PROFESSOR MAX CHENG (Orcid ID : 0000-0002-4702-0573)

Article type : Original Article

The emergence and evolution of intron-poor and intronless genes in intron-rich plant gene families

Hui Liu[1], Hai-Meng Lyu[1], Kaikai Zhu[2], Yves Van de Peer[1,3,4,5], Zong-Ming (Max) Cheng[1,6,*]

[1]College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China

[2]College of Forestry, Nanjing Forestry University, Nanjing 210037, China

[3]Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium

[4]VIB Center for Plant Systems Biology, Ghent, Belgium

[5]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, South Africa

[6]Department of Plant Sciences, University of Tennessee, Knoxville 37996, USA

[*]Corresponding author (E-mail: zmc@njau.edu.cn, zcheng@utk.edu)

Running title: Intronless gene sub-families' origin and evolution.

Summary

Eukaryotic genes can be classified into intronless (no introns), intron-poor (three or fewer introns per gene), or intron-rich. Early eukaryotic genes were mostly intron-rich, and their alternative splicing into multiple transcripts, giving rise to different proteins, might have played pivotal roles in adaptation and evolution. Interestingly, extant plant genomes contain many gene families with one or sometimes few sub-families with genes that are intron-poor or intronless, and it remains unknown when and how these intron-poor or intronless genes have originated and evolved and what their possible functions are. In this study, we identified 33 such gene families that contained intronless and intron-poor sub-families. Intronless genes seemed to have first emerged in early land plant evolution, while intron-poor sub-families seemed first to have appeared in green algae. In contrast to intron-rich genes, intronless genes in intron-poor sub-families occurred later, and were subject to stronger functional constraints. Based on RNA-seq analyses in *Arabidopsis* and rice, intronless or intron-poor genes in *AP2*, *EF-hand_7, bZIP*, *FAD_binding_4*, *STE_STE11*, *CAMK_CAMKL-CHK1*, and *C2* gene families were more likely to play a role in response to drought and salt stress, compared to intron-rich genes in the same gene families, whereas intronless genes in the *B_lectin* and *S_locus_glycop* gene family were more likely to participate in epigenetic processes and plant development. Understanding the origin and evolutionary trajectory, as well as the potential functions, of intronless and intron-poor sub-families provides further insight into plant genome evolution and the functional divergence of genes.

Keywords:

Intronless gene, Intron-poor sub-family, Adaptation, Duplication, Abiotic stresses

Introduction

An intron is "a transcription unit containing regions which could be lost from the mature messenger" (Gilbert, 1978). Introns in pre-mRNAs are spliced by conserved sequences at their 5' and 3' borders in a massive complex ribonucleoprotein, called spliceosome (Jurica and Moore, 2003, Nilsen, 2003, Rogozin, 2012). Genes with multiple introns offer opportunities of alternative splicing and provide variant proteins which may play different roles in biological processes (Min *et al.*, 2015). The recently determined genome sequences of red and green algae showed that introns were genome-wide abundant prior to the origin of land plants (Baier *et al.*, 2018).

Based on the presence or absence of introns, eukaryotic genes can be divided into two categories: intronless and intron-containing genes. The latter can be further divided arbitrarily into intron-rich (having more than 3 introns per gene) or intron-poor (with only 1-3 introns). It is commonly known that genes in a given large gene family are usually scattered over the whole genome, and mostly consist of multi-intron genes, but some gene families contain one, and sometimes more, sub-family that possesses no introns (Liu *et al.*, 2014, Zhu *et al.*, 2016, Zhu *et al.*, 2018). Since the most early-diverging plant species in the plant kingdom contain few intronless genes (Sakharkar *et al.*, 2007), it is reasonable to assume that these sub-families of intronless genes emerged after intron-rich genes. Furthermore, a few reports suggested that these intronless genes could actually be processed pseudogenes, which also lack a 5' promoter sequence (Tutar, 2012). These processed pseudogenes are generated through retro-transposition and have usually no function, although some of them have been shown to be expressed and may play a role in gene regulation (Pink *et al.*, 2011, McDonell and Drouin, 2012, Tutar, 2012). Because few systematic studies have been performed on these intronless gene sub-families in plants, several intriguing questions remain, such as, in a given species, how many gene families may contain such intronless or intron-poor sub-families? When did these originate? Did they evolve once or multiple times? How did they evolve? What are their potential functions? What were the mechanisms responsible for their origin and expansion? Did intronless and/or intron poor genes face the same or different selection pressures as the intron-rich genes?

