Adaptive and degenerative evolution of the *S-Phase Kinase-Associated Protein 1-Like* family in *Arabidopsis thaliana*

Zhihua Hua [Corresp., 1], **Zhenyu Gao** [1, 2]

[1] Department of Environmental and Plant Biology and Interdisciplinary Program in Molecular and Cellular Biology, Ohio University, Athens, Ohio, United States

[2] State Key Laboratory of Rice Biology, China National Rice Research Institute, Hangzhou, China

Corresponding Author: Zhihua Hua
Email address: hua@ohio.edu

Genome sequencing has uncovered tremendous sequence variation within and between species. In plants, in addition to large variations in genome size, a great deal of sequence polymorphism is also evident in several large multi-gene families, including those involved in the ubiquitin-26S proteasome protein degradation system. However, the biological function of this sequence variation is yet not clear. In this work, we explicitly demonstrated a single origin of retroposed *Arabidopsis Skp1-Like* (*ASK*) genes using an improved phylogenetic analysis. Taking advantage of the 1,001 genomes project, we hereinto provide several lines of polymorphism evidence showing both adaptive and degenerative evolutionary processes in *ASK* genes. Yeast two-hybrid quantitative interaction assays further confirmed that recent neutral changes in the *ASK2* coding sequence weakened its interactions with some F-box proteins. The trend that highly polymorphic upstream regions of *ASK1* yield high levels of expression implied negative expression regulation of *ASK1* by an as-yet-unknown transcriptional suppression mechanism, which may contribute to the polymorphic roles of Skp1-CUL1-F-box complexes. Taken together, this study provides new evolutionary evidence to guide future functional genomic studies of SCF-mediated protein ubiquitylation.

1 **Adaptive and Degenerative Evolution of the *S-Phase Kinase-Associated Protein 1-Like***
2 **Family in *Arabidopsis thaliana***
3

5 Zhihua Hua[1] and Zhenyu Gao[1,2]
6
7
8
9 *[1]Department of Environmental and Plant Biology and Interdisciplinary Program in*
10 *Molecular and Cellular Biology, Ohio University, Athens, Ohio 45701, USA;*
11 *[2]State Key Laboratory of Rice Biology, China National Rice Research Institute,*
12 *Hangzhou 310006, Zhejiang, China*

14
15
16
17
18 **Running title:** Evolution of the *Arabidopsis Skp1-like* family
19
20
21
22

24 ***Corresponding author:** Zhihua Hua, Department of Environmental and Plant Biology,
25 Ohio University, Athens, Ohio 45701 USA. Tel: +01 740 593 1123, Email: hua@ohio.edu

28   **Abstract**

29   Genome sequencing has uncovered tremendous sequence variation within and between

30   species. In plants, in addition to large variations in genome size, a great deal of sequence

31   polymorphism is also evident in several large multi-gene families, including those involved in the

32   ubiquitin-26S proteasome protein degradation system. However, the biological function of this

33   sequence variation is yet not clear. In this work, we explicitly demonstrated a single origin of

34   retroposed *Arabidopsis Skp1-Like* (*ASK*) genes using an improved phylogenetic analysis.

35   Taking advantage of the 1,001 genomes project, we hereinto provide several lines of

36   polymorphism evidence showing both adaptive and degenerative evolutionary processes in

37   *ASK* genes. Yeast two-hybrid quantitative interaction assays further confirmed that recent

38   neutral changes in the *ASK2* coding sequence weakened its interactions with some F-box

39   proteins. The trend that highly polymorphic upstream regions of *ASK1* yield high levels of

40   expression implied negative expression regulation of *ASK1* by an as-yet-unknown

41   transcriptional suppression mechanism, which may contribute to the polymorphic roles of Skp1-

42   CUL1-F-box complexes. Taken together, this study provides new evolutionary evidence to

43   guide future functional genomic studies of SCF-mediated protein ubiquitylation.

44 **Introduction**

45 Proteins play fundamental roles in driving life processes by sensing diverse environmental cues,

46 catalyzing biochemical reactions, monitoring the stability of genetic materials, and combating

47 abiotic and biotic stresses. In addition, they are believed to be the only molecules capable of

48 mechanical movement in any organism. To accomplish these diverse roles, not only is protein

49 synthesis precisely controlled, but the structure, activity, and turnover of each protein is also

50 sophisticatedly regulated in a temporal and spatial manner. The ubiquitin-26S proteasome

51 system (UPS) is the primary degradative machinery for rapidly modulating protein content in

52 eukaryotic cells. Given the power of its selective turnover of numerous intracellular proteins, the

53 UPS plays an essential regulatory role in controlling cell cycle progression, signal transduction,

54 gene expression regulation, genome stability, and many other cellular processes (Finley et al.

55 2012; Vierstra 2009; Yau and Rape 2016). This function has been demonstrated to be

56 particularly important in plants, as evidenced by the extremely large expansion of several gene

57 superfamilies that encode plant UPS members (Vierstra 2009).

58

59 Among these members, S-Phase Kinase-Associated Protein 1 (Skp1), F-box proteins, Cullin 1

60 and RING box protein 1 (RBX1) compose the largest group of ubiquitin-ligase complexes in

61 plants, termed SCF complexes, which target protein substrates for ubiquitylation and

62 subsequent turnover by the 26S proteasome (Hua and Vierstra 2011). Genetic, genomic,

63 evolutionary, and biochemical analyses have shown that the *F-box* multi-gene superfamily

64 encodes a substrate receptor that determines the specificity of the SCF complex, while the Skp1

65 protein family functions as an adaptor to bridge the variable F-box proteins to the N-terminus of

66 Cullin 1 to assemble a holo-ubiquitin ligating enzyme (Gagne et al. 2002; Hua and Vierstra

67 2011; Yang et al. 1999; Zheng et al. 2002). Although a handful of F-box proteins have been

68 functionally shown to target the degradation of proteins involved in the cell cycle, circadian

69 rhythms, photomorphogenesis, pathogen defense, hormone signaling, and plant reproduction,

70 many recently duplicated and species- or lineage-specific members remain uncharacterized

71 (Hua et al. 2011). Given that the size of the *F-box* superfamily is species-specific and is often

72 not correlated with the complexity of plant species, a genomic drift evolutionary mechanism has

73 been postulated to explain the random size drift of the *F-box* gene superfamily in plants (Hua et

74 al. 2011; Nozawa et al. 2007; Xu et al. 2009). The high sequence polymorphism of lineage

75 specific *F-box* genes and their enrichment of transcriptional suppression-related epigenetic

76 modifications further support this hypothesis (Hua et al. 2013). However, this does not preclude

77 the existence of some young *F-box* genes that play a lineage specific role in plant adaptation

78   (Gagne et al. 2002; Shabek and Zheng 2014; Yang et al. 2008). Unfortunately, it remains

79   difficult to find these members both experimentally and theoretically, in part due to the large size

80   of this group and the low/no expression of most lineage specific *F-box* genes (Hua et al. 2013).

81

82   Similar to the *F-box* gene superfamily, the *Skp1* family has also expanded significantly in land

83   plants. While there is only one single Skp1 protein encoded in yeast and human genomes, the

84   genomes of *Arabidopsis thaliana* and *Oryza sativa* contain 21 and 32 annotated *Skp1* loci,

85   respectively (Kong et al. 2007). Cross-kingdom evolutionary studies have suggested that the

86   plant *Skp1* genes are also rapidly evolving through a birth-and-death evolutionary mechanism

87   (Kong et al. 2004; 2007). However, unlike many inactive *F-box* genes (Hua et al. 2011; Kuroda

88   et al. 2012), 20 out of 21 *A. thaliana Skp1-Like* (*ASK*) genes are transcribed in at least one out

89   of 6 tissues/organs examined, including seedlings, roots, stems, leaves, inflorescences, and

90   siliques (Kong et al. 2004; Zhao et al. 2003), suggesting that most *ASK* genes are active.

91   Phylogenetic analysis further implied that all plant *Skp1* genes shared one common ancestor,

92   although evolutionary rates of individuals are highly heterogeneous. Therefore, it has been

93   inferred that some moderately and rapidly evolving members might have lost their original

94   functions and/or gained new functions (Kong et al. 2004). Despite the sequence diversity of

95   ASK proteins, a recent biochemical study showed that all ASK proteins retained the biochemical

96   function of their ancestor Skp1 protein for interacting with F-box proteins (Kuroda et al. 2012).

97

98   The fact that all ASK proteins interact with an F-box protein implied that their sequences are not

99   sufficiently diverged from their ancestor sequence for their original biochemical function to have

100   been lost. However, this rapid evolution has dramatically diversified the sequences of plant

101   *Skp1* genes, making it challenging to uncover the true phylogenetic relationships among

102   distantly related plant species. Indeed, to avoid the effects of long-branch attraction, type II

103   *Skp1* genes, which carry multiple introns in various positions, unlike type I *Skp1* genes that

104   contain only one or none introns, were excluded in a previous phylogenetic analysis of *Skp1*

105   genes (Kong et al. 2007). Improved understanding of the evolutionary mechanisms of the *Skp1*

106   gene family may aid further exploration of the functions of many unknown SCF complexes. To

107   date, the genome sequences of three Arabidopsis species, *A. thaliana*, *A. lyrata*, and *A. halleri*,

108   which split 5-10 million years ago (mya) (Hu et al. 2011; Koch and Kiefer 2005), have been

109   obtained [The Arabidopsis Information Resource (TAIR), V10; Phytozome V12], and >1,000

110   individual *A. thaliana* accessions have been sequenced (Consortium 2016). These datasets

111   can allow us to further fine-tune the phylogenetic relationships and fixation processes of rapidly-

112   evolving genes in plants, which may help better define their functional constraints. Because the

113   diverse functions of the SCF complexes are primarily determined by the large *F-box* gene family

114   along with the *Skp1* gene family, in this work we analyzed the short evolutionary history of the

115   *Skp1* genes within and between Arabidopsis species in order to uncover important evolutionary

116   patterns in SCF regulatory pathways. Our new evidence suggests that the *ASK* genes are

117   under both adaptive and degenerative evolutionary processes.

