Molecular Biology and Evolution, 2025, **42**, 1–14
https://doi.org/10.1093/molbev/msae262
Advance access publication 24 December 2024
Discoveries

MBE OXFORD

A de novo Gene Promotes Seed Germination Under Drought Stress in Arabidopsis

Guang-Teng Jin [1,2,3,†] **Yong-Chao Xu** [1,2,†] **Xing-Hui Hou** [1,2] **Juan Jiang** [1,2,3] **Xin-Xin Li** [1,2,3] **Jia-Hui Xiao** [1,2,3] **Yu-Tao Bian** [1,2,3] **Yan-Bo Gong** [1,2,3] **Ming-Yu Wang** [4] **Zhi-Qin Zhang** [1,2,3] **Yong E. Zhang** [3,5] **Wang-Sheng Zhu** [4] **Yong-Xiu Liu** [2,3,6] **Ya-Long Guo** [1,2,3,*]

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
[2]China National Botanical Garden, Beijing 100093, China
[3]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[4]State Key Laboratory of Maize Bio-breeding/College of Plant Protection, China Agricultural University, Beijing 100193, China
[5]State Key Laboratory of Integrated Management of Pest Insects and Rodents and Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[6]Key Laboratory of Plant Molecular Physiology, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

[†]These authors contributed equally to this work.
***Corresponding author:** E-mail: yalong.guo@ibcas.ac.cn.
Associate editor: Emily Josephs

Abstract

The origin of genes from noncoding sequences is a long-term and fundamental biological question. However, how de novo genes originate and integrate into the existing pathways to regulate phenotypic variations is largely unknown. Here, we selected 7 genes from 782 de novo genes for functional exploration based on transcriptional and translational evidence. Subsequently, we revealed that *Sun Wu-Kong* (*SWK*), a de novo gene that originated from a noncoding sequence in *Arabidopsis thaliana*, plays a role in seed germination under osmotic stress. *SWK* is primarily expressed in dry seed, imbibing seed and silique. *SWK* can be fully translated into an 8 kDa protein, which is mainly located in the nucleus. Intriguingly, *SWK* was integrated into an extant pathway of hydrogen peroxide content (folate synthesis pathway) via the upstream gene *cytHPPK/DHPS*, an Arabidopsis-specific gene that originated from the duplication of *mitHPPK/DHPS*, and downstream gene *GSTF9*, to improve seed germination in osmotic stress. In addition, we demonstrated that the presence of *SWK* may be associated with drought tolerance in natural populations of Arabidopsis. Overall, our study highlights how a de novo gene originated and integrated into the existing pathways to regulate stress adaptation.

Keywords: adaptive evolution, *Arabidopsis thaliana*, de novo gene, regulatory networks

Introduction

De novo genes are new genes that originated from the noncoding sequence (Levine et al. 2006; Cardoso-Moreira and Long 2012; Carvunis et al. 2012), which is one of the most important processes to produce new genes. After being discovered in Drosophila, de novo genes have been found in multiple species (Levine et al. 2006; Guo et al. 2007; Guo 2013; Zhao et al. 2014; Ruiz-Orera et al. 2015; Li et al. 2016; Blevins et al. 2021). However, understanding the function of de novo genes and how they integrate into pre-existing molecular pathways is a major challenge (McLysaght and Hurst 2016).

The function of de novo genes is the first crucial question to be addressed. Previously reported de novo genes were mainly involved in reproduction, resistance, toxin response, and metabolic regulation (Singh and Syrkin Wurtele 2020). For instance, the first de novo gene with functional study, *Poldi*, is a positive regulator of sperm motility and testicular weight in the house mouse (Heinen et al. 2009). In plants, *QQS* was the first studied de novo gene, which plays a role in regulating both carbon and nitrogen allocation and pest resistance in Arabidopsis (Li et al. 2009, 2015; Qi et al. 2019). Until now, more than twenty de novo genes have been functionally studied in plants (Jiang et al. 2022). For example, *TaFROG* gene in wheat regulates resistance to Fusarium head blight disease (Perochon et al. 2015), and *Xio1* gene in rice contributes to immune responses to bacterial pathogens (Moon et al. 2022).

How de novo genes are integrated into pre-existing molecular regulatory networks represents another major challenge (McLysaght and Hurst 2016). This involves 2 questions: (i) What are the interacting proteins of the de novo proteins? (ii) What are their upstream and downstream regulatory genes? The first question has attracted considerable attention in previous studies, and interacting partners for several de novo proteins have been identified (Li et al. 2015; Ni et al. 2017; Perochon et al. 2019; Qi et al. 2023). In addition, it has been suggested that proteins of new duplicated genes and de novo genes tend to interact with each other (Capra et al. 2010). However, there are few relevant studies on the upstream and downstream regulatory genes of de novo genes, except that the upstream regulatory gene *Atss3* of *QQS* has been identified (Li et al. 2009). Therefore, it is crucial to understand how de novo genes integrate into existing regulatory pathways.

Received: July 3, 2024. **Revised:** October 29, 2024. **Accepted:** December 6, 2024

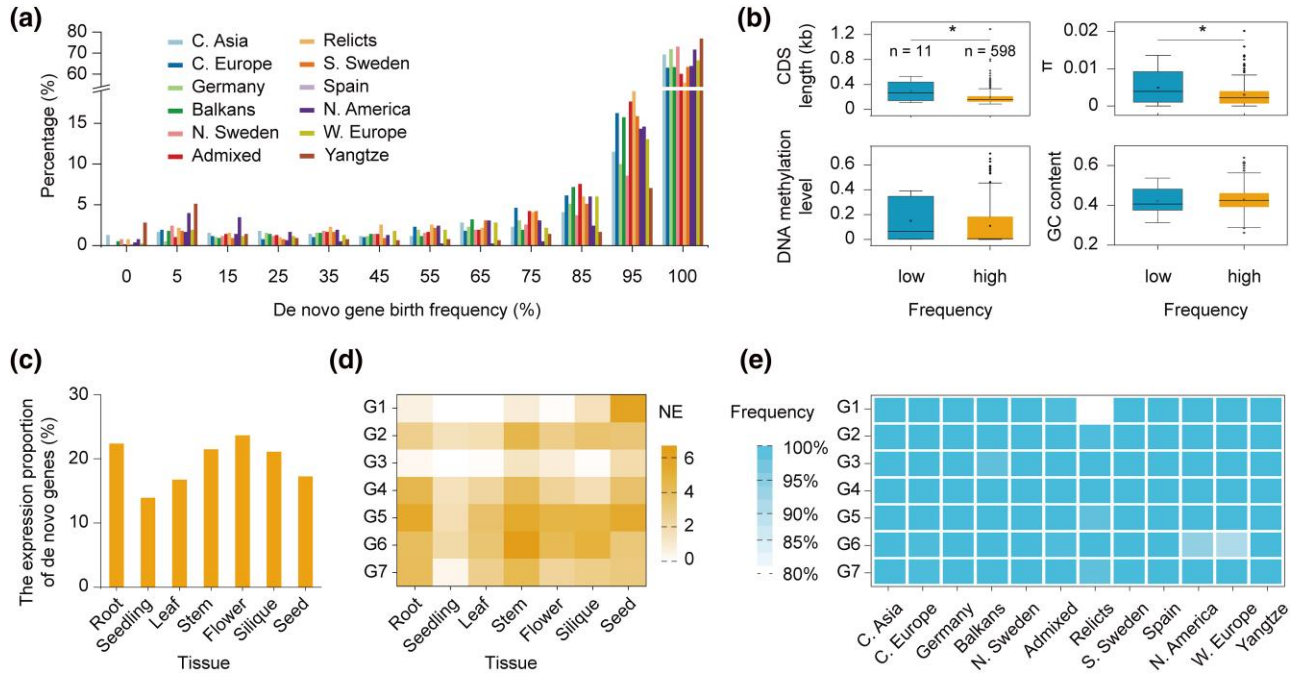

Fig. 1. Characteristics of de novo genes and their representatives. a) Frequency distribution of 782 de novo genes in 12 populations. 0: frequency = 0; 100: frequency = 100%; 5: 0 < frequency <10%; 15: 10% ≤ frequency < 20%, with other numbers following the same pattern. C. Asia, Central Asia; C. Europe, Central Europe; N. Sweden, North Sweden; S. Sweden, South Sweden; N. America, North America; W. Europe, Western Europe; Yangtze, Yangtze River. b) Comparison of gene features between low-frequency (frequency < 10%) and high-frequency (frequency ≥ 90%) de novo genes in 1,115 accessions. Nucleotide polymorphism (π), CDS length, GC (Guanine-Cytosine) content, and DNA methylation level were calculated based on the Col-0 reference. Wilcoxon rank-sum test was used for testing significance, *, $P < 0.05$. c) The percentage of 782 de novo genes expressed in 7 different tissues. Expression is defined as TPM > 1. d) The expression patterns in different tissues of 7 representative genes for functional exploration. G1, AT1G03106; G2, AT1G58150; G3, AT2G07000; G4, AT2G13550; G5, AT2G14460; G6, AT4G36515; G7, AT5G67245. NE, Normalized expression, $\log_2(\text{TPM} + 1)$. e) Frequency of the 7 representative de novo genes in 12 populations.

How de novo genes reach a high-frequency or even become fixed in natural populations, whether adaptive or neutrally, are a fundamental question. The "adaptation following neutrality" theory proposes that positive selection can drive new genes to rapidly evolve important functions (Wu and Zhang 2013). In Drosophila, it has been proposed that natural selection plays an important role in the fixation of de novo genes in natural populations (Zhao et al. 2014). In contrast, many de novo genes may be retained through neutral processes. De novo genes can evolve from "frozen accidents" that survived initial purging (Schmitz et al. 2018). In yeast, there is evidence of hundreds of translated open reading frames that do not show evolutionary conservation or selective constraint (Ruiz-Orera et al. 2018). In the house mouse, similarly, there is no signal of positive selection in the de novo gene *Gm13030*, indicating that a de novo gene can directly exert biological functions without undergoing adaptive mutation (Xie et al. 2019). More broadly, the retention of new mutations is driven by a combination of different factors, such as selection and genetic drift (Barrett and Schluter 2008). The fixation of a new mutation in a population depends on its fitness effects and the strength of selection (Orr 2005; Eyre-Walker and Keightley 2007). Thus, the question of whether functional de novo genes are subject to natural selection requires in-depth evidence.

Although it is widely acknowledged that de novo genes represent an important mechanism for the formation of new genes (Begun et al. 2006; Levine et al. 2006; Chen et al. 2007), the potential role of de novo genes in adaptation remains largely unknown. In Drosophila, a study shows that de novo genes are crucial for survival (Chen et al. 2010). Given the birth of de novo genes occurs at the population level, it is thus important to understand their importance at the population level (Schlötterer 2015). However, whether de novo genes can help natural populations adapt to changing environments is still largely unknown.

Here, we revealed the function of a de novo gene in Arabidopsis, showing how it integrated into pre-existing networks and demonstrating its significance in adaptive evolution. In particular, we demonstrated that this de novo gene, AT1G03106 (*SWK*), enhances *Arabidopsis thaliana* resistance to osmotic stress by regulating hydrogen peroxide signaling. Overall, this study reveals that how a de novo gene originated and integrated into the existing pathways to regulate phenotypic variation and affect adaptation.

