Plant Physiology Preview. Published on June 14, 2016, as DOI:10.1104/pp.15.01207

Running head: Expressologs identify functional orthologs

Corresponding authors:

Malay Das
Department of Biological Sciences,
Presidency University
86/1 College Street
Kolkata- 700073, India
Phone: 91-8017966849, Email: **malay.dbs@presiuniv.ac.in**

Anton R. Schäffner
Institute of Biochemical Plant Pathology
Helmholtz Zentrum München
German Research Center for Environmental Health (GmbH)
Ingolstädter Landstr. 1
85764 Neuherberg Germany
Phone: 49-(0)89 3187 2930, E. mail: schaeffner@helmholtz-muenchen.de

Research area: Breakthrough technologies

Expression pattern similarities support the prediction of orthologs retaining common functions after gene duplication events

Malay Das[1,2], Georg Haberer[3], Arup Panda[4], Shayani Das Laha[2], Tapas Chandra Ghosh[4], Anton R. Schäffner[1]

[1]Institute of Biochemical Plant Pathology, Helmholtz Zentrum München, München, Germany
[2]Department of Biological Sciences, Presidency University, Kolkata, India
[3]Plant Genome and Systems Biology Group, Helmholtz Zentrum München, München Germany
[4]Bioinformatics Center, Bose Institute, Centenary Campus, Kolkata, India

SUMMARY

Expressologs identify functional orthologs and will be a powerful tool in future orthology assignment.

Footnotes:

This work was initially supported by an Alexander von Humboldt Foundation Postdoctoral fellowship, Germany to MD. Also financial supports from University Grant Commission (MRP-MAJOR-BIOT-2013-18380), Council of Scientific and Industrial Research [38(1386)/14/EMR-II] and Department of Biotechnology, India (BT/PR10778/PBD/16/1070/2014) to MD laboratory are gratefully acknowledged.

Author contributions:

MD, GH and ARS conceptualized the research and data evaluation. MD performed majority of the stress, microarray, real-time PCR and genetic complementation experiments. GH performed the ortho MCL, microarray and related bioinformatics analyses. AP performed the promoter analyses under the supervision of TCG. ARS performed the complementation assay of AL.TSO2A. SDL helped in the characterization of few complemented lines. MD and ARS wrote the manuscript with help from all co-authors.

ABSTRACT

Identification of functionally equivalent, orthologous genes (functional orthologs) across genomes is necessary for accurate transfer of experimental knowledge from well-characterized organisms to others. This frequently relies on automated, coding sequence-based approaches such as OrthoMCL, Inparanoid, KOG, which usually work well for one-to-one homologous states. However, this strategy does not reliably work for plants due to the occurrence of extensive gene/genome duplication. Frequently, for one query gene multiple orthologous genes are predicted in the other genome and it is not clear a priori from sequence comparison and similarity which one preserves the ancestral function. We have studied eleven organ-dependent and stress-induced gene expression patterns of 286 *A. lyrata* duplicated gene groups and compared them to the respective *A. thaliana* genes to predict putative expressologs and non-expressologs based on gene expression similarity. Promoter sequence divergence as an additional tool to substantiate functional orthology only partially overlapped with expressolog classification. By cloning eight *A. lyrata* homologs and complementing them in the respective four *A. thaliana* loss-of-function mutants we experimentally proved that predicted expressologs are indeed functional orthologs, while non-expressologs or non-functionalized orthologs are not. Our study demonstrates that even a small set of gene expression data in addition to sequence homologies are instrumental in the assignment of functional orthologs in the presence of multiple orthologs.

INTRODUCTION

With the rapid advancements of next generation sequencing technologies, sequencing a transcriptome/genome is highly feasible nowadays within decent time and at low-cost. One important bottle neck for downstream analysis is the annotation, i.e. how accurately we can transfer gene function information from well characterized reference genomes and model plants to these newly sequenced genomes and/or crop plants. The major reason of this uncertainty is the occurrence of multiple homologous sequences, as a result of gene family expansions and polyploidization events. Orthologs are defined as genes in different species that have emerged as a result of an evolutionary speciation event. Since they are derived from a single gene in the last common ancestor, orthologs frequently share the same function in the newly evolved species. However, gene duplications after the speciation may result in a functional divergence where the ancestral function either is split between such co-orthologs, or the functions are otherwise transformed (see below). Thus, a multiple orthology situation has arisen in such cases and the congruence of evolutionary relationship and conserved function may have been lost (Remm et al. 2001; Bandyopadhyay et al., 2006). In accordance with these previous studies, we define *functional orthologs* as those co-orthologs that have retained highly similar functions in the two species in such a multiple orthology situation. Therefore, correct identification of functional orthologs is critical for gene annotations by extrapolating functions across species barriers.

Genes that arose following duplication events (whole genome, segmental or tandem duplications) are called paralogs. Paralogs, which are also orthologs, i. e. which have been formed after a speciation event, are called in-paralogs (or co-orthologs) in contrast to out-paralogs, which are derived from a gene duplication before an evolutionary speciation (Remm et al., 2001). The older the duplication event, the higher the chances will be that the (in) paralogs will undergo functional divergence. The possible fates of such gene copies and the gene groups are (1) non-functionalization, pseudogenization: one ortholog retains the ancestral function, while the other ortholog(s) lose(s) the function by acquiring deleterious mutations and (2) neo-functionalization: one ortholog acquires a new function by beneficial mutations, whereas the other one retains the original function. In the course of its adaption to a distinct environment, an ortholog in one species may also undergo neo-functionalization resulting in a species-specific function for this gene. A subsequent duplication of this gene actually results in two in-paralogous copies that significantly differ in their function from the ortholog of the other species. We therefore define a group leading to (3) species-specific functionalization: the whole orthologous group in one species differs the other species and does not retain the function, i.e. either (a) all in-paralogs acquire new roles or (b) one ortholog has a new function, while the other(s) lost their original role (non-functionalized). An extreme case of such a development is (4) species-specific non-functionalization: all orthologs are pseudogenized and lose their function in one species. A fifth possible fate is (5) sub-functionalization: the ancestral gene function is split among duplicated copies. Finally, there is (6) genetic redundancy: all co-orthologs still share the ancestral function. However, in an existing, already further evolved species, genetic redundancy and sub-functionalization or neo-functionalization will overlap and depend on the depth of phenotypic analysis. Thus, in most cases genetic redundancy may not define an independent evolutionary category of genes per se, but rather point to a lack of detailed knowledge about divergent functions of these genes.

Several automated cluster methods with varying degrees of selectivity and sensitivity have been developed to assign orthologous relationships across genomes (COG, Tatusov et al., 1997; KOG, Tatusov et al., 2003; OrthoMCL, Li et al., 2003; Inparanoid, O’Brien et al., 2005). These sequence-based methods are appropriate to cluster genes with high similarity and possible common ancestry, but they cannot unambiguously identify functional orthologs. One way to track the functionality of the homologous genes after species split is to dissect their expression patterns under a range of spatio-temporal and/or environmental conditions. In yeast, regulatory neo-functionalization events were identified for 43 duplicated gene pairs based on their asymmetric expression profiles, which the sequence data analysis had failed to detect (Tirosh and Barkai, 2007). In plants most attention was paid to study how polyploidy has fueled expression divergence of duplicated gene pairs in a single species (Blanc and Wolfe, 2004; Duarte et al., 2006; Ha et al., 2007; Throude et al., 2009; Whittle and Krochko, 2009). With the availability of multiple genome sequences, cross species comparisons has been gaining momentum. Publicly available gene expression data were used to conduct a cross species comparison between rice and poplar in order to identify transcription factors associated with leaf development (Street et al., 2008). Gene co-expression network analysis was performed on 3182 DNA microarrays from human, flies, worms, and yeast to identify core biological functions that are evolutionarily conserved across the animal kingdom and yeast (Stuart et al., 2003). A similar study conducted on six evolutionarily divergent species, *S. cerevisiae*, *C. elegans*, *E. coli*, *A. thaliana*, *D. melanogaster* and *H. sapiens*, concluded that functionally related genes are often co-expressed across species barriers (Bergmann et al., 2004). Taken together, all these studies indicate that combining sequence and expression data may increase the prediction ability of gene function annotation. However, such co-expression approaches are only possible, if large-scale transcriptome analyses are available for both (or more) species to be compared. Thereby, less well studied and/or newly sequenced species are not (immediately) amenable to such comparisons. Furthermore, none of these studies could experimentally prove the success rate of such prediction at the level of individual gene functions.

An alternative strategy to predict functional orthologs was established by Patel et al. (2012). These authors ranked genes from homology clusters of seven plant species based on extensive gene expression profiles obtained from comparable tissues among these species. The top ranking homolog based on expression pattern similarity was termed "expressolog", which should indicate the functional ortholog. Bandyopadhyay et al. (2006) employed protein-protein interaction data to identify functional orthologs among large *Saccharomyces cerevisiae* and *Drosophila melanogaster* paralogous gene families; in about half of the studied cases, the most conserved functions were not favored by sequence analyses.

The two well annotated, but biologically divergent Brassicaceae species *A. thaliana* and *A. lyrata* included in this study have diverged approximately 10 million years ago (Hu et al., 2011). Both species substantially differ in several biological traits that are crucial differences in their life style: life cycle (annual *A. thaliana vs*. perennial *A. lyrata*), mating system (selfing *A. thaliana vs*. out-crossing *A. lyrata*), geographical distribution (continuous distribution of *A. thaliana vs*. scattered distribution of *A. lyrata*) and genome size (125 Mb *A. thaliana vs.* 207 Mb *A. lyrata*). Furthermore, the *Arabidopsis* lineage has undergone three rounds of whole genome duplication followed by differential loss of gene/s in the different species. Therefore we aimed at identifying genes that exist as a single copy gene in one species, but as multiple copies in the other species, and thus define as a 'one-to-many' situation. Due to the lack of large-scale expression data for *A. lyrata*, a co-expression-based approach was not possible. Instead, we studied expression pattern correlation based on a small set of eleven, yet diverse experimental scenarios involving expression in organs (leaf, root, flower bud) and under different stress conditions. Using such gene expression similarities we predicted the 'expressolog' for individual gene clusters and thereby candidates of functional orthology in the new, to be analyzed species *A. lyrata*. Importantly, we could prove that predicted expressologs were indeed functionally equivalent, while non-expressologs or non-functionalized genes were not, by using genetic complementation experiments.

RESULTS

OrthoMCL analysis to identify one-to-many situations between *Arabidopsis thaliana* and *A. lyrata*

OrthoMCL analysis between *A. thaliana* and *A. lyrata* transcriptomes identified 2850 gene clusters where either one-to-many or many-to-many situations were present. Of these 2850 clusters 613 were one *A. thaliana* gene : multiple *A. lyrata* genes, 366 one *A. lyrata* gene : multiple *A. thaliana* genes, and 1871 multiple *A. thaliana* genes : multiple *A. lyrata* genes. One of the major aims of this study is to experimentally check the efficiency of predicted expressologs in terms of their function. Gene-specific loss-of-function mutants are currently available for *A. thaliana* but not for *A. lyrata* and therefore we focused our studies on the 'one *A. thaliana* gene : multiple *A. lyrata* genes' group.

Microarray studies on *A. thaliana* and *A. lyrata* plants to dissect organ-dependent and stress-responsive expression patterns of duplicated genes

Genome wide expression analyses were performed on *A. thaliana* and *A. lyrata* plants to determine gene expression similarity or divergence between closely related homologs (Table S1). Gene expression data were collected from three different tissues (shoot, root and flower bud) and from plants subjected to salt, drought and UV-B stress regimes to measure gene expression patterns in different organs and for time courses of diverse stress situations. Pearson correlation analysis was performed by analyzing all organ, control and stress-induced gene expression data obtained from *A. thaliana* and *A. lyrata*. The mean correlation value of all-against-all comparisons between *A. thaliana* and *A. lyrata* transcriptomes was 0.019 (Table 1). On the contrary, when syntenic *A. thaliana* - *A. lyrata* orthologous or OrthoMCL one-to-one gene groups were analyzed, much higher correlation values of 0.329 and 0.320 were obtained, which is in accordance with the expectation that the majority of orthologous gene copies still share similar functions. We excluded 263 out of 613 candidates based on probes with a cross-hybridization potential in order to avoid ambiguous measurements due to high sequence similarities of *A. lyrata* paralogs. An average expression threshold value of 9.0 ($\log_2$ scale) was introduced to exclude such gene groups, where all members are only lowly expressed close to the detection limit of our system in shoots, roots and flower buds of both *A. thaliana* and *A. lyrata* (Methods). The final gene set comprised 272 *A. thaliana* genes each having two or more *A. lyrata* homologs.

Functional categorization based on gene expression data and prediction of expressologs and non-expressologs

Pearson correlation coefficients for each At:Al pair present within an OrthoMCL gene group were calculated based on microarray data collected from all stress and control experiments. The differential expression patterns of each of the duplicated *A. lyrata* genes along with the related *A. thaliana* copies were measured under salt, drought and UV-B stresses conditions. The normalized expression levels of these genes were calculated in shoot, root and flower bud tissues. Based on these analyses we predicted functionally related (expressolog/s) and functionally diverged homolog/s for each of the 272 OrthoMCL gene groups (Table S2). An *A. lyrata* ortholog was classified as an expressolog, (i) if it was detected in the same pattern in rosette leaves, roots and flowers like the *A. thaliana* gene, and (ii) if its correlation regarding the stress responsiveness across all eight tested scenarios was bigger than 0.3. If the genes were not stress-responsive in our conditions, the stress response correlation was not taken into account. All other cases showing detectable gene expression were denoted as non-expressologs (for details, Table S2).

If the normalized organ expression value of any single member of an OrthoMCL gene group is below 9.0 in all organs or any of the stress scenarios studied here we predict that the gene is non-functionalized under the studied conditions, since such a level is close to the detection limit. This classification cannot exclude the possibility that the gene is expressed in yet another scenario, which would indicate a neo-functionalization of the respective ortholog.

This strategy identified 34 out of 272 (12.5%) OrthoMCL gene groups, where one *A. lyrata* ortholog retains the original function (expressolog), while the other ortholog(s) is (are) non-functional (Table S2; group 1). One example is constituted by the three members of the chloroplast TIC complex (*A. thaliana* AT1G06950, *A. lyrata* scaffold_100703.1 and *A. lyrata* fgenesh2_kg.1__669__AT1G06950.1). Normalized organ expression level of the *A. thaliana* and *A. lyrata* fgenesh2_kg.1__669__AT1G06950.1 gene were in a range of 12-15, while *A. lyrata* scaffold_100703.1 gene copy had very low expression levels of 6.44, 3.63 and 4.05 in shoots, roots and flower buds, respectively (Table S2). The pair-wise correlation analysis between the two highly expressed genes is 0.60, while it drops to -0.48 between *A. thaliana* AT1G06950 and the putatively non-functionalized *A. lyrata* copy.

In 49 (~18%) gene groups, one *A. lyrata* homolog maintained a similar expression pattern like the *A. thaliana* gene, while the other homolog showed a differential expression pattern at a significant expression level (non-expressolog); therefore, we classified them as neo-functionalized (Table S2; group 2). For example, two *A. lyrata* (fgenesh2_kg.1__967__AT1G09240.1, fgenesh2_kg.1__4760__AT1G56430.1) and one *A. thaliana* (AT1G09240) members were detected in a gene group encoding *NICOTIANAMINE SYNTHASE 3*. While the *A. thaliana* AT1G09240 and *A. lyrata* fgenesh2_kg.1__967__AT1G09240.1 genes are positively correlated under drought (r = 0.84) and salt (r = 0.45) stressed conditions with a total stress correlation of r = 0.59, the *A. lyrata* fgenesh2_kg.1__4760__AT1G56430.1 gene was negatively correlated under drought (r = -0.90) and salt (r = -0.85) stressed conditions with a total stress correlation of r = -0.47. When the expression of these genes in different organs were studied, the loss of expression of *A. lyrata* fgenesh2_kg.1__4760__AT1G56430.1 gene in flower bud further differentiates it from the *A. thaliana* and the other *A. lyrata* genes (Table S2). This clearly indicates that *A. lyrata* fgenesh2_kg.1__967__AT1G09240.1 is the predicted expressolog to *A. thaliana* AT1G09240, while *A. lyrata* fgenesh2_kg.1__4760__AT1G56430.1 has acquired a new expression pattern and is likely neo-functionalized.

A total of 115 (~42%) gene groups were categorized as species-specific functionalization since the expression pattern of all functional *A. lyrata* genes in a OrthoMCL cluster were different from that of the *A. thaliana* gene. Two types of divergences were recorded: (a) either all *A. lyrata* orthologs are neo-functionalized (non-expressologs, 74 gene groups) or (b) one *A. lyrata* ortholog is a non-expressolog, while the other(s) lost the original function (non-functionalized, 41 gene groups) (Table S2; groups 3a and 3b). For instance, the members of *UDP-XYLOSE TRANSPORTER1/UXT1* cluster consist of *A. thaliana* AT2G28315/*UXT1*, *A. lyrata* scaffold_8500004.1 and *A. lyrata* fgenesh1_pm.C_scaffold_4000618. The two *A. lyrata* genes acquired salt- and drought responsiveness and are negatively correlated to the *A. thaliana* gene under salt and drought stresses (r = -0.8, Table S2). A small group of six gene clusters showed an extreme form of species-specific functionalization, where all the *A. lyrata* genes present in a cluster are non-functionalized (Table S2; group 4).