118

119   **Materials and Methods**

120   **Identification of *Skp1* genes in *A. lyrata* and *A. halleri***

121   The full set of Skp1 seed sequences that encompass a 70 to 86 amino acid core Skp1 domain

122   were retrieved from Pfam (PF01466, Version 27, http://pfam.xfam.org) and used as query in a

123   BLASTp search (Altschul et al. 1990) against the annotated proteome of each species, which

124   was retrieved from Phytozome (http://phytozome.jgi.doe.gov/; *A. lyrata* V2 and *A. halleri* V1.1).

125   The presence of Skp1 and any additional protein-protein interacting domains in each full-length

126   hit sequence were further confirmed by hmmscan (http://hmmer.org) against the Pfam-A

127   database (Pfam 27, http://pfam.xfam.org). To identify a complete list of *Skp1* genes in each

128   species, a previously developed sequence similarity-based annotation algorithm, called Closing

129   Target Trimming (Hua and Early 2018; Hua et al. 2011), was also used to search the genomes

130   for any new *Skp1* loci that may not have been annotated.

131

132   **Sequence alignment and phylogenetic analysis**

133   Instead of manual adjustment and artificial deletion of ambiguous alignment as reported in the

134   previous studies (Kong et al. 2004; 2007; Zhao et al. 2003), two Skp1 protein sequence

135   alignments were obtained by MUSCLE (Edgar 2004) and MAFFT (Katoh et al. 2017), and then

136   used to make a consensus alignment by Trimal (-conthreshold 0.5) (Capella-Gutierrez et al.

137   2009). The resulting alignment was used to generate a maximum likelihood phylogenetic tree

138   by RAxML (Version 8.1; Stamatakis 2014) with the PROTGAMMAJTT substitution model. The

139   statistical significance was evaluated with 1,000 bootstrap replicates using a rapid bootstrap

140   analysis.

141

142   **Birth and death of the *Skp1* genes in the Arabidopsis genus**

143   Gene duplication and loss events were inferred by reconciling the maximum likelihood gene tree

144   with the species tree using Notung (version 2.9) (Chen et al. 2000).

145

Gene structure analysis and reference sequence retrieval

147 The number of introns in each *Skp1* gene was counted based on the Generic Feature Format
148 (GFF3) file from each genome project. According to the chromosomal coordinates, the upstream
149 and downstream regions of a *Skp1* gene, which are 500 nucleotides upstream of the start codon
150 and downstream of the stop codon, respectively, were retrieved from the genomes of *A. thaliana*
151 (TAIR V10; www.arabidopsis.org) and *A. lyrata* (V2; Rawat et al. 2015). The coding sequence of
152 an *ASK* gene within the Col-0 reference genome and its *A. lyrata* ortholog was retrieved from the
153 annotated transcriptomes of *A. thaliana* (TAIR V10) and *A. lyrata* (V2; Rawat et al. 2015),
154 respectively.

155

Sequence assembly for polymorphism analysis

157 To assemble an *ASK* allelic sequence, single-nucleotide polymorphic (SNP) alleles (Phred quality
158 score ≥ 25) within the coding and non-coding flanking sequences were first retrieved from each
159 Arabidopsis accession (http://1001genomes.org). In total, 774 accessions were selected (Cao et
160 al. 2011; Consortium 2016; Long et al. 2013; Schmitz et al. 2013). Based on their co-ordinates,
161 both variant and the Col-0 reference SNP alleles were used to substitute the nucleotides in a
162 reference sequence to assemble two allelic sequences. Only when the new Col-0 allelic
163 sequence was 100% identical to the reference sequence was the variant allele considered to be
164 assembled correctly. To assemble an outgroup sequence, an amino acid sequence alignment of
165 an ASK protein and its *A. lyrata* ortholog was obtained by MAFFT (Katoh et al. 2017) and used to
166 guide the assembly of a nucleotide sequence of the outgroup. The sites introducing gaps in the
167 reference *ASK* sequence were removed.

168

Determination of tandemly duplicated *Skp1* genes

170 Two *Skp1* genes were determined to be tandemly duplicated if they were both separated by ≤5
171 genes and located within 10 kb.

172

Clustering analysis

174 Sequences were clustered using Heatmap.2 (dist method = "manhattan", hclust method =
175 "word.D") in R (http://www.r-project.org) to demonstrate similar evolutionary constrains of
176 mutations as described previously (Hua et al. 2018).

177

Expression data resources

179   The RNA-Seq data for leaf or seedling transcriptomes from 144 and 19 different *A. thaliana*
180   accessions were retrieved from the projects published by Schmitz et al. (2013) and Gan et al.
181   (2011), respectively. Microarray expression data of 79 samples collected from 8 different
182   tissues/organs throughout the *A. thaliana* Col-0 life cycle were downloaded from
183   http://jsp.weigelworld.org/expviz/expviz.jsp.

184

185   **Cross species test of neutral evolution**
186   The orthologous pairing of a *Skp1* gene between *A. thaliana* and *A. lyrata* was determined by
187   OrthoMCL (Li et al. 2003) and used to examine its neutral evolutionary process as previously
188   described (Nekrutenko et al. 2002), with minor modifications. The nucleotide sequences of each
189   pair were aligned based on the protein sequence alignment obtained by T-Coffee (Taly et al.
190   2011) and used as an input file to run the codeml program from the PAML4 package (Yang 2007)
191   twice, with the *Ka/Ks* ratio either fixed at 1 or free. The maximum likelihood (ML) values ML1 and
192   ML2 from the two runs were collected to calculate the likelihood ratio as LR = 2(lnML1–lnML2). If
193   LR is less than 2.71 (5% significance for $\chi 2$ distribution with one degree of freedom) (Yang 2007),
194   the *Ka/Ks* ratio is considered not significantly different from 1, i.e. the *Skp1* gene is under a neutral
195   evolutionary process.

196

197   **Molecular cloning and yeast two-hybrid analysis**
198   The coding sequences of 15 selected known *F-box* genes and *ASK1/2* were PCR amplified
199   from cDNA clones that were obtained from the Arabidopsis Biological Resource Center
200   (https://abrc.osu.edu) and ligated in-frame to the 3'-end of GAL4-BD and GAL4-AD coding
201   sequences present in the yeast two-hybrid vectors, pGBK-T7 (bait) and pGAD-T7 (prey),
202   respectively. The resulting bait and prey vectors were separately transformed into the haploid
203   yeast strains, AH109 and Y187, respectively, which were subsequently mated according to the
204   pairwise interaction combinations.

205

206   For yeast growth interaction assays, the number of mated diploid yeast cells were normalized
207   and diluted with sterile water in series to an OD_{600} of 0.8, 0.4, 0.2 and 0.1. 5 μL of yeast cells
208   from each concentration were then spotted on either a quadruple synthetic dropout medium
209   (SD-Leu-Trp-Ade-His) containing X-α-gal (40 μg/mL) for interaction assays, or on a double
210   synthetic dropout medium (SD-Leu-Trp) as a growth control. To quantify the interaction
211   strength, the intensity of yeast growth from the scanned images was calculated using

212 ImageQuant version 5.2 (GE Healthcare). Each interaction signal on SD-Leu-Trp-Ade-His+X-α-

213 gal medium was normalized to that detected on SD-Leu-Trp medium.

214

215 For the β-galactosidase activity assay, six to ten mated yeast colonies grown on SD-Leu-Trp

216 medium were freshly harvested and resuspended in 0.5 mL of Z buffer (50 mM sodium

217 phosphate, 10 mM potassium chloride, 2 mM magnesium sulfate, pH 7.0) in a 2 mL deep-well

218 plate. 0.1 mL of resuspended yeast cells were further aliquoted into a new 2 mL deep-well plate

219 and used for β-galactosidase activity according to Miller (Miller 1972). Relative β-galactosidase

220 activities were calculated based on the method described previously (Hua and Kao 2006). In

221 total, two independent replicates were assayed.

222

223 **Results**

224 **Identification of *Skp1* genes in three closely related Arabidopsis species**

225 The list of 21 *A. thaliana ASK* genes, which contain 19 Type I (*ASK1-19*) and 2 Type II (*ASK20*

226 and *21*) *Skp1* genes, has been well annotated in previous studies (Farras et al. 2001; Kong et

227 al. 2007; Kong et al. 2004; Zhao et al. 2003). However, a full list of *Skp1* genes has not been

228 reported in *A. lyrata* or *A. halleri*. To identify a comprehensive list of *Skp1* genes in these two

229 Arabidopsis genomes, BLASTp (Altschul et al. 1990) and hmmscan (http://hmmer.org) were first

230 applied to search the available genome annotations [*A. lyrata* V2 (Rawat et al. 2015), *A. halleri*

231 V1.1 (Phytozome V12)]. In total, 17 and 11 loci were identified that encode a Type I Skp1

232 protein, and 4 and 3 loci were discovered to encode a Type II Skp1 protein, in *A. lyrata* and *A.*

233 *halleri*, respectively (Supplemental File S1). After a subsequent sequence-similarity based

234 Closing Target Trimming search (Hua and Early 2018; Hua et al. 2011), no additional hits were

235 identified in each genome. The relatively long length of the Skp1 domain, the low number or

236 absence of introns in the *Skp1* loci, and the small size of the *Skp1* family, may facilitate the

237 annotation of *Skp1* genes in genomes. Therefore, in total 21, 21, and 14 *Skp1* loci are present

238 in *A. thaliana*, *A. lyrata*, and *A. halleri*, respectively. The size variation among these closely

239 related Arabidopsis species indicates that the *Skp1* family is, like the *F-box* gene superfamily,

240 under a rapid birth-and-death evolutionary process.