Here, we conducted a genome-wide survey in *Arabidopsis* on these intronless sub-families of genes, and attempted to address some of the above questions. We identified 33 gene families that clearly contained intronless or intron-poor sub-families. Next, we characterized these gene families in seven genomes representative for the tree of plant life (the green alga *Chlamydomonas reinhardtii,* the moss *Physcomitrella patens,* the the early vascular plant *Selaginella moellendorffii,* the gymnosperm *Ginkgo biloba,* the early diverging angiosperm *Amborella trichopoda,* the monocot *Oryza sativa,* and the eudicot *Arabidopsis thaliana*). We contrasted the origin, evolution, expansion, and selective pressures of these intronless and intron-poor sub-families with their intron-rich homologs. To learn more about the function of intronless and intron-poor genes, and to see whether their function differs with respect to their homologs, we analyzed *Arabidopsis* and rice expression data, and considered gene ontology (GO) information.

Results

Identification and classification of gene families containing intron-poor sub-families

We surveyed GFF3 files of *Arabidopsis*, which resulted in the identification of 6308 intronless genes (Table S1). To identify the gene families that contain at least one intron-poor sub-family, we constructed unrooted phylogenetic trees for 149 gene families in *Arabidopsis* (Table S2). By evaluating the exon/intron structure of all of the 149 gene families, we identified a total of 33 gene families that contained at least one intronless, or an intron-poor sub-family with at least five genes in a single intronless clade, and intron-poor clade (see a schematic representative gene family in Fig 1-A).

The 33 gene families (Table S3) consisted of eight transcription factor (TF) gene families (*bZIP*, *AP2*, *B3*, *zf-CCCH*, *MYB*, *NAC*, *SRF-TF*, and *HLH*), one transcription regulator gene family (*SET*), one transporter gene family (*MFS-1*), four protein kinase gene families (*CAMK_CAMKL-CHK1*, *STE_STE11*, *STE_STE7*, *TKL-PI-4*), and 19 other gene families. The identified intronless genes from the different gene sub-families were distributed over all

chromosomes in both *Arabidopsis* and rice (Figure S1).

The origin and evolution of intron-poor and intronless sub-families

To determine the origin of intronless and intron-poor genes in intronless or intron-poor sub-families, we first determined the origin of intron-poor sub-families in the plant tree of life (Figure 1-B). The intron-poor sub-families in the 33 gene families originated at different times (Table S3). Nine out of 33 (27.3%) intron-poor sub-families originated in green algae, 18/33 (54.5%) intron-poor sub-families first appeared in *P. patens* (land plant), and 4/33 (12.1%) originated in seed plants (Figure 2). Only one intron-poor sub-families had a unique origin in *S. moellendorffii*.

For the 33 gene families, none of the intronless genes in intron-poor and intronless sub-families seemed to have emerged first in green algae or *G. biloba* (gymnosperm), whereas intronless genes in 13 and 11 gene families first appeared in *P. patens* and *S. moellendorffii* (land plants), respectively (Figure 3). Furthermore, all intronless genes seemed to have originated from intron-poor genes, and not from intron-rich genes (or sub-families).