Sub-functionalization of genes would be indicated by a complementary expression of the co-orthologs which covers the whole expression pattern of the corresponding gene in the other species (group 5). Possibly due to the limited number of eleven tested scenarios in the expression analyses, there were no clear indications for such a sub-functionalization. Instead, in 68 (25%) gene groups both *A. lyrata* homologs maintained similar organ and stress expression patterns like the *A. thaliana* genes and were interpreted as a group composed of genetically redundant genes based on our experimental assays. This is also reflected in the comparable correlation values between individual *A. lyrata* and *A. thaliana* pairs residing in the same cluster. One such gene group consists of *A. thaliana* AT1G06680, *A. lyrata* fgenesh2_kg.1__643__AT1G06680.1 and *A. lyrata* scaffold_401578.1. All three genes are well expressed in the three organs studied (Table S2). They were upregulated in the late time-point of salt and drought treatment, while no response was found in UV-B. Consistently, the overall stress correlation value between *A. thaliana* AT1G06680 and *A. lyrata* fgenesh2_kg.1__643__AT1G06680.1 is 0.939 and between *A. thaliana* AT1G06680 and *A. lyrata* scaffold_401578.1 is 0.916.

Nucleotide substitution rate calculation and comparison between expressologs, non-expressologs and non-functionalized genes in four different functional categories

The transcription of a gene is largely controlled by its promoter. Therefore, we first tested if the promoter sequences of expressologs were more conserved than those of the predicted non-expressologs. Such a correlation could initially support the identification of expressologs in newly sequenced species even in the absence of expression data. Shared motif divergence (d_{SM}) method was employed to quantify the nucleotide changes in the upstream regions of *A. lyrata* gene groups with respect to the orthologous *A. thaliana* genes. This analysis revealed that the upstream sequences of an expressologous gene group were on average less divergent as compared to the divergence of non-expressologous genes such as neo- or non-functionalized groups (Fig. 1A). The promoter sequences divergence score of the one-to-one gene group were comparable to that of the expressologous gene group (Fig. 1A). To avoid complication in data analyses arising due to presence of too many *A. lyrata* homologs within a gene cluster or unavailability of sufficiently long promoter sequences, a few genes were discarded from the analysis. Therefore the number of gene groups compared in the current analyses was 32 for the non-functionalized group, 35 for the neo-functional group, 57 for the genetically redundant group, and 81 for the species-specific group, respectively.

In addition to the overall analyses comparing promoters of expressolog vs. non-expressologs, the promoter divergence of the genes within the different evolutionary gene groups was assessed. Within the non-functionalized gene groups 78% of the cases (25 out of 32 gene groups) the *A. lyrata* expressologs revealed less promoter divergence compared to the non-functionalized genes (Figure 1B). Similarly, in 77% of the neo-functionalized gene groups (27 out of 35 gene groups) the expressologs possessed less promoter divergence than those of the non-expressologs (Figure 1C). In contrast to these two groups, where one member showed a conserved and one member exhibited a non-conserved expression pattern, the all *A. lyrata* genes contained in the genetically redundant gene groups and in the species-specific functionalization groups either show a similar (genetic redundancy group) or divergent

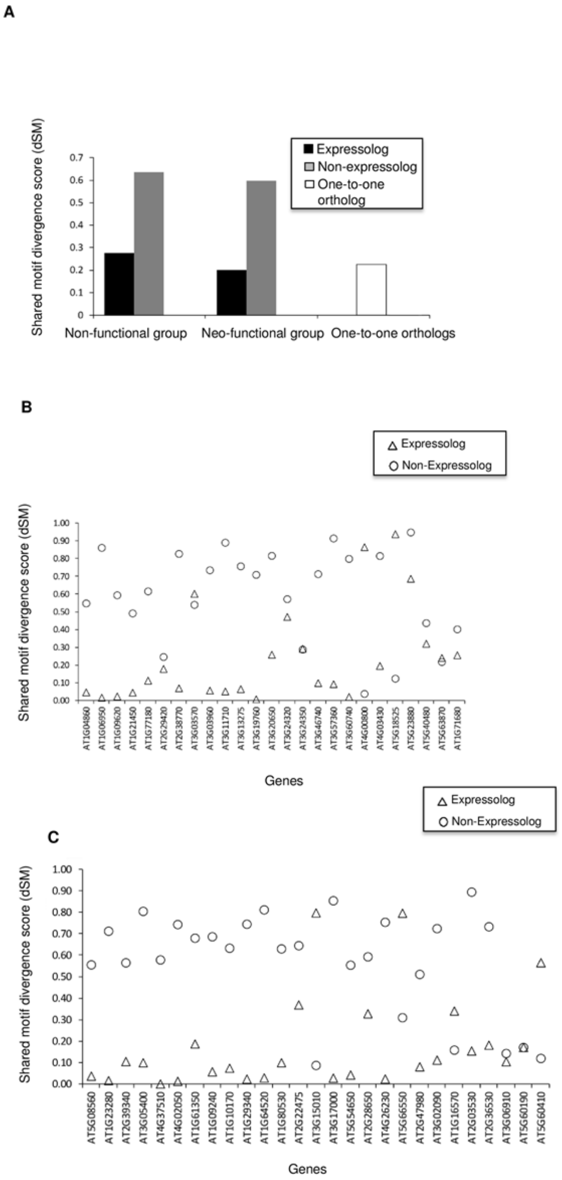


Figure1. Promoter sequences divergence analysis between expressologs and non-expressologs in two different functional categories.

(A) Shared motif divergence scores (d_{SM}) of expressologs, non-expressologs and one-to-one orthologs. The first panel compares between the promoter sequence divergence scores of *A. lyrata* expressologs and neo-functionalized non-expressologous genes (as predicted by our gene expression analysis), the second panel compares between the expressologs and non-functionalized non-expressologous genes, the third panel compares between the promoter sequence divergence scores of *A. lyrata* genes having single orthologous copy of *A. thaliana* genes (as predicted by Ortho-MCL).

(B) Promoter analyses of the gene group, where at least one *A. lyrata* gene has been non-functionalized as predicted by gene expression analyses. For each *A. thaliana* and *A. lyrata* orthologous gene pair within a gene group, we calculated the promoter sequence divergence scores (d_{SM}) of *A. lyrata* genes with reference to the promoter sequence of their *A. thaliana* orthologous gene (in x-axis) by shared motif divergence method. Here, "o" represents the promoter sequence divergence score (d_{SM}) of *A. lyrata* gene copy predicted as expressolog by the gene expression analyses, "Δ" stands for that of non-expressologs.

(C) Promoter analyses of the gene group, where at least one *A. lyrata* gene has been predicted to be neo-functionalized. The other parameters used were same as described in Fig. 1A.

(species-specific group) expression pattern in organs and/or stress conditions with respect to the *A. thaliana* gene. Therefore, in these cases it was interesting to analyze whether this differential behavior was also obvious among the promoters of the two members present within such a gene group in comparison to the corresponding *A. thaliana* gene. To assess this question the average promoter divergences of all *A. lyrata* genes compared to the respective *A. thaliana* genes in the genetically redundant and the species-specific functionalization groups were calculated separately. Indeed the average d_{SM} of species-specific group is almost two fold (0.391) than that of the genetically redundant group (0.219). To address the promoter divergence of these two groups also at the individual gene group level, the difference of the promoter divergence between *A. thaliana*: *A. lyrata* 1 (d_{SM}1) and *A. thaliana*: *A. lyrata* 2 (d_{SM}2) [delta d_{SM} = d_{SM}1- d_{SM}2] present within the same gene group was calculated. If there would be an overlap with the expression-based classification, lower delta d_{SM} values would be expected for the genetically redundant than for the species-specific gene groups. If we consider a conservative delta d_{SM} cut-off of <0.2 meaning high promoter similarity, then in 53% (43 out of 81) of the species-specific groups the two *A. lyrata* promoter sequences are not comparable with respect to their sequence divergence from the *A. thaliana* promoter. Thus, in about one half of the cases the promoters of the species-specific groups have undergone a strong change in agreement with their changing expression pattern, whereas in the other half the promoter divergences were not indicative of the expression patterns (Figure S1). In case of the genetically redundant gene pairs 40% (23 out of 57) of the gene groups also showed a high differential divergence of promoters of the co-orthologs compared to the *A. thaliana* gene in contrast to the similar and conserved expression patterns observed.

One such example from the genetic redundancy group consists of *A. thaliana* AT1G06680, *A. lyrata fgenesh2_kg.1__643__AT1G06680.1* and *A. lyrata* scaffold_401578.1 genes. While the two *A. lyrata* genes are highly correlated to the *A. thaliana* gene with respect to their organ expression and their stress-responsive gene expression pattern (r = 0.98), the promoters of the two *A. lyrata* genes reveal a differential sequence divergence from the *A. thaliana* gene with a delta d_{SM} = 0.38 (AT1G06680: *A. lyrata fgenesh2_kg.1__643__AT1G06680.1* dSM = 0.003, AT1G06680: *A. lyrata* scaffold_401578.1 dSM = 0.382).

The gene group *A. thaliana* AT2G31160, *A. lyrata* fgenesh1_pg.C_scaffold_4001226 and *A. lyrata* fgenesh2_kg.163__1__AT2G31160.1 provides an example from the species-specific category, which shows a high promoter conservation of the *A. lyrata* gene promoters in comparison to the *A. thaliana* gene despite the changed expression pattern. Both *A. lyrata* co-orthologs were induced by salt stress in contrast to the *A. thaliana* copy contributing to the low correlation of the total stress responses (r = -0.0581 and r = -0.1191). Furthermore, the two *A. lyrata* copies were different among themselves with one copy being expressed at very low level in all organs (Table S2). Nevertheless, delta d_{SM} was 0 and the d_{SM} levels for both AL : AT comparisons were very low ($d_{SM} = 0.007$).

It is evident from our analyses that while promoter divergence analysis can be used as an additional tool for annotation purposes, experimental classification as expressologs/non-expressologs provides more accurate functional information and mode of functional divergence such as non-, neo- or species-specific functionalization and genetic redundancy, which the promoter analyses cannot fully offer.

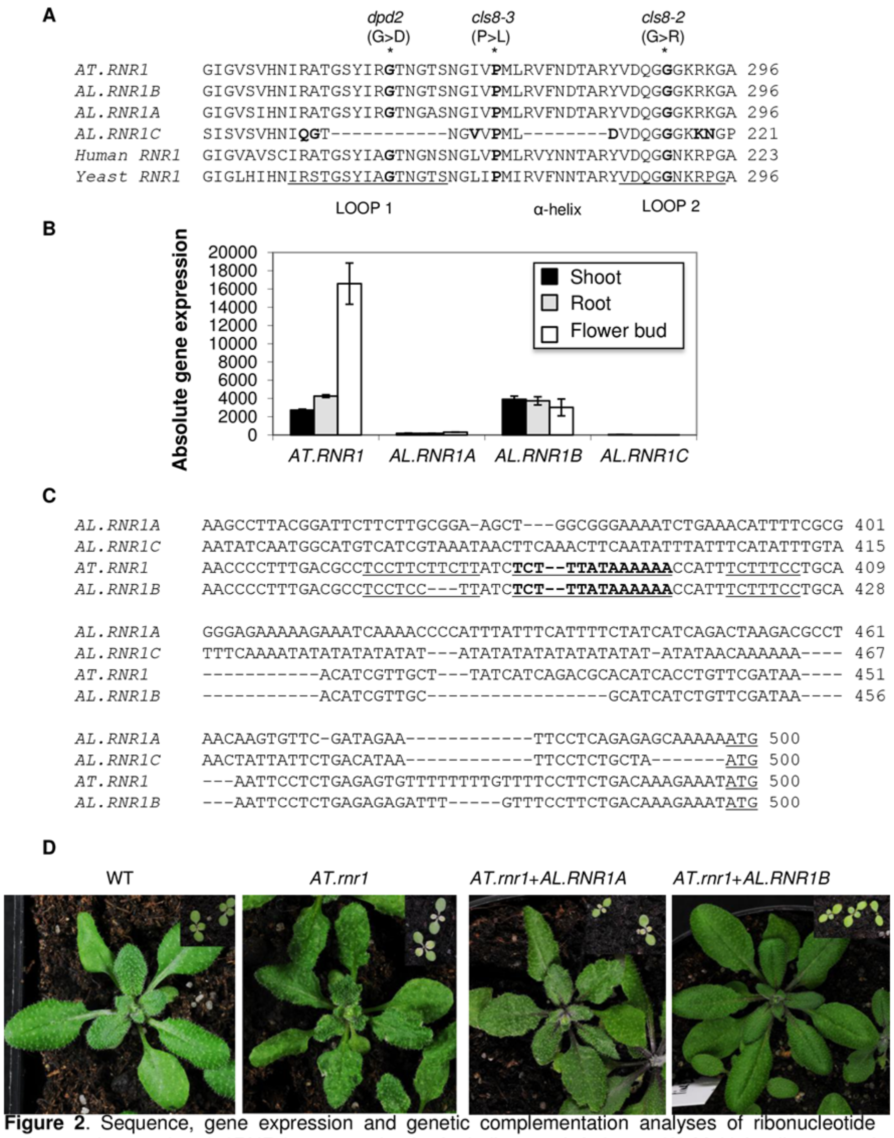
Identification of genetic mutants for experimental validation of predicted expressologs

To confirm functionality of our predicted expressologs, we applied genetic complementation assays using the *A. lyrata* gene variants transformed to *A. thaliana* loss-of-function mutants for the group of one *A. thaliana* : multiple *A. lyrata* candidate genes. We scanned the insertion mutant repositories to identify mutant lines corresponding to our list of 272 genes. Additionally, we checked the available literature for appropriate mutants. Out of 272 queried one-to-many *A. thaliana* genes, homozygous mutant SALK lines were obtained for 147 genes. All these 147 insertion lines were grown under green-house conditions, but no obvious morphological phenotypes could be observed for any of these lines studied.

However, four published *A. thaliana* mutants, *cls8-1*, *tso2-1*, *sta1-1* and *mtp11* could be used for our analyses (Table S3). Their mutant phenotypes could be clearly reproduced and the corresponding *A. lyrata* homologous gene copies along with their native promoters were amplified for genetic complementation assay. Based on our expressolog classification, the corresponding AT : AL gene groups represented one case of a possible neo-functionalization (*CLS8*/*RNR1*) and one case of pseudogenization (*STA1*). Two cases (*TSO2* and *MTP11*) were indicative of genetic redundancy.

Example 1: Potential neo-functionalization by acquiring changes in the regulatory region of the gene and in the coding region

Neo-functionalization is predicted for the *A. lyrata* genes encoding the large subunit of ribonucleotide reductase, which catalyzes the reduction of ribonucleoside diphosphates to deoxyribonucleotides, the rate limiting step in the *de novo* synthesis of dNTPs (Sauge-Merle et al., 1999). In *A. thaliana* the large subunit is encoded by a single copy gene *AT.RNR1* (AT2G21790, *CLS8*), while in *A. lyrata* three homologous copies exist, *AL.RNR1A* (Al_scaffold_0007_128/AL7G01310), *AL.RNR1B* (fgenesh2_kg.4__104__AT2G21790.1/AL4G01010) and *AL.RNR1C*



```
                                     *          *                   *
AT.RNR1      GIGVSVHNIRATGSYIRGTNGTSNGIVPMLRVFNDTARYVDQGGGKRKGA 296
AL.RNR1B     GIGVSVHNIRATGSYIRGTNGTSNGIVPMLRVFNDTARYVDQGGGKRKGA 296
AL.RNR1A     GIGVSIHNIRATGSYIRGTNGASNGIVPMLRVFNDTARYVDQGGGKRKGA 296
AL.RNR1C     SISVSVHNIQGT-----------NGVVPML--------DVDQGGGKKNGP 221
Human RNR1   GIGVAVSCIRATGSYIAGTNGNSNGLVPMLRVYNNTARYVDQGGNKRPGA 223
Yeast RNR1   GIGLHIHNIRSTGSYIAGTNGTSNGLIPMIRVFNNTARYVDQGGNKRPGA 296
```

```
AL.RNR1A     AAGCCTTACGGATTCTTCTTGCGGA-AGCT---GGCGGGAAAATCTGAAACATTTTCGCG 401
AL.RNR1C     AATATCAATGGCATGTCATCGTAAATAACTTCAAACTTCAATATTTATTTCATATTTGTA 415
AT.RNR1      AACCCCTTTGACGCCTCCTTCTTCTTATCTCT--TTATAAAAAACCATTTCTTTCCTGCA 409
AL.RNR1B     AACCCCTTTGACGCCTCCTCC---TTATCTCT--TTATAAAAAACCATTTCTTTCCTGCA 428

AL.RNR1A     GGGAGAAAAAGAAATCAAAACCCCATTTATTTCATTTTCTATCATCAGACTAAGACGCCT 461
AL.RNR1C     TTTCAAAATATATATATATAT---ATATATATATATATATAT-ATATAACAAAAAA---- 467
AT.RNR1      -----------ACATCGTTGCT---TATCATCAGACGCACATCACCTGTTCGATAA---- 451
AL.RNR1B     -----------ACATCGTTGC-----------------GCATCATCTGTTCGATAA---- 456

AL.RNR1A     AACAAGTGTTC-GATAGAA------------TTCCTCAGAGAGCAAAAAATG 500
AL.RNR1C     AACTATTATTCTGACATAA------------TTCCTCTGCTA-------ATG 500
AT.RNR1      ---AATTCCTCTGAGAGTGTTTTTTTTGTTTTCCTTCTGACAAAGAAATATG 500
AL.RNR1B     ---AATTCCTCTGAGAGAGATTT-----GTTTCCTTCTGACAAAGAAATATG 500
```

Figure 2. Sequence, gene expression and genetic complementation analyses of ribonucleotide reductase large sub-unit (*RNR1*) gene copies in *A. thaliana* and *A. lyrata*. (A) Multiple alignment of the *Arabidopsis* RNR1 amino acid sequences along with those of human and yeast sequences. Biologically important LOOP1 and LOOP2 regions are depicted. Part of the highly conserved LOOP1 region is missing and two non-synonymous amino acid changes were detected in the LOOP2 region of *AL.RNR1C*. However, the *AL.RNRB* coding sequence is identical to that of *AT.RNR1*. These regions play important roles in the enzymatic function by controlling specificity of the incoming dNTP. The biological importance of this region is emphasized by the identification of three mutations that caused severe developmental defects (indicated by * on top). Another allele, *cls8-1* affects a distant region leading to an amino acid change G718E, however showing the same mutant phenotype (Tab. S3). (B) Study of the expression patterns of the four *Arabidopsis RNR1* genes in root, shoot and flower bud. Background corrected and multiplicatively de-trended signal intensities were imported to Gene Spring (G3784AA, version 2011) to calculate normalized gene expression values (see Methods for details). (C) Comparison of core promoter regions (250 bp upstream from ATG) indicates the loss of the *AT.RNR1*-like TATA box (bold, underlined) and Y patch (underlined) in the case of *AL.RNR1A* and *AL.RNR1C* homologs. This analysis was done in plant promoter database (http://133.66.216.33/ppdb/cgi-bin/index.cgi#Homo). (D) Genetic complementation of *A. thaliana rnr1/cls8-1* with *AL.RNR1B* and *AL.RNR1A* gene copies. The phenotype of the *AL.RNR1B* (predicted expressolog) complemented plants resemble wild type. However, the plants complemented by *AL.RNR1A* (predicted pseudogene) show the mutant phenotype such as the yellowish, first true leaf (in the inset) and the crinkled, matured leaves.