241

242 **Phylogenetic analysis of the *Skp1* genes in a short evolutionary history**

243 Since the sequences of Type II *Skp1* genes are significantly diverged from Type I *Skp1* genes,

244 and most *Skp1* genes are Type I (Kong et al. 2007), hereafter we focused on the evolutionary

245 study of the Type I group. To understand the birth-and-death process of the *Skp1* family in a

246   short evolutionary history, we performed a phylogenetic analysis using an improved sequence

247   alignment approach. Since manual adjustment and artificial deletion of ambiguous alignment

248   sites are not always reproducible, Trimal (Capella-Gutierrez et al. 2009) was used to remove

249   poorly aligned regions automatically. In addition, a consensus result was obtained from

250   MUSCLE (Edgar 2004) and MAFFT (Katoh et al. 2017) sequence alignments to improve

251   accuracy (see Materials and Methods). The resulting alignment not only significantly reduced

252   gaps and mis-matched sites, but also retained a reproducible result with 95 ± 5% of the full

253   length Skp1 protein sequences being aligned (Figure S1), maximizing the sequence length and

254   variable sites essential for a good phylogenetic analysis (Nei and Kumar 2000). As a proof of

255   concept, an ASK2-rooted maximum likelihood tree generated based on the 19 aligned ASK

256   protein sequences showed a compatible topology to the one reported previously (Kong et al.

257   2007) (Figure 1A). However, unlike the previous tree where intronless and intronic *ASK* genes

258   are intermingled (Kong et al. 2007), all intron-containing *ASK* genes (except for *ASK15*, whose

259   intron was gained after duplication (Kong et al. 2007)) were clustered at the base of the tree

260   (Figure 1A). This result better explains a single origin of the intronless *ASK* genes, which were

261   duplicated through retroposition from a highly expressed *ASK* gene, most likely the ancestor of

262   *ASK1*.

263

264   Subsequently, a maximum likelihood tree rooted to ASK2 was generated based on the

265   consensus protein sequence alignment of the 47 Arabidopsis Skp1 protein sequences by

266   RAxML (Stamatakis 2014). Encouragingly, the resulting phylogenetic tree showed a similar

267   topology to that obtained for the *ASK* genes, and the 47 *Skp1* genes from the three Arabidopsis

268   species were intermingled in 11 clusters (Figure 1B). Therefore, most Arabidopsis *Skp1* genes

269   were duplicated at least 5-10 mya, at the time when the three Arabidopsis species split (Hu et

270   al. 2011; Koch and Kiefer 2005). Similarly to the *ASK* genes, all intronic Arabidopsis *Skp1*

271   genes were clustered at the base of the tree, while the remaining intronless *Skp1* genes were

272   clustered into one big clade, suggesting a common role of retroposition in the expansion of the

273   *Skp1* family in Arabidopsis. Among these, three ancestor loci (Figure 1B, nodes *a*, *c*, and *d*)

274   were likely duplicated through retroposition, with each likely further undergoing tandem

275   duplication events to yield the current *Skp1* members. Two *Skp1* clades (Figure 1B, clades 2

276   and 6) were likely the direct product of a retroposition event. Therefore, the mRNAs produced

277   by the highly expressed *ASK1* ancestor were likely retrotransposed to 5 ancestor loci in total

278   (Figure 1B, nodes *a-e*). This phylogenetic tree also showed a clear duplication event between

279   the *ASK1* clade and the *ASK4* clade before the split of three Arabidopsis species. Although

280   Kong et al. (2007) first reported the contribution of segmental duplication in duplicating the

281   *ASK1* and *ASK4* loci, their phylogenetic tree did not reflect the direct connection between these

282   two genes (Kong et al. 2007).

283

284   We further reconciled a gene tree based on this maximum likelihood tree. Along with a species

285   tree, we detected significant variance of birth/death rates between each species (Figure 1C).

286   While *A. halleri* lost 7 *Skp1* loci after a recent split from *A. lyrata*, *A. thaliana* has gained and lost

287   5 loci each from the 19 common Arabidopsis *Skp1* gene ancestors. The birth/death rate of

288   *Skp1* genes in *A. lyrata* was intermediate among the three species; it gained and lost 2 and 4

289   *Skp1* genes, respectively, following the split from *A. thaliana*. Such a significant size variation

290   even within a short evolutionary history implied that some ancestral retroposed *Skp1* loci

291   resided in a hot spot of tandem duplications, which contributed to the differential sizes of the

292   *Skp1* family among the three Arabidopsis species. For example, at the ancestor "*a*" locus, 4

293   and 3 *Skp1* genes were gained through tandem duplications within the past 5-10 million years in

294   *A. thaliana* and *A. lyrata*, respectively (Figure 1B).

295

296   **Low evolutionary constraints of intronless *ASK* genes**

297   The phylogenetic tree revealed a clear orthology relationship between the *Skp1* genes of the

298   three Arabidopsis species. To further demonstrate this relationship, we applied OrthoMCL (Li et

299   al. 2003) to identify 14 *Skp1* orthologous groups, among which 18 *ASK* genes have been

300   partnered with one *A. lyrata Skp1* orthologous gene (Supplemental Table S1). This clear

301   orhology relationship allowed us to examine evolutionary constraints on the sequence

302   divergence of *ASK* genes. We primarily applied the method of Nekrutenko *et al.* (2002) to test

303   whether the *Ka/Ks* ratio (i.e., ω) of an *ASK* gene is significantly diverged from 1, which indicates

304   between-species neutral nucleotide divergence. Surprisingly, 9 out of 14 (64%) intronless *ASK*

305   genes were detected to be under a neutral evolutionary process, while none of the 4 intronic

306   *ASK* genes belongs to this category (Table 1), suggesting that the former group has lower

307   functional constraints than the latter. Interestingly, using GeneWise reannotation (Birney et al.

308   2004), we identified a frame-shift mutation in the *ASK6*, *AlySkp02*, and *AhaSkp11* loci that is

309   characteristic of pseudogenes (Figures S2-S4).

310

311   **Expression variation of the *ASK* genes**

312   Low functional constraints do not necessarily mean no function (Nei and Kumar 2000). To

313   understand the functional differences between *ASK* genes, previous studies examined their

314   expression patterns in different tissues/organs by semi-quantitative reverse transcription-PCR

315   (RT-PCR) and *in situ* hybridization (Kong et al. 2004; Zhao et al. 2003). The results suggested

316   that all *ASK* genes could be expressed in at least one of six samples examined (mostly in

317   siliques). However, only the highly-expressed *ASK1* and *ASK2* genes showed a consistent

318   result in both studies, while other *ASK* genes with low expression levels were not stably

319   detected by RT-PCR, probably due to technical limitations (Kong et al. 2004; Zhao et al. 2003).

320   Since then, a number of high throughput transcriptomic analyses, including microarray and

321   RNA-Seq experiments, have been performed in *A. thaliana*, providing good resources to further

322   examine the expression patterns of *ASK* genes in a more statistical manner.

323

324   Microarray experiments from 79 *A. thaliana* samples collected from different developmental

325   stages confirmed that *ASK1* and *ASK2* are the two most highly expressed *ASK* genes, followed

326   by *ASK3*. Unlike the previous studies (Kong et al. 2004; Zhao et al. 2003), not all *ASK* genes

327   were detected in these microarray transcriptomic analyses, probably due to the low sensitivity

328   and precision of microarray technologies (Wang et al. 2009) (Figure 2A). After comparing the

329   expression levels of *ASK* genes in 8 different tissues/organs (79 samples in total), we found that

330   several *ASK* genes, including *ASK1*, *2*, *3*, *7*, *8*, *9*, and *10*, were expressed most highly in seeds

331   (Figure 2B), suggesting an important role for SCF-mediated protein ubiquitylation in seed

332   development.

333

334   We further analyzed the expression variance of *ASK* genes at the population level. Based on

335   one RNA-Seq experiment (Schmitz et al. 2013), which provided a much more precise and

336   sensitive transcriptomic analysis than microarrays (Wang et al. 2009), the transcripts of each

337   *ASK* gene could be detected in leaf tissues of 144 *A. thaliana* accessions, but with dramatic

338   variance (Figure 2C). Consistent with the previous studies and the microarray data, *ASK1* has

339   the highest expression level, followed by *ASK2*. However, the remaining *ASK* genes were only

340   expressed at an average of 1.3 Reads Per Kilobase of transcript per Million mapped reads

341   (RPKM), 121- and 42-fold below the mean expression of *ASK1* and *ASK2*, respectively.

342   Therefore, while *ASK1* and *ASK2* likely have important function(s), the role(s) of the other *ASK*

343   genes in leaves appears minor. Interestingly, the four *ASK* genes (*ASK1*, *2*, *4*, and *18*) whose

344   mean expression levels are significantly higher than the other *ASK* genes (Wilcoxon rank- sum

345   test, p < 2.2e-16,) are all under strong evolutionary constraints, and three of them are intronic,

346   further supporting our conclusion regarding the low functional constraints of intronless *ASK*

347   genes.