Results

Sequence Characteristics of Representative de novo Genes

In a previous study, we identified 782 de novo genes in Arabidopsis (Col-0) (Li et al. 2016). Based on the published resequencing data of 1,115 natural accessions from 12 populations (Cao et al. 2011; The 1001 Genomes Consortium 2016; Durvasula et al. 2017; Zou et al. 2017; Jiang et al. 2024), between 74.8% and 86.2% of the 782 de novo genes were nearly fixed (frequency >90%) in at least one of the 12 populations (Fig. 1a). Among these populations, the North America population has the largest proportion (86.2%) of nearly fixed de novo genes (frequency >90%) (Fig. 1a). In contrast, Yangtze River basin population has the highest proportion of low-frequency (frequency <10%) de novo genes (5.1%) (Fig. 1a).

The comparison of gene features between low-frequency (frequency <10%) and high-frequency (frequency ≥90%) de novo genes showed that the coding sequence (CDS) length and nucleotide polymorphism (π) of low-frequency de novo genes were significantly higher than those of high-frequency de novo genes (Wilcoxon rank-sum test, $P < 0.05$, Fig. 1b). This indicates that low-frequency de novo genes are longer and more polymorphic. However, other gene features, such as gene length, GC (Guanine-Cytosine) content, DNA methylation level, the ratio of nonsynonymous to synonymous nucleotide diversity (π_n/π_s), codon usage bias, and the density of accessible chromatin regions, were not significantly different (Fig. 1b and supplementary fig. S1a, Supplementary Material online). Our results are consistent with reports in Drosophila indicating that the exon lengths of segregating de novo genes are significantly longer than those of fixed de novo genes (Zhao et al. 2014), supporting the preadaptation hypothesis of de novo gene origin (Wilson et al. 2017). According to RNA sequencing data from different tissues, de novo genes are mainly expressed in flower, root, stem, and silique, followed by the seed (Fig. 1c). This pattern differs slightly from previous studies in rice (callus, shoot, and flower) and mangrove (root, fruit, and flower) (Guo et al. 2007; Ma et al. 2021). Notably, the percentage of de novo genes expressed in the flowers, is exceptionally high, which is similar to the higher percentage of de novo genes expressed in animal testis (Levine et al. 2006; An et al. 2023).

According to the transcriptional and translational evidence in Col-0, these 782 de novo genes were divided into 3 groups (Li et al. 2016): 398 de novo genes without any evidence of transcription or translation (TN); 239 de novo genes with only transcriptional evidence (TC); 145 de novo genes with translation evidence (TL). Given that polyadenylation plays an important role in translation initiation (Wells et al. 1998), TL-type de novo genes could be classified into 2 types according to the presence or absence of polyadenylation: 80 TL type genes without poly(A) tails (TL^-); 65 TL type with poly(A) tails (TL^+) (Li et al. 2016). In total, we selected 6 TL^+ type and 1 TC type de novo gene that can be expressed (Transcripts Per Million (TPM) >1) in multiple tissues to perform in-depth study (supplementary fig. S1b, Supplementary Material online). In detail, the 7 genes are mainly expressed in root, stem, flower, silique, or seed (Fig. 1d). Among the 7 de novo genes, except for AT1G03106 (G1) in the Relicts population (frequency = 79.0%) and AT4G36515 (G6) in the Western Europe population (frequency = 89.6%), the other 5 genes were nearly fixed in each population (frequency >90%, Fig. 1e).

Functional Characterization of Representative de novo Genes

In order to explore the function of these 7 de novo genes, we obtained the T-DNA insertion mutants and constructed overexpression (OE) lines in the Col-0 background (supplementary fig. S1c and S1d, Supplementary Material online). We measured 8 phenotypes throughout the life cycle, including seed germination, rate of increase in rosette leaf number, rosette fresh weight, chlorophyll content, flowering time, silique length, plant height and number of secondary branches, using the mutant, Col-0 (WT), and OE lines (Fig. 2a). Furthermore, 5 phenotypes under stress conditions were assessed after treatment in Petri dishes: seed germination and root elongation under salt stress, seed germination and root elongation under osmotic stress, and root elongation under oxidative stress (Fig. 2a).

Among all the phenotypes examined for these 7 de novo genes, only seed germination under osmotic stress for the comparison of AT1G03106 mutant (G1) and its WT or OE lines showed a significant difference (two-sided Student's *t*-test, Benjamini–Hochberg correction, $P < 0.05$, Fig. 2b). Under osmotic stress (400 mM mannitol), the seed germination of the OE and WT plants was significantly higher than that of the mutants (two-sided Student's *t*-test, Benjamini–Hochberg correction, $P < 0.05$, Fig. 2c and 2d, and supplementary fig. S2a, Supplementary Material online). Hereafter, we referred to AT1G03106 as *Sun Wu-Kong* (*SWK*, the Monkey King born from a stone in "*Journey to the West*"). Given that ABA is a key hormone affecting germination, we estimated seed germination under exogenous ABA stress, and found that the seed germination of mutants was significantly higher than that of an OE line (two-sided Student's *t*-test, Benjamini–Hochberg correction, $P < 0.05$, supplementary fig. S2b, Supplementary Material online). The effect of the 2 different stresses on the phenotype suggested that *SWK* probably regulates seed germination under either osmotic stress or ABA stress.

To validate the function of *SWK*, besides the available T-DNA insertion mutant, we generated 3 *SWK* mutants using the CRISPR/Cas9 system in Col-0 background (supplementary fig. S2c and S2d, Supplementary Material online). The knockout mutants do not affect the protein structure of adjacent genes (supplementary fig. S2c, Supplementary Material online), and adjacent genes are not involved in regulating germination under osmotic stress according to previous reports. Based on the 3 different homozygous edited lines with predicted truncated proteins, we found that seed germination of the knockout mutants under osmotic stress was significantly lower than in the wild type, further confirming the function of *SWK* (Fig. 2e).

Expression Pattern and Subcellular Localization of *SWK*

To reveal the expression pattern of *SWK*, we estimated its expression level in 8 different tissues, including root, rosette leaf, stem, cauline leaf, flower, silique, dry seed, and imbibing seed. The results indicated that *SWK* is expressed in all these 8 tissues, including leaf tissue, but highly expressed in dry seed, imbibing seed, and silique (Fig. 2f), indicating that *SWK* likely plays a role in seed development processes.

To verify whether the de novo CDS of *SWK* can produce a complete protein, we constructed a SWK-GFP (GFP, Green fluorescent protein) vector by fusing the CDS of GFP to the C-terminus of SWK, which is predicted to produce a GFP-tagged SWK fusion protein when transformed into Col-0. Western blots indicated that the mass of the fusion protein was increased by 8 kDa (total 35 kDa) as compared to that of GFP (27 kDa), consistent with the predicted size of the SWK protein (Fig. 2g).

SWK is composed of an alpha helix and a disordered sequence based on protein structure modeling using AlphaFold2 (Fig. 2h), as expected for a younger protein. In addition, the GFP-tagged SWK overexpressing plants showed that SWK mainly localizes at the nucleus, but is also present in the cytoplasm (exclude chloroplasts) and the cell membrane (Fig. 2i and supplementary fig. S2e, Supplementary Material online). Localization of SWK in the nucleus is consistent with the characteristics of transcription factors (Puertollano et al. 2018),