(scaffold_200715.1/AL2G07030). Nonsense point mutations in *A. thaliana* caused visible early and late developmental phenotypes such as bleached first true leaves and crinkled rosette leaves with white pits on the surface (Garton et al., 2007; Table S3). All *Arabidopsis* RNR1 sequences were aligned to the yeast and human RNR proteins to analyze whether any sequence alteration could be observed in the two catalytically important 10-15 amino acids-sized stretches called LOOP1 and LOOP 2 (Xu et al., 2006). Single amino acid, non-synonymous mutations located at LOOP 1/ LOOP 2 region cause the phenotypic defects in *A. thaliana* (*dpd2*, *cls8-2*, *cls8-3*). Therefore, we focused our analysis mostly on this region. A stretch of 11 amino acids was missing in the LOOP1 region of *AL.RNR1C*, although no such change was noticed in *AT.RNR1*, *AL.RNR1A*, *AL.RNR1B*, human and yeast copies (Figure 2A). In addition, *AL.RNR1C* was not detected in any of the expression analyses and therefore it was also denoted as a non-functional copy based on the expression data (Table S2). Correlation analysis indicated that *AL.RNR1B* is most closely related to *AT.RNR1* (r = 0.83) based on its stress-responsive gene expression pattern. Since it was also expressed in all organs like the *A. thaliana* gene, *AL.RNR1B* was predicted as the expressolog (Table 2; Table S2). *AL.RNR1A* also reported a good, albeit lower stress-related correlation (r = 0.64). However, its organ expression level was close to or below the detection level of the microarray analysis and a detailed examination of all three types of stress experiments indicated that only salt responsiveness was partially retained by *AL.RNR1A* leading to an expression above the detection threshold (Figure 2B; Table S2). Thus, *AL.RNR1A* could be a neo-functionalized co-ortholog which is only active in certain stress scenarios.

Since the low expression level of *AL.RNR1A* in unstressed conditions is an important signature for possible promoter mutations, we checked the presence/absence of important transcriptional regulators in the promoter regions of the *RNR1* genes. While overlapping, intact *AT.RNR1*-like TATA element and Y patches were predicted for *AL.RNR1B*, these were disrupted both in the *AL.RNR1A* and *AL.RNR1C* copies (Figure 2C). Finally, to check the reliability of expression-based prediction about gene functionality we had cloned the expressologous (*AL.RNR1B*) and non-expressologous (*AL.RNR1A*) gene copies and tested for complementation of the *AT.rnr1*/*AT.cls8-1* mutant (Table S3). Recovery of wild-type phenotype was observed in the case of *AL.RNR1B* complemented plants. However, *AL.RNR1A* complemented plants did not revert the mutant phenotype, which indicate that the *AL.RNR1A* homolog does not retain the RNR1 function (Fig. 2D). Although three independent transgenic lines each clearly differentiated the complementing from the non-complementing ortholog, we confirmed the presence of the transgene insertion of *AL.RNR1A* by PCR (Fig. S2); expression of the *AL.RNR1A* transgene was not detected by RT-PCR probably due to its low expression level as observed in *A. lyrata*.

Example 2: Event 1 of genetic redundancy

Interestingly, the gene(s) encoding the small subunit of ribonucleotide reductase (RNR2) were also among the genes of the one *A. thaliana*: multiple *A. lyrata* in addition to the genes encoding its large subunit (see above). The small subunit-related genes are *AT.TSO2* (AT3G27060), *AT.RNR2A* (AT3G23580) and *AT.RNR2B* (AT5G40942). However, among these three subunits, *TSO2* is biologically the most active copy. In *A. lyrata TSO2* is found to be duplicated resulting in *AL.TSO2A* and *AL.TSO2B*. The phenotype of *AT.tso2-1* revealed similar developmental defects like *AT.rnr1,* such as irregular leaves and homeotic transformations (Wang and Liu, 2006). Multiple sequence alignment of *AT.TSO2*, *AL.TSO2A* and *AL.TSO2B* reveals only one non-synonymous change between *AT.TSO2* and *AL.TSO2A*, while 28 non-synonymous changes were noticed between *AT.TSO2* and *AL.TSO2B* outside the region of important enzymatic function (Figure S3). The two *A. lyrata* copies are well expressed in different organs like the *A. thaliana* gene (Figure 3A). Correlation analysis based on its stress response pattern indicated that *AL.TSO2B* is closest to *AT.TSO2* (r = 0.85) and therefore predicted as the expressolog. However, *AL.TSO2A* also showed a reasonably good correlation (r = 0.55) (Table S2). This indicates that *AL.TSO2A* and *2B* are possibly redundant to each other within the resolution provided by our expression study. The promoter comparisons revealed that the TATA box and the Y patch were preserved in both *AL.TSO2A* and *AL.TSO2B*. Both *AL.TSO2A* and *AL.TSO2B* copies were cloned along with their native promoter and transformed into the *AT.tso2-1* plants. The transformed plants restored the wild-type phenotype in both cases and thus proved that *AL.TSO2A* and *AL.TSO2B* are functionally redundant in the analyzed context and orthologous to *AT.TSO2* (Figure 3B).

Example 3: Pseudogenization by acquiring changes in the coding region of the gene

STABILIZED 1 (STA1) is a pre-mRNA splicing factor. The gene function is similar to the human U5 small ribonucleoprotein and to the yeast pre-mRNA splicing factors Prp1p and Prp6p (Lee et al., 2006). *A. thaliana* harbors a single gene (AT4G03430), while in *A. lyrata* two copies, *AL.STA1A* and *AL.STA1B*, have been identified by our OrthoMCL analysis. The *A. thaliana* loss-of-function mutant shows many developmental and stress-related phenotypes, such as smaller plant height, smaller leaf size and higher sensitivity to ABA as compared to the wild type (Lee et al., 2006). The expression level of *AL.STA1B* was below the detection limit of our microarray analysis in all the three organs and in all stress scenarios (Table S2). On the contrary, *AL.STA1A* was expressed above the detection limit and was similarly regulated under diverse stress conditions like *AT.STA1* (r = 0.75) (Figure 4A; Table S2). Therefore, we predicted that while *AL.STA1A* was the expressolog, *AL.STA1B* had

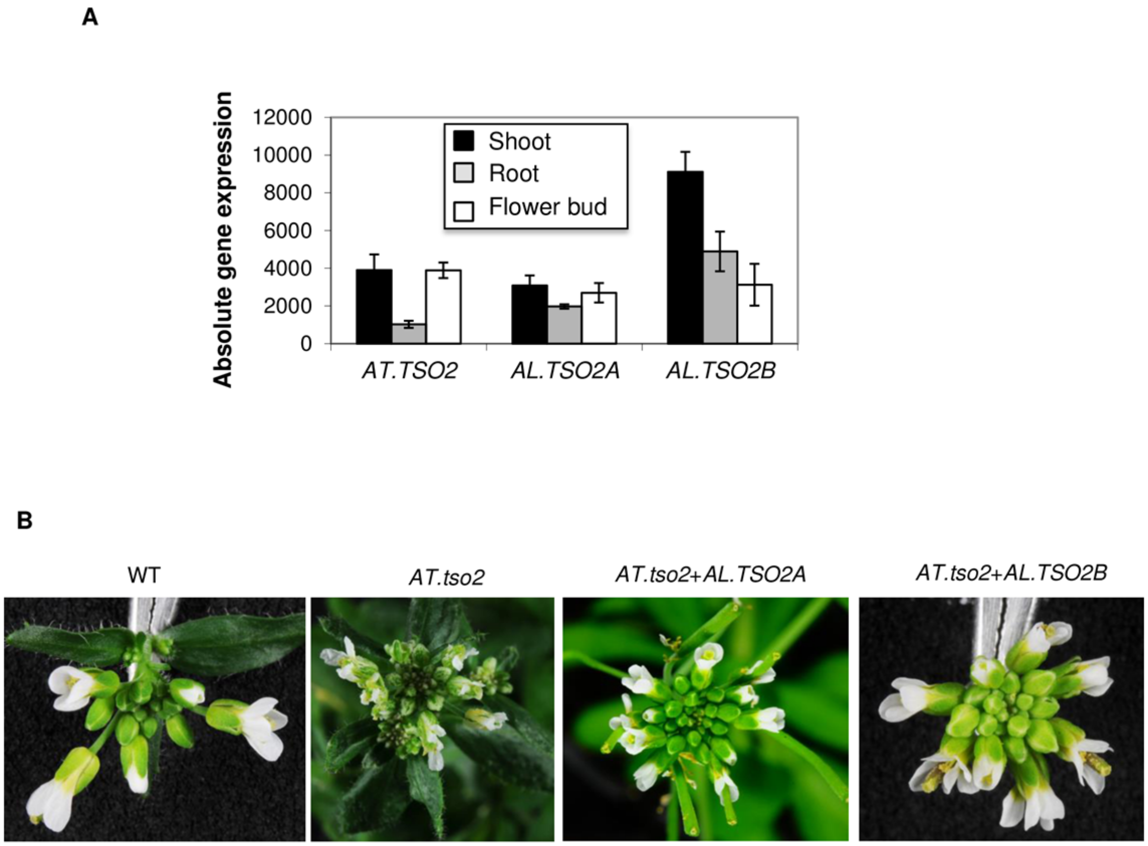


Figure 3. Gene expression and genetic complementation analyses of ribonucleotide reductase small sub-unit (*TSO2*) gene copies in *A. thaliana* and *A. lyrata*. (A) Expression of the three *Arabidopsis TSO2* genes in root, shoot and flower bud. All the copies are expressed well above background. (B) Genetic complementation of *A. thaliana tso2-1* with *AL.TSO2A* and *AL.TSO2B* gene copy. Both *AL.TSO2A* and *AL.TSO2B* were predicted as expressologs by our analysis, although the sequence analysis indicated several non-synonymous changes in the *AL.TSO2B* copy as compared to the *A. thaliana* gene (Fig. S3A). The result of the complementation assay supported this prediction.

presumably been pseudogenized (Table 2). We checked the coding regions of *AL.STA1B* for additional indications of its pseudogenization. Although *AT.STA1* does not contain any intron, a 43 nucleotide long intron was predicted for the *AL.STA1A* gene model, while three introns of 50, 44 and 324 nucleotides length were predicted for the *AL.STA1B* gene model. Therefore, we sequenced the *AL.STA1B* cDNA to verify such splicing events in this *A. lyrata* gene. However, the *AL.STA1B* cDNA sequence indicated that it was also an intronless gene like *AT.STA1*. Additionally we detected the insertion of one A nucleotide at position 1352 of the *AL.STA1B* CDS, which causes a premature stop codon and possible pseudogenization of this gene copy (Fig. S4). To check the accuracy of this prediction we cloned both *AL.STA1A* and *B* copies and transformed them in *At.sta1-1* plants (Table S3). The wild-type phenotype could

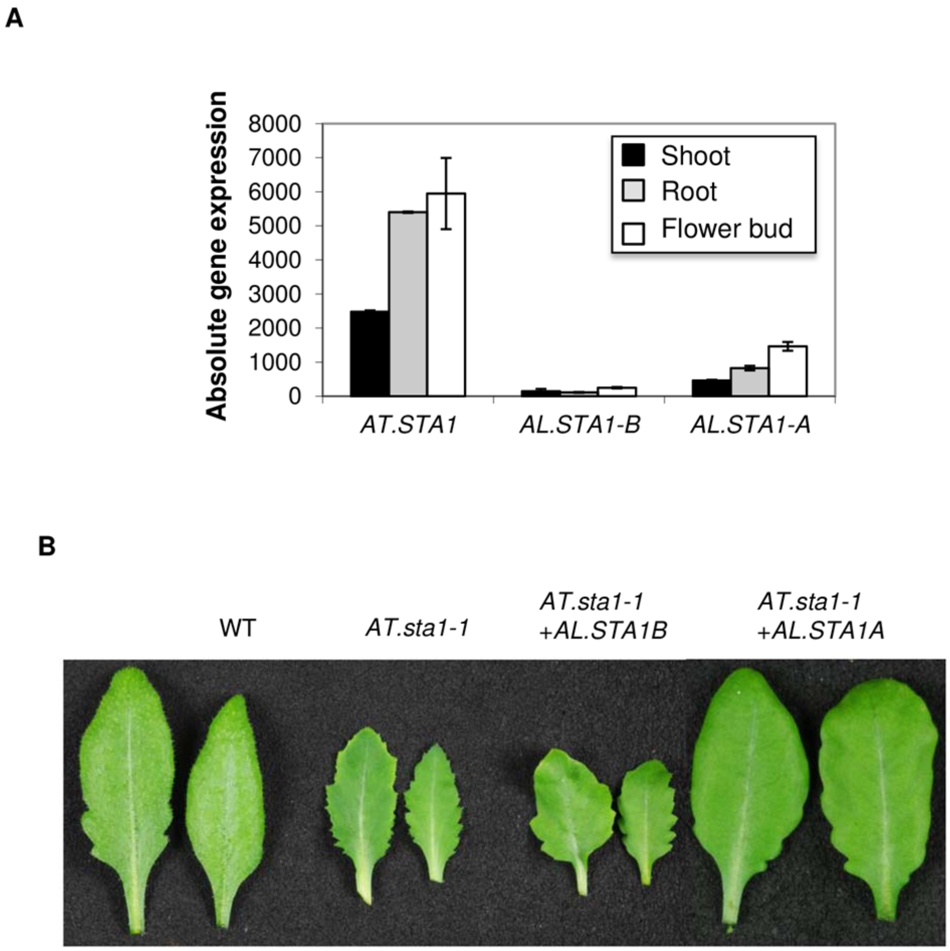


Figure 4. Gene expression and genetic complementation analyses of *STABILIZED 1* (*STA1*) gene copies in *A. thaliana* and *A. lyrata*. (A) Study of the expression patterns of the three *Arabidopsis STA1* genes in root, shoot and flower bud. (B) Comparison of leaf morphology of *A. thaliana sta1-1* plants with *AT.sta1-1+AL.STA1B* and *AT.sta1-1+AL.STA1A* complemented lines. While the leaf size and margins of *AL.STA1A* complemented plants look like the wild type, the *AL.STA1B* transformed lines resemble the mutant.

be recovered for *AL.STA1A*-transformed plants, while plants harboring the *AL.STA1B* construct still exhibited the mutant phenotype (Figure 4B). Three independent transgenic lines each clearly differentiated the complementing from the non-complementing ortholog. Furthermore, the presence of the transgene insertion of the non-complementing *AL.STA1B* was confirmed by PCR (Fig. S5); expression of the *AL.STA1B* transgene was not detected by RT-PCR probably due to its low expression level as observed in *A. lyrata*.