348

349 In addition to the expression variation among different *ASK* genes, the expression of *ASK1* and

350 *ASK2* varied dramatically among different individuals. For example, the highest and the lowest

351 expression levels of *ASK1* were detected in the Gr-1 (480 RPKM, Longitude/Latitude/Altitude =

352 15.5/47/300) and Co-1 (50 RPKM, Longitude/Latitude/Altitude = -8.3/40.1/100) accessions,

353 respectively, varying by 430 RPKM, while the highest and the lowest expression levels for *ASK2*

354 were detected in Ven-1 (210 RPKM, Longitude/Latitude/Altitude = 5.6/52/parking lot) and Ann-1

355 (10 RPKM, Longitude/Latitude/Altitude = 6.1/45.9/garden), respectively, differing by 200 RPKM

356 (Figure 2C). To confirm this result, we also examined the differential expression of *ASK1* and

357 *ASK2* in another RNA-Seq experiment (Gan et al. 2011). The expression of *ASK1* varied

358 significantly among Col-0, Ler-0, and Ws-0, which are from distinct geographic regions, while

359 *ASK2* showed only mild changes (Figure 2D). Therefore, although *ASK1* has the highest

360 expression level among all the *ASK* genes, its expression varies the most among individuals,

361 reflecting a possible adaptation of its expression regulation.

362

363 **Differential sequence polymorphism of the *ASK* genes**

364 We further examined sequence diversity and polymorphism of the *ASK* genes within the family

365 by comparing number of segregating sites per nucleotide site (θ) and nucleotide diversity (π)

366 values in the regions 500 bp upstream of the start codon, within the coding sequence, and 500

367 bp downstream of the stop codon of an *ASK* gene among 774 *A. thaliana* accessions based on

368 their single nucleotide polymorphic (SNP) data (Consortium 2016). We first calculated the allele

369 frequency distribution and noticed that minor allele frequency (MAF) alleles were significantly

370 enriched. In the total populations analyzed, 34% of non-synonymous alleles and 35%

371 synonymous alleles are only present once (i.e., singleton) in the *ASK* family (Figure 3A),

372 suggesting that many mutations are rare.

373

374 Consistent with the enrichment of low MAF alleles, the π values are significantly lower than the

375 θ values in coding sequences (Figure 3B), because the former is determined by allele frequency

376 and the latter is not (Nei and Kumar 2000). Similar to the coding sequences, the π values are

377 also significantly lower than the θ values in both upstream and downstream regions, indicating

378 that low MAF alleles are also high in these two regions. As expected, these two regions have

379 higher π and θ values than the coding sequences due to their low functional constraints.

380

381   The variance of π and θ values within and between different regions of the *ASK* genes indicates

382   different extents of polymorphism. To further examine the evolutionary constraints of

383   polymorphic mutations, a Tajima's D value (Tajima 1989) was calculated for the three regions of

384   an *ASK* gene (upstream, coding sequence (CDS) and downstream). According to the variance

385   of this value, the 18 *ASK* genes as described in Table 1 were clustered into 3 groups (Figure

386   4C). In Group I, which contains *ASK4*, *6*, *10*, *14*, and *19*, a high Tajima's D value was observed

387   in the coding sequences, suggesting a balancing or neutral evolutionary process. Group II,

388   which includes *ASK5*, *8*, *9*, *12*, *13*, *16*, and *17*, shows intermediate Tajima's D values, some of

389   which are significantly smaller than the Tajima's critical values of neutral mutations ($p < 0.05$)

390   (Tajima 1989), suggestive of purifying selection. The remaining 6 *ASK* genes (*ASK1*, *2*, *3*, *7*,

391   *11*, and *18*) are clustered into Group III where most have a Tajima's D value below the Tajima's

392   critical values of neutral mutations in all three regions ($p < 0.05$). Therefore, mutations in this

393   group are rare, and most sequences are under strong purifying selection. Interestingly, this

394   group enriched 5 out of 6 *ASK* genes that were detected to be under non-neutral changes by

395   orthology comparison (Table 1), further confirming their high evolutionary constrains. It is worth

396   noting that the Tajima's D value of the coding sequence of *ASK2* is within the range of Tajima's

397   critical values, suggesting that most polymorphic SNPs in *ASK2* are neutral.

398

399   **Variation association of *ASK1* expression and polymorphism**

400   To associate sequence polymorphism with gene expression variation, we compared the

401   expression patterns and pair-wise nucleotide diversity (π) of 144 *A. thaliana* accessions from

402   the RNA-Seq experiment done by Schmitz et al. (2013). Based on the wide latitudinal

403   distribution of accessions, 5 subgroups were separated (Figure 4A). Although no linear

404   regression is observed between individual expression levels and latitudes, accessions within

405   latitudes 45-50°N have significantly higher levels of *ASK1* gene expression than the two flanking

406   regions 5° to the north or south (Wilcoxon rank-sum test, $p < 0.01$). Since 116 out of 144

407   accessions (81%) reside in these three sub-regions, the populations in these areas were further

408   analyzed. To better understand the relationship between nucleotide diversity and expression

409   levels of *ASK1*, the entire 2.5 kb intergenic sequence upstream of the transcription start site

410   (TSS) of *ASK1* was compared. A window-sliding analysis (200 bp window and 100 bp slide)

411   showed that the windowed π values from nucleotide -1600 to -500 are significantly higher in

412   accessions within latitudes 45-50°N than those in the other two regions (Figure 4B, Wilcoxon

413   rank-sum test, $p < 0.05$). Such a relationship between high polymorphism and high expression

414   suggests that nucleotide variance in the promoter may prevent the binding of a putative

415   transcriptional repressor that may suppress the expression of *ASK1*. To further demonstrate

416   that the accessions within latitudes 45-50°N are highly polymorphic in the promoter region, we

417   also compared accessions from the projects in Cao et al. (2011) and Long et al. (2013). In total,

418   29, 20, and 12 accessions resided in latitudes 40-45, 45-50, and 50-60°N, respectively.

419   Consistently, the accessions from latitudes 45-50°N also have the highest sequence

420   polymorphism from nucleotide -1600 to -500 among the three groups compared (Figure 4C).

421   We also applied a different grouping method by separating the aforementioned 144 accessions

422   into low (49 accessions with *ASK1* expression value in the 50-137 RPKM range), medium (50

423   accessions with *ASK1* expression of 137-170 RPKM), and high (45 accessions with *ASK1*

424   expression of 170-480 RPKM) expression groups. Among these three groups, high nucleotide

425   polymorphism in the -1000 to -500 region is also most evident in the group with high *ASK1*

426   expression (Figure 4D).

427

428   **Variance of biochemical interactions of ASK1 and ASK2 with known F-box proteins**

429   The different extent of sequence variation between *ASK1* and *ASK2* coding sequences (Figure

430   3C) led us to speculate that their encoded proteins might show differential strength of interaction

431   with F-box proteins. To address this question, we tested the interactions of ASK1 and ASK2 by

432   pair-wise yeast two-hybrid assay with 15 randomly-selected F-box proteins whose functions

433   have been identified (Figure 5 and Table S2). Due to the identification of neutral polymorphic

434   mutations in the *ASK2* CDS and purifying selection in the *ASK1* CDS, we hypothesized that

435   ASK2 might have lost or reduced its interactions with a number of F-box proteins. To provide a

436   starting point to examine the potential biochemical differences between ASK1 and ASK2, we

437   performed both yeast growth assay on quadruple synthetic dropout medium (SD-Leu-Trp-Ade-

438   His) containing X-α-gal (Figure 5A and 5B) and β-galactosidase activity analysis (Figure 5C).

439   Interestingly, we detected that two F-box proteins (JMJ22 and DIF) interacted with ASK1 with a

440   strength >2 fold more than with ASK2 (Figure 5B and 5C).   However, the remaining F-box

441   proteins showed an average of 15% variation in their interactions with ASK1 and ASK2,

442   suggesting that the ASK2 protein, albeit carrying recently neutral mutations, still retains the

443   capability in binding many functional F-box proteins. It is worth mentioning that the interaction

444   strengths of ASK1 or ASK2 with different F-box proteins also vary dramatically. This is

445   consistent with the high sequence divergence of *F-box* genes (Hua et al. 2013).

446

447   **Discussion**

448   Gene duplication has been a long-standing topic of interest in genome evolution (Ohno 1970).

449   In eukaryotic genomes, this process plays an essential role in the expansion of many multi-gene

450   families including those involved in the ubiquitin-26S proteasome system (Hua et al. 2011;

451   2013; 2018; Hurles 2004; Li et al. 2016; Panchy et al. 2016). While gene duplication provides

452   the raw genetic material for genome innovation, the large size of multi-gene families has been a

453   hurdle in exploring genome function. For example, it is not yet clear whether and how multiple

454   *ASK* members contributed to evolutionary innovation through sub-functionalization or neo-

455   functionalization, or whether the expansion of the *ASK* family is simply due to the selective

456   advantage of gene dosage or is a result of genomic drift. In this study, we applied several novel

457   approaches to address these questions, as outlined below.

458

459   **Retroposed *ASK* genes originated from one single ancestor locus**

460   Instead of cross kingdom long-distant phylogenetic studies (Kong et al. 2004; 2007), we focused

461   on a short evolutionary history within the Arabidopsis genus, so that the orthology relationships

462   and duplication history of individual *ASK* members could be more clearly illustrated. In addition,

463   more advanced sequence alignment tools have been adopted to improve the phylogenetic

464   analysis (Figure 1). For example, although Kong et al. (2007) discovered that *ASK1* and *ASK4*

465   were duplicated through a segmental duplication event by comparing whole genome duplication

466   blocks, their phylogenetic analysis did not reflect this duplication event. Here, we provided

467   improved phylogenetic evidence not only showing the segmental duplication relationship

468   between *ASK1* and *ASK4*, but also demonstrating a single origin of all intronless *ASK* members

469   from the ancestor of *ASK1* through retroposition (Figure 1). This new discovery is also

470   consistent with the high expression of *ASK1* (Figure 2) and the young evolutionary history of

471   intronless *ASK* members. Similarly, a previous study on the placental mammalian Gli-Kruppel

472   type zinc finger transcription factor YY1 family revealed that two retroposed intronless

473   subfamilies, YY2 and Reduced Expression 1 (REX1), were separately clustered but not mingled

474   with the intronic YY1 members (Kim et al. 2007).