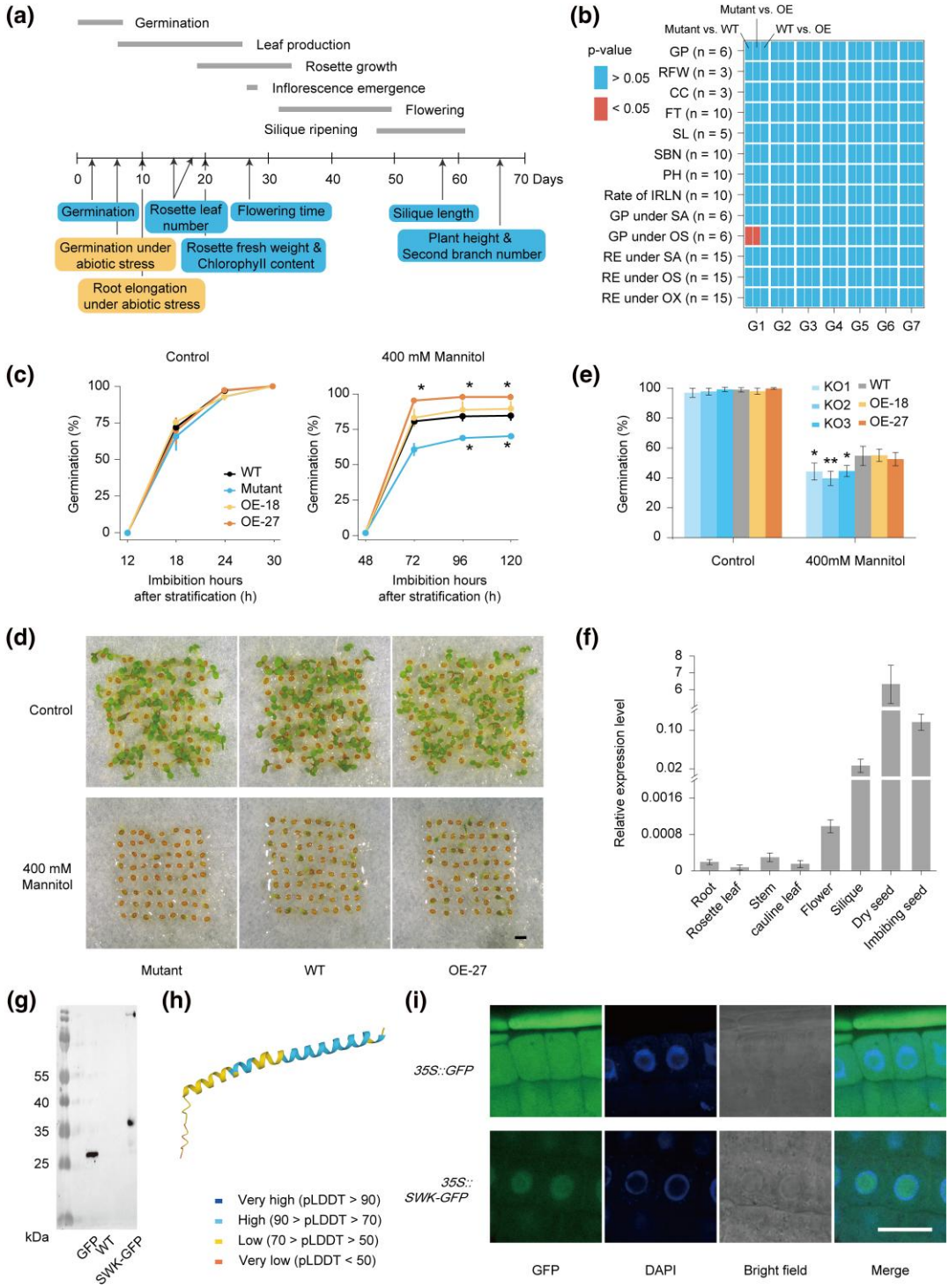

Fig. 2. Phenotyping of representative de novo genes. a) Traits measured in this study and timings. The numbers under the line represent days after sowing. The boxes represent phenotypes throughout the life cycle (blue) and under abiotic stress (yellow). The gray horizontal line represents the developmental period marked by the same horizontal line. b) Pairwise phenotypic comparison of mutant and Col-0 (WT) or OE. GP, seed germination percentage; RFW, Rosette fresh weight; CC, chlorophyll content; FT, flowering time; SL, silique length; SBN, number of secondary branches; PH, plant height; Rate of IRLN, rate of increase in rosette leaf number; GP under SA, seed germination percentage under salt stress; GP under OS, seed germination percentage under osmotic stress; RE under SA, root elongation under salt stress; RE under OS, root elongation under osmotic stress; RE under OX, root elongation under oxidative stress. *n*, the number of plants used for phenotyping. c) Seed germination percentage under mannitol stress. Percentages of seed germination are means (±SD) based on the seeds from 6 individual plants. Either mutant or OE was compared to WT at the same time point. d) Images showing seed germination of *SWK* mutants, WT, and OE-27 under control and 400 mM mannitol stress after 4 d of growth. Scale bar, 1 mm. e) Seed germination percentage of knockout mutants (KOs). Percentages of seed germination are means (±SD) based on the seeds from 6 individual plants. The statistical time of the control group was after 1 d of growth, while the statistical time of the mannitol stress treatment group was after 5 d of growth. All KOs were compared to WT. f) RT-qPCR analysis of *SWK* expression in 8 tissues of WT. Expression values are means (±SD) based on 3 replicates per line. g) Western blotting showing the expression of SWK constructs in seed. Lanes from left to right are as follows: markers, seeds transfected with GFP, WT, SWK-GFP fusion protein of OE (SWK-GFP). h) SWK protein structure predicted by AlphaFold2. pLDDT, predicted local distance difference test. i) Subcellular localization of SWK in the root tip. The images were obtained from the GFP channel, DAPI channel, Bright channel, and a merged image of the 3 channels. Scale bar, 20 μm. Two-sided Student's *t*-test was used for significance test. *, $P < 0.05$; **, $P < 0.01$.

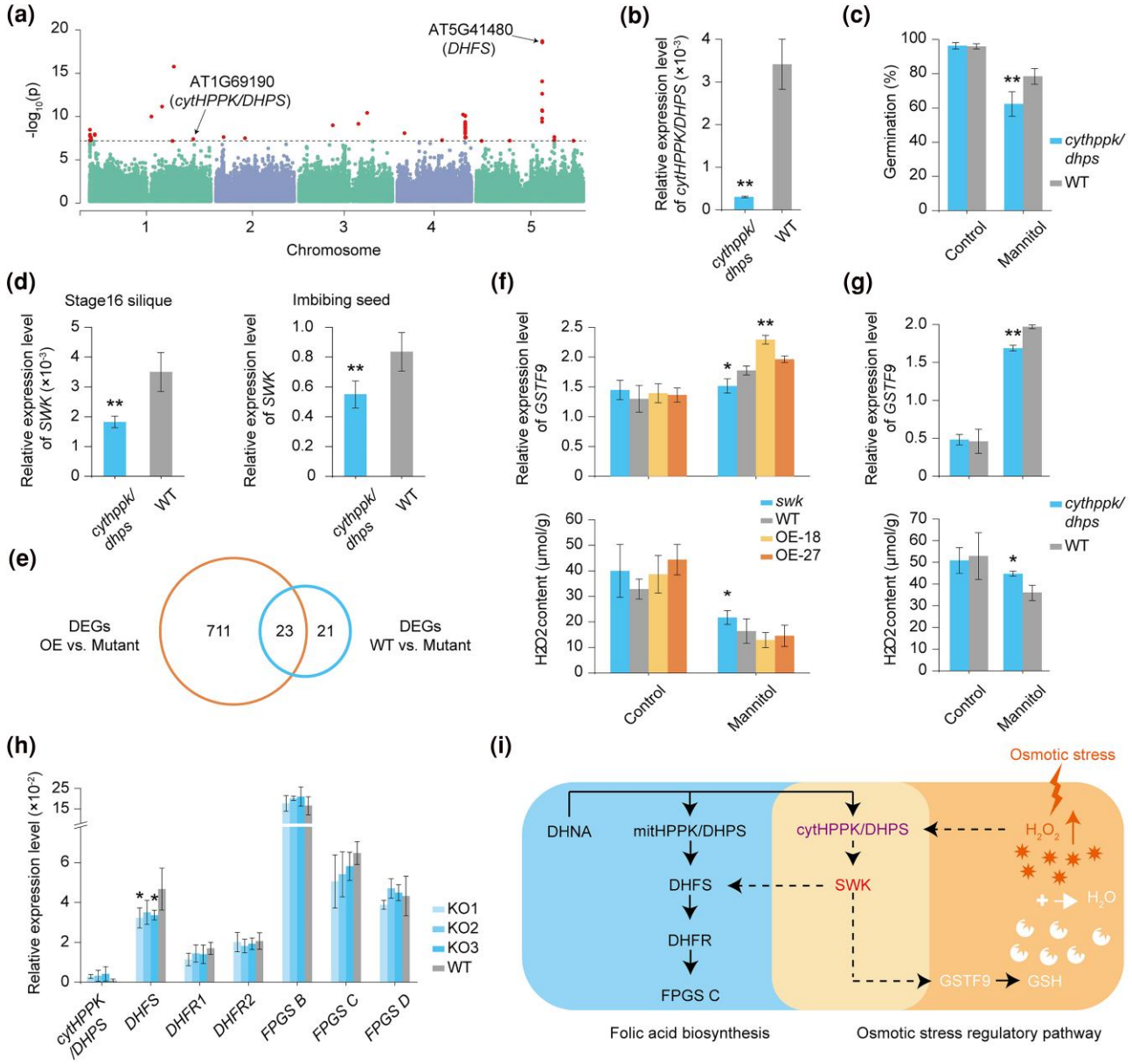

Fig. 3. Upstream and downstream regulatory genes of *SWK*. a) eGWAS analysis based on *SWK* expression levels and whole-genome variation in 414 *A. thaliana* accessions. The dashed line represents the threshold line [$-\log_{10}(0.05/1{,}133{,}651)$]. The 2 genes marked are the genes in the folate biosynthesis pathway. b) Expression difference between the *cytHPPK/DHPS* mutant and WT. Values are means (±SD) based on 3 replicates per line. c) Differences in seed germination percentages between the *cytHPPK/DHPS* mutants and WT under mannitol stress. Percentages of seed germination are means (±SD) based on the seeds from 6 individual plants. d) *SWK* expression level in the *cytHPPK/DHPS* mutant and WT. The imbibing seeds are under the stress of 400 mM mannitol. Values are means (±SD) based on 3 replicates per line. e) Number of DEGs in the comparison of the *SWK* mutant with WT or OE by RNA-seq analysis. f) Expression levels of *GSTF9* and H_2O_2 content differences in *SWK* mutant, WT and OE under 400 mM mannitol stress. Values are means (±SD) based on 3 replicates per line. Either mutant or OE was compared to WT. g) Expression levels of *GSTF9* and H_2O_2 content differences in the *cytHPPK/DHPS* mutant and WT under mannitol stress. Values are means (±SD) based on 2 replicates per line. h) Comparison of expression levels of folate biosynthesis genes in the *SWK*-knockout mutants (KOs), WT, and OE lines. Values are means (±SD) based on three replicates per line. All KOs were compared to WT. i) Schematic diagram of the 2 pathways involving *cytHPPK/DHPS* and *SWK*. The black solid arrows indicate direct regulation; the black dotted arrows suggest regulation, but it is not confirmed whether there are intermediate genes; the white arrow and plus sign denote chemical reaction; the red lightning symbolizes osmotic stress stimulation, and the arrow indicates an increase in the signaling factor. Two-sided Student's *t*-test was used for significance test. *, $P<0.05$; **, $P<0.01$.

while the localization of SWK on the cell membrane supports the tendency of new genes to encode putative membrane domains (Cui et al. 2015; Vakirlis et al. 2020).

Identification of Upstream Genes of *SWK*

In order to find the upstream regulatory genes of *SWK*, we used RNA-seq data of leaf tissues from 414 (including Col-0) natural accessions (Kawakatsu et al. 2016) to perform a genome-wide association analysis for the trait of *SWK* expression level (eGWAS, Fig. 3a). To enhance the reliability of the results based on SNPs, we used 2 different software programs based on both SNPs and INDELs at 2 different minor allele frequency (MAF) thresholds (1% and 5%) (supplementary fig. S3, Supplementary Material online), and found that the results are consistent using different genotype

data sets, programs or MAF thresholds. There are 66 significant SNPs in the eGWAS for *SWK*, which involve 61 genes (supplementary tables S1 and S2, Supplementary Material online). In contrast, for the other 6 de novo genes studied, the significant signals in eGWAS were mainly located near the genes themselves, except for the gene AT2G07000 (supplementary table S1, Supplementary Material online, supplementary fig. S3e, Supplementary Material online). Taken together, these results suggested that *SWK* may have built up interactions with other genes, while the other 6 de novo genes not. This is consistent with the fact that the other 6 genes do not show any phenotypic effect on the traits measured in this study.

Among the 61 candidate genes associated with *SWK* expression variation, it has been reported that AT1G69190 (*cytHPPK/DHPS*) could increase the seed germination of Arabidopsis under mannitol stress (Storozhenko et al. 2007). Here in this study, the mutant analysis of *cytHPPK/DHPS* confirmed that this gene, indeed, could increase seed germination under mannitol stress (Fig. 3b and c and supplementary fig. S4a, Supplementary Material online). In addition, the expression level of *SWK* is significantly reduced in the *cytHPPK/DHPS* mutant background, which implied that *cytHPPK/DHPS* is an upstream regulator of *SWK* (Fig. 3d and supplementary fig. S4b, Supplementary Material online). In contrast, *DHFS*, which displays the strongest eGWAS signal, does not alter *SWK* expression following *DHFS* knockdown (Fig. 3a and supplementary fig. S4c, S4d, and S4e, Supplementary Material online), suggesting that it is not likely an upstream regulatory gene of *SWK*.

In addition to eGWAS, the RNA-seq database provides another way to identify the upstream regulatory genes of a target gene (http://ipf.sustech.edu.cn/pub/athrna/). Based on the RNA-seq data of different mutants in the database, we screened genes that could change the expression of *SWK* by more than 2-fold in knockout mutants, which are potential candidate regulatory genes upstream of *SWK*.

In total, *SWK* expression was significantly changed in mutants of 71 genes (supplementary table S3, Supplementary Material online). Four of them are involved in regulating the seed germination under osmotic stress, and 22 genes are involved in regulating the seed germination under other conditions. The KEGG pathway enrichment results showed that these 71 genes were significantly enriched in 4 KEGG pathways ($P < 0.05$), including the circadian rhythm-plant, plant hormone signal transduction, lysine degradation, phosphonate and phosphinate metabolism (supplementary fig. S4f, Supplementary Material online, supplementary table S4, Supplementary Material online).