Also, some of the intronless sub-family genes seem to have been lost in some lineages (Table S4). For example, the intronless genes in the *zf-CCCH, HLH* and *SET* gene families first appeared in *P. patens* (land plants), but were not found in *S. moellendorffii* (vascular plants), while they seem to exist in *G. biloba* (seed plants). For the *B3* and *TKL-PI-4* gene families, intronless genes were not detected in *G. biloba*, and were uncovered in *A. trichopoda* (angiosperm).

Expansion of intronless genes

To determine how intronless genes expanded in both intron-poor and intronless sub-family, each intronless member of all gene families in the seven species was assigned to one of five different duplication categories: singleton, tandem, proximal, WGD and segmental, and dispersed (Wang *et al.*, 2013). The combined WGD and segmental duplication events of intronless genes contributed to over 80% of the gene expansion in *P. patens* (85.3%), *O. sativa* (89.3%), and *A. thaliana*

(86.1%), respectively, for all 33 gene families combined (Figure S2). Remarkably, the proportion of intronless tandem-duplicated genes was 10.2%, 13.2%, 14.6%, 39.9%, and 40.4% in *A. thaliana, P. patens*, *S. moellendorffii*, *G. biloba*, and *A. trichopoda*, respectively, which was more than 10% (Figure 4A, S2). These results illustrate that both WGD/segmental and tandem duplications played an important role in the expansion of intronless genes in the intron-poor sub-families.

The WGD-generated intronless genes had significantly higher *Ks* values than tandem duplication-generated intronless genes (*t*-test, $P<0.01$) (Figure 4B). Therefore, the tandem duplicated intronless genes seem generally younger than the WGD duplicated intronless genes in the intron-poor sub-families. In general, the *Ks* values were significantly different among intronless, intron-poor and intron-rich genes (Figure 5A). The intronless sub-family genes had significantly lower *Ks* values than intron-poor and intron-rich sub-family genes (*t*-test, $P<0.01$). Moreover, the intron-poor sub-family genes also had lower *Ks* values than intron-rich sub-family genes. Therefore, the intron-poor genes seem younger than the intron-rich genes. In turn, the intronless genes seem younger than both intron-poor and intron-rich genes in the seven species (Figure 5A).

To further understand the origin and expansion of intronless genes in intron-poor sub-families, we separated orthologous and paralogous clades in 33 gene families (Table S5). Almost all intron-poor sub-families had orthologous and paralogous gene pairs, except *FAD_binding_4*, *NAC*, *Peptidase_C1*, and *SRF-TF* intron-poor sub-families, which only contained paralogous gene pairs. Therefore, these intronless paralogs are defined as species-specific duplicates. This is also confirmed by average *Ks* values between the paralogs, which are significantly lower than those of orthologs in all gene families.

Selection pressures on intron-rich, -poor and -less gene sub-families

To screen for possible signs of selection on genes in the intronless and intron-poor sub-families, we calculated the *Ka/Ks* ratios of intron-rich sub-families and intron-poor sub-families (intronless

and intron-poor gene pairs) for each gene family (Table S6). The *Ka/Ks* values for gene pairs among 33 gene families of intron-poor sub-families and intron-rich sub-families were less than 1, except for the *EF-hand_7* and *NAC* gene family. Therefore, most genes in intronless and intron-poor sub-families seem to have undergone purifying selection. In contrast to intron-rich sub-families, intronless gene pairs have a narrower range of *Ka/Ks* values (*t*-test, $P<0.01$), thus, intronless genes seem to have been subject to stronger functional constraints than intron-rich genes. Intron-poor gene pairs had *Ka/Ks* values between those of intron-rich and intronless gene pairs (Figure 5B).

GC-content variation

The GC content of coding sequences in the seven species for intronless, intron-poor, and intron-rich sub-family genes is shown in Figure 6. The GC content varied among different genomic regions. The *C. reinhardtii* and *O. sativa* genes had a GC content of 69% for coding sequences and 61%, respectively, which was higher than those for the other species. *A. thaliana* showed the lowest GC content (45%) for coding sequences. Based on the box plots (Figure 6), the correlation between the GC contents in intronless genes was significantly higher than those in intron-rich genes for the seven species, except for *S. moellendorffii* (*t*-test, $P<0.01$).