475

476   **Degenerative processes of *ASK* genes**

477   Since the first draft genome of *A. thaliana* was released, the evolutionary process of both the *F-*

478   *box* and *ASK* families has been a hot topic in plant biology due to the importance of SCF-

479   mediated protein ubiquitylation and the unequal expansion of the two families. Although it has

480   been hypothesized that variant ASK proteins might contribute to interactions with a specific

481   group of F-box proteins (Gagne et al. 2002; Kuroda et al. 2012), *in vivo* data is currently lacking.

482 Indeed, most functionally characterized F-box proteins physically interact with ASK1 (Hua and
483 Vierstra 2011). Furthermore, recently evolutionary studies have suggested that the number of
484 active *F-box* members is much fewer than the size of the family, due to a genomic drift
485 evolutionary process (Hua et al. 2011; Nozawa et al. 2007; Xu et al. 2009). This raised a
486 question as to whether all ASK proteins are involved in the assembly of active SCF complexes.
487
488 In this study, we integrated multiple levels of evidence to better describe the functional
489 constraints of individual *ASK* members. Orthology comparisons revealed that 64% of intronless
490 *ASK* genes were under neutral changes indicative of non-functionalization (Table 1). Both
491 developmental and population expression comparisons suggested that most intronless *ASK*
492 members have a very low expression level (Figure 2). In addition, sequence polymorphism
493 analysis showed a significant enrichment of intronless *ASK* members in the groups whose
494 mutations were under neutral changes (Figure 3C, clades I and II). Collectively, these data
495 suggest that most, if not all, retroposed *ASK* members are under low functional constraints.
496 This is indeed not surprising, since retroposed genes have a much high rate of
497 pseudogenization, as suggested in the duplication studies of the human genome (Torrents et al.
498 2003). Consistently, our previous evolutionary studies of the *F-box* gene superfamily also
499 discovered a significant enrichment of intronless genes in the pseudogene group (Hua et al.
500 2011). It will be of interest to explore the role of retroposition on the expansion of intronless *F-*
501 *box* genes.
502
503 **The large expression variance of *ASK1* may indicate diverse functions of SCF complexes**
504 Expression comparisons revealed a greater than 430 RPKM (~10 fold) variance in *ASK1*
505 transcript levels among individual *A. thaliana* accessions (Figure 2). Sequence analyses further
506 suggested the presence of a putative transcriptional repressor that might contribute to such
507 large variations in expression. Upstream sequences with more variable sites may prevent the
508 binding of this transcriptional repressor, thus increasing expression. The finding that more
509 polymorphic upstream sequences result in higher expression may support this model (Figure 4).
510 Such large variation in *ASK1* expression among natural variants suggests that ASK1 is either
511 very effective in promoting the assembly of SCF complexes or is involved in an as-yet-unknown
512 mechanism to regulate the polymorphic functions of SCF complexes. It will be noteworthy to
513 further investigate the proteomic variance of SCF-mediated protein ubiquitylation in the future.
514
515 **Conclusions**

516   In this study, our improved phylogenetic analysis resolved the inconsistency between the

517   phylogeny of *ASK* genes and the single origin of retroposed *ASK* members (Kong et al. 2007).

518   Through evolutionary selection analysis and sequence polymorphism comparison, we

519   discovered both adaptive and degenerative evolutionary processes in the *ASK* family. Yeast

520   two-hybrid quantitative interaction assay and expression analysis across different accessions

521   further indicated that recent neutral changes in the *ASK2* coding sequence likely weakened its

522   interactions with F-box proteins and that highly polymorphic upstream regions of *ASK1* may

523   contribute to adaptive roles of SCF complexes in Arabidopsis, respectively.

524

525   **Supplemental Information**

526   **Table S1.** OrthoMCL groups of Skp1 proteins in *A. thaliana*, *A. lyrata*, and *A. halleri*

527   **Table S2.** Gene identifications of 15 known F-box proteins selected for testing the differential

528   interactions with ASK1 and ASK2

529   **Table S3.** Raw data for β-galactosidase activity assay

530   **File S1.** Amino acid sequences of the Skp1 proteins annotated in *A. thaliana*, *A. lyrata*, and *A.*

531   *halleri*

532   **Figure S1.** An improved amino acid sequence alignment of 47 Arabidopsis Skp1 protein

533   sequences

534   **Figure S2.** A GeneWise reannotation uncovers a frame-shift mutation in *ASK6*.

535   **Figure S3.** A GeneWise reannotation uncovers a frame-shift mutation in *AlySkp02*.

536   **Figure S4.** A GeneWise reannotation uncovers a frame-shift mutation in *AhaSkp11*.

537   **Figure Legends**

538   **Figure 1.** A short evolutionary history of the Arabidopsis *Skp1* genes within the Arabidopsis

539   genus. A) Phylogenetic relationships of the *ASK* members. Intronic genes are highlighted in

540   red. Scale bar, average substitutions per site. B) An improved phylogenetic analysis reveals

541   one single origin of retroposed *Skp1* genes in the Arabidopsis genus. *a-e*, 5 ancient retroposed

542   loci produced by the transcripts of the highly expressed ancestor locus of *ASK1*. 1-11, 11

543   clades shaded with dark and gray color showing clear orthologous relationships among the

544   three Arabidopsis species. The tree was generated by a maximum likelihood method.

545   Statistical significance equal to or greater than 50% of 1,000 bootstrap re-samplings is indicated

546   in each node. S: segmental duplication; T: tandem duplication; R: retroposed duplication. The

547   ID number from the original annotation is indicated if a *Skp1* gene is duplicated through tandem

548   duplications or described in File S1. Intronic *Skp1* genes and scale bar are described as in A.

549   Species abbreviations: *A. halleri* (*Aha*), *A. lyrata* (*Aly*), and *A. thaliana* (*Ath*). C) Birth-and-death

550   history of *Skp1* genes in the Arabidopsis genus. + and – indicate gain and loss of *Skp1*

551   members, respectively. Species abbreviations are as in B.

552

553   **Figure 2.** Expression variation of *ASK* genes. A) Variation of absolute expression throughout a

554   developmental time course. B) A heatmap analysis of mean expression variation of *ASK* genes

555   across 8 tissues/organs. C) Expression variation of *ASK* genes in leaves of 144 accessions.

556   The width of each boxplot is proportional to the number of accessions having a non-zero

557   expression data of the corresponding *ASK* gene. D) Expression comparison of *ASK1* and *ASK2*

558   in Col-0, Ler-0, and Ws-0.

559

560   **Figure 3.** Sequence polymorphic comparison among different *ASK* genes. A) Frequency

561   distribution of rare *ASK* alleles with non-synonymous and synonymous mutations. B) Variation

562   of sequence polymorphisms in three regions of an *ASK* gene. UP: 500 bp upstream of start

563   codon, CDS: coding sequence, DN: 500 bp downstream of the stop codon. C) A heatmap

564   representation of Tajima's D values demonstrating the differential evolutionary constraints of

565   polymorphic mutations among *ASK* genes. Asterisks indicate a significantly low Tajima's D

566   value that deviates from neutral mutations ($p < 0.05$).

567

568   **Figure 4.** Contribution of sequence polymorphism to gene expression. A) Latitudinal variation

569   of *ASK1* expression. Asterisks indicate that the mean expression of the indicated group is

570   significantly lower than that of the group within latitudes 45-50°N (Wilcoxon rank-sum test, $p <$

571   0.01). B) A window-slide polymorphic comparison of a 2.5 kb region upstream of the

572   transcriptional start site of *ASK1* among populations from 3 latitudinal regions as in A. C) A

573   window-slide polymorphic comparison of a group of test populations from 3 latitudinal regions,

574   performed as in B. D) A window-slide polymorphic comparison of three groups of populations

575   with low, medium, and high expression levels of *ASK1* that was performed as in B.

576

577   **Figure 5.** Quantitative interaction assay of ASK1 and ASK2 with 15 selected known F-box

578   proteins. The accession IDs of F-box proteins are listed in Table S2. A) Growth of yeast cells

579   expressing each indicated pair of bait and prey proteins on SD-Leu-Trp-Ade-His+X-α-gal

580   medium. B) Control growth of the corresponding yeast cells in A on SD-Leu-Trp medium.

581   Yeast cells were grown on media at 30°C for an indicated time period and their growth was

582   subsequently recorded by scanning the plates with a Canon 9000F Mark II scanner. B)

583   Quantification of pairwise ASK1 and ASK2 interactions with 15 F-box proteins as shown in A.

584   The interaction strength of each pair of bait and prey proteins shown in A was normalized by the

585   control growth strength of the corresponding yeast cells in B. C) β-galactosidase activity assay.

586   The mated yeast cells expressing the indicated pair of bait and prey proteins were grown on

587   SD-Leu-Trp medium and used for the assay. The β-galactosidase activities shown are mean

588   values ± SD measured from two independent assays.

