSD significance group (see Supplemental Table SII). The significance groups are color coded according to the mean expansion rate: orange: >2.25 fold expansion; red: 1.75-2.25 fold expansion; purple: 1.40-1.75 fold expansion; blue: 0.80-1.40 fold expansion (i.e. no expansion). The outlier SGs are labeled for each species. See Table II for species IDs.

**Figure 4:** Phylogenetic tree of the 33 species studied here. Five-digit species identifiers are given in parenthesis next to the species name. Species which diverged <15 mya were merged for the LSE analysis (see Materials and Methods): ANDRO, ORYZA, SOLAN, CUCUM,
BRASS. Polyploidy events and their estimated ages are indicated on the tree: circles on the branches represent WGD, dark circles represent WGT. The numbers in the circles refer to details on the polyploidization events given in Table I. Species divergence and their estimated age are indicated by grey squares on the nodes. The numbers in the squares refer to details on the divergence times given in Supplemental Table SI. Dots and asterisks on the branches indicate SG expansions. Dots, 2-fold; small asterisks, between 2 and 4-fold; large asterisks, equal or more than 4-fold. SG I* (brown), SG_IV (dark green), SG_V (grey), SG_VIIa (orange), SG_VIIb (yellow), SG_VIII-1 (dark brown), SG_VIII-2 (green), SG_IX (light blue), SG_Xa (dark blue), SG_XIIa (pink), SG_XIIb (purple), SG_XIIia (red), SG_XIV (black), SG_XV* (white). The asterisks and dots do not indicate the exact age.

**Figure 5:** Distribution of UP and SO clusters and sequences over all SGs. Frequency of all extracted UP (dark blue) and SO clusters (dark orange) for each SG; Frequency of all extracted UP (light blue) and SO sequences (light orange) for each SG.

**Figure 6:** (A) Hit frequency (i.e. frequency of codons under selection vs. total number of sites) for each domain of the LRR-RLK genes. (B) Schematic structure of the LRR-RLK genes, here with SG_I* gene structure as an example. The absence/presence and size of the domains varies between SGs, please see text for details. N-term, N-terminal end; SP (dark grey), signal peptide; Cys-pair 1 (blue), first cysteine pair; NC1, N-terminal of Cys-pair 1; CC1, C-terminal of Cys-pair 1; other (green), other domains; LRR (red), leucine-rich repeat; Cys-pair 2 (blue), first cysteine pair; NC2, N-terminal of Cys-pair 2; CC2, C-terminal of Cys-pair 2; TM (black), trans-membrane domain; JM, juxta-membrane domain; KD (yellow), kinase domain; C-term, C-terminal end; inter, other inter-domain regions. (C) Frequency of amino acids in the LRR domain under positive selection. L, leucine; N, asparagine; G, glycine; I, isoleucine; P, proline; x, variable; is, island between LRRs.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Arya P, Kumar G, Acharya V, Singh AK (2014) Genome-wide identification and expression analysis of NBS-encoding genes in Malus x domestica and expansion of NBS family genes in Rosaceae. PLOS ONE 9: e107987

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Fawcett JA, Maere S, Van de Peer Y (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci USA 106: 5737-5742

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Author and Title

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title