Example 4: Event 2 of genetic redundancy

Manganese transporter 11 (MTP11) is a member of the large cation diffusion family and is involved in Mn^{2+} transport and tolerance (Gustin et al., 2011). It exists as a single copy gene

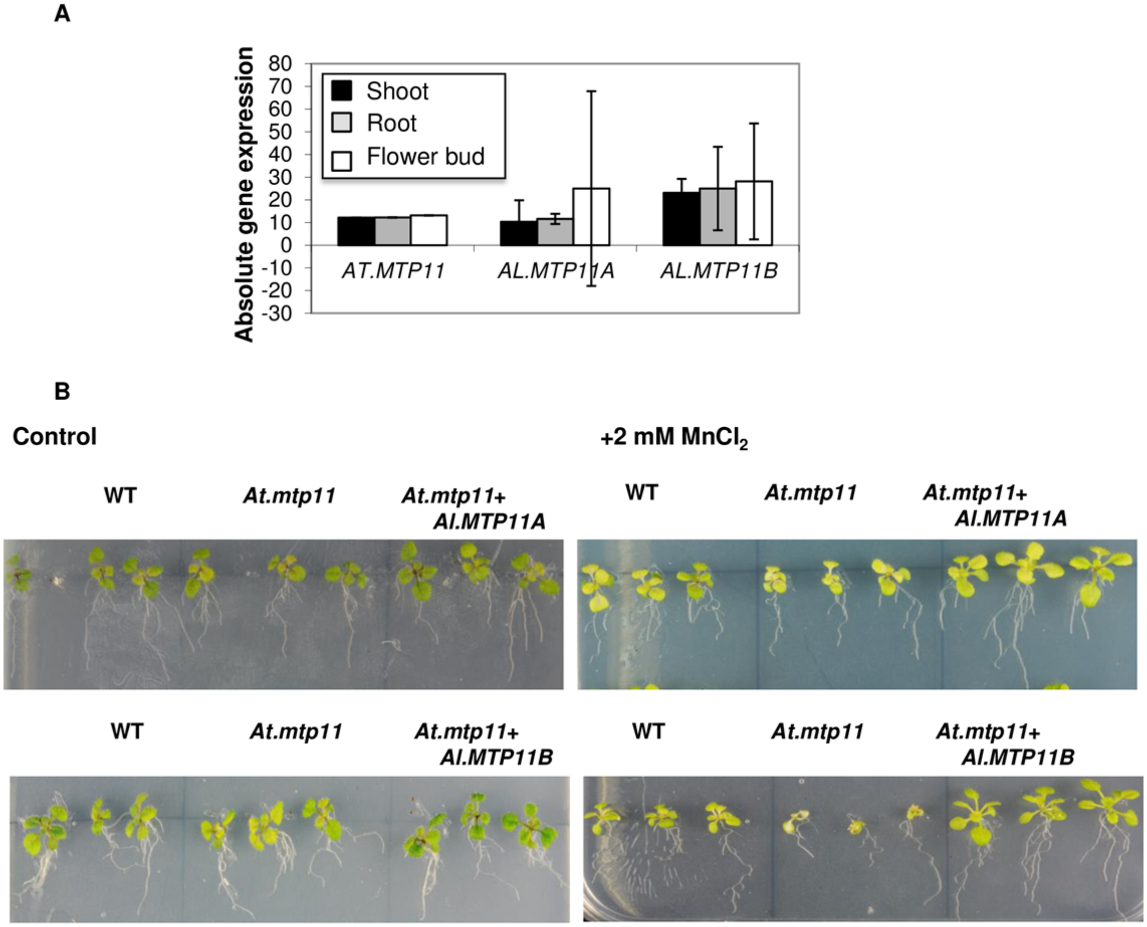


Figure 5. Gene expression and genetic complementation analyses of *MANGANESE TRANSPORTER 11* (*MTP11*) gene copies in *A. thaliana* and *A. lyrata*. (A) Organ expression patterns of the two *A. lyrata MTP11* homologous genes assessed by RT-qPCR analysis. (B) Genetic complementation of *A. thaliana mtp11* loss-of-function mutant with *AL.MTP11A* and *AL.MTP11B* gene copies. Seedlings were grown on MS medium (1% sucrose, 1X MS, 1.2% phytoagar) for eight days and then transferred to agarose medium supplemented with 2 mM Mn^{2+}. As predicted by the expression data, both the homologs could complement the mutant phenotype.

in *A. thaliana* (AT2G39450), while duplicated copies of *AL.MTP11A* and *AL.MTP11B* were identified in *A. lyrata*. The loss-of-function *AT.mtp11* plants are more sensitive to Mn^{2+} ions. Under Mn^{2+}-stressed condition, they grow less vigorously as compared to the wild-type plants (Delhaize et al., 2007). The study of phylogenetic relationship and high sequence homology indicate that *AL.MTP11A* and *AL.MTP11B* share a recent origin. Since the two *A. lyrata* homologs are highly similar (98% identity at the CDS level), it was not possible to design gene-specific microarray probes. Therefore, we assessed their gene expression pattern by RT-qPCR analysis. Both homologs were well expressed in different organs and the expression levels were comparable to that of the *AT.MTP11* gene; thus they are predicted to be genetically redundant with respect to this data set (Figure 5A). To confirm the functional equivalence of *AL.MTP11A* and *AL.MTP11B*, we transformed the full-length genes along with their native promoters into the *A. thaliana* mutant plants. Phenotypic assay of the knockout and transformed lines revealed that while the growth of the mutant line was compromised on plates containing 2 mM $MnCl_2$, growth of both transgenic lines was similar to *A. thaliana* wild-type plants (Figure 5B, Figure S6), indicating genetic redundancy in this context and functional equivalence of both *A. lyrata* gene copies.

DISCUSSION

Plants are sessile and are subject to varying environmental stresses. Gene duplication is the event by which a plant can gain novel adaptive genes that enable them to meet their specific ecological needs (Conant and Wolfe, 2008; Ha et al., 2009; Van de Peer et al., 2009). All plant genomes sequenced to date have undergone at least one round of whole genome duplication (Fischer et al., 2014). While gene duplication is evolutionarily advantageous for the polyploidized plants, it imposes a challenge to transfer gene function annotation across species barriers by simple sequence comparisons. Since prediction of a correct annotation is key to any genome sequencing project and translational approaches, a number of sequence homology based methods have been developed (Gabaldon, 2008; Kuzniar et al., 2008). While these tools are effective for single copy genes and for genome-wide comparisons, additional support is required for large multi-gene families. OrthoMCL is one such tool, which is commonly used in genome-wide comparisons. Therefore, this method was employed to analyze *A. thaliana* and *A. lyrata* CDS identifying 2850 genes (6.5% of *A. thaliana* transcripts) that exist as one-to-many or many-to-many copies between these two species. Such uncertainty in predicting functional orthologs may be even worse in crop species of the *Brassica* lineage, which have undergone one round of whole genome triplication in addition to whole genome duplication events shared with the *Arabidopsis* lineage. Therefore, in such a situation where coding sequence-based analyses are limited with respect to assigning functional orthology, additional support is required.

Two major approaches have been proposed to address this issue. The most popular is gene co-expression analysis that has been successfully used for well characterized genomes for which large scale expression data are already available (Stuart et al., 2003, Bergmann et al., 2004, Mutwil et al., 2011, Movahedi et al., 2011). The second method also relies on extensive gene expression profiles obtained from comparable tissues among the species compared and subsequent implementation of a ranking system of genes based on expression similarity with top ranked genes called expressologs (Patel et al., 2012). However, both methods are dependent on the availability of large sets of highly comparable expression data obtained from diverse tissues and conditions for all of the species of interest, which is not feasible for newly sequenced genomes. Similarly, protein-protein interaction network data, which can assist the identification of functional orthologs among large paralogous gene families, would not be available in these cases (Bandyopadhyay et al., 2006). Therefore, in the current study we have tested the utility of a relatively small set of gene expression data for the prediction of functionally related orthologs. We compared the expression pattern in three organs and conducted Pearson correlation analysis based on data obtained from stress gene expression experiments in *A. thaliana* and *A. lyrata*. In contrast to a very low basic mean correlation value of 0.019 obtained from all-against-all comparisons between *A. thaliana* and *A. lyrata* transcriptome, a much higher correlation value of 0.329 was obtained when syntenic *A. thaliana* - *A. lyrata* orthologous gene groups were analyzed. This finding proves that though our gene expression dataset is small in size, it is appropriate to identify functionally related genes across species. The correlation analyses also revealed that OrthoMCL one-to-one gene group holds a higher correlation value (r_{mean} = 0.354) than that of the OrthoMCL all (r_{mean} = 0.313) or OrthoMCL many-to-many gene groups (r_{mean} = 0.300, Table 1). This illustrates some limitations of OrthoMCL analyses to predict functional orthologs in one-to-many or many-to-many situations.

The analyses of transcriptional expression patterns and correlations in this study show that functional categorization and prediction of expressologs based on gene expression patterns are possible for one-to-many orthologous relationships. In all tested gene groups we could identify the functional fate of duplicated *A lyrata* genes. In 12.5% of the gene groups at least one *A. lyrata* gene copy had been putatively non-functionalized. It should be noted, however, that the putative non-functionalization is based on very low expression levels for a limited set of expression data and that these genes might well be expressed under other, untested conditions. These limitations may well apply to the other categories as well. Approximately 18% of the gene groups suggest a functional divergence on the basis of at least one conserved co-ortholog, since at least one *A. lyrata* gene has undergone neo-functionalization. The biggest group (42%) is composed of genes that show species-specific gene expression patterns and three forms of such expression patterns were recorded. The two *Arabidopsis* species are phylogenetically very close and have diverged only 10 Myr ago (Hu et al., 2011). However, they have adapted to distinct environments and therefore show many differences in terms of their life cycles as well as in their reproductive and ecological habitats (Mitchell-Olds, 2001; Clauss and Koch, 2006). Therefore such species-specific gene expression patterns may have evolved in relation to these different life styles.

In 25% of the cases our analyses consists of duplicated clusters for which both copies exhibit similar expression patterns and therefore were assumed to be genetically redundant within the tested conditions. Although no clear evidence of sub-functionalization was noticed in the current study, it is possible that more extended gene expression analyses may identify such candidates among the currently classified genetically redundant group (MacCarthy and Bergman, 2007).

Taken together, our finding indicates that expressologs strongly reduce the existing uncertainty associated with the coding sequence homology-based methods to assign functional orthologs in the presence of multiple orthologs. However, there are still limitations. One major challenge of this approach is to identify comparable biological tissues or experimental conditions (severity of the applied stresses, time points etc.) to measure comparable expression patterns of the targeted genes across species. For phylogenetically and/or ecologically distant species, it might be challenging to find comparable conditions and further studies are required to test the efficiency of predicted expressologs in these situations.

We also examined whether comparative analysis of promoter sequences could be used as a satisfying alternative to predict functional orthology, when comparative gene expression data are not or not yet available for a species, e.g. in case of newly sequenced genomes. Therefore promoter divergence analyses were performed to check in how many instances the promoters of predicted expressologs were less divergent than those of the non-expressologs or non-functionalized genes. If the gene predicted as expressolog based on our gene expression analyses harbors less promoter divergence (d_{SM}) than that of the non-expressolog, then we would assume that the functional orthology prediction based on gene expression patterns and promoter divergence analyses overlapped with each other. We investigated individual OrthoMCL gene clusters and found that in 78% of the studied cases for the non-functional group and 77% of the neo-functionalized gene group the *A. lyrata* expressologs indeed had less promoter divergence than that of the non-expressologs or non-functionalized genes. The overlap between predictions made by expressolog and promoter divergence is even much lower for the genetic redundancy (60%) and species-specific categories (47%). Thus, in the absence of gene expression data, determination of promoter divergence can be complementary to the limitations of CDS-based methods such as OrthoMCL. However, a considerable number of genes could still not be correctly annotated. Furthermore, there are some other serious limitations that have to be considered in the case of promoter sequence analysis: (1) Determination of the boundaries of the promoter regions is a critical issue, since cis elements have been reported in *Arabidopsis* to be located several kilobases upstream of the transcription start (Rombauts et al., 2003). Moreover, small 5'-exons or divergent UTR sizes can result in the comparison of completely unrelated sequences. (2) With the increase in the phylogenetic distance of the species compared, an altered sequence composition of the cis elements may jeopardize their classification and the deduction of promoter divergence scores.

(3) Unlike gene expression analysis, promoter analyses cannot study the mode of gene function diversification. For example, the promoter divergence analyses in the case of genetic redundancy or species-specific categories could not reveal any distinction between these two divergent categories. However, detailed gene expression analyses revealed that in the case of genetic redundancy the two or more *A. lyrata* genes were regulated in the same direction as the *A. thaliana* gene, whereas in the event of species-specific expression the two or more *A. lyrata* genes were regulated in a diverse manner from the *A. thaliana* gene. Therefore, expressologs even based on a small set of expression analyses like in our study are a reliable and superior tool that can supplement the genome annotation pipeline for a more accurate transfer of gene functions.

Although previous studies and our data support the hypothesis that large and even small scale gene expression data could provide clues about gene functionality, no studies have been conducted so far to check the reliability of such prediction *in planta*. Here we show that this concept is valid by testing it experimentally for the evolutionary categories non-functionalization and neo-functionalization, and for two genes of the genetic redundancy class. We provide *in planta* evidence that the two genes of the two *Arabidopsis* species having closest expression patterns (expressologs) are functionally comparable (functional orthologs).

Two of the case studies were implying genetic redundancy based on the expressolog classification. Coding sequence-based prediction, promoter analysis and expressolog prediction for *AL.MTP11A* and *AL.MTP1B* had pointed toward their possible functional similarity. On the contrary, identification of 28 non-synonymous nucleotide changes and a high promoter divergence between *AT.TSO2* and *AL.TSO2B* indicate their possible functional divergence, although our gene expression analysis predicted *AL.TSO2A* and *AL.TSO2B* as expressologs. The functional equivalence as predicted by the expressolog classification was eventually corroborated by the genetic complementation analysis. Two further cases referred to pseudogenization and to neo-functionalization events as deduced from the gene expression pattern. While *A. lyrata* genome annotation project did not identify these pseudogenes and non-expressologs, our promoter sequence and expression analyses indicate possible mechanisms and directions by which these genes have been evolving. Again, we could experimentally verify the authenticity of this prediction by transforming the *A. thaliana* loss-of-function mutants with the corresponding *A. lyrata* pseudogenized and neo-functionalized genes, which did not lead to complementation.

In conclusion, we could experimentally verify functional orthologs among the one *A. thaliana* : two (many) *A. lyrata* gene groups. These annotations could not be deduced using sequence-based algorithms only; instead, they were predicted based on comparative expression analyses. This success emphasizes the strength and added value of an expressolog/ non-expressolog classification based on an even limited set of expression data in order to predict the functional orthologs in such one : many gene groups.

MATERIALS AND METHODS

OrthoMCL analysis

OrthoMCL version 1.4 with an initial cutoff e-value 1e^{-05} was used for the BLASTP comparisons between the transcriptomes of *A. thaliana* and *A. lyrata*. Inflation parameter was set to 1.5 and all other parameters were set to default values as recommended by the developers. When exactly one *A. thaliana* and exactly one *A. lyrata* identifiers were identified in an OrthoMCL cluster, this was defined as a one-to-one situation. In case of one-to-many situations, one *A. thaliana* and multiple *A. lyrata* identifiers or *vice versa* were present in a cluster. For all further analyses we focused on one *A. thaliana* to multiple *A. lyrata* groups to tap the possibility of experimental verification by utilizing *A. thaliana* loss of function mutants.

Collection of tissues from stress induced plants and from different organs of *A. thaliana* and *A. lyrata*

Four-week-old *A. thaliana* Col-0 and six-week-old *A. lyrata* ssp. *lyrata* soil-grown plants were treated either with 250 mM or 500 mM NaCl solutions by flush flooding (to soak the soil for a short period), while the control group was watered. Leaf tissues were harvested at 3 h and 27 h post-treatment. To assess the effective salt exposure to the plants, the raw soil electrical conductivity (EC) was measured before tissue harvesting by the use of 5TE sensor attached to Procheck handheld datalogger (Decagon, USA). Since direct EC measurement in soil was not reproducible because of the presence of particles and air pockets, the soil EC in solution was measured by mixing a ratio of 1:5 soil: water (Fig. S7). The effective salt concentrations that the plants were subjected to was within the moderate range of soil salinity as per the recommendations made by the soil and water salinity testing protocol, Government of South Australia, fact sheet no.- No: 66/00 (www.pir.sa.gov.au/factsheets). For drought treatment, leaf samples were collected 8 d and 11 d after withdrawal of regular watering to the soil. Plants were exposed to UV-B radiation plus PAR (400-700 nm) of 140 µmol m^{-2} s^{-1}. The biological effective UV-B radiation, weighted after generalized plant action spectrum

(Caldwell, 1971) and normalized at 300 nm were 1.31 kJ m^{-2} and 2.62 kJ m^{-2} for 4 h and 8 h time points respectively. For collecting root tissues, *A. thaliana* and *A. lyrata* plants were hydroponically grown on a raft following standard procedures (Conn et al., 2013).

Microarray design

The *A. thaliana* array was customized by printing biological (43603) and replicated probe groups (50X5) commercially available from Agilent Technologies (Id: 029132). The design of *A. lyrata* probes was done by uploading total transcriptome (32670) to the Agilent e-array facility (https://earray.chem.agilent.com/earray/). One probe per target sequence was generated for 32386 transcripts, while no probes were reported for 284 sequences. These sequences were either repeat masked-out or did not pass the required quality check. The specificity of the designed probes was further confirmed by blasting against the *A. lyrata* transcriptome. In addition to the main probe group, a replicated probe-group of selected 477 genes was printed on the array for multiplicative de-trending. The mean probe length was 60. Both *A. thaliana* and *A. lyrata* arrays were printed in 8X60K format (Table S1).

RNA extraction, array hybridization and scanning

RNA was extracted by using a combination of Trizol (Invitrogen) and RNAeasy kit (Qiagen; Hilden, Germany; Das et al., 2010). Quality was checked by Bioanalyzer analysis (Agilent Technologies). Approximately 100 ng of total RNA was used for cRNA synthesis and subsequent Cy3 labeling by using the one color low-amp quick amplification labeling kit (Agilent Technologies). Array hybridization, washing and scanning was done according to the recommended procedures by Agilent Technologies.