589

590

591

592

593   **References**

594

595   Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search

596   tool. *J Mol Biol* 215:403-410. 10.1016/S0022-2836(05)80360-2

597   Birney E, Clamp M, and Durbin R. 2004. GeneWise and Genomewise. *Genome Res* 14:988-

598   995. 10.1101/gr.1865504

599   Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O,

600   Lippert C, Wang X, Ott F, Muller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, and

601   Weigel D. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana*

602   populations. *Nat Genet* 43:956-963. 10.1038/ng.911

603   Capella-Gutierrez S, Silla-Martinez JM, and Gabaldon T. 2009. trimAl: a tool for automated

604   alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972-1973.

605   10.1093/bioinformatics/btp348

606   Chen K, Durand D, and Farach-Colton M. 2000. NOTUNG: a program for dating gene

607   duplications and optimizing gene family trees. *J Comput Biol* 7:429-447.

608   10.1089/106652700750050871

609   Consortium G. 2016. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis*

610   *thaliana*. *Cell* 166:481-491. 10.1016/j.cell.2016.05.063

611   Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high

612   throughput. *Nucleic Acids Res* 32:1792-1797. 10.1093/nar/gkh340

613   Farras R, Ferrando A, Jasik J, Kleinow T, Okresz L, Tiburcio A, Salchert K, del Pozo C, Schell

614   J, and Koncz C. 2001. SKP1-SnRK protein kinase interactions mediate proteasomal

615   binding of a plant SCF ubiquitin ligase. *EMBO J* 20:2742-2756.

616   10.1093/emboj/20.11.2742

617   Finley D, Ulrich HD, Sommer T, and Kaiser P. 2012. The ubiquitin-proteasome system of

618   *Saccharomyces cerevisiae*. *Genetics* 192:319-360. 10.1534/genetics.112.140467

619   Gagne JM, Downes BP, Shiu SH, Durski AM, and Vierstra RD. 2002. The F-box subunit of the

620   SCF E3 complex is encoded by a diverse superfamily of genes in Arabidopsis. *Proc Natl*

621   *Acad Sci USA* 99:11519-11524. 10.1073/pnas.162339999

622   Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ,

623   Osborne EJ, Sreedharan VT, Kahles A, Bohnert R, Jean G, Derwent P, Kersey P,

624   Belfield EJ, Harberd NP, Kemen E, Toomajian C, Kover PX, Clark RM, Ratsch G, and

625   Mott R. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*.

626   *Nature* 477:419-423. 10.1038/nature10414

627   Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J,
628   Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA,
629   Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington
630   JC, Gaut BS, Schmutz J, Mayer KF, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel
631   D, and Guo YL. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid
632   genome size change. *Nat Genet* 43:476-481. 10.1038/ng.807
633   Hua Z, Doroodian P, and Vu W. 2018. Contrasting duplication patterns reflect functional
634   diversities of ubiquitin and ubiquitin-like protein modifiers in plants. *Plant J* 95:296-311.
635   10.1111/tpj.13951
636   Hua Z, and Early MJ. 2018. Closing Target Trimming: a Perl package for discovering hidden
637   superfamily loci in genomes. *bioRxiv*:490490. 10.1101/490490
638   Hua Z, and Kao TH. 2006. Identification and characterization of components of a putative
639   petunia *S*-locus F-box-containing E3 ligase complex involved in S-RNase-based self-
640   incompatibility. *Plant Cell* 18:2531-2553. 10.1105/tpc.106.041061
641   Hua Z, Pool JE, Schmitz RJ, Schultz MD, Shiu SH, Ecker JR, and Vierstra RD. 2013.
642   Epigenomic programming contributes to the genomic drift evolution of the F-Box protein
643   superfamily in Arabidopsis. *Proc Natl Acad Sci USA* 110:16927-16932.
644   10.1073/pnas.1316009110
645   Hua Z, and Vierstra RD. 2011. The cullin-RING ubiquitin-protein ligases. *Annu Rev Plant Biol*
646   62:299-334. 10.1146/annurev-arplant-042809-112256
647   Hua Z, Zou C, Shiu SH, and Vierstra RD. 2011. Phylogenetic comparison of F-Box (FBX) gene
648   superfamily within the plant kingdom reveals divergent evolutionary histories indicative of
649   genomic drift. *PLoS One* 6:e16219. 10.1371/journal.pone.0016219
650   Hurles M. 2004. Gene duplication: the genomic trade in spare parts. *PLoS Biol* 2:E206.
651   10.1371/journal.pbio.0020206
652   Katoh K, Rozewicki J, and Yamada KD. 2017. MAFFT online service: multiple sequence
653   alignment, interactive sequence choice and visualization. *Brief Bioinform*.
654   10.1093/bib/bbx108
655   Kim JD, Faulk C, and Kim J. 2007. Retroposition and evolution of the DNA-binding motifs of
656   YY1, YY2 and REX1. *Nucleic Acids Res* 35:3442-3452. 10.1093/nar/gkm235
657   Koch MA, and Kiefer M. 2005. Genome evolution among cruciferous plants: a lecture from the
658   comparison of the genetic maps of three diploid species - *Capsella rubella*, *Arabidopsis*
659   *lyrata subsp. petraea*, and *A. thaliana*. *Am J Bot* 92:761-767. 10.3732/ajb.92.4.761

660 Kong H, Landherr LL, Frohlich MW, Leebens-Mack J, Ma H, and dePamphilis CW. 2007.
661   Patterns of gene duplication in the plant *SKP1* gene family in angiosperms: evidence for
662   multiple mechanisms of rapid gene birth. *Plant J* 50:873-885. 10.1111/j.1365-
663   313X.2007.03097.x

664 Kong H, Leebens-Mack J, Ni W, dePamphilis CW, and Ma H. 2004. Highly heterogeneous rates
665   of evolution in the *SKP1* gene family in plants and animals: functional and evolutionary
666   implications. *Mol Biol Evol* 21:117-128. 10.1093/molbev/msh001

667 Kuroda H, Yanagawa Y, Takahashi N, Horii Y, and Matsui M. 2012. A comprehensive analysis
668   of interaction and localization of Arabidopsis SKP1-like (ASK) and F-box (FBX) proteins.
669   *PLoS One* 7:e50009. 10.1371/journal.pone.0050009

670 Li L, Stoeckert CJ, Jr., and Roos DS. 2003. OrthoMCL: identification of ortholog groups for
671   eukaryotic genomes. *Genome Research* 13:2178-2189. 10.1101/gr.1224503

672 Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, and De Smet R. 2016. Gene
673   duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* 28:326-
674   344. 10.1105/tpc.15.00877

675 Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, Zhang Q, Vilhjalmsson BJ, Korte
676   A, Nizhynska V, Voronin V, Korte P, Sedman L, Mandakova T, Lysak MA, Seren U,
677   Hellmann I, and Nordborg M. 2013. Massive genomic variation and strong selection in
678   *Arabidopsis thaliana* lines from Sweden. *Nat Genet* 45:884-890. 10.1038/ng.2678

679 Miller JH. 1972. Experiments in Molecular Genetics. *Cold Spring Harbor, NY: Cold Spring
680   Harbor Laboratory Press*.

681 Nei M, and Kumar S. 2000. Molecular evolution and phylogenetics. *Oxford University Press,
682   Oxford*.

683 Nekrutenko A, Makova KD, and Li WH. 2002. The K_A/K_S ratio test for assessing the protein-
684   coding potential of genomic regions: an empirical and simulation study. *Genome Res*
685   12:198-202. 10.1101/gr.200901

686 Nozawa M, Kawahara Y, and Nei M. 2007. Genomic drift and copy number variation of sensory
687   receptor genes in humans. *Proc Natl Acad Sci USA* 104:20421-20426.
688   10.1073/pnas.0709956104

689 Ohno S. 1970. Evolution by gene duplication. *Springer-Verlag, New York*.

690 Panchy N, Lehti-Shiu M, and Shiu SH. 2016. Evolution of gene duplication in plants. *Plant
691   Physiol* 171:2294-2316. 10.1104/pp.16.00523

692   Rawat V, Abdelsamad A, Pietzenuk B, Seymour DK, Koenig D, Weigel D, Pecinka A, and
693   Schneeberger K. 2015. Improving the annotation of *Arabidopsis lyrata* using RNA-Seq
694   data. *PLoS One* 10:e0137391. 10.1371/journal.pone.0137391
695   Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen
696   H, Schork NJ, and Ecker JR. 2013. Patterns of population epigenomic diversity. *Nature*
697   495:193-198. 10.1038/nature11968
698   Shabek N, and Zheng N. 2014. Plant ubiquitin ligases as signaling hubs. *Nat Struct Mol Biol*
699   21:293-296. 10.1038/nsmb.2804
700   Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
701   large phylogenies. *Bioinformatics* 30:1312-1313. 10.1093/bioinformatics/btu033
702   Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA
703   polymorphism. *Genetics* 123:585-595.
704   Taly JF, Magis C, Bussotti G, Chang JM, Di Tommaso P, Erb I, Espinosa-Carrasco J, Kemena
705   C, and Notredame C. 2011. Using the T-Coffee package to build multiple sequence
706   alignments of protein, RNA, DNA sequences and 3D structures. *Nat Protoc* 6:1669-
707   1682. 10.1038/nprot.2011.393
708   Torrents D, Suyama M, Zdobnov E, and Bork P. 2003. A genome-wide survey of human
709   pseudogenes. *Genome Res* 13:2559-2567. 10.1101/gr.1455503
710   Vierstra RD. 2009. The ubiquitin-26S proteasome system at the nexus of plant biology. *Nat Rev*
711   *Mol Cell Biol* 10:385-397. 10.1038/nrm2688
712   Wang Z, Gerstein M, and Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics.
713   *Nat Rev Genet* 10:57-63. 10.1038/nrg2484
714   Xu G, Ma H, Nei M, and Kong H. 2009. Evolution of F-box genes in plants: different modes of
715   sequence divergence and their relationships with functional diversification. *Proc Natl*
716   *Acad Sci USA* 106:835-840. 10.1073/pnas.0812043106
717   Yang M, Hu Y, Lodhi M, McCombie WR, and Ma H. 1999. The *Arabidopsis SKP1-LIKE1* gene is
718   essential for male meiosis and may control homologue separation. *Proc Natl Acad Sci*
719   *USA* 96:11416-11421.
720   Yang X, Kalluri UC, Jawdy S, Gunter LE, Yin T, Tschaplinski TJ, Weston DJ, Ranjan P, and
721   Tuskan GA. 2008. The *F-box* gene family is expanded in herbaceous annual plants
722   relative to woody perennial plants. *Plant Physiol* 148:1189-1200.
723   10.1104/pp.108.121921
724   Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586-
725   1591. 10.1093/molbev/msm088