Functional preference of differentially expressed genes in *Arabidopsis* and rice of intronless, intron-poor, and intron–rich sub-family genes in response to drought and salinity

Large-scale RNA-seq data from different tissues and developmental stages of *Arabidopsis* and rice (Table S7) showed intronless and intron-rich genes from the same gene families to be differentially expressed in response to drought and salt stress (see Methods). In the 13 *Arabidopsis* studies with drought treatments, 136 out of 406 (33.5%) intronless, 67 out of 254 (26.4%) intron-poor, and 206 out of 859 (24.0%) intron-rich DEGs were discovered, respectively (Figure 7A). The number of intronless DEGs was significantly greater than those of intron-rich DEGs (*t*-test, $P<0.01$). The 136 *Arabidopsis* intronless DEGs, of which 54, 9, 8, 7, 7, and 6 *AP2*,

FAD_binding_4, *EF-hand_7*, *C2*, *STE_STE11*, and *CAMK_CAMKL-CHK1*, respectively, account for 60.0 (54/90), 45.0 (9/20), 61.5 (8/13), 33.3 (7/21), 30.4% (7/23), and 46.2 (6/13) percent of the genes in the respective family (Table S8). Similarly, in nine studies in rice, a total of 127 out of 441 (28.8%) intronless and 210 out of 878 (23.9%) intron-rich DEGs were detected in response to drought stress (Figure 7A). The percentages of intronless DEGs were significantly greater than those of intron-rich DEGs (*t*-test, $P<0.01$): the 127 intronless DEGs belonged to different gene families, with 40 (43.5%), 13 (25.0%), 9 (39.1%), 7 (46.7%), 7 (41.2%), and 6 (46.2%) intronless genes in the *AP2*, *B_lectin*, *EF-hand_7*, *Myb_DNA-binding*, *bZIP*, and *CAMK_CAMKL-CHK1* gene family, respectively (Table S8).

In six salt-response studies in *Arabidopsis*, a total of 139 out of 406 (34.2%) intronless and 272 out of 859 (31.6%) intron-rich DEGs were detected (Figure 7B). Among the 139 intronless DEGs, 49 were *AP2* genes, nine were *FAD_binding_4* genes, and *STE_STE11* genes, eight each were *EF-hand_7* and *CAMK_CAMKL-CHK1* genes, and seven were *C2* genes, the others were listed in Table S8. In the seven salt-response studies in rice, there were 98 out of 441 (22.2%) intronless, 53 out of 304 (17.4%) intron-poor, and 152 out of 878 (17.3%) intron-rich DEGs, respectively (Figure 7B). In the 98 intronless DEGs, 20 were *AP2* rice genes, eight were *EF-hand_7* genes, six were *bZIP* genes, and five were *CAMK_CAMKL-CHK1* genes (Table S8). The percentages of intronless DEGs were significantly greater than those in intron-rich DEGs in both *Arabidopsis* and rice studies (*t*-test, $P<0.01$). The intronless DEGs in the *AP2*, *EF-hand_7, bZIP*, *FAD_binding_4*, *STE_STE11*, *CAMK_CAMKL-CHK1*, and *C2* gene family, played more significant roles in responding to drought and salt stresses than other gene family members (Table S8).