Array data analysis

Data were extracted by using an Agilent scanner and an Agilent Feature Extraction program. Background corrected and multiplicatively de-trended hybridization signals were imported to GeneSpring (G3784AA, version 2011) for log$_2$ transformation and data normalization. The normalization conditions used were: threshold raw signals to- 1.0, normalization algorithm-scale, percentile target- 75. For stress data the normalized signal intensity values were baseline corrected to the median of all samples. However, to get normalized expression data in different organs, baseline transformation was turned off. To get information about the differential expression of genes in diverse stressed conditions a Z-score, i.e. the number of standard deviation changes between control and respective treatment were calculated. To know how tightly orthologous *A. thaliana* and *A. lyrata* genes were related in terms of gene expression, we have calculated the total Pearson correlation values for OrthoMCL all, OrthoMCL one-to-one, OrthoMCL multiple and syntenic gene groups (Table 1). Also Pearson correlation for individual At: Al Ortho MCL pairs were calculated to predict possible expressologs and non-expressologs based on expression similarity or divergence.

Real-time RT-qPCR analysis

Since AL.MTP11A and AL.MTP11B homologs are highly identical, the designed array probes were cross hybridizing to each other. Therefore, the expression levels of these two homologs were measured in shoots, roots and flower buds of *A. lyrata* by quantitative RT-PCR analyses. First strand cDNA was synthesized by using QuantiTect Reverse Transcription Kit (Qiagen, Hilden, Germany) and SYBR green fluorescence was used to measure the expression level of the targeted genes in *A. lyrata*. Transcript abundance of *AL.MTP11A* and *AL.MTP11B* homologs were calculated in geNORM by using *AL.UBQ5* and *AL.S16* as reference genes (Vandesompele et al., 2002, Table S4).

To design gene-specific primers we targeted two SNPs (there are only 20 SNPs over the entire coding sequence region among *AL.MTP11A* and *AL.MTP11B*) and designed gene-specific real-time RT-qPCR primers only based on one nucleotide sequence divergence at the 3' end. Indeed by restriction enzyme digestion and subsequent sequencing of the amplified PCR product we could confirm the identity of the amplified gene products (Fig. S4A and B). Gene-specific primers for *AL.STA1A* were designed from the CDS region located after the insertion of premature stop codon to avoid the possibility of getting amplification from truncated mRNA.

Promoter divergence analysis

DNA distance matrix for upstream sequences of *A. lyrata* genes (neo-/non- versus expressologs) with respect to the upstream sequence of their *A. thaliana* orthologs were calculated based on 1000 bp upstream sequences from the start codon. We obtained the divergence score (d_{SM}) for the upstream sequences of *A. lyrata* genes based on motif divergence method SSM (Castillo-Davis et al., 2004). For a pair of sequences, SSM calculates functional regulatory changes within the sequences and provides the divergence score (d_{SM}) that quantifies the fraction of unaligned regions between the sequences

Gene amplification, GATEWAY cloning, plant transformation and selection

Gene-specific loss-of-function mutants were either obtained from NASC (http://arabidopsis.info/) or from individual laboratories (Scholl et al., 2000; Table S3). High fidelity Phusion polymerase (New England Biolabs) was used to amplify genes plus native promoters of approximately 2-2.5 kb upstream 5'-region and 0.5 kb 3'-downstream region of *A. lyrata*. GATEWAY recombination sequences were always tagged to the 5'-end of the gene-specific primers (Table S4). Amplified PCR products were eluted from gels, cloned in pDONR221 vector and subsequently recombined to a modified, promotor-less pAlligator2 vector (35S promoter deleted by restriction with *HindIII* and *Eco RV*, blunting with T4 DNA polymerase and religation) (Benshimen et al., 2004; Wei Zhang and ARS for 35S promoter deletion). Since the promoter and 3'-UTR regions of *AL.MTP11A* and *B* were highly identical, the full length gene sequence for genetic complementation of both genes was amplified by identical primer pairs; the cloned fragments were analyzed by restriction digest and sequencing to distinguish *AL.MTP11A* and *B* isolates.

Cloning of the correct sequences was always confirmed by sequencing the entire insert. Finally the expression clones were mobilized to competent *Agrobacterium* pGV3101/pMP90 strains and *Arabidopsis* plants were transformed by the floral dipping method (Clough and Bent, 1998). Transformed T1 seeds were selected by observing green fluorescence of the GFP reporter gene; at least three independent T1 plants were subsequently phenotyped.

ACCESSION NUMBERS

Expression data from Agilent microarray hybridization are deposited at GEO (http://www.ncbi.nlm.nih.gov/geo/)

GSE80099- *A. thaliana* transcriptomic responses against drought stress

GSE80100- *A. thaliana* root and flower bud transcriptomes

GSE80108- *A. lyrata* ssp. *lyrata* root and flower bud transcriptomes

GSE80110- *A. lyrata* ssp. *lyrata* transcriptomic responses against drought stress

GSE80111- *A. thaliana* transcriptomic responses against UV-B stress

GSE80112- *A. lyrata* transcriptomic responses against UV-B stress

GSE80114 - *A. thaliana* transcriptomic responses against salt stress

GSE80115- *A. lyrata* transcriptomic responses against salt stress

SUPPLEMENTAL MATERIAL

Supplemental Table S1. Summary of *Arabidopsis thaliana* and *A. lyrata* array design features.

Supplemental Table S2. Classification of gene expression patterns and categorization of AT : AL gene groups.

Supplemental Table S3. *A. thaliana* mutants used in this study for genetic complementation with *A. lyrata* homologs.

Supplemental Table S4. Oligonucleotides used in this study for different purposes.

Supplemental Figure S1. Promoter sequences divergence analysis between expressologs and non-expressologs in the genetic redundant and species-specific functional categories.

Supplemental Figure S2. Confirmation of the presence of the transgene insertion of *AL.RNR1A* in three independent transgenic plants by PCR analysis.

Supplemental Figure S3. Sequence alignments of *AT.TSO2* and *AL.TSO2* homologous genes.

Supplemental Figure S4. Pseudogenization due to insertion of one A nucleotide at position 1352.

Supplemental Figure S5. Confirmation of the presence of the transgene insertion of *AL.STA1B* in three independent transgenic plants each by PCR analysis.

Supplemental Figure S6. Distinction of *ALMTP11A* and *B* homologs.

Supplemental Figure S7. Measurements of soil salinity for the stress assays used in the microarray based gene expression analyses.

Supplemental Table S1. Summary of *Arabidopsis thaliana* and *A. lyrata* array design features.

Supplemental Table S2. Classification of gene expression patterns and categorization of AT : AL gene groups. Gene groups consisting of one *A. thaliana* gene and multiple *A. lyrata* genes (One-to-two, one-to-many cases) are listed with their expression values in three organs (rosette leaves, roots, flowers) and stress conditions (salt, drought, UV-B) analyzed. These data were used to classify the *A. lyrata* genes as expressologs, non-expressologs or as non-functional copies. This classification is further used to assign the gene groups to different evolutionary categories.

Specific information to columns:

(B) Gene groups are categorized (see also Introduction): Group 1 Non-functionalization: one ortholog retains the original function (expressolog), while the other ortholog(s) is (are) non-functional; Group 2 Neo-functionalization: one ortholog is a non-expressolog, while the other ortholog retains the original function (expressolog); Group 3 Species-specific functionalization: a) all orthologs are neo-functionalized (non-expressologs), or b) one ortholog is a non-expressolog, while the other(s) lost the original function (non-functionalized); Group 4 Species-specific non-functionalization: all orthologs are non-functionalized/ pseudogenized; Group 6 Redundancy: all co-orthologs retain the function. There were not any clear indications of Group 5 Sub-functionalization (see Introduction & Discussion), where the original functions would be split among the orthologs. However, the Group 3a and Group 6 could include such members, which are not resolved within the 11 scenarios analyzed in this study.

(C) Classification of orthologs: *A. lyrata* genes were classified as expressolog, non-expressolog according to the combined expression code (column D); if an ortholog was not detected above a level of log2 = 9 for the normalized expression values in any of the 11 scenarios analyzed, it was denoted 'non-functional'.

(D) Combined expression code (sum of columns F + H): Expressologs: +2 = expressolog, +1 = non-stress-resp. expressolog; Non-expressologs: 0 = non-expressolog (stress), -1 = non-stress-resp. non-expressolog, -2 = non-expressolog (both organ & stress).

(E) Stress expressolog: Genes showing a Pearson correlation for the total stress responses (columns AD to AK) >0.3 were regarded as expressologs according to our stress experiments; however, in case there was no stress response (i.e. a change below log2 = I0.9I) observed in any condition (Total stress index = 0; column J, see also columns K,M,O), the corresponding genes were classified as not stress-responsive.

(G) Organ expressolog: Organ expressions were considered to be conserved, if the expression was present (log2 value >= 9) or absent in the same pattern for rosette leaves, roots and flowers. The absolute value was not considered. To account for the variability of the measurements, a log2 value of at least 8.40 was accepted as a detectable expression, if the expression in the particular organ was recorded in other members of the group.

(I) Pearson correlation of stress response of *A. thaliana* gene and individual *A. lyrata* co-orthologs.

(J) Total stress response index. Sum of individual stress indices (salt, drought, UV; columns K, M, O) to indicate any stress-response in our experiments.

(K, M, O) Indices for stress response [1, response; 0 no response] in the related stress experiments. A log2-fold change above |0.9| was regarded as stress-response.

(Q,R,S) Log2 normalized expression level in the respective organs (Methods).

(U - AB) Log2 normalized expression level upon the indicated stress experiments.

(AD - AK) Log2-fold changes in response to the indicated stresses with respect to the control condition (Methods).

(AM) Synteny: genes present in syntenic regions or not, "-" represents absence of data.

(AN) d_{SM} scores of promoter sequences analyses (Methods). "-" represents absence of data.

Supplemental Table S3. *A. thaliana* mutants used in this study for genetic complementation with *A. lyrata* homologs. The complemented plants were phenotyped according to the conditions described in the original reference.

Supplemental Table S4. Oligonucleotides used in this study for different purposes.

*Since, the promoter and 3' UTR region of *AL.MTP11A* and *B* are sequentially highly similar, we used the same oligonucleotide pairs to amplify both homologs.

Supplemental Figure S1. Promoter sequences divergence analysis between expressologs and non-expressologs in the genetic redundant and species-specific functional categories (A) Promoter analyses of the gene group, where all the *A. lyrata* genes present within a gene group are genetically redundant to *A. thaliana* gene as predicted by gene expression analyses. For each *A. thaliana* and *A. lyrata* orthologous gene clusters we calculated the differences in promoter sequence divergence scores (delta d_{SM} scores) between the *A. lyrata* gene copies. For clusters with more than two *A. lyrata* gene copies we considered any two *A. lyrata* copies by random choice. In order to define the clusters that are closer in their d_{SM} values we have considered a threshold value of delta d_{SM} which has been shown by dotted line (<0.2). (B) Promoter analyses of the gene group, where all the *A. lyrata* genes present within a gene group are depicting species-specific difference compared to *A. thaliana* gene as predicted by gene expression analyses. The other parameters used were the same as described in Figure S1A.

Supplemental Figure S2. Confirmation of the presence of the transgene insertion of *AL.RNR1A* in three independent transgenic plants by PCR analysis. An *AL.RNR1A*-specific primer pair (Table S4) was used to amplify a diagnostic fragment from genomic DNA, which was absent from an untransformed control plant (WT).

Supplemental Figure S3. Sequence alignments of *AT.TSO2* and *AL.TSO2* homologous genes. (A) Multiple sequence alignment of *AT.TSO2*, *AL.TSO2A*, *AL.TSO2B*, human (HS.RRM2, NP_001025) and *Saccharomyces cerevisiae* (SC.RNR2, NP_012508) sequences. The nucleotides highlighted in red are identified non-synonymous changes between *AT.TSO2* and *AL.TSO2A*/*AL.TSO2B*. Twenty eight non-synonymous changes were noticed between *AT.TSO2* and *AL.TSO2B*. Regions containing residues of important for enzymatic function are underlined (Philipps et al., 1995). Twenty-seven out of 28 amino acid changes for TSO2B are outside these regions indicating that gene function of TSO2B was probably not affected by these changes. The bold, underlined amino acids are three known TSO2 alleles in *Arabidopsis thaliana* (Wang and Liu, 2006). (B) Multiple alignment of 1000 bp upstream region of *AT.TSO2*, *AL.TSO2A* and *AL.TSO2B*. The bold, underlined sequence (CTCCTATATAAATA) is the TATA box in the core promoter region of AT2G21790; while underlined region (TCTCTTCTTC) is the Y patch. Y Patch is a direction-sensitive plant core promoter element that appears around TSS.

Supplemental Figure S4. Pseudogenization due to insertion of one A nucleotide at position 1352 (marked in red font and underlined), which causes premature insertion of the stop codon (-) in the *AL.STA1B* gene copy.

Supplemental Figure S5. Confirmation of the presence of the transgene insertion of *AL.STA1B* in three independent transgenic plants each by PCR analysis. An *AL.STA1B*-specific primer pair (Table S4) was used to amplify a diagnostic fragment from genomic DNA, which was absent from an untransformed control plant (WT).

Supplemental Figure S6. Distinction of *ALMTP11A* and *B* homologs. (A) Pairwise sequence comparison of the two *A. lyrata* MTP11A and B homologs to design gene specific primers (bold, underlined) for quantitative real time RT-PCR analyses. Presence of a restriction enzyme (*Aci* I) cut site (CCGC) was detected and underlined in the *AL.MTP11B* sequence, which is absent in the *AL.MTP11A* sequence. (B) To confirm specificity of the amplified PCR products both the amplified fragments were digested with *Aci* I. Only one fragment was noticed for *AL.MTP11A* (lane 1), while two fragments were noticed for *AL.MTP11B* (lane 2).

This primer pair was used in the real-time RT-qPCR analyses to calculate expression patterns of these two homologs. M, DNA size marker (pUC19 digested with *Msp* I).

Supplemental Figure S7. Measurements of soil salinity for the stress assays used in the microarray based gene expression analyses. The abbreviations used: At.C.E- *A. thaliana*, control, early time point (3h); At.S2.E- *A. thaliana*, 250 mM NaCl, early time point (3h); At.S1.E- *A. thaliana*, 500 mM NaCl, early time point (3h); At.C.L- *A. thaliana*, control, late time point (27h); At.S2.L- *A. thaliana*, 250 mM NaCl, late time point (27h), At.S1.L- *A. thaliana*, 500 mM NaCl, late time point (27h); Al.C.E- *A. lyrata*, control, early time point (3h); Al.S2.E- *A. lyrata*, 250 mM NaCl, early time point (3h), Al.S1.E- *A. lyrata*, 500 mM NaCl; Al.C.L- *A. lyrata*, control, late time point (27h), Al.S2.L- *A. lyrata*, 250 mM NaCl, late time point (27h), Al.S1.L- *A. lyrata*, 500 mM NaCl, late time point (27h).

.

ACKNOWLEDGMENTS

We thank Glenn Thorlby, Zhongchi Liu and Byeong-ha Lee for providing us mutant seeds, which were used in the current study. Generous technical assistance from Birgit Geist and Elisabeth Becker and helpful discussions with Andreas Albert, Werner Heller, Soumita Poddar and Debarun Acharya are also gratefully acknowledged. We also thank Jörg Durner for his continuous encouragement during the course of this study.

FIGURE LEGENDS

Figure 1. Promoter sequences divergence analysis between expressologs and non-expressologs in two different functional categories.

(A) Shared motif divergence scores (d_{SM}) of expressologs, non-expressologs and one-to-one orthologs. The first panel compares between the promoter sequence divergence scores of *A. lyrata* expressologs and neo-functionalized non-expressologous genes (as predicted by our gene expression analysis), the second panel compares between the expressologs and non-functionalized genes, the third panel compares between the promoter sequence divergence scores of *A. lyrata* genes having single orthologous copy of *A. thaliana* genes (as predicted by Ortho-MCL).

(B) Promoter analyses of the gene group, where at least one *A. lyrata* gene has been non-functionalized as predicted by gene expression analyses. For each *A. thaliana* and *A. lyrata* orthologous gene pair within a gene group, we calculated the promoter sequence divergence scores (d_{SM}) of *A. lyrata* genes with reference to the promoter sequence of their *A. thaliana* orthologous gene (in x-axis) by shared motif divergence method. Here, "o" represents the promoter sequence divergence score (d_{SM}) of *A. lyrata* gene copy predicted as expressolog by the gene expression analyses, "Δ" stands for that of non-expressologs.

(C) Promoter analyses of the gene group, where at least one *A. lyrata* gene has been predicted to be neo-functionalized. The other parameters used were same as described in Fig. 1A.

Figure 2. Sequence, gene expression and genetic complementation analyses of ribonucleotide reductase large sub-unit (*RNR1*) gene copies in *A. thaliana* and *A. lyrata*. (A) Multiple alignment of the *Arabidopsis* RNR1 amino acid sequences along with those of human and yeast sequences. Biologically important LOOP1 and LOOP2 regions are depicted. Part of the highly conserved LOOP1 region is missing and two non-synonymous amino acid changes were detected in the LOOP2 region of *AL.RNR1C*. However, the *AL.RNRB* coding sequence is identical to that of *AT.RNR1*. These regions play important roles in the enzymatic function by controlling specificity of the incoming dNTP. The biological importance of this region is emphasized by the identification of three mutations that caused severe developmental defects (indicated by * on top). Another allele, *cls8-1* affects a distant region leading to an amino acid change G718E, however showing the same mutant phenotype (Tab. S3). (B) Study of the expression patterns of the four *Arabidopsis RNR1* genes in root, shoot and flower bud. Background corrected and multiplicatively de-trended signal intensities were imported to Gene Spring (G3784AA, version 2011) to calculate normalized gene expression values (see Methods for details). (C) Comparison of core promoter regions (250 bp upstream from ATG) indicates the loss of the *AT.RNR1*-like TATA box (bold, underlined) and Y patch (underlined) in the case of *AL.RNR1A* and *AL.RNR1C* homologs. This analysis was done in plant promoter database (http://133.66.216.33/ppdb/cgi-bin/index.cgi#Homo). (D) Genetic complementation of *A. thaliana rnr1*/*cls8-1* with *AL.RNR1B* and *AL.RNR1A* gene copies. The phenotype of the *AL.RNR1B* (predicted expressolog) complemented plants resemble wild type. However, the plants complemented by *AL.RNR1A* (predicted pseudogene) show the mutant phenotype such as the yellowish, first true leaf (in the inset) and the crinkled, matured leaves.