726   Yau R, and Rape M. 2016. The increasing complexity of the ubiquitin code. *Nat Cell Biol*
727   18:579-586. 10.1038/ncb3358
728   Zhao D, Ni W, Feng B, Han T, Petrasek MG, and Ma H. 2003. Members of the *Arabidopsis-*
729   *SKP1-like* gene family exhibit a variety of expression patterns and may play diverse
730   roles in Arabidopsis. *Plant Physiol* 133:203-217.
731   Zheng N, Schulman BA, Song L, Miller JJ, Jeffrey PD, Wang P, Chu C, Koepp DM, Elledge SJ,
732   Pagano M, Conaway RC, Conaway JW, Harper JW, and Pavletich NP. 2002. Structure
733   of the Cul1-Rbx1-Skp1-F box^{Skp2} SCF ubiquitin ligase complex. *Nature* 416:703-709.
734   10.1038/416703a
735

Maximum likelihood test of neutral evolution by comparing neutral evolutionary model (ML1: dN/dS=1) and non-neutral evolutionary model (ML2: free dN/dS value) of *ASK* genes.

1   **Table 1.** Maximum likelihood test of neutral evolution by comparing neutral evolutionary model
2   (ML1: dN/dS=1) and non-neutral evolutionary model (ML2: free dN/dS value) of *ASK* genes

Ath Skps	Aly Skps	dS	dN	dN/dS	*ln*ML1	*ln*ML2	2ΔlnML	Selection*
ASK2	AlySkp05	0.2	0.0	0.1	-799.9	-779.5	40.8	Non-neutral
ASK4	AlySkp04	0.3	0.0	0.1	-810.3	-792.4	35.8	Non-neutral
ASK3	AlySkp04	0.3	0.0	0.1	-793.3	-778.7	29.1	Non-neutral
ASK1	AlySkp01	0.2	0.0	0.1	-759.6	-745.3	28.6	Non-neutral
ASK18	AlySkp08	0.2	0.1	0.3	-916.5	-908.8	15.5	Non-neutral
ASK6	AlySkp02	0.3	0.1	0.3	-463.8	-460.0	7.5	Non-neutral
ASK11	AlySkp06	0.2	0.1	0.4	-776.3	-773.8	5.0	Non-neutral
ASK12	AlySkp06	0.2	0.1	0.5	-779.2	-777.2	4.0	Non-Neutral
ASK13	AlySkp13	0.2	0.1	0.6	-831.1	-829.8	2.6	Neutral
ASK14	AlySkp15	0.3	0.2	0.6	-836.6	-835.5	2.2	Neutral
ASK8	AlySkp09	0.2	0.3	1.6	-909.0	-908.0	2.0	Neutral
ASK10	AlySkp09	0.2	0.2	1.4	-880.8	-880.4	0.9	Neutral
ASK19	AlySkp03	0.2	0.2	0.8	-1110.3	-1109.9	0.8	Neutral
ASK9	AlySkp09	0.2	0.2	0.8	-892.6	-892.5	0.3	Neutral
ASK16	AlySkp12	0.1	0.1	0.9	-885.8	-885.8	0.1	Neutral
ASK17	AlySkp07	0.1	0.1	1.1	-779.6	-779.6	0.1	Neutral
ASK5	AlySkp10	0.1	0.1	0.9	-804.5	-804.4	0.1	Neutral
ASK7	AlySkp09	0.2	0.2	1.1	-740.1	-740.1	0.1	Neutral

3   *x^2 (p=0.05, df=1)* = 2.71 (Yang 2007). Abbreviations: *A. thaliana* (*Ath*) and *A. lyrata* (*Aly*).

A short evolutionary history of the Arabidopsis *Skp1* genes within the Arabidopsis genus.

A) Phylogenetic relationships of the *ASK* members. Intronic genes are highlighted in red. Scale bar, average substitutions per site. B) An improved phylogenetic analysis reveals one single origin of retroposed *Skp1* genes in the Arabidopsis genus. *a-e*, 5 ancient retroposed loci produced by the transcripts of the highly expressed ancestor locus of *ASK1*. 1-11, 11 clades shaded with dark and gray color showing clear orthologous relationships among the three Arabidopsis species. The tree was generated by a maximum likelihood method. Statistical significance equal to or greater than 50% of 1,000 bootstrap re-samplings is indicated in each node. S: segmental duplication; T: tandem duplication; R: retroposed duplication. The ID number from the original annotation is indicated if a *Skp1* gene is duplicated through tandem duplications or described in File S1. Intronic *Skp1* genes and scale bar are described as in A. Species abbreviations: *A. halleri*(*Aha*), *A. lyrata* (*Aly*), and *A. thaliana*(*Ath*). C) Birth-and-death history of *Skp1* genes in the Arabidopsis genus. + and – indicate gain and loss of *Skp1* members, respectively. Species abbreviations are as in B.

A

- 100 — ASK7
- 66 — ASK8
- 100 — ASK9
- ASK10
- 100 — ASK12
- ASK11
- ASK18
- ASK19
- 76 — ASK15
- 100 — 53 — ASK16
- ASK17
- 77 — ASK14
- ASK5
- 98 — ASK6
- ASK13
- 67 (R) — ASK1
- ASK4
- 100 — ASK3
- ASK2

0.2

B

Group 1 (grey):
- 100 — AlySkp14 (AL3G35990)
- 84 (T) — AlySkp16 (AL3G36000)
- 86 — AlySkp09 (AL3G36010)
- AhaSkp05
- 99 a — ASK8 (AT3G21830)
- 100 (T) — ASK7 (AT3G21840)
- 75 (T) — ASK9 (AT3G21850)
- 95 (T) — ASK10 (AT3G21860)

Group 2 (yellow):
- 100 — ASK12
- 62 b — ASK11
- AlySkp06

Group 3 (grey):
- 83 — AlySkp10 (AL5G42200)
- 97 — AhaSkp09
- 100 — ASK5 (AT3G60020)
- 99 — AhaSkp11
- 96 c — AlySkp02
- ASK6 (T)

Group 4 (yellow):
- 97 — AlySkp13 (AL5G42190)
- 100 — AhaSkp06
- ASK13 (AT3G60010)

Group 5 (grey):
- 96 — AlySkp07
- ASK17
- 78 — AhaSkp08 (25219s0006)
- 66 — AlySkp12 (AL5G12840)
- ASK16 (AT2G03190)

Group 6 (yellow):
- 94 — AhaSkp07
- 100 e — AlySkp08
- ASK18

Group 7 (grey):
- 54 — AlySkp03 (AL5G12820)
- 63 — ASK19 (AT2G03160)
- AhaSkp02
- ASK15

Group 8 (yellow):
- 100 — AlySkp15 (AL5G12830)
- 51 (T) — AhaSkp10 (25219s0005)
- ASK14 (AT2G03170)

76
80 (R) d (T)
100
100 (S)

Group 9 (yellow):
- 99 — AlySkp01
- 68 — AhaSkp01
- ASK1

Group 10 (grey):
- AhaSkp03
- 72 — AlySkp04
- 81 — 93 — ASK4
- ASK3
- 99 — AlySkp11
- AlySkp17

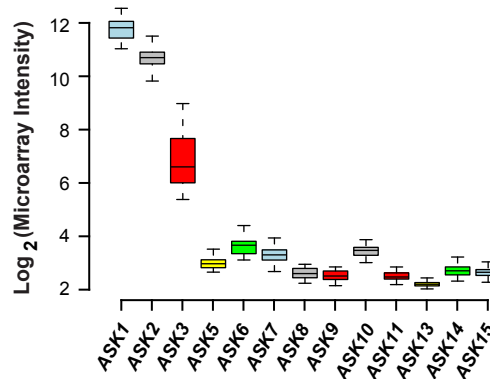
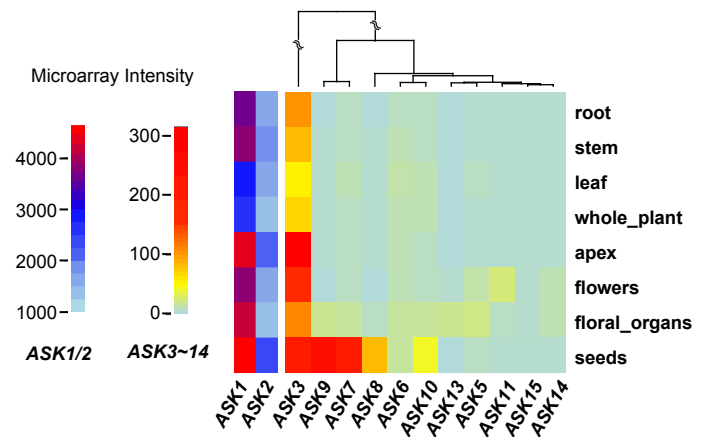
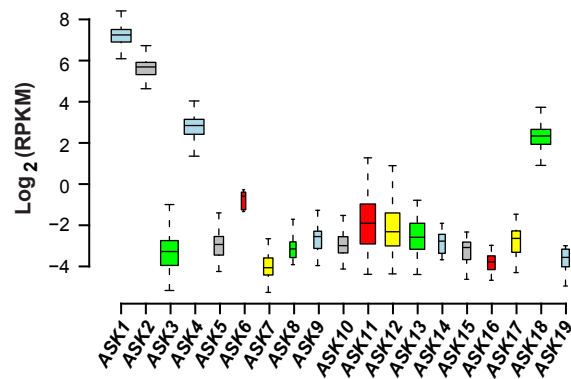
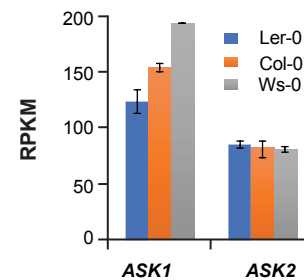
Group 11 (yellow):
- 97 — AlySkp05
- AhaSkp04
- 100 — ASK2