Identification of Downstream Genes of *SWK*

To identify the downstream genes of *SWK*, we extracted RNA from imbibing seeds of the mutant, Col-0 and OE-27 under both control and osmotic stress treatments after 12 h of lighting for RNA-seq. Given the phenotype of the *swk* mutant is significantly different from that of the WT and OE, we took the shared differentially expressed genes (DEGs) between the mutant and the WT, and between the mutant and the OE under stress treatment, as candidate downstream genes (Fig. 3e and supplementary table S5, Supplementary Material online). There are 23 candidate *SWK* downstream genes, of which 2 are involved in the regulation of osmotic stress responses, 2 are involved in the regulation of germination, and 4 genes are involved in the regulation of other abiotic stress responses (supplementary table S5, Supplementary Material online). The KEGG pathway enrichment results showed that these 23 genes were significantly enriched in 6 KEGG pathways ($P < 0.05$), including the glucosinolate biosynthesis, tryptophan metabolism, 2-oxocarboxylic acid metabolism, metabolic pathways, biosynthesis of secondary metabolites, and glutathione metabolism (supplementary fig. S4g, Supplementary Material online, supplementary table S6, Supplementary Material online).

Of these 23 genes, we focused on 2 candidate genes, *GSTF9* (Nguyen et al. 2020) and *GSTF10* (Ryu et al. 2009), which are known to be involved in the glutathione metabolic pathway and the osmotic stress response. To confirm whether these 2 candidate genes were downstream genes of *SWK*, we estimated their expression levels in the *SWK* mutant, WT and OE. The results suggest that OE showed higher *GSTF9* expression than WT, and WT showed higher expression than mutant under stress conditions (two-sided Student's *t*-test, Benjamini–Hochberg correction, $P < 0.05$), but *GSTF10* expression did not differ significantly under stress conditions (Fig. 3f and supplementary fig. S4h, Supplementary Material online). Given that *GSTF9* increases plant resistance to mannitol stress by degrading hydrogen peroxide (Nguyen et al. 2020), we measured the hydrogen peroxide content of *SWK* mutant, WT and OE under mannitol stress and found that the mutant has significantly higher levels than either WT or OE (two-sided Student's *t*-test, Benjamini–Hochberg correction, $P < 0.05$, Fig. 3f).

In addition, given that *cytHPPK/DHPS* regulates *SWK* expression, we detected a significant difference in the expression of *SWK* in the *cytHPPK/DHPS* mutant. In *cytHPPK/DHPS* mutant background, *GSTF9* is expressed in a similar pattern to *SWK*, suggesting that *cytHPPK/DHPS* acts as an upstream regulatory gene of *GSTF9* (Fig. 3g). *cytHPPK/DHPS* is an Arabidopsis-specific gene that was generated by the duplication of the folic acid biosynthetic pathway gene *mitHPPK/DHPS* (Storozhenko et al. 2007; Gorelova et al. 2019). Therefore, we estimated the expression level of these genes in the folate biosynthesis pathway in both the *SWK*-knockout and WT backgrounds. Intriguingly, the expression of *DHFS* in the *SWK*-knockout lines was significantly lower than that in WT, indicating that *SWK* is involved in the folate biosynthesis pathway and acts as the upstream regulatory gene of *DHFS* (Fig. 3h). The expression of genes downstream of *DHFS* did not change, possibly due to the strong feedback inhibition by downstream enzymes (Mouillon et al. 2002; Hanson and Gregory 2011). However, the *cytHPPK/DHPS* mutant can affect the folate content (Storozhenko et al. 2007). It appears that *SWK* is not the only downstream gene of *cytHPPK/DHPS* involved in folate biosynthesis. Taken together, these functional analyses illustrate an instance where a de novo gene and a duplication-generated new gene collaborate to integrate into a regulatory network in plants (Fig. 3i). Subsequently, we found that none of the 3 upstream and downstream proteins directly interact with SWK through yeast two-hybrid experiments (supplementary fig. S4i, Supplementary Material online). Overall, these findings highlight the complex interactions between de novo and duplication-generated genes, demonstrating their collaboration during the integration process into plant regulatory networks despite the lack of direct protein-protein interactions.

The Origin and Evolution of *SWK*

To trace the evolutionary history of *SWK*, we performed rigorous blast searches against a wide array of genomic databases (Phytozome 13, MycoCosm, PhycoCosm, IMG/M, and Ensembl 109). The homologous sequence of the *SWK* ORF was only found in 4 species of Brassicaceae family, including Arabidopsis, *Arabidopsis lyrata*, *Arabidopsis halleri*, and *Rorippa islandica* (supplementary fig. S5a, Supplementary Material online). Based on orthologous sequences of *SWK* from Arabidopsis, *A. lyrata*, and *A. halleri*, we used FastML (Ashkenazy et al. 2012) to infer the common ancestral sequence of *Arabidopsis* genus (supplementary fig. S5b, Supplementary Material online). Then, we inferred the evolutionary process of *SWK*, based on the alignment of *SWK* from 3 species of *Arabidopsis* genus, the common ancestral sequence of *Arabidopsis* genus, and *R. islandica*, with *R. islandica* as outgroup. We identified 2 stop codons that were lost during the transition from *R. islandica* to the common ancestor of the *Arabidopsis* genus, using *R. islandica* as outgroup (supplementary fig. S5c, Supplementary Material online). The loss of 2 stop codons due to base substitutions extended the length of the ORF (Fig. 4a and b and supplementary fig. S5b, Supplementary Material online). Subsequently, 19 nonsynonymous and 2 synonymous substitutions occurred in Arabidopsis, giving rise to *SWK* (Fig. 4a and b and supplementary fig. S5b, Supplementary Material online). However, the formation of new premature stop codons in *A. lyrata*, and *A. halleri* resulted in a truncated ORF (Fig. 4a and b and supplementary fig. S5b, Supplementary Material online).

In order to further reveal the evolutionary history of *SWK*, we extracted the orthologous sequences flanking *SWK* in the representative species of Brassicaceae family with a genome assembly. The 5 genes, located either upstream or downstream of *SWK*, are largely syntenic in these species, but there is no homolog sequence of *SWK* at the syntenic region, except in Arabidopsis (supplementary fig. S5d, Supplementary Material online). Multiple Helitron and MuDR type transposable elements are found within 2 kb upstream and downstream of the homolog in Arabidopsis (supplementary fig. S5e, Supplementary Material online). Except for Helitron, transposable elements could form a target site duplication (TSD) of generally 4 to 10 bp on both sides of the insertion site (Wicker et al. 2007; Tan et al. 2016). Here, we found remnants of TSD (CATGGC) at 167 and 232 bp upstream of the *SWK* start codon in Arabidopsis, which can also be observed in closely related species, *A. lyrata* and *A. halleri* (supplementary fig. S5f, Supplementary Material online). These findings suggest that copy number changes of *SWK* homologs may have been caused by ancient transposition events. Based on the phylogenetic tree of Brassicaceae (Hendriks et al. 2023), Arabidopsis is closely related with other species (e.g. *Capsella rubella*, *Malcolmia maritima*, *Nevada holmgrenii*, and *Nerisyrenia camporum*) that lack homologous *SWK* sequences or remnants, than that of *R. islandica*. Given the complete absence of *SWK* in these species, horizontal gene transfer (HGT) between *R. islandica* and Arabidopsis is more likely the explanation according to the principle of parsimony. In addition, previous studies have documented HGT between distantly related species, such as between tomato and Arabidopsis (Cheng et al. 2009).

Based on sequence analysis of the *SWK* CDS in different populations using published genomic data of 1,115 natural accessions of 12 different populations (Cao et al. 2011; The 1001 Genomes Consortium 2016; Durvasula et al. 2017; Zou et al. 2017; Jiang et al. 2024), we found that *SWK* was present in nearly all of the natural accessions except for 8 accessions located in the Mediterranean region (Fig. 4c). Every *SWK*-absent natural accession contains only 1 or 2 variants (SNPs/Indels) compared to Col-0 (supplementary fig. S6, Supplementary Material online). These mutations either disrupt stop codons, introduce premature stop codons, or lead to the loss of start codon, while the rest of the sequence remains largely conserved. Based on the limited number of variant sites within accessions and the small number of variant accessions (8/1115), we speculate that these accessions with nonfunctional allele of *SWK* more probably represent a subsequent loss of *SWK* following its formation.

Given the African population is one of the oldest natural populations and represents the early polymorphism of Arabidopsis (Durvasula et al. 2017), we chose the African population to explore the potential role of *SWK* in adaptation. There are 60 natural accessions in the African population, 55 of which have formed *SWK*, and the other 5 accessions not (Fig. 4c and supplementary fig. S6, Supplementary Material online). Comparison of the 19 environmental factors between the habitat of *SWK*-present accessions and *SWK*-absent accessions in these 60 accessions from Africa (supplementary fig. S7a, Supplementary Material online, supplementary tables S7 and S8, Supplementary Material online) suggested that the precipitation of the driest quarter (bio17) in the *SWK*-present accessions was significantly lower than those in the *SWK*-absent accessions (two-sided Student's *t*-test, $P < 0.001$, supplementary fig. S7b, Supplementary Material online). To avoid false positives caused by sample number, we randomly selected bio17 from 5 samples, repeated 1,000 times from 55 *SWK*-present accessions, and compared bio17 with 5 SWK-absent accessions; the results suggested that the bio17 values were still significantly different from those of the *SWK*-absent accessions ($P < 0.05$, Fig. 4d). These results align with the functional characteristics of *SWK* in enhancing the drought stress resistance of Arabidopsis induced by mannitol. Additionally, we performed Redundancy Analysis (RDA) to investigate the association between genomic variation in natural accessions from African populations and environmental factors of their habitats (Capblancq and Forester 2021). However, the results indicated that *SWK* is not included in the candidate adaptive outliers (supplementary fig. S7c, Supplementary Material online), most probably due to the limited number of absent SWK accessions. In summary, we hypothesize that the emergence of *SWK* may have enhanced Arabidopsis's adaptation to drought stress, although further evidence is required to support this conclusion.