Functional classification of intronless sub-family genes in *Arabidopsis* and rice

By using GO term analysis, we examined whether the intronless genes in the 33 gene families were overrepresented for certain functional classes of genes. Intronless genes in *Arabidopsis* and rice were mainly enriched in terms related to regulation, signaling, modification, and metabolism,

especially in *AP2*, *EF-hand_7, bZIP*, *FAD_binding_4*, *STE_STE11*, *CAMK_CAMKL-CHK1*, and *C2* gene family (Figure S3, 4). In particular, the enrichment of these genes was found to be associated with the biological process "transcriptional regulation". Besides, the GO terms of intronless genes in different genes families varied greatly. Intronless genes in *B_lectin* and *S_locus_glycop* gene family were enriched in GO terms of epigenetic processes, such as "protein phosphorylation", and in GO terms of development, "recognition of pollen". In addition, intronless genes in *Inhibitor_I9*, *PA*, *Peptidase_S8* gene family were enriched in GO terms, "negative regulation of catalytic activity", and "proteolysis".

Discussion

Since their discovery (Gilbert, 1978), intron evolution has been widely studied (Da *et al.*, 2013, Verhelst *et al.*, 2013). Within the green plant lineage, it has been clearly shown that introns are abundant in algal genes (Baier *et al.*, 2018). On the other hand, in extant land plants, many gene families contain intron-poor or even intronless sub-families (Zhu *et al.*, 2016, Liu *et al.*, 2017). However, there is little information on when and how these intronless and intron-poor sub-families originated and evolved and what their potential functions are.

Intron-poor sub-families originated early in the evolution of green plants

After surveying genomes of representative species, it is clear that intronless or intron-poor sub-families of genes exist in the plant kingdom. Here, we analyzed in greater detail 33 gene families possessing intron-poor and intronless sub-families with at least five intronless gene members, which is a small number relatively to all gene families (33/10,412) in *Arabidopsis*. These 33 gene families belong to different categories of genes, including transcription factors, transcription regulators, transporters, protein kinase, and others, suggesting their potentially broad involvement in plant growth and development.

Over half of the gene families containing intron-poor sub-families originated in the land plants (e.g. *P. patens*), which evolved in the Ediacaran to middle Ordovician at 559.3–459.9 Ma

(Foster *et al.*, 2016, Morris *et al.*, 2018, Yuan *et al.*, 2019). There are several hypotheses on how plants transitioned to life on land and adapted to terrestrial environments and different biotic and abiotic stresses (Pierrehumbert *et al.*, 2011, Ruhfel *et al.*, 2014, Prave *et al.*, 2016). The origination of intron-poor and intronless sub-families might have been a factor in adaptation as the intron-poor and intronless genes can be transcribed with fewer and without splicing, offering fast responses to drastic climatic changes, although at this point this is purely speculative. The evolution of intron-poor and intronless genes may also have facilitated responding to changing developmental processes as their appearance may be also associated with vascular system and seed development, which allowed the evolution of plants with higher plant complexity (Rensing *et al.*, 2008). For instance, the *AP2* gene family, containing intron-poor sub-families, first appeared in the green algae. Previous reports suggested that *AP2* genes play broad roles in developmental regulation in reproductive, vegetative organs, and lateral organ development by influencing cell number and growth (Riechmann and Meyerowitz, 1998, Shigyo *et al.*, 2006) and stress responses (Mizoi *et al.*, 2012). However, the *CIPK* intron-poor genes first appeared in seed plants, which has been reported to participate in the Ca^{2+} signaling process (Zhu *et al.*, 2016).

Evolution of intron-poor and intronless sub-families

Despite the mechanism still being unknown, it is quite clear that intron-poor gene sub-families originated from intron-rich sub-family genes. The question is whether intronless genes derived directly from intron-rich genes or progressively from an intron-poor gene in the intron-poor sub-family. Our results clearly suggest the latter, based on the smaller *Ks* values of intronless sub-families than those of intron-poor sub-families, as also shown previously (Zhong *et al.*, 2018, Zhu *et al.*, 2018) . It is interesting to note that intronless genes in intron-poor sub-families evolved multiple times and also got lost in several lineages (Table S4). Although intron gain and loss occurred often during the evolution of plants, especially in land plants (Roy and Penny, 2007), the evolution of entire sub-families of intronless genes has not been examined in detail before**.**