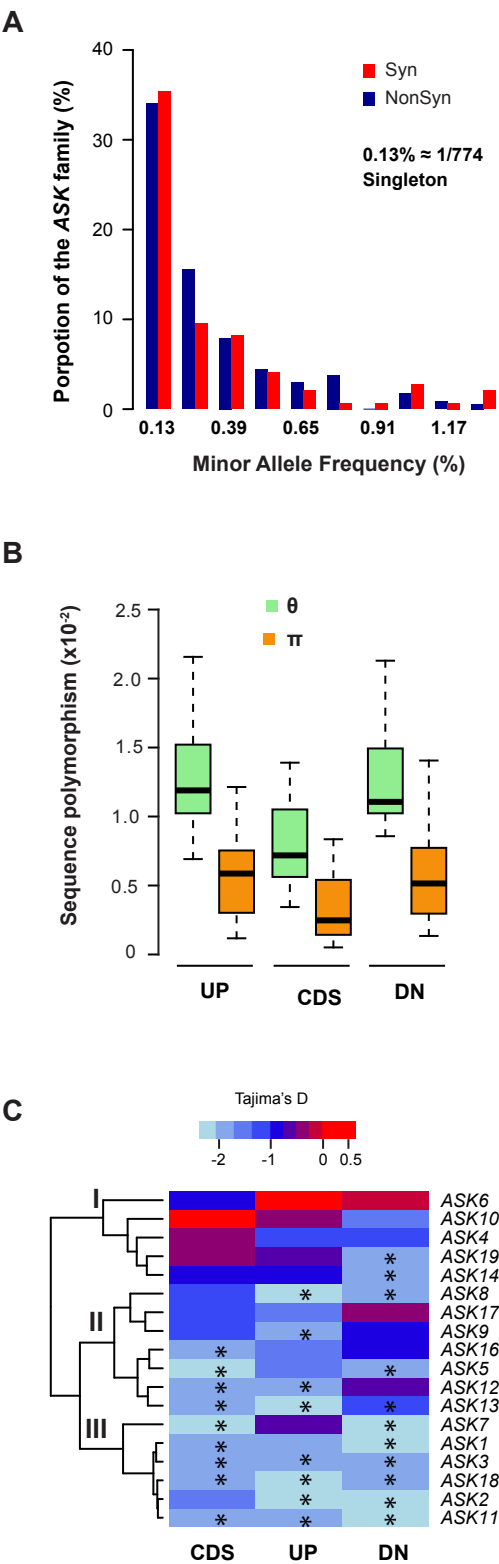
0.2

C

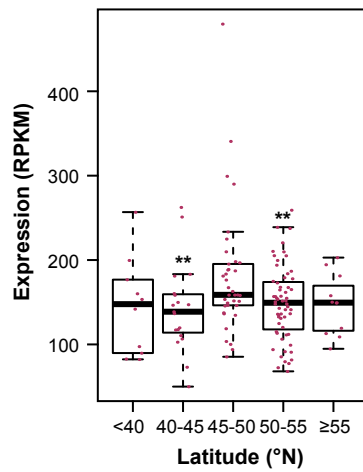
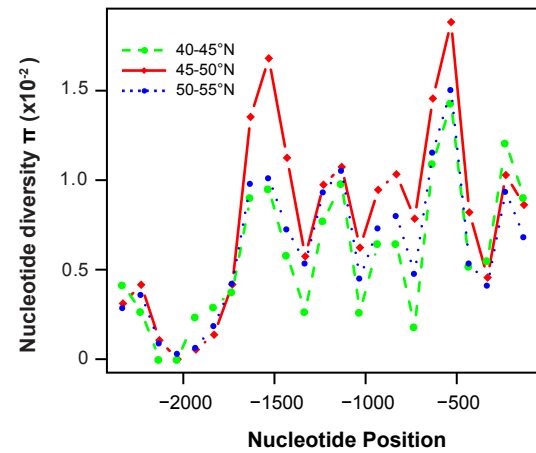
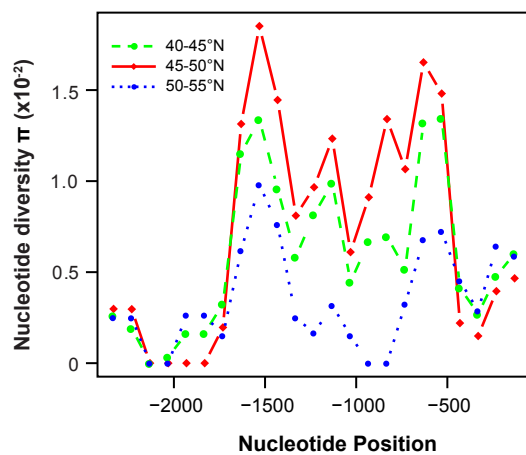
- +1 / -2 — (19)
- +0 / -7 — **Aha** (11)
- (18) — +1 / -2 — **Aly** (17)
- +5 / -5 — **Ath** (19)

Expression variation of *ASK* genes.

A) Variation of absolute expression throughout a developmental time course. B) A heatmap analysis of mean expression variation of *ASK* genes across 8 tissues/organs. C) Expression variation of *ASK* genes in leaves of 144 accessions. The width of each boxplot is proportional to the number of accessions having a non-zero expression data of the corresponding *ASK* gene. D) Expression comparison of *ASK1* and *ASK2* in Col-0, Ler-0, and Ws-0.

Sequence polymorphic comparison among different *ASK* genes.

A) Frequency distribution of rare *ASK* alleles with non-synonymous and synonymous mutations. B) Variation of sequence polymorphisms in three regions of an *ASK* gene. UP: 500 bp upstream of start codon, CDS: coding sequence, DN: 500 bp downstream of the stop codon. C) A heatmap representation of Tajima's D values demonstrating the differential evolutionary constraints of polymorphic mutations among *ASK* genes. Asterisks indicate a significantly low Tajima's D value that deviates from neutral mutations (*p*< 0.05).

Contribution of sequence polymorphism to gene expression.

A) Latitudinal variation of *ASK1* expression. Asterisks indicate that the mean expression of the indicated group is significantly lower than that of the group within latitudes 45-50°N (Wilcoxon rank-sum test, *p*< 0.01). B) A window-slide polymorphic comparison of a 2.5 kb region upstream of the transcriptional start site of *ASK1* among populations from 3 latitudinal regions as in A. C) A window-slide polymorphic comparison of a group of test populations from 3 latitudinal regions, performed as in B. D) A window-slide polymorphic comparison of three groups of populations with low, medium, and high expression levels of *ASK1* that was performed as in B.

A

B

C

D

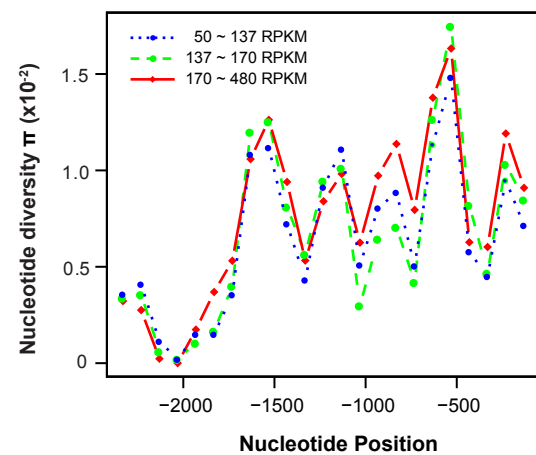



Figure 5

Quantitative interaction assay of ASK1 and ASK2 with 15 selected known F-box proteins.

The accession IDs of F-box proteins are listed in Table S2. A) Growth of yeast cells expressing each indicated pair of bait and prey proteins on SD-Leu-Trp-Ade-His+X-alpha-gal medium. B) Control growth of the corresponding yeast cells in A on SD-Leu-Trp medium. Yeast cells were grown on media at 30°C for an indicated time period and their growth was subsequently recorded by scanning the plates with a Canon 9000F Mark II scanner. B) Quantification of pairwise ASK1 and ASK2 interactions with 15 F-box proteins as shown in A. The interaction strength of each pair of bait and prey proteins shown in A was normalized by the control growth strength of the corresponding yeast cells in B. C) Beta-galactosidase activity assay. The mated yeast cells expressing the indicated pair of bait and prey proteins were grown on SD-Leu-Trp medium and used for the assay. The beta-galactosidase activities shown are mean values ±SD measured from two independent assays.