Figure 3. Gene expression and genetic complementation analyses of ribonucleotide reductase small sub-unit (*TSO2*) gene copies in *A. thaliana* and *A. lyrata*. (A) Expression of the three *Arabidopsis TSO2* genes in root, shoot and flower bud. All the copies are expressed well above background. (B) Genetic complementation of *A. thaliana tso2-1* with *AL.TSO2A* and *AL.TSO2B* gene copy. Both *AL.TSO2A* and *AL.TSO2B* were predicted as expressologs by our analysis, although the sequence analysis indicated several non-synonymous changes in the *AL.TSO2B* copy as compared to the *A. thaliana* gene (Fig. S3A). The result of the complementation assay supported this prediction.

Figure 4. Gene expression and genetic complementation analyses of *STABILIZED 1* (*STA1*) gene copies in *A. thaliana* and *A. lyrata*. (A) Study of the expression patterns of the three *Arabidopsis STA1* genes in root, shoot and flower bud. (B) Comparison of leaf morphology of *A. thaliana sta1-1* plants with *AT.sta1-1+AL.STA1B* and *AT.sta1-1+AL.STA1A* complemented lines. While the leaf size and margins of *AL.STA1A* complemented plants look like the wild type, the *AL.STA1B* transformed lines resemble the mutant.

Figure 5. Gene expression and genetic complementation analyses of *MANGANESE TRANSPORTER 11* (*MTP11*) gene copies in *A. thaliana* and *A. lyrata*. (A) Organ expression patterns of the two *A. lyrata MTP11* homologous genes assessed by RT-qPCR analysis. (B) Genetic complementation of *A. thaliana mtp11* loss-of-function mutant with *AL.MTP11A* and *AL.MTP11B* gene copies. Seedlings were grown on MS medium (1% sucrose, 1X MS, 1.2% phytoagar) for eight days and then transferred to agarose medium supplemented with 2 mM Mn^{2+}. As predicted by the expression data, both homologs could complement the mutant phenotype.

Table 1. Pearson correlation analysis of stress induced gene co-expression data between different groups of orthologous and non-orthologous genes of *A. thaliana* and *A. lyrata*.

Gene groups	**Mean**	**Median**
OrthoMCL all	0.272	0.313
OrthoMCL one-to-one	0.320	0.354
OrthoMCL multiple	0.262	0.300
Syntenic	0.329	0.369
A. thaliana vs A. lyrata	0.019	0.036

Table 2. Gene expression similarity (r) and promoter sequence divergence (d_{SM}) for selected genes analyzed by genetic complementation assay.

*Because of very high sequence similarity microarray probes were not gene specific and hence expression similarity was assessed by RT-qPCR analyses.

Gene name	Gene identifier	Promoter divergence score (dsm)	Correlation based on stress expression data (r)	Syntenic gene	Predicted expressolog	Functional Ortholog based on genetic complementation
AT.RNR1	AT2G21790	-	-	-	-	-
AL.RNR1A	Al_scaffold_0007_128	0.685	0.64	No	No	No
AL.RNR1B	fgenesh2_kg.4__104__AT2G21790.1	0.333	0.83	Yes	Yes	Yes
AL.RNR1C	scaffold_200715.1	0.676	0.65	No	No	Not tested
AT.STA1	AT4G03430	-	-		-	-
AL.STA1A	fgenesh2_kg.6__3353__AT4G03430.1	0.197	0.75	Yes	Yes	Yes
AL.STA1B	scaffold_700051.1	0.815	-0.50	No	No	No
AT.TSO2	AT3G27060	-	-		-	-
AL.TSO2A	fgenesh2_kg.5__483__AT3G27060.1	0.155	0.56	Yes	No	Yes
AL.TSO2B	scaffold_703867.1	0.828	0.86	No	Yes	Yes
AT.MTP11	AT2G39450	-	-	-	-	-
AL.MTP11A	fgenesh2_kg.4__2026__AT2G39450.1	0.307	-	Yes	Yes*	Yes
AL.MTP11B	fgenesh2_kg.463__5__AT2G39450.1	0.297	-	No	Yes*	Yes

Parsed Citations

Bandyopadhyay S, Sharan R, Ideker T (2006) Systematic identification of functional orthologs based on protein network comparison. Genome Res 16: 428-35

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

Bergmann S, Ihmels J, Barkai N (2004) Similarities and differences in genome-wide expression data of six organisms. PLoS Biol 2: E9

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

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16: 1679-1691

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

Bensmihen S, To A, Lambert G, Kroj T, Giraudat J, Parcy F (2004) Analysis of an activated ABI5 allele using a new selection method for transgenic Arabidopsis seeds. FEBS Lett 561: 127-131

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

Caldwell MM (1971) Solar UV irradiation and the growth and development of higher plants. In AC Giese eds, Vol 6, Chapter 4, Academic Press, New York, pp 131-177

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

Castillo-Davis CI, Hartl DL, Achaz G (2004) cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res 14: 1530-1536

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

Clauss MJ, Koch MA (2006) Poorly known relatives of Arabidopsis thaliana. Trends Plant Sci 11: 449-459

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

Clough SJ, Bent AF (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. Plant J 16: 735-743

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

Conant GC, Wolfe KH (2008) Turning a hobby into a job: How duplicated genes find new functions. Nature Rev Genet 9: 938-950

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

Conn SJ, Hocking B, Dayod M, Xu B, Athman A, Henderson S, Aukett L, Conn V, Shearer MK, Fuentes S, Tyerman SD, Gilliham M (2013) Protocol: optimising hydroponic growth systems for nutritional and physiological analysis of Arabidopsis thaliana and other plants. Plant Methods 9: 4

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

Das M, Reichman JR, Haberer G, Welzl G, Aceituno FF, Mader MT, Watrud LS, Pfleeger TG, Gutiérrez R, Schäffner AR, Olszyk D (2010) A composite transcriptional signature differentiates responses towards closely related herbicides in Arabidopsis thaliana and Brassica napus. Plant Mol Biol 72: 545-556

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

Delhaize E, Gruber BD, Pittman JK, White RG, Leung H, Miao Y, Jiang L, Ryan PR, Richardson AE (2007) A role for the AtMTP11 gene of Arabidopsis in manganese transport and tolerance. Plant J 51: 198-210

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

Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. Mol Biol Evol 23: 469-478

Pubmed: Author and Title
CrossRef: Author and Title

Google Scholar: Author Only Title Only Author and Title

Fischer I, Dainat J, Ranwez V, Glémin S, Dufayard JF, Chantret N (2014) Impact of recurrent gene duplication on adaptation of plant genomes. BMC Plant Biol 14: 151
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Garton S, Knight H, Warren GJ, Knight MR, Thorlby GJ (2007) crinkled leaves 8 - A mutation in the large subunit of ribonucleotide reductase - leads to defects in leaf development and chloroplast division in Arabidopsis thaliana. Plant J 50: 118-127
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Gabaldón T (2008) Large-scale assignment of orthology: back to phylogenetics? Genome Biology 9: 235

Gustin JL, Zanis MJ, Salt DE (2011) Structure and evolution of the plant cation diffusion facilitator family of ion transporters. BMC Evol Biol 11: 76
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ha M, Li WH, Chen ZJ (2007) External factors accelerate expression divergence between duplicate genes. Trends Genet 23: 162-166
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ha M, ED Kim, Chen ZJ (2009) Duplicate genes increase expression diversity in closely related species and allopolyploids. Proc Natl Acad Sci USA 106: 2295-2300
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KF, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo YL (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet 43: 476-481
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kuzniar A, van Ham RC, Pongor S, Leunissen JA (2008) The quest for orthologs: finding the corresponding gene across genomes. Trends Genet 24: 539-551
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lee B, Kapoor A, Zhu J, Zhu JK (2006) STABILIZED1, a Stress-Upregulated Nuclear Protein, Is Required for Pre-mRNA Splicing, mRNA Turnover, and Stress Tolerance in Arabidopsis. Plant Cell 18: 1736-1749
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

MacCarthy T, Bergman A (2007) The limits of subfunctionalization. BMC Evol Biol 7: 213
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Mitchell-Olds T (2001) Arabidopsis thaliana and its wild relatives: a model system for ecology and evolution. Trends Ecol and Evol 16: 693-700
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Movahedi S, Van de Peer Y, Vandepoele K (2011) Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice. Plant Physiol 156: 1316-1330
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S (2011) PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. Plant Cell 23: 895-910
Pubmed: Author and Title
CrossRef: Author and Title

Google Scholar: Author Only Title Only Author and Title

O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33: D476-480
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Patel RV, Nahal HK, Breit R, Provart NJ (2012) BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species. Plant J 71: 1038-50
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. 314: 1041-52
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Rombauts S, Florquin K, Lescot M, Marchal K, Rouzé P, Van de Peer Y (2003) Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol 132: 1162-1176
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Sauge-Merle S, Falconet D, Fontecave M (1999) An active ribonucleotide reductase from Arabidopsis thaliana. Eur J Biochem 266: 62-69
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Scholl RL, May ST, Ware DH (2000) Seed and molecular resources for Arabidopsis. Plant Physiol 124:1477-1480
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Street NR, Sjodin A, Bylesjo M, Gustafsson P, Trygg J, Jansson S (2008) A cross-species transcriptomics approach to identify genes involved in leaf development. BMC Genomics 9: 589
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302: 249-255
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631-637
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Throude M, Bolot S, Bosio M, Pont C, Sarda X, Quraishi UM, Bourgis F, Lessard P, Rogowsky P, Ghesquiere A, Murigneux A, Charmet G, Perez P, Salse J (2009) Structure and expression analysis of rice paleo duplications. Nucleic Acids Res 37: 1248-1259
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Tirosh I, Barkai N (2007) Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. Genome Biol 8(4): R50
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Vandesompele J, Preter KD, Pattyn F, Poppe B, Roy ND, Paepe AD, Speleman F (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 3: 34.1-34.11
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K (2009) The flowering world: a tale of duplications. Trends Plant Sci 14: 680-688
Pubmed: Author and Title
CrossRef: Author and Title

Google Scholar: Author Only Title Only Author and Title

Wang C, Liu Z (2006) Arabidopsis ribonucleotide reductases are critical for cell cycle progression, DNA damage repair, and plant development. Plant Cell 18: 350-365

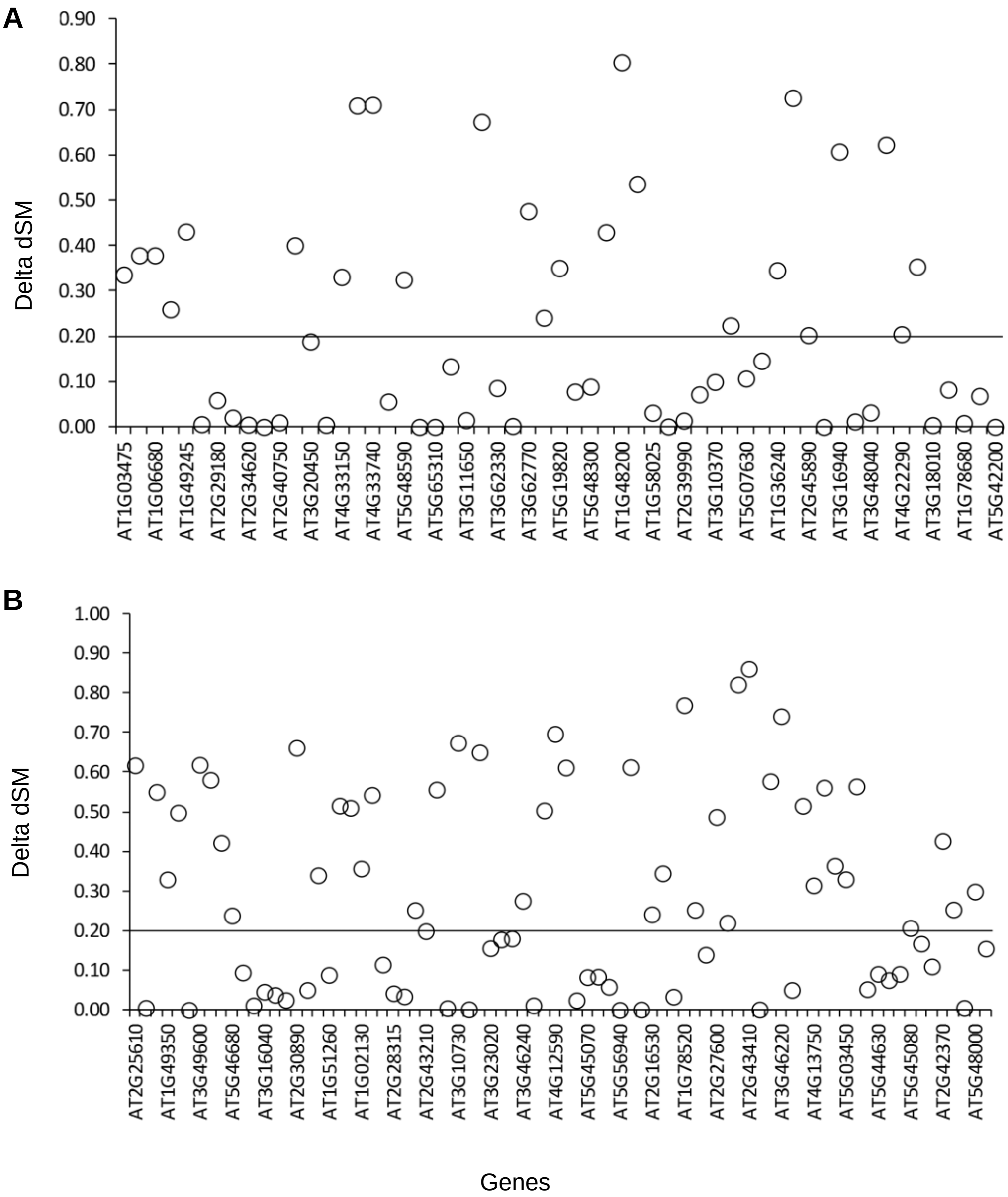
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Whittle CA, Krochko JE (2009) Transcript profiling provides evidence of functional divergence and expression networks among ribosomal protein gene paralogs in Brassica napus. Plant Cell 21: 2203-2219

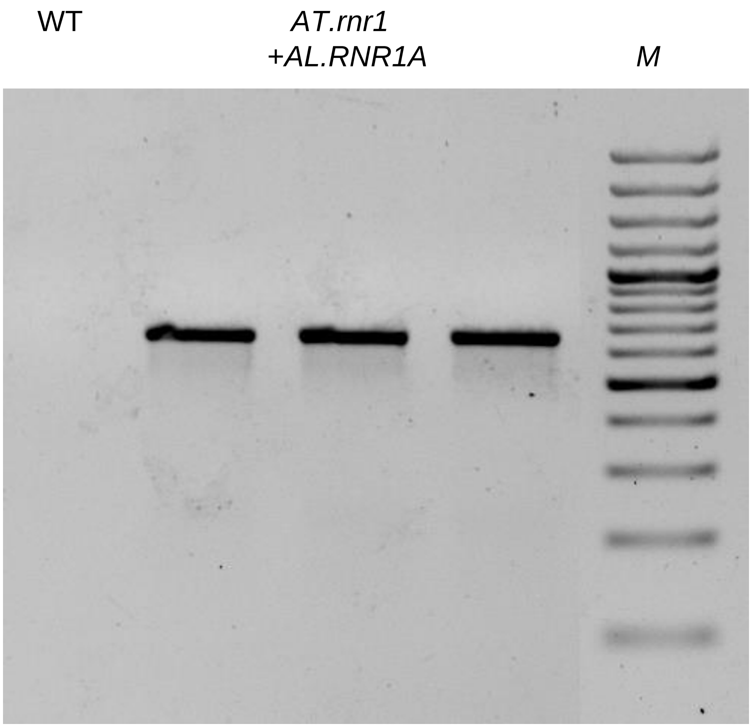
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Xu H, Faber C, Uchiki T, Fairman JW, Racca J, Dealwis C (2006) Structures of eukaryotic ribonucleotide reductase I provide insights into dNTP regulation. Proc Natl Acad Sci USA 103: 4022-4027

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



Supplemental Figure S1. Promoter sequences divergence analysis between expressologs and non-expressologs in the genetic redundant and species-specific functional categories (A) Promoter analyses of the gene group, where all the *A. lyrata* genes present within a gene group are genetically redundant to *A. thaliana* gene as predicted by gene expression analyses. For each *A. thaliana* and *A. lyrata* orthologous gene clusters we calculated the differences in promoter sequence divergence scores (delta d_{SM} scores) between the *A. lyrata* gene copies. For clusters with more than two *A. lyrata* gene copies we considered any two *A. lyrata* copies by random choice. In order to define the clusters that are closer in their d_{SM} values we have considered a threshold value of delta d_{SM} which has been shown by dotted line (<0.2). (B) Promoter analyses of the gene group, where all the *A. lyrata* genes present within a gene group are depicting species specific difference compared to *A. thaliana* gene as predicted by gene expression analyses. The other parameters used were the same as described in Figure S1A.