To infer whether *SWK* is under positive selection in the natural populations, we performed a selection scan for each natural population, and found a total of 171 (21.87%) de novo genes are located at the selected regions in at least one population (Fig. 4e). Among them, the African population has the fewest de novo genes under selection (13, 1.66%), the northern Swedish population has the most de novo genes under selection (52, 6.65%) (Fig. 4e). The proportion of selected genes unique to each population ranges from 0 (African and Western European populations) to 50% (Yangtze River population) (Fig. 4e). Selection analysis based on the sequence polymorphic level suggested that *SWK* is not under positive

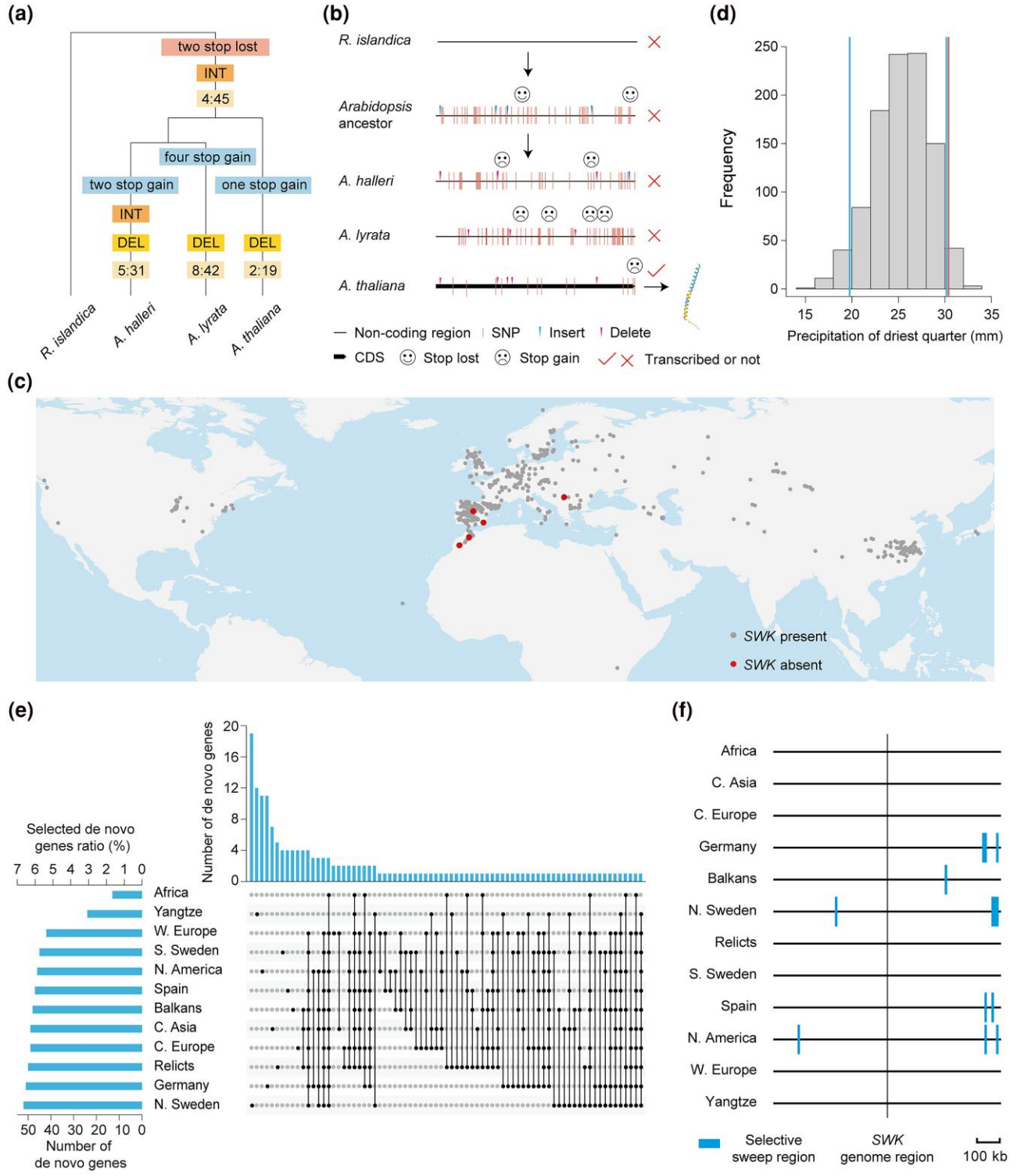

Fig. 4. Evolutionary analysis of *SWK*. a) The sequences with the highest similarity to the CDS of *SWK* in *A. lyrata*, *A. halleri*, and *R. islandica* were compared to analyze the evolutionary history of *SWK*. Numbers refer to synonymous: nonsynonymous substitutions. INT, insertion; DEL, deletion. b) The evolutionary history of *SWK*, with types of sequence variations are annotated according to the alignment results in supplementary fig. S5b, Supplementary Material online. c) Distribution of *SWK* in natural accessions worldwide. d) A 1,000-permutation test was used to generate the expected distribution of the difference in bio17 (precipitation of the driest quarter) between *SWK*-present and *SWK*-absent accessions in Africa. The area between the 2 blue vertical lines represents the 95% confidence interval. The red vertical line represents the bio17 average of 5 *SWK*-absent accessions. e) This panel shows the number of selected de novo genes in different populations and the number of selected de novo genes shared by different populations. f) Selective sweep regions of *SWK* and its upstream and downstream regions (500 kb each) in 12 natural populations based on linkage disequilibrium.

selection in any population even global population (Fig. 4f and supplementary figs. S8 and S9, Supplementary Material online). Taken together, our results support that 21.87% de novo genes are under natural selection either directly or indirectly by hitchhiking effect in natural populations and *SWK* is not under positive selection.

Discussion

De novo genes are important fuel for evolution (Levine et al. 2006; Cardoso-Moreira and Long 2012; Carvunis et al. 2012). Characterizing their function is the most important aspect to understand the origin of de novo genes. The functions of de novo genes that have been reported are mainly concentrated in 4 biological processes: reproduction, stress resistance, toxicity, and metabolic regulation (Singh and Syrkin Wurtele 2020). Among the 782 de novo genes in Arabidopsis (Li et al. 2016), only *QQS* has been functionally characterized (Li et al. 2009, 2015; Jiang et al. 2022). A previous study has shown that approximately half of the randomly selected gene mutants could affect Arabidopsis flowering time (Chong and Stinchcombe 2019). In contrast, here in our study, only 1 out of 7 de novo gene mutants showed a phenotypic effect under osmotic stress. This emphasizes that phenotypic effects of de novo genes only show up in very few specific develop stages or environmental conditions.

Here, we studied the function of *SWK* and revealed its role in response to osmotic stress and folate biosynthesis, which is consistent with the evolutionary characteristics of de novo genes that initially perform only regulatory functions (Capra et al. 2010; Zhang et al. 2015). The 71 genes that significantly altered *SWK* expression are significantly enriched in pathways of circadian rhythm-plant, plant hormone signal transduction, lysine degradation, phosphonate, and phosphinate metabolism based on DEGs between the wild type and mutant of *SWK* (supplementary fig. S4f, Supplementary Material online, supplementary table S4, Supplementary Material online), which implies that *SWK* responds to environmental, developmental signals.

De novo genes are one of the important sources of new genes besides gene duplication (Levine et al. 2006; Carvunis et al. 2012). The most intriguing question is how a de novo gene is maintained and integrates into existing regulatory pathways. For those new genes maintained in populations, certain biological or evolutionary functions are expected (He and Zhang 2005; Long et al. 2013; Kuzmin et al. 2022). It has been demonstrated that duplicate genes are easier to be integrated into gene networks than de novo genes of the same age, but they do not gain function as quickly as de novo genes (Capra et al. 2010). *cytHPPK/DHPS* is a young duplicate gene unique to Arabidopsis (Storozhenko et al. 2007). Our findings illustrate the phenomenon where a de novo gene could collaborate with a recently duplicated gene to integrate into 2 distinct pathways, the osmotic stress regulatory pathway and the folate biosynthesis pathway. In particular, our study illustrates that the origins of new genes through duplication and de novo origin are not independent events, but rather could be a mutually reinforcing process of genetic innovation. This study provides a new perspective on adding a de novo gene to existing gene regulatory networks.

In this study, we found that *SWK* can enhance the resistance of Arabidopsis to osmotic stress caused by mannitol, which is the main way to simulate drought stress (Zhu 2002; Verslues et al. 2006; Ozturk et al. 2021). Therefore, when we observe that the precipitation in the driest quarter in the distribution area of the *SWK*-present natural accessions is significantly lower than that of the *SWK*-absent natural accessions in the African population, we speculated that a de novo gene may improve the adaptability of Arabidopsis natural populations to drought environments. However, we did not detect any signal of natural selection for *SWK*, even though 21.87% of de novo genes were under natural selection in either one or a few populations. It is possible that *SWK* has been nearly fixed in globally distributed populations, and the signal of selection could not be detected any more (McLysaght and Hurst 2016). In addition, it is possible that it fixed directly after emergence during the speciation process without going through natural selection. However, in the Cape Verde Islands population, *SWK* was lost in 99.1% (332/335) accessions (supplementary fig. S10, Supplementary Material online).

Apparently, the exact mechanism of *SWK* fixation remains unclear, and further evidence is needed to better understand the evolutionary dynamics of de novo genes. In addition, the functional analysis of *SWK* in-depth will be insightful for revealing its function, and also will provide the broader implications of the findings for plant breeding, particularly in the context of climate change and the need for drought-resistant crops. Overall, this study highlights how a de novo gene originated and integrated into the existing pathways to regulate phenotypic variation and affect adaptation to environmental challenges at the population level.

Materials and Methods

Plant Material and Growth Condition

All Arabidopsis plants used in this study were Col-0 ecotype. The following mutants were used, including Salk_145188 (AT1G03106), Salk_037192C (AT1G58150), Salk_118656C (AT2G07000), Salk_024197C (AT2G13550), Salk_010561C (AT2G14460), Salk_089713C (AT4G36515), Salk_131289C (AT5G67245), Salk_093782 (AT1G69190), and Salk_076 763C (AT5G41480). These T-DNA insertion mutants were obtained from the Arabidopsis Biological Resource Center (ABRC, http://abrc.osu.edu/).

Seeds were stratified at 4 °C for 3 d in 0.1% agar after surface-sterilized by 75% ethanol, and sowed into the soil. Plants were grown at 20 °C under long-day conditions (16-h day/8-h night).

Characterization and Frequency Identification of de novo Genes

The presence of de novo genes in natural accessions was determined by the absence of loss-of-function mutations in these genes (Xu et al. 2019; Zhang et al. 2019a). Gene length, exon number, CDS length and GC content are calculated based on Col-0 (TAIR10); meanwhile, the number of transcripts, nucleotide polymorphism (π), DNA methylation level, the ratio of nonsynonymous to synonymous nucleotide diversity (π_n/π_s), codon usage bias, and the density of accessible chromatin regions are downloaded from previous studies (Hamala and Tiffin 2020).

Vector Construction and Plant Transformation

In order to get the transgenic overexpressing plants, the full length coding sequence without stop codon of 7 de novo genes were cloned into the modified pCAMBIA 2300 vector, with a GFP in the C terminal, respectively. To generate *SWK*-knockout mutants, 4 guide RNAs targeting the exon and the 3′UTR of *SWK* (supplementary fig. S2c, Supplementary Material online) were designed and constructed into *pYLCRISPR–Cas9Pubi-H* according to the user's manual (Ma et al. 2015). The constructs were then transformed into *Agrobacterium tumefaciens* GV3101 and

subsequently introduced into Col-0 plants by using the floral dip method (Clough and Bent 1998). The primers used for plasmid construction are listed in supplementary table S9, Supplementary Material online.

Expression Profiles and Protein Structure Prediction of de novo Genes

The RNA-seq data of 6 tissues were downloaded from NCBI: root (accession number GSE116553) (Deforges et al. 2019), seedling (GSE98045) (Fukudome et al. 2017), stem (GSE121407) (Zhang et al. 2019b), leaf (GSE85653) (Albihlal et al. 2018), flower (GSE106943) (Yan et al. 2018), silique (GSE156491) (Li et al. 2021). In addition, the RNA-seq of seed (GSA: CRA014531) was completed in this study. RNA-seq data of 7 different tissues were normalized to TPM by StringTie (Pertea et al. 2015). Protein structure prediction was performed using the online tool AlphaFold2 (https://www.alphafold.ebi.ac.uk/) (Jumper et al. 2021).