Supplemental Figure S2. Confirmation of the presence of the transgene insertion of AL.RNR1A in three independent transgenic plants by PCR analysis. An AL.RNR1A-specific primer pair (Table S4) was used to amplify a diagnostic fragment from genomic DNA, which was absent from an untransformed control plant (WT).

A

```
AT.TSO2     -------------------MPSMPEEPLLTPTPDRFCMFPIHYPQIWEMY 31
AL.TSO2A    ----------------------MPEEPLLTPTPDRFCMFPIHYPQIWEMY 28
AL.TSO2B    -------------------MPSMPEEPILTPTPDRFCMFPIQYPQIWEMY 31
HS.RRM2     SKTARRIFQEPTEPKTKAAAPGVEDEPLLRENPRRFVIFPIEYHDIWQMY 94
SC.RNR2     DAENHKAYLKSHQVHRHKLKEMEKEEPLLNEDKERTVLFPIKYHEIWQAY 100

                             TSO2-1 (D49>N)
AT.TSO2     KKAEASFWTAEEVDLSQDNRDWENSLNDGERHFIKHVLAFFAASDGIVLE 81
AL.TSO2A    KKAEASFWTAEEVDLSQDNRDWENNLNDGERHFIKHVLAFFAASDGIVLE 78
AL.TSO2B    KKAEASFWTAEEVDLSQDNRDWENSLSNDERHFIKHVLAFFAASDGIVLE 81
HS.RRM2     KKAEASFWTAEEVDLSKDIQHWES-LKPEERYFISHVLAFFAASDGIVNE 143
SC.RNR2     KRAEASFWTAEEIDLSKDIHDWNNRMNENERFFISRVLAFFAASDGIVNE 150

                             TSO2-3 (R97>S)
AT.TSO2     NLASRFMSDVQVSEARAFYGFQIAIENIHSEMYSLLLDTYIKDNKERDHL 131
AL.TSO2A    NLASRFMSDVQVSEARAFYGFQIAIENIHSEMYSLLLDTYIKDNKERDHL 128
AL.TSO2B    NLSTRFMSDVQISEARAFYGFQIAIENIHSEMYSLLLDTYIKDNKERDHL 131
HS.RRM2     NLVERFSQEVQITEARCFYGFQIAMENIHSEMYSLLIDTYIKDPKEREFL 193
SC.RNR2     NLVENFSTEVQIPEAKSFYGFQIMIENIHSETYSLLIDTYIKDPKESEFL 200

                                             TSO2-2 (G170>S)
AT.TSO2     FRAIETIPCVAKKAQWAMKWIDG-SQTFAERIIAFACVEGIFFSGSFCSI 180
AL.TSO2A    FRAIETIPCVAKKAQWAMKWIDG-SQTFAERIIAFACVEGIFFSGSFCSI 177
AL.TSO2B    FRAIETIPCVTKKAEWAMKWING-SQSFAERIVAFACVEGIFFSGSFCSI 180
HS.RRM2     FNAIETMPCVKKKADWALRWIGDKEATYGERVVAFAAVEGIFFSGSFASI 243
SC.RNR2     FNAIHTIPEIGEKAEWALRWIQDADALFGERLVAFASIEGVFFSGSFASI 250

AT.TSO2     FWLKKRGLMPGLTFSNELISRDEGLHCDFACLLYTLLKTKLSEERVKSIV 230
AL.TSO2A    FWLKKRGLMPGLTFSNELISRDEGLHCDFACLLYTLLKTKLSEERVKSIV 227
AL.TSO2B    FWLKKRGLMPGLTFSNELISRDEGLHCDFACLIYSLLRTKLDEERLKSIV 230
HS.RRM2     FWLKKRGLMPGLTFSNELISRDEGLHCDFACLMFKHLVHKPSEERVREII 293
SC.RNR2     FWLKKRGMMPGLTFSNELICRDEGLHTDFACLLFAHLKNKPDPAIVEKIV 300

AT.TSO2     CDAVEIEREFVCDALPCALVGMNRDLMSQYIEFVADRLLGALGYGKVYGV 280
AL.TSO2A    CDAVEIEREFVCDALPCALVGMNRDLMSQYIEFVADRLLGALGYGKVYGV 277
AL.TSO2B    CDAVEIEREFVCDALPCALVGMNRELMSQYIEFVADRLLTALGCGKVYGV 280
HS.RRM2     INAVRIEQEFLTEALPVKLIGMNCTLMKQYIEFVADRLMLELGFSKVFRV 343
SC.RNR2     TEAVEIEQRYFLDALPVALLGMNADLMNQYVEFVADRLLVAFGNKKYYKV 350

AT.TSO2     TNPFDWMELISLQGKTNFFEKRVGDYQKASVMSSVNGNGAF-DNHVFSLD 329
AL.TSO2A    TNPFDWMELISLQGKTNFFEKRVGDYQKASVMSSVNGNGAF-DNHVFSLD 326
AL.TSO2B    SNPFDWMELISLQGKTNFFEKRVGEYQKASVMSSVHGNAAFNDDHVFKLD 330
HS.RRM2     ENPFDFMENISLEGKTNFFEKRVGEYQRMGVMSSP-------TENSFTLD 386
SC.RNR2     ENPFDFMENISLAGKTNFFEKRVSDYQKAGVMSKS----TKQEAGAFTFN 396
```

B

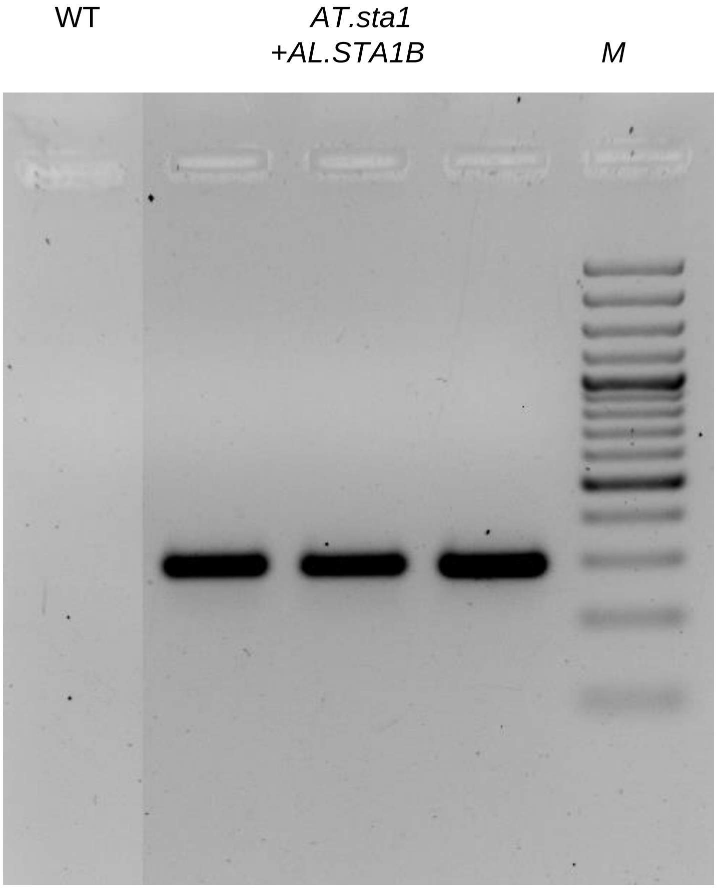
```
AT.TSO2      TCA----CGTCAAAA-ATTCAAAAA---CCCC--AAAACCCTATAATCTCCTATATAAAT 410
AL.TSO2A     CCT----CGACAAAA-ATTCAAAAA---CCCCCAAAAACCCTCTAATCTCCAGTATAAAT 411
AL.TSO2B     CTAGGAGCGGGAAAATATTTTCACATTTCCCTCCTATATCCCCAAATTTTCAAGATAAAT 430

AT.TSO2      -ATTCAGCCCTAGATC--TTATAATTCATCAATCAAACAATCTCTTCAATCAAATCTCTT 467
AL.TSO2A     -TCTCATCCCTAGATC--TCATAATTCATCAATCAAATCATCTCTTCAATCACTTCTCTT 468
AL.TSO2B     ATCTCAATCACAGATCCATAATAATTCACACAAAAAAAAAACTCATAAATC-------TT 483

AT.TSO2      CTTCAATCAAATCTTCAAA--TCCCTTCAAAGATG 500
AL.TSO2A     CTTCAATCAAATCTTCAAAGATGCCTTCAATG--- 500
AL.TSO2B     CTTCA-TAGAA-----AAA---------AATG--- 500
```

Supplemental Figure S3. Sequence alignments of *AT.TSO2* and *AL.TSO2* homologous genes. (A) Multiple sequence alignment of *AT.TSO2*, *AL.TSO2A*, *AL.TSO2B*, human (HS.RRM2, NP_001025) and *Saccharomyces cerevisiae* (SC.RNR2, NP_012508) sequences. The nucleotides highlighted in red are identified non-synonymous changes between *AT.TSO2* and *AL.TSO2A*/*AL.TSO2B*. Twenty eight non-synonymous changes were noticed between *AT.TSO2* and *AL.TSO2B*. Regions containing residues of important for enzymatic function are underlined (Philipps et al., 1995). Twenty-seven out of 28 amino acid changes for TSO2B are outside these regions indicating that gene function of TSO2B was probably not affected by these changes. The bold, underlined amino acids are three known TSO2 alleles in *Arabidopsis thaliana* (Wang and Liu, 2006). (B) Multiple alignment of 1000 bp upstream region of *AT.TSO2*, *AL.TSO2A* and *AL.TSO2B*. The bold, underlined sequence (CTCCTATATAAATA) is the TATA box in the core promoter region of AT2G21790; while underlined region (TCTCTTCTTC) is the Y patch. Y Patch is a direction-sensitive plant core promoter element that appears around TSS.

```
atggtgttcctctcgattccaaacgtgaagaccatgtcgatcaatgtgaaccctagtgca
 M  V  F  L  S  I  P  N  V  K  T  M  S  I  N  V  N  P  S  A
accaccatctccgccttcgaacaattggtccatcaacgcactcatcttcctcaacctctc
 T  T  I  S  A  F  E  Q  L  V  H  Q  R  T  H  L  P  Q  P  L
cttcgttactcgctctgtctccgcaaccctagtcttgattcctccgatcctgccctgtta
 L  R  Y  S  L  C  L  R  N  P  S  L  D  S  S  D  P  A  L  L
tcggatctaggttttggccctttgtctacggtacttgttcatgtccctctaatcggtgga
 S  D  L  G  F  G  P  L  S  T  V  L  V  H  V  P  L  I  G  G
gcggctccgcctcagcccctcttcaattcaaactatgtcgctggtttgggtcgtggggct
 A  A  P  P  Q  P  L  F  N  S  N  Y  V  A  G  L  G  R  G  A
acagggtttactacccgctccgatattggtcctgctcgtgctgatggcgatgacgtgaat
 T  G  F  T  T  R  S  D  I  G  P  A  R  A  D  G  D  D  V  N
cacaagtttgatgactttgaagggaatgatgcgggattgttcgctaatgccgagtgtgat
 H  K  F  D  D  F  E  G  N  D  A  G  L  F  A  N  A  E  C  D
gacgaagacaaagaggctgacgccattgataggaggaggaaagacagaagagacatcgag
 D  E  D  K  E  A  D  A  I  D  R  R  R  K  D  R  R  D  I  E
aattacagagcctccaaccctaaagtttctgagcagtttgtggatctgaagagaaagttg
 N  Y  R  A  S  N  P  K  V  S  E  Q  F  V  D  L  K  R  K  L
catactttgtctgaggatgaatgggatagtattccagagattgggaattactcgcatcgg
 H  T  L  S  E  D  E  W  D  S  I  P  E  I  G  N  Y  S  H  R
agcaagaagaagaggtttgagagctttgtgcctgttcctgacacgcttttgcaggaaaaa
 S  K  K  K  R  F  E  S  F  V  P  V  P  D  T  L  L  Q  E  K
gggatcgtctcggccttaggcccaaatagcagagccgctggtggatcggagacgccatgg
 G  I  V  S  A  L  G  P  N  S  R  A  A  G  G  S  E  T  P  W
atagacttgacttcagtcggtgagggaagaggttttctgttgtctctgaagcttgagagg
 I  D  L  T  S  V  G  E  G  R  G  F  L  L  S  L  K  L  E  R
ttatcagattctctttcagggcaaactgttgtggatcctaaaggctacttaactgacctt
 L  S  D  S  L  S  G  Q  T  V  V  D  P  K  G  Y  L  T  D  L
aagaataaggaactcaccaacgatgcagacatttttcatattaatagagctagacccttta
 K  N  K  E  L  T  N  D  A  D  I  F  H  I  N  R  A  R  P  L
ttaaagagtattacacagtcgaatcccaagaatcccaatggctggattgctgctgcgaga
 L  K  S  I  T  Q  S  N  P  K  N  P  N  G  W  I  A  A  A  R
ctcgaggagagggctggtaaaataaaagccgctagaactcagattcagaagggatgcaat
 L  E  E  R  A  G  K  I  K  A  A  R  T  Q  I  Q  K  G  C  N
gagtgcccaaaacatgaggatgtttgggttgaggcttgtatgctggccacaccggaggat
 E  C  P  K  H  E  D  V  W  V  E  A  C  M  L  A  T  P  E  D
gccaaggcggtgattgcaatgggagttaagcaaatacccaactcggtgaagctatggttg
 A  K  A  V  I  A  M  G  V  K  Q  I  P  N  S  V  K  L  W  L
gaggctgcaaagttggaacatgatgaggataacaagagtagggtgttgagaaaaggactg
 E  A  A  K  L  E  H  D  E  D  N  K  S  R  V  L  R  K  G  L
gagcatattccagactcggttaggctatggaagactgttaaggacatggctaataaagaa
 E  H  I  P  D  S  V  R  L  W  K  T  V  K  D  M  A  N  K  E
gatgcagtggttttgcttcacagagctgtggaatgctgccctctgcatccggagctatgg
 D  A  V  V  L  L  H  R  A  V  E  C  C  P  L  H  P  E  L  W
atggcgcttgcgaggcttgaaacatacgaaaaacacaaagaaagtgttgaacagagcgag
 M  A  L  A  R  L  E  T  Y  E  K  H  K  E  S  V  E  Q  S  E
agagaagctccccaaggagcggggatttggatcaccgctgctaagctagaggaagataa
 R  E  A  P  Q  G  A  G  D  L  D  H  R  C  -  A  R  G  R  -
tgggaatactactaaggttggaaagatcattgagaagggtataaatgctctgcagagaga
 W  E  Y  Y  -  G  W  K  D  H  -  E  G  Y  K  C  S  A  E  R
agaggttgtcattgaccgggaaaagtggaggtctctgagagagccgggtatgtaacaacc
 R  G  C  H  -  P  G  K  V  E  V  S  E  R  A  G  Y  V  T  T
tgccaggcaattattaagatcattattggttttgaagtcgatgaagaggatagaaagaaa
 C  Q  A  I  I  K  I  I  I  G  F  E  V  D  E  E  D  R  K  K
acttgggttgctgatgcagaggagtgcaagaagaggggttccatcgagactgcaagagca
 T  W  V  A  D  A  E  E  C  K  K  R  G  S  I  E  T  A  R  A
atatacgcacatgctcttaccgtgttctttactaagaaaagtatctggctgcgcagttag
 I  Y  A  H  A  L  T  V  F  F  T  K  K  S  I  W  L  R  S  -
```

```
agaagagtcatggtagtatggagtctcttgatgccgtgttgcgtaaggctgtgacatacc
 R  R  V  M  V  V  W  S  L  L  M  P  C  C  V  R  L  -  H  T
tccctcaggctgaggttctctggctcatgtgtgccaaggagaagtggcttgctggagatg
 S  L  R  L  R  F  S  G  S  C  V  P  R  R  S  G  L  L  E  M
ttccagcagcccgtggcattctacaagaggctcatgccgcagttccaaactccgaggaaa
 F  Q  Q  P  V  A  F  Y  K  R  L  M  P  Q  F  Q  T  P  R  K
tctggcttgctgcttttaagctagagtttgagagcagggaggtggagagggcgaggatga
 S  G  L  L  L  L  S  -  S  L  R  A  G  R  W  R  G  R  G  -
ttctcgcaaaagcaagggaaagaggaactactgggagggtgtggatgaaatcagccattg
 F  S  Q  K  Q  G  K  E  E  L  L  G  G  C  G  -  N  Q  P  L
ttgagagggaactaggcaacgtagaggaggagaggagattgcttgaagaaggcgtgaaga
 L  R  G  N  -  A  T  -  R  R  R  G  D  C  L  K  K  A  -  R
aattcccagcattcttcaagctttggttgatgcttgggcagcttggggaaaggtttaggc
 N  S  Q  H  S  S  S  F  G  -  C  L  G  S  L  G  K  G  L  G
atctggaacaggccaagaaagcttacacatctggtttgaggcactgtcccgagtgcacac
 I  W  N  R  P  R  K  L  T  H  L  V  -  G  T  V  P  S  A  H
cattgtggctctcgctcgctgatattgaagagaaagtgaatgggctcaacaaagctcgtg
 H  C  G  S  R  S  L  I  L  K  R  K  -  M  G  S  T  K  L  V
tagttctcactctggccaggaagaaaaaccctaaggcggatgagctatggctagctgctg
 -  F  S  L  W  P  G  R  K  T  L  R  R  M  S  Y  G  -  L  L
ttcgtgttgaaattagacatggcaacaagagagaagcagagcgcttgatgtcaaaggccc
 F  V  L  K  L  D  M  A  T  R  E  K  Q  S  A  -  C  Q  R  P
tgcaagagtctcccaaaagtggtcttctcttggctgctgacatcgagatggcaccgccat
 C  K  S  L  P  K  V  V  F  S  W  L  L  T  S  R  W  H  R  H
gtctgctcccgcaaacgaagattgatgatgctctgaagaagtgtgtgaagaaggaggcgg
 V  C  S  R  K  R  R  L  M  M  L  -  R  S  V  -  R  R  R  R
cgcatgtcactgcaatggtcgccaagatctcctggcaagataggaaggtggataaagcca
 R  M  S  L  Q  W  S  P  R  S  P  G  K  I  G  R  W  I  K  P
gattgtggtttcaacggaccgtgaacgtcgacccagataatggagatttctgggccttgt
 D  C  G  F  N  G  P  -  T  S  T  Q  I  M  E  I  S  G  P  C
actacaaatttgaacttgaacatggctctgaggagaagcagaaggaggtgctgaccaaat
 T  T  N  L  N  L  N  M  A  L  R  R  S  R  R  R  C  -  P  N
gtgtggcgtctgagccaaagcacggtgagaagtggcaagccatatccaaagcgttggaga
 V  W  R  L  S  Q  S  T  V  R  S  G  K  P  Y  P  K  R  W  R
atgcccaccagcctgttgaagtcatcttgaagagagtggtggttgcattgacaagggaag
 M  P  T  S  L  L  K  S  S  -  R  E  W  W  L  H  -  Q  G  K
agcgtaacaaactctaa
 S  V  T  N  S
```

Supplemental Figure S4. Pseudogenization due to insertion of one A nucleotide at position 1352 (marked in red font and underlined), which causes premature insertion of the stop codon (-) in the *AL.STA1B* gene copy.