Measurement of Phenotypes and Hydrogen Peroxide Content

We counted the number of rosette leaves, with 10 replicates per sample ($n = 10$), on the 12th and 18th days after growth, respectively (Li et al. 1998). The fresh weight ($n = 3$) (El-Lithy et al. 2004) and chlorophyll content of rosette leaves ($n = 3$) were measured on the 19th day after growth. Chlorophyll content was assessed using acetone extraction (Arnon 1949). On the 20th day after growth, we noted the time when the first flower bud emerged ($n = 10$) (Atwell et al. 2010). The length of siliques (for the first 10 siliques per plant, $n = 5$) was measured on the 49th day after growth (Atwell et al. 2010). On the 60th day after sowing, the number of secondary branches ($n = 10$) and plant height ($n = 10$) were recorded (Boyes et al. 2001). The schedule for these phenotypic measurements was aligned with the developmental stages of Arabidopsis (Boyes et al. 2001).

The experiment to count the seed germination was conducted on Petri dishes. The experimental procedures and climate room conditions were consistent with previous studies (Wang et al. 2016), and the seeds were not stratified as described here, nor did we use 75% alcohol to treat the seeds. The treatment solution included: water, 150 or 200 mM NaCl solution, 300 or 400 mM mannitol solution. Each germination experiment was performed with 6 replicates. Radicle outgrowth is defined as the criterion for judging germination (Alonso-Blanco et al. 2003).

The root length experiment was carried out in a square Petri dish with a side length of 13 cm. Four-day-old seedlings grown on MS plates were transferred to new MS plates or those supplemented with 100 or 150 mM NaCl, 200 or 300 mM mannitol, and 0.01 or 0.05 μM paraquat, ensuring that the root tips were aligned at the same level. Each germination experiment was performed with 15 replicates. After culturing the root tip of the seedlings vertically downward for 7 d (20 °C, 16-h day/8-h night), the length of the root growth was measured using a ruler (Verslues et al. 2006).

The hydrogen peroxide content was determined using a hydrogen peroxide quantitative test kit (Boxbio; AKAO009C) according to the manufacturer's instructions (Qiu et al. 2022).

Subcellular Localization and Yeast Two-Hybrid Assay

To examine the subcellular localization of SWK, we used *35S::SWK-GFP* and *35S::GFP* seedlings to observe. Root tips and leaves of 4-day-old seedlings were observed under a super-resolution confocal microscope (Zeiss LSM 980 with Elyra7). GFP fluorescence was detected at 488 nm (excitation) and 509 nm (emission). Nuclei were labeled with DAPI staining. DAPI staining fluorescence was detected at 353 nm (excitation) and 465 nm (emission). Autofluorescence of the chloroplasts was detected at 650 to 750 nm. The yeast two-hybrid assay was carried out based with the GAL4-based two-hybrid system according to the manufacturer's instructions (Clontech, USA) (Li et al. 2023).

Expression Analysis

We used RNAprep Pure Plant Plus Kit (Cat.#DP441, TIANGEN) to extract total RNA from dry seeds or imbibing seeds for 12 h under light, and Micro Elute total Plant RNA Kit (Cat.#R6831-01, Omega) to extract total RNA from other plant materials. We reverse-transcribed 1 μg of RNA into cDNA using the RevertAid First Strand cDNA Synthesis Kit (Cat.#K1622, Thermo Scientific). We conducted RT-qPCR using TB Green Premix Ex Taq (Cat. #RR420A, TaKaRa) on Analytik Jena AG (Jena). Each RT-qPCR experiment included 3 biological replicates and 3 technical replicates per sample. Expression levels were calculated using the $2^{-\Delta\Delta CT}$ method and normalized to *Actin8* (AT1G49240) (Livak and Schmittgen 2001). The primers used for the gene expression analyses are listed in supplementary table S9, Supplementary Material online.

eGWAS Analysis

To identify potential causal variants associated with *SWK* gene expression variation across natural accessions, we performed GWAS using EMMAX (Kang et al. 2010) and GEMMA (Zhou and Stephens 2012), both based on the linear mixed model, with the expression levels of *SWK* from leaves of 414 natural accessions (Kawakatsu et al. 2016). A total of 1,133,651 SNPs, with a missing rate of less than 10% and a MAF >5% were used as genotype. Principal components and a kinship matrix were used to control for population structure. The significant threshold was set based on Bonferroni correction [$-\log_{10}(0.05/1{,}133{,}651)$].

High-throughput RNA Sequencing and Differential Expression Analysis

Total RNA was extracted from the control and 400 mM mannitol-stressed imbibing seeds after 12 h of light exposure. Three biological replicates were utilized for each sample. A sequencing library was constructed for paired-end 150 bp read length sequencing on the Illumina HiSeq 2500 platform. Sequence alignment was performed using TopHat v2.0.12 (Kim et al. 2013). The reads were counted by HTSeq (Anders et al. 2015). Subsequently, StringTie v 2.2.0 assembled reads into transcripts and calculated TPM values for each gene (Pertea et al. 2015). DESeq2 v1.18.1 conducted statistical analysis of DEGs (Anders and Huber 2010). Given the de novo genes originated recently and their interaction with other genes might be weak, we used the value of $\log_2$-FoldChange >0.2 and FDR <0.05 for differential expression gene analysis. We verified the identified candidate downstream genes by RT-qPCR.

KEGG Pathway Analysis and Evolutionary Analysis

In this study, we used online KOBAS v3.0 (http://39.103.204.200/genelist/) to perform KEGG pathway analysis (Bu et al. 2021). The default set of genes used to test for enrichment was all annotated genes in the *A. thaliana* genome (TAIR10).

We performed blastn analysis using several databases: Phytozome 13 (304 plant genomes) (Goodstein et al. 2012), MycoCosm (2,418 fungus genomes) (Grigoriev et al. 2014), PhycoCosm (178 algae genomes) (Grigoriev et al. 2021), IMG/M (12,387 microorganism genomes) (Chen et al. 2023), and Ensembl 109 (369 animal genomes) (Cunningham et al. 2022). For all blastn searches, we used a stringent E-value threshold of 1e-5 to ensure that only highly significant hits were considered. The analysis was performed using the default parameters provided by the blastn online genome databases, with further manual curation to remove redundant or low-quality hits. We filtered the results based on sequence identity and alignment length, ensuring that only sequences with a minimum identity of 60% and alignment covering at least 70% of the query sequence were retained for further analysis.

MEGA X was employed for alignment and phylogenetic analysis (Kumar et al. 2018). All nucleotide sequences were aligned to the Col-0 reference genome, and amino acids were translated according to the Col-0 reference reading frame (note that this implies frameshifts in individual sequences are not represented). All compilation types were determined based on the comparison. Ancestral sequence reconstruction was performed using FastML with default parameters (Ashkenazy et al. 2012). MCScanX, implemented in TBtools, was utilized for syntenic analysis (Chen et al. 2020). The genome *Aethionema arabicum* was downloaded from PLAZA (https://bioinformatics.psb.ugent.be/plaza/), while all other genome data were obtained from Phytozome (https://phytozome-next.jgi.doe.gov/). TSD was searched for using the online tool CENSOR (https://www.girinst.org/censor/index.php) (Kohany et al. 2006). The sequencing data for the accessions of the Cape Verde Islands were obtained from the study by Fulgione et al. (2022).

Nineteen climate factors were derived from the WORLDCLIM database (www.worldclim.org). The comparison of environmental factor differences in the natural geographic distribution of *SWK* presence or absence in African populations was conducted using the two-sided Student's *t*-test. The RDA analysis was performed using previously reported software (Capblancq and Forester 2021). Four environmental factors, including bio2 (mean diurnal range), bio4 (temperature seasonality), bio17 (Precipitation of driest quarter), and bio19 (precipitation of coldest quarter) were used after filtering high correlated factors (Pearson's correlation >0.7 were filtered).

The selective sweep regions were identified using OmegaPlus (version 3.0.3) (Alachiotis et al. 2012), where the ω statistic was computed at 10 kb intervals with the parameters "-minwin 10,000" and "-maxwin 100,000." The top 5% regions with the high ω value were defined as regions under positive selection.

Statistical Analyses

All the statistical analyses were performed in R (http://www.r-project.org/). Benjamini–Hochberg corrected *P*-values were applied for multiple testing.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We would like to thank Song Ge and Bo Xu for valuable comments about this work, and Dr. Yi Han for providing technical assistance during the experiments. This work was supported by the National Natural Science Foundation of China (31925004 and 32221001) and CAS Project for Young Scientists in Basic Research (YSBR-078).

Author Contributions

Y.-L.G. conceived the study. G.-T.J., Y.-C.X., X.-H.H., J.J., X.-X.L., J.-H.X., Y.-T.B., Y.-B.G., M.-Y.W., Z.-Q.Z., Y.E.Z., W.-S.Z., Y.-X.L., and Y.-L.G. performed the experiments and analyzed and interpreted the data. G.-T.J. and Y.-L.G. wrote the paper with contribution from all authors.

Conflict of Interest

The authors declare no competing interests.

Data Availability

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive in National Genomics Data Center (CNCB-NGDC Members and Partners 2022), China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA014531) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa. The data underlying this article are available in the article and in its online supplementary material.

References

Alachiotis N, Stamatakis A, Pavlidis P. OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics*. 2012:28(17):2274–2275. https://doi.org/10.1093/bioinformatics/bts419.

Albihlal WS, Obomighie I, Blein T, Persad R, Chernukhin I, Crespi M, Bechtold U, Mullineaux PM. Arabidopsis *HEAT SHOCK TRANSCRIPTION FACTORA1b* regulates multiple developmental genes under benign and stress conditions. *J Exp Bot*. 2018:69(11):2847–2862. https://doi.org/10.1093/jxb/ery142.

Alonso-Blanco C, Bentsink L, Hanhart CJ, Blankestijn-de Vries H, Koornneef M. Analysis of natural allelic variation at seed dormancy loci of *Arabidopsis thaliana*. *Genetics*. 2003:164(2):711–729. https://doi.org/10.1093/genetics/164.2.711.

An NA, Zhang J, Mo F, Luan X, Tian L, Shen QS, Li X, Li C, Zhou F, Zhang B, *et al*. De novo genes with an lncRNA origin encode unique human brain developmental functionality. *Nat Ecol Evol*. 2023:7(2):264–278. https://doi.org/10.1038/s41559-022-01925-6.

Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010:11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106.

Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015:31(2):166–169. https://doi.org/10.1093/bioinformatics/btu638.

Arnon DI. Copper enzymes in isolated chloroplasts. Polyphenoloxidase in *Beta vulgaris*. *Plant Physiol*. 1949:24(1):1–15. https://doi.org/10.1104/pp.24.1.1.

Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. FastML: a web server for probabilistic

reconstruction of ancestral sequences. *Nucleic Acids Res*. 2012:40(W1):W580–W584. https://doi.org/10.1093/nar/gks498.
Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*. 2010:465(7298):627–631. https://doi.org/10.1038/nature08800.
Barrett RD, Schluter D. Adaptation from standing genetic variation. *Trends Ecol Evol*. 2008:23(1):38–44. https://doi.org/10.1016/j.tree.2007.09.008.
Begun DJ, Lindfors HA, Thompson ME, Holloway AK. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics*. 2006:172(3):1675–1681. https://doi.org/10.1534/genetics.105.050336.
Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, Díez J, Carey LB, Albà MM. Uncovering de novo gene birth in yeast using deep transcriptomics. *Nat Commun*. 2021:12(1):604. https://doi.org/10.1038/s41467-021-20911-3.
Boyes DC, Zayed AM, Ascenzi R, McCaskill AJ, Hoffman NE, Davis KR, Görlach J. Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. *Plant Cell*. 2001:13(7):1499–1510. https://doi.org/10.1105/tpc.010011.
Bu D, Luo H, Huo P, Wang Z, Zhang S, He Z, Wu Y, Zhao L, Liu J, Guo J, *et al.* KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res*. 2021:49(W1):W317–W325. https://doi.org/10.1093/nar/gkab447.
Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet*. 2011:43(10):956–963. https://doi.org/10.1038/ng.911.
Capblancq T, Forester BR. Redundancy analysis: a Swiss army knife for landscape genomics. *Methods Ecol Evol*. 2021:12(12):2298–2309. https://doi.org/10.1111/2041-210X.13722.
Capra JA, Pollard KS, Singh M. Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biol*. 2010:11(12):R127. https://doi.org/10.1186/gb-2010-11-12-r127.
Cardoso-Moreira M, Long M. The origin and evolution of new genes. *Methods Mol Biol*. 2012:856:161–186. https://doi.org/10.1007/978-1-61779-585-5_7.
Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, *et al.* Proto-genes and de novo gene birth. *Nature*. 2012:487(7407):370–374. https://doi.org/10.1038/nature11184.
Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol Plant*. 2020:13(8):1194–1202. https://doi.org/10.1016/j.molp.2020.06.009.
Chen IA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, Hajek P, Ritter SJ, Webb C, Wu D, *et al.* The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res*. 2023:51(D1):D723–D732. https://doi.org/10.1093/nar/gkac976.
Chen S, Zhang YE, Long M. New genes in Drosophila quickly become essential. *Science*. 2010:330(6011):1682–1685. https://doi.org/10.1126/science.1196380.
Chen ST, Cheng HC, Barbash DA, Yang HP. Evolution of *hydra*, a recently evolved testis-expressed gene with nine alternative first exons in *Drosophila melanogaster*. *PLoS Genet*. 2007:3(7):e107. https://doi.org/10.1371/journal.pgen.0030107.
Cheng X, Zhang D, Cheng Z, Keller B, Ling HQ. A new family of Ty1-copia-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. *Genetics*. 2009:181(4):1183–1193. https://doi.org/10.1534/genetics.108.099150.
Chong VK, Stinchcombe JR. Evaluating population genomic candidate genes underlying flowering time in *Arabidopsis thaliana* using T-DNA insertion lines. *J Hered*. 2019:110(4):445–454. https://doi.org/10.1093/jhered/esz026.
Clough SJ, Bent AF. Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J*. 1998:16(6):735–743. https://doi.org/10.1046/j.1365-313x.1998.00343.x.
CNCB-NGDC Members and Partners. Database resources of the national genomics data center, China national center for bioinformation in 2022. *Nucleic Acids Res*. 2022:50(D1):D27–D38. https://doi.org/10.1093/nar/gkab951.
Cui X, Lv Y, Chen M, Nikoloski Z, Twell D, Zhang D. Young genes out of the male: an insight from evolutionary age analysis of the pollen transcriptome. *Mol Plant*. 2015:8(6):935–945. https://doi.org/10.1016/j.molp.2014.12.008.
Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, *et al.* Ensembl 2022. *Nucleic Acids Res*. 2022:50(D1):D988–D995. https://doi.org/10.1093/nar/gkab1049.
Deforges J, Reis RS, Jacquet P, Sheppard S, Gadekar VP, Hart-Smith G, Tanzer A, Hofacker IL, Iseli C, Xenarios I, *et al.* Control of cognate sense mRNA translation by cis-natural antisense RNAs. *Plant Physiol*. 2019:180(1):305–322. https://doi.org/10.1104/pp.19.00043.
Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, Tsuchimatsu T, Burbano HA, Pico FX, Alonso-Blanco C, *et al.* African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A*. 2017:114(20):5213–5218. https://doi.org/10.1073/pnas.1616736114.
El-Lithy ME, Clerkx EJ, Ruys GJ, Koornneef M, Vreugdenhil D. Quantitative trait locus analysis of growth-related traits in a new Arabidopsis recombinant inbred population. *Plant Physiol*. 2004:135(1):444–458. https://doi.org/10.1104/pp.103.036822.
Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. *Nat Rev Genet*. 2007:8(8):610–618. https://doi.org/10.1038/nrg2146.
Fukudome A, Sun D, Zhang X, Koiwa H. Salt stress and CTD PHOSPHATASE-LIKE4 mediate the switch between production of small nuclear RNAs and mRNAs. *Plant Cell*. 2017:29(12):3214–3233. https://doi.org/10.1105/tpc.17.00331.
Fulgione A, Neto C, Elfarargi AF, Tergemina E, Ansari S, Göktay M, Dinis H, Döring N, Flood PJ, Rodriguez-Pacheco S, *et al.* Parallel reduction in flowering time from de novo mutations enable evolutionary rescue in colonizing lineages. *Nat Commun*. 2022:13(1):1461. https://doi.org/10.1038/s41467-022-28800-z.
Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. 2012:40(D1):D1178–D1186. https://doi.org/10.1093/nar/gkr944.
Gorelova V, Bastien O, De Clerck O, Lespinats S, Rébeillé F, Van Der Straeten D. Evolution of folate biosynthesis and metabolism across algae and land plant lineages. *Sci Rep*. 2019:9(1):5731. https://doi.org/10.1038/s41598-019-42146-5.
Grigoriev IV, Hayes RD, Calhoun S, Kamel B, Wang A, Ahrendt S, Dusheyko S, Nikitin R, Mondo SJ, Salamov A, *et al.* PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res*. 2021:49(D1):D1004–D1011. https://doi.org/10.1093/nar/gkaa898.
Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res*. 2014:42(D1):D699–D704. https://doi.org/10.1093/nar/gkt1183.
Guo WJ, Li P, Ling J, Ye SP. Significant comparative characteristics between orphan and nonorphan genes in the rice (*Oryza sativa L.*) genome. *Comp Funct Genomics*. 2007:2007:21676. https://doi.org/10.1155/2007/21676.
Guo YL. Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes. *Plant J*. 2013:73(6):941–951. https://doi.org/10.1111/tpj.12089.
Hamala T, Tiffin P. Biased gene conversion constrains adaptation in *Arabidopsis thaliana*. *Genetics*. 2020:215(3):831–846. https://doi.org/10.1534/genetics.120.303335.

Hanson AD, Gregory JF III. Folate biosynthesis, turnover, and transport in plants. *Annu Rev Plant Biol*. 2011:62(1):105–125. https://doi.org/10.1146/annurev-arplant-042110-103819.

He X, Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*. 2005:169(2):1157–1164. https://doi.org/10.1534/genetics.104.037051.

Heinen TJ, Staubach F, Häming D, Tautz D. Emergence of a new gene from an intergenic region. *Curr Biol*. 2009:19(18):1527–1531. https://doi.org/10.1016/j.cub.2009.07.049.

Hendriks KP, Kiefer C, Al-Shehbaz IA, Bailey CD, van Huysduynen AH, Nikolov LA, Nauheimer L, Zuntini AR, German DA, Franzke A, *et al*. Global Brassicaceae phylogeny based on filtering of 1,000-gene dataset. *Curr Biol*. 2023:33(19):4052–4068.e6. https://doi.org/10.1016/j.cub.2023.08.026.

Jiang J, Xu YC, Zhang ZQ, Chen JF, Niu XM, Hou XH, Li XT, Wang L, Zhang YE, Ge S, *et al*. Forces driving transposable element load variation during Arabidopsis range expansion. *Plant Cell*. 2024:36(4):840–862. https://doi.org/10.1093/plcell/koad296.

Jiang M, Li X, Dong X, Zu Y, Zhan Z, Piao Z, Lang H. Research advances and prospects of orphan genes in plants. *Front Plant Sci*. 2022:13:947129. https://doi.org/10.3389/fpls.2022.947129.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021:596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*. 2010:42(4):348–354. https://doi.org/10.1038/ng.548.

Kawakatsu T, Huang SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y, *et al*. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell*. 2016:166(2):492–505. https://doi.org/10.1016/j.cell.2016.06.044.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013:14(4):R36. https://doi.org/10.1186/gb-2013-14-4-r36.

Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and censor. *BMC Bioinformatics*. 2006:7(1):474. https://doi.org/10.1186/1471-2105-7-474.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 2018:35(6):1547–1549. https://doi.org/10.1093/molbev/msy096.

Kuzmin E, Taylor JS, Boone C. Retention of duplicated genes in evolution. *Trends Genet*. 2022:38(1):59–72. https://doi.org/10.1016/j.tig.2021.06.016.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A*. 2006:103(26):9935–9939. https://doi.org/10.1073/pnas.0509809103.

Li B, Suzuki JI, Hara T. Latitudinal variation in plant size and relative growth rate in *Arabidopsis thaliana*. *Oecologia*. 1998:115(3):293–301. https://doi.org/10.1007/s004420050519.

Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES. Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *Plant J*. 2009:58(3):485–498. https://doi.org/10.1111/j.1365-313X.2009.03793.x.

Li L, Zheng W, Zhu Y, Ye H, Tang B, Arendsee ZW, Jones D, Li R, Ortiz D, Zhao X, *et al*. *QQS* orphan gene regulates carbon and nitrogen partitioning across species via NF-YC interactions. *Proc Natl Acad Sci U S A*. 2015:112(47):14734–14739. https://doi.org/10.1073/pnas.1514670112.

Li X, Martín-Pizarro C, Zhou L, Hou B, Wang Y, Shen Y, Li B, Posé D, Qin G. Deciphering the regulatory network of the NAC transcription factor FvRIF, a key regulator of strawberry (*Fragaria vesca*) fruit ripening. *Plant Cell*. 2023:35(11):4020–4045. https://doi.org/10.1093/plcell/koad210.

Li YJ, Yu Y, Liu X, Zhang XS, Su YH. The Arabidopsis MATERNAL EFFECT EMBRYO ARREST45 protein modulates maternal auxin biosynthesis and controls seed size by inducing *AINTEGUMENTA*. *Plant Cell*. 2021:33(6):1907–1926. https://doi.org/10.1093/plcell/koab084.

Li ZW, Chen X, Wu Q, Hagmann J, Han TS, Zou YP, Ge S, Guo YL. On the origin of de novo genes in *Arabidopsis thaliana* populations. *Genome Biol Evol*. 2016:8(7):2190–2202. https://doi.org/10.1093/gbe/evw164.

Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods*. 2001:25(4):402–408. https://doi.org/10.1006/meth.2001.1262.

Long M, VanKuren NW, Chen S, Vibranovski MD. New gene evolution: little did we know. *Annu Rev Genet*. 2013:47(1):307–333. https://doi.org/10.1146/annurev-genet-111212-133301.

Ma D, Ding Q, Guo Z, Zhao Z, Wei L, Li Y, Song S, Zheng HL. Identification, characterization and expression analysis of lineage-specific genes within mangrove species *Aegiceras corniculatum*. *Mol Genet Genomics*. 2021:296(6):1235–1247. https://doi.org/10.1007/s00438-021-01810-0.

Ma X, Zhang Q, Zhu Q, Liu W, Chen Y, Qiu R, Wang B, Yang Z, Li H, Lin Y, *et al*. A robust CRISPR/Cas9 system for convenient, high-efficiency multiplex genome editing in monocot and dicot plants. *Mol Plant*. 2015:8(8):1274–1284. https://doi.org/10.1016/j.molp.2015.04.007.

McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. *Nat Rev Genet*. 2016:17(9):567–578. https://doi.org/10.1038/nrg.2016.78.

Moon H, Jeong AR, Kwon OK, Park CJ. Oryza-specific orphan protein triggers enhanced resistance to *Xanthomonas oryzae* pv. *oryzae* in rice. *Front Plant Sci*. 2022:13:859375. https://doi.org/10.3389/fpls.2022.859375.

Mouillon JM, Ravanel S, Douce R, Rébeillé F. Folate synthesis in higher-plant mitochondria: coupling between the dihydropterin pyrophosphokinase and the dihydropteroate synthase activities. *Biochem J*. 2002:363(2):313–319. https://doi.org/10.1042/bj3630313.

Nguyen PN, Tossounian MA, Kovacs DS, Thu TT, Stijlemans B, Vertommen D, Pauwels J, Gevaert K, Angenon G, Messens J, *et al*.Dehydrin ERD14 activates glutathione transferase Phi9 in *Arabidopsis thaliana* under osmotic stress. *Biochim Biophys Acta Gen Subj*. 2020:1864(3):129506. https://doi.org/10.1016/j.bbagen.2019.129506.

Ni F, Qi J, Hao Q, Lyu B, Luo MC, Wang Y, Chen F, Wang S, Zhang C, Epstein L, *et al*. Wheat Ms2 encodes for an orphan protein that confers male sterility in grass species. *Nat Commun*. 2017:8(1):15121. https://doi.org/10.1038/ncomms15121.

Orr HA. The genetic theory of adaptation: a brief history. *Nat Rev Genet*. 2005:6(2):119–127. https://doi.org/10.1038/nrg1523.

Ozturk M, Turkyilmaz Unal B, García-Caparrós P, Khursheed A, Gul A, Hasanuzzaman M. Osmoregulation and its actions during the drought stress in plants. *Physiol Plant*. 2021:172(2):1321–1335. https://doi.org/10.1111/ppl.13297.

Perochon A, Jianguang J, Kahla A, Arunachalam C, Scofield SR, Bowden S, Wallington E, Doohan FM. TaFROG encodes a pooideae orphan protein that interacts with SnRK1 and enhances resistance to the mycotoxigenic fungus *Fusarium graminearum*. *Plant Physiol*. 2015:169(4):2895–2906. https://doi.org/10.1104/pp.15.01056.

Perochon A, Kahla A, Vranic M, Jia J, Malla KB, Craze M, Wallington E, Doohan FM. A wheat NAC interacts with an orphan protein and enhances resistance to Fusarium head blight disease. *Plant Biotechnol J*. 2019:17(10):1892–1904. https://doi.org/10.1111/pbi.13105.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015:33(3):290–295. https://doi.org/10.1038/nbt.3122.

Puertollano R, Ferguson SM, Brugarolas J, Ballabio A. The complex relationship between TFEB transcription factor phosphorylation and subcellular localization. *EMBO J*. 2018:37(11):e98804. https://doi.org/10.15252/embj.201798804.

Qi J, Mo F, An NA, Mi T, Wang J, Qi JT, Li X, Zhang B, Xia L, Lu Y, *et al*. A human-specific de novo gene promotes cortical expansion and folding. *Adv Sci (Weinh)*. 2023:10(7):e2204140. https://doi.org/10.1002/advs.202204140.

Qi M, Zheng W, Zhao X, Hohenstein JD, Kandel Y, O'Conner S, Wang Y, Du C, Nettleton D, MacIntosh GC, *et al*. QQS orphan gene and its interactor NF-YC4 reduce susceptibility to pathogens and pests. *Plant Biotechnol J*. 2019:17(1):252–263. https://doi.org/10.1111/pbi.12961.

Qiu J, Liu Z, Xie J, Lan B, Shen Z, Shi H, Lin F, Shen X, Kou Y. Dual impact of ambient humidity on the virulence of *Magnaporthe oryzae* and basal resistance in rice. *Plant Cell Environ*. 2022:45(12):3399–3411. https://doi.org/10.1111/pce.14452.

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabido E, Kondova I, Bontrop R, Marques-Bonet T, Alba MM. Origins of de novo genes in human and chimpanzee. *PLoS Genet*. 2015:11(12):e1005721. https://doi.org/10.1371/journal.pgen.1005721.

Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Canas JL, Messeguer X, Alba MM. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat Ecol Evol*. 2018:2(5):890–896. https://doi.org/10.1038/s41559-018-0506-6.

Ryu HY, Kim SY, Park HM, You JY, Kim BH, Lee JS, Nam KH. Modulations of AtGSTF10 expression induce stress tolerance and BAK1-mediated cell death. *Biochem Biophys Res Commun*. 2009:379(2):417–422. https://doi.org/10.1016/j.bbrc.2008.11.156.

Schlötterer C. Genes from scratch–the evolutionary fate of de novo genes. *Trends Genet*. 2015:31(4):215–219. https://doi.org/10.1016/j.tig.2015.02.007.

Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat Ecol Evol*. 2018:2(10):1626–1632. https://doi.org/10.1038/s41559-018-0639-7.

Singh U, Syrkin Wurtele E. How new genes are born. *eLife*. 2020:9:e55136. https://doi.org/10.7554/eLife.55136.

Storozhenko S, Navarrete O, Ravanel S, De Brouwer V, Chaerle P, Zhang GF, Bastien O, Lambert W, Rebeille F, Van Der Straeten D. Cytosolic hydroxymethyldihydropterin pyrophosphokinase/dihydropteroate synthase from *Arabidopsis thaliana*: a specific role in early development and stress response. *J Biol Chem*. 2007:282(14):10749–10761. https://doi.org/10.1074/jbc.M701158200.

Tan S, Cardoso-Moreira M, Shi W, Zhang D, Huang J, Mao Y, Jia H, Zhang Y, Chen C, Shao Y, *et al*. LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. *Genome Res*. 2016:26(12):1663–1675. https://doi.org/10.1101/gr.204925.116.

The 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*. 2016:166(2):481–491. https://doi.org/10.1016/j.cell.2016.05.063.

Vakirlis N, Acar O, Hsu B, Castilho Coelho N, Van Oss SB, Wacholder A, Medetgul-Ernar K, Bowman RW II, Hines CP, Iannotta J, *et al*. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat Commun*. 2020:11(1):781. https://doi.org/10.1038/s41467-020-14500-z.

Verslues PE, Agarwal M, Katiyar-Agarwal S, Zhu J, Zhu JK. Methods and concepts in quantifying resistance to drought, salt and freezing, abiotic stresses that affect plant water status. *Plant J*. 2006:45(4):523–539. https://doi.org/10.1111/j.1365-313X.2005.02593.x.

Wang Z, Chen F, Li X, Cao H, Ding M, Zhang C, Zuo J, Xu C, Xu J, Deng X, *et al*. Arabidopsis seed germination speed is controlled by SNL histone deacetylase-binding factor-mediated regulation of AUX1. *Nat Commun*. 2016:7(1):13412. https://doi.org/10.1038/ncomms13412.

Wells SE, Hillner PE, Vale RD, Sachs AB. Circularization of mRNA by eukaryotic translation initiation factors. *Mol Cell*. 1998:2(1):135–140. https://doi.org/10.1016/S1097-2765(00)80122-7.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, *et al*. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2007:8(12):973–982. https://doi.org/10.1038/nrg2165.

Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat Ecol Evol*. 2017:1(6):0146–0146. https://doi.org/10.1038/s41559-017-0146.

Wu DD, Zhang YP. Evolution and function of de novo originated genes. *Mol Phylogenet Evol*. 2013:67(2):541–545. https://doi.org/10.1016/j.ympev.2013.02.013.

Xie C, Bekpen C, Künzel S, Keshavarz M, Krebs-Wheaton R, Skrabar N, Ullrich KK, Tautz D. A de novo evolved gene in the house mouse regulates female pregnancy cycles. *eLife*. 2019:8:e44392. https://doi.org/10.7554/eLife.44392.

Xu YC, Niu XM, Li XX, He W, Chen JF, Zou YP, Wu Q, Zhang YE, Busch W, Guo YL. Adaptation and phenotypic diversification in arabidopsis through loss-of-function mutations in protein-coding genes. *Plant Cell*. 2019:31(5):1012–1025. https://doi.org/10.1105/tpc.18.00791.

Yan W, Chen D, Smaczniak C, Engelhorn J, Liu H, Yang W, Graf A, Carles CC, Zhou DX, Kaufmann K. Dynamic and spatial restriction of Polycomb activity by plant histone demethylases. *Nat Plants*. 2018:4(9):681–689. https://doi.org/10.1038/s41477-018-0219-5.

Zhang W, Gao Y, Long M, Shen B. Origination and evolution of orphan genes and de novo genes in the genome of *Caenorhabditis elegans*. *Sci China Life Sci*. 2019a:62(4):579–593. https://doi.org/10.1007/s11427-019-9482-0.

Zhang W, Landback P, Gschwend AR, Shen B, Long M. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol*. 2015:16(1):202. https://doi.org/10.1186/s13059-015-0772-4.

Zhang Y, Yin B, Zhang J, Cheng Z, Liu Y, Wang B, Guo X, Liu X, Liu D, Li H, *et al*. Histone deacetylase HDT1 is involved in stem vascular development in Arabidopsis. *Int J Mol Sci*. 2019b:20(14):3452. https://doi.org/10.3390/ijms20143452.

Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science*. 2014:343(6172):769–772. https://doi.org/10.1126/science.1248286.

Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*. 2012:44(7):821–824. https://doi.org/10.1038/ng.2310.

Zhu JK. Salt and drought stress signal transduction in plants. *Annu Rev Plant Biol*. 2002:53(1):247–273. https://doi.org/10.1146/annurev.arplant.53.091401.143329.

Zou YP, Hou XH, Wu Q, Chen JF, Li ZW, Han TS, Niu XM, Yang L, Xu YC, Zhang J, *et al*. Adaptation of *Arabidopsis thaliana* to the Yangtze River basin. *Genome Biol*. 2017:18(1):239. https://doi.org/10.1186/s13059-017-1378-9.