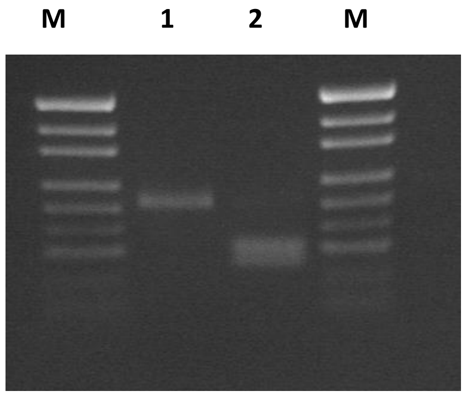


Supplemental Figure S5. Confirmation of the presence of the transgene insertion of AL.STA1B in three independent transgenic plants each by PCR analysis. An AL.STA1B-specific primer pair (Table S4) was used to amplify a diagnostic fragment from genomic DNA, which was absent from an untransformed control plant (WT).

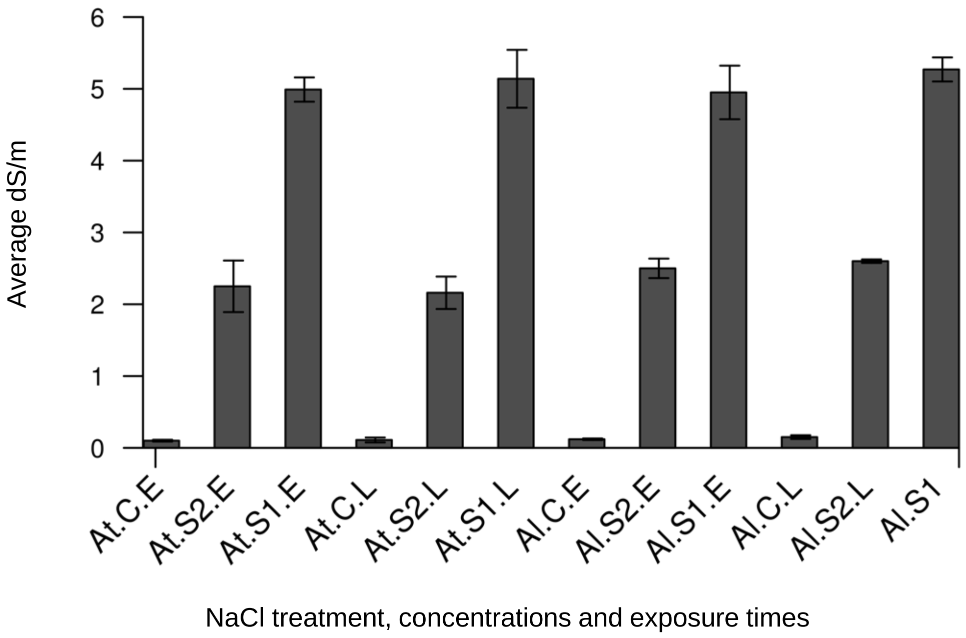
A

```
AL.MTP11ACDS        151 TGTCTTGGTTGTTTGGGTCCGGAAGACAATGTGGCAGATTATTACCAGCA    200
                        ||||||||||||||||||||||||||||||||.|||||||||||||||||
AL.MTP11BCDS        151 TGTCTTGGTTGTTTGGGTCCGGAAGACAATGTAGCAGATTATTACCAGCA    200

AL.MTP11ACDS        201 GCAAGTAGAGATGCTTGAGGGATTTACTGAAATGGATGAACTTGCAGAAC    250
                        |||||||||||||||||||||.||.|||||||||||||||||||||||||
AL.MTP11BCDS        201 GCAAGTAGAGATGCTTGAGGGCTTCACTGAAATGGATGAACTTGCAGAAC    250

AL.MTP11ACDS        251 GTGGCTTTGTTCCTGGAATGTCAAAGGAAGAGCAGGATAATTTGGCTAAA    300
                        |.||||||||||||||||||||||||||||||||||||||||||||||||
AL.MTP11BCDS        251 GCGGCTTTGTTCCTGGAATGTCAAAGGAAGAGCAGGATAATTTGGCTAAA    300

AL.MTP11ACDS        301 AGCGAGACATTGGCGATTAGAATATCAAACATTGCAAACATGCTTCTTTT    350
                        ||.|||||||||||||||||||||||||||||||||||||||.|||||||
AL.MTP11BCDS        301 AGTGAGACATTGGCGATTAGAATATCAAACATTGCAAACATGGTTCTTTT    350

AL.MTP11ACDS        351 TGCTGCTAAAGTCTATGCTTCTGTCACAAGTGGCTCTTTAGCAATCATTG    400
                        |||||||||||||||.||||||||||||||||||||||||||||||||||
AL.MTP11BCDS        351 TGCTGCTAAAGTCTACGCTTCTGTCACAAGTGGCTCTTTAGCAATCATTG    400
```

B



Supplemental Figure S6. Distinction of *ALMTP11A* and *B* homologs. (A) Pairwise sequence comparison of the two *A. lyrata* MTP11A and B homologs to design gene specific primers (bold, underlined) for quantitative real time RT-PCR analyses. Presence of a restriction enzyme (*Aci* I) cut site (CCGC) was detected and underlined in the *AL.MTP11B* sequence, which is absent in the *AL.MTP11A* sequence. (B) To confirm specificity of the amplified PCR products both the amplified fragments were digested with *Aci* I. Only one fragment was noticed for *AL.MTP11A* (lane 1), while two fragments were noticed for *AL.MTP11B* (lane 2). This primer pair was used in the real-time RT-qPCR analyses to calculate expression patterns of these two homologs. M, DNA size marker (pUC19 digested with *Msp* I).



Supplemental Figure S7. Measurements of soil salinity for the stress assays used in the microarray based gene expression analyses. The abbreviations used: At.C.E- *A. thaliana*, control, early time point (3h); At.S2.E- *A. thaliana*, 250 mM NaCl, early time point (3h); At.S1.E- *A. thaliana*, 500 mM NaCl, early time point (3h); At.C.L- *A. thaliana*, control, late time point (27h); At.S2.L- *A. thaliana*, 250 mM NaCl, late time point (27h), At.S1.L- *A. thaliana*, 500 mM NaCl, late time point (27h); Al.C.E- *A. lyrata*, control, early time point (3h); Al.S2.E- *A. lyrata*, 250 mM NaCl, early time point (3h), Al.S1.E- *A. lyrata*, 500 mM NaCl; Al.C.L- *A. lyrata*, control, late time point (27h), Al.S2.L- *A. lyrata*, 250 mM NaCl, late time point (27h), Al.S1.L- *A. lyrata*, 500 mM NaCl, late time point (27h).

.

Supplemental Table S1. Summary of *Arabidopsis thaliana* and *A. lyrata* array design features

Array features	***A. thaliana***	***A. lyrata ssp. lyrata***
Agilent design Id	029132	030951
Design format	8X60K	8X60K
Number of biological probes	43603	32386
Number of replicated probes	50 X 5	477 X 10
Mean probe length (bp)	60	60
Agilent controls on array	1319	1319
% filled by selected probe group	71.64	61.09
Total number of features on array	62976	62976
Total % filled	100	100

Supplemental Table S3. *A. thaliana* mutants used in this study for genetic complementation with *A. lyrata* homologs. The complemented plants were phenotyped according to the conditions described in the original reference.

Gene	**Annotation**	**Mutant name**	**Nature of mutation**	**Phenotype**	**Genetic background**	**Reference**	**Utility in our complementation assay**
AT2G21790	Ribonucleotide reductase 1/RNR1	*crinkled leaves 8 (cls8-1)*	Point mutation. Missense G>A substitution/ G718>E.	First developing true leaves emerge bleached, subsequent leaves emerge curled with bleached edges, matured rosette leaves become crinkled and show patches of white pits on the surface	Columbia	Garton et al. (2007) Plant Journal 50: 118–127	Used in complementation assay.
AT3G27060	Ribonucleaotide reductase 2/TSO2	*tso2-1*	Point mutation. Missense change: D49>N.	White sectors in green organs, uneven thickness, rough surfaces, irregular margins of leaves or floral organs, sepals rough and uneven, stamens occasionally exhibited carpel characteristics indicating homeotic transformation.	Landsberg *erecta*	Wang and Liu (2006) Plant Cell 18: 350-365	Used in complementation assay.
AT2G39450	Manganese transporter 11/MTP11	*N859636, SALK_025271*	T-DNA insertion	On nutrient agar supplied with Mn^{2+} concentrations ranged from basal to toxic levels, the mutant was more sensitive to Mn^{2+} than the wild type, as determined by significantly reduced shoot dry weights.	Columbia	Delhaize et al. (2007) The Plant Journal 51: 198–210	Used in complementation assay.
AT4G03430	Stabilized 1/STA1	*sta1-1*	In-frame deletion of two amino acids: 1249 to 1254 bp from the translation initiation site/ 417C, 418P	*sta1-1* plants showed many developmental and stress-related phenotypes, smaller in size and heights than the wild-type plants. Mutant leaves were more serrated with a pointed leaf tip. The mutant was more sensitive to ABA.	Columbia gl1	Lee et al. (2006) Plant Cell 18: 1736–1749	Used in complementation assay.

Supplemental Table S4. Oligonucleotides used in this study for different purposes.

*Since, the promoter and 3' UTR region of *AL.MTP11A* and *B* are sequentially highly similar, we used the same oligonucleotide pairs to amplify both homologs.

Target gene	Gene identifier	Oligo name	5'-3' sequence	Amplified product length	Purpose
AL.RNR1A	Al_scaffold_0007_128	AL.CLSA.GW.F	GGGGACAAGTTTGTACAAAAAAGCAGGCTaaagacgacaaaacaaaacg	6.3 Kb	Gene amplification- GATEWAY cloning
AL.RNR1A	Al_scaffold_0007_128	AL.CLSA.GW.R	GGGGACCACTTTGTACAAGAAAGCTGGGtctgagatttgaggatgagg	-	Gene amplification- GATEWAY cloning
AL.RNR1B	fgenesh2_kg.4__104__AT2G21790.1	AL.CLSB.GW.F	GGGGACAAGTTTGTACAAAAAAGCAGGCTaagaggtgcgttgaagtcta	6.5 Kb	Gene amplification- GATEWAY cloning
AL.RNR1B	fgenesh2_kg.4__104__AT2G21790.1	AL.CLSB.GW.R	GGGGACCACTTTGTACAAGAAAGCTGGGTaagttccacaaaatcctcct	-	Gene amplification- GATEWAY cloning
AL.TSO2A	fgenesh2_kg.5__483__AT3G27060.1	AL.TSO2A.GW.F	GGGGACAAGTTTGTACAAAAAAGCAGGCTgttcacaaacatggcttagg	3.73 Kb	Gene amplification- GATEWAY cloning
AL.TSO2A	fgenesh2_kg.5__483__AT3G27060.1	AL.TSO2A.GW.R	GGGGACCACTTTGTACAAGAAAGCTGGGtccaatctataaaacacaaaaca	-	Gene amplification- GATEWAY cloning
AL.TSO2B	scaffold_703867.1	AL.TSO2B.GW.F	GGGGACAAGTTTGTACAAAAAAGCAGGCTcatctgaatcatggtccttt	3.58 Kb	Gene amplification- GATEWAY cloning
AL.TSO2B	scaffold_703867.1	AL.TSO2B.GW.R	GGGGACCACTTTGTACAAGAAAGCTGGGTaactcggccatatcaactta	-	Gene amplification- GATEWAY cloning
AL.STA1A	fgenesh2_kg.6__3353__AT4G03430.1	AL.STA1A.GW.F2	GGGGACAAGTTTGTACAAAAAAGCAGGCTggtcttggtaataacgtcca	5.78 Kb	Gene amplification- GATEWAY cloning
AL.STA1A	fgenesh2_kg.6__3353__AT4G03430.1	AL.STA1A.GW.R2	GGGGACCACTTTGTACAAGAAAGCTGGGTcaacatatcccgttgtttct	-	Gene amplification- GATEWAY cloning
AL.STA1B	scaffold_700051.1	AL.STA1B.GW.F	GGGGACAAGTTTGTACAAAAAAGCAGGCTagaattgggggacttaaca	5.8 Kb	Gene amplification- GATEWAY cloning
AL.STA1B	scaffold_700051.1	AL.STA1B.GW.R	GGGGACCACTTTGTACAAGAAAGCTGGGTaaactcaagttcgatccgta	-	Gene amplification- GATEWAY cloning
AL.MTP11*	-	AL.MTP11GW.F	GGGGACAAGTTTGTACAAAAAAGCAGGCTgatggagtggaaacagaaga	4.9 Kb	Gene amplification- GATEWAY cloning

AL.MTP11*	-	AL.MTP11GW.R	GGGGACCACTTTGTACAAGAAAGCTGGGTggtgagaatcagagtgagga	-	Gene amplification- GATEWAY cloning
AL.MTP11A	fgenesh2_kg.4__2026__AT2G39450.1	AL.MTP11A.RT.F1	GGTCCGGAAGACAATGTG	199 bp	RT-qPCR- gene expression assay
AL.MTP11A	fgenesh2_kg.4__2026__AT2G39450.1	AL.MTP11A.RT.R	AGACTTTAGCAGCAAAAAGAAG		RT-qPCR- gene expression assay
AL.MTP11B	fgenesh2_kg.463__5__AT2G39450.1	AL.MTP11B.RT.F1	GGTCCGGAAGACAATGTA	199 bp	RT-qPCR- gene expression assay
AL.MTP11B	fgenesh2_kg.463__5__AT2G39450.1	AL.MTP11B.RT.R	AGACTTTAGCAGCAAAAAGAAC		RT-qPCR- gene expression assay
AL.UBQ5	fgenesh2_kg.5__2722__AT3G62250.1.1	AL.UBQ5_fnew	GATGGATCTGGAAAAGTTCAG	168 bp	RT-qPCR -reference gene for *A. lyrata*
AL.UBQ5	fgenesh2_kg.5__2722__AT3G62250.1.1	AL.UBQ5_rnew	AGCGGTTGCTAGAACAGATC	-	RT-qPCR reference gene for *A. lyrata*
AL.S16	fgenesh2_kg.6__1842__AT2G09990.1.1	AL.S16qRT_f	TTTACGCCATCCGGCAGAGTAT	186 bp	RT-qPCR reference gene for *A. lyrata*
AL.S16	fgenesh2_kg.6__1842__AT2G09990.1.1	AL.S16qRT_r	GGAAACGAGCACGAGCAC	-	RT-qPCR reference gene for *A. lyrata*
At.mtp11 TDNA line	-	SALK_025271.LP	AATCTGCAATCCAAGTGTTGC		Genotyping
At.mtp11 TDNA line	-	SALK_025271.RP	CTGCTCGAGTTTCACGGTAAC		Genotyping
AL.RNR1A	Al_scaffold_0007_128	AL_CLSA_F	ATGGTTCTATCGTGAATGTCAAG	650 bp	PCR assay to confirm transgene insertion
AL.RNR1A	Al_scaffold_0007_128	AL_CLSA_R	TTGTCTCGTTGTCTTCTTCTGTTG	-	PCR assay to confirm transgene insertion
AL.STA1B	scaffold_700051.1	AL_STA1-B_F	AGTTAGAGAAGAGTCATGGTAGTAT	300 bp	PCR assay to confirm transgene insertion
AL.STA1B	scaffold_700051.1	AL_STA1-B_R	TTCATCCACACCCTCCCAGTAGT	-	PCR assay to confirm transgene insertion