How the intron-poor and intronless gene sub-families originated remains unclear. Intron gain

and loss might be correlated with transposable element activity (Scott William and Walter, 2005). The loss of introns is likely dependent on recombination with reverse transcribed copies of spliced mRNAs. This pattern is based on reverse-transcribed mRNA-mediated loss. The intron loss event supports a model where intron loss is mediated via germline recombination with an intronless cDNA of the gene. The intron loss events occur as a result of recombination with nearly complete cDNA (Coulombe-Huntington and Majewski, 2007).

In general, intron loss tends to be particularly the case in genes with housekeeping functions and that are expressed at relatively high levels (Zhang *et al.*, 2003, Zou *et al.*, 2011). The preferential loss of introns in highly expressed housekeeping genes is also consistent with selection for transcription efficiency favoring the resulting short transcript (Castillo-Davis *et al.*, 2002). Selection pressure and recombination might both contribute to the association of intron loss and expression levels. Selection alone would favor the loss of longer introns (Coulombe-Huntington and Majewski, 2007). The stronger selective pressures on intron-rich genes, compared to intronless genes demonstrated that intron-rich genes were much more likely to lose introns. Moreover, the recombination events that resulted in intron losses and that were inherited to the next generation had the chance to increase in frequency in the population (Coulombe-Huntington and Majewski, 2007).

Since the 4/1-like protein in *Chara braunii* (class Charophyceae) contains no introns, it was proposed that a presumed precursor 4/1-like gene in basal charophytes initially originated by capturing an alpha-helical myosin-like cistron by a retrotransposon and further retrotransposon-dependent transfer to the algal genome (Morozov and Solovyev, 2019). This suggested the activity of retrotransposons in the ancestral species. We detected 22.2% to 38.9% transposed duplicated genes in intronless genes in the seven representative species. However, it remains unclear whether these transposable elements in intronless genes caused the initial emergence of intronless genes.

The intron-poor and intronless sub-families initially appeared as a single intron-reduced, or intronless gene in the early stages of plant life, then expanded through different mechanisms.

Polyploidization (WGD) is widely known to give rise to gene duplicates during evolution (De Bodt *et al.*, 2005, Zwaenepoel and Van de Peer, 2019). A large number of gene duplication events has driven the expansion of intron-poor sub-families. The genome of *P. patens* has undergone at least two WGD events (Clark and Pcj, 2017, Lang *et al.*, 2018). Most sequenced seed plants have undergone one or two WGD events during their evolution (Ruprecht *et al.*, 2017). According to our study, WGD events played an important role in the expansion of intron-poor and intronless genes. Tandem duplicated genes form the second largest group of duplicated intronless genes (Figure 4A).

An apparent stronger purifying selection pressure in intronless duplicated gene pairs, compared to intron-rich duplicated gene pairs, suggests that intronless duplicates are subject to higher selection pressure than intron-rich genes. This, together with the functions ascribed to intronless genes suggests that intronless genes may play some important roles in plant growth, development or response to biotic or abiotic stresses.

Significant numbers of intronless sub-family genes respond to drought and salt stress in *Arabidopsis* and rice

Previous studies revealed that intronless genes are highly induced by stress (Kousuke *et al.*, 2008). It is unknown which (or any) decisive events have driven the origination of intron loss or the generation of some sub-families of intronless genes, or whether such generation of intronless sub-family genes were the outcome of responses to some other environmental or developmental changes. Our meta-analysis of current RNA-seq data in multiple studies in both *Arabidopsis* and rice clearly showed that the intronless sub-family genes might play significant roles in responding to drought and salt stresses. The most significant DEGs, both in drought and salt related studies in *Arabidopsis* and rice were found in *AP2*, *EF-hand_7, bZIP*, *FAD_binding_4*, *STE_STE11*, *CAMK_CAMKL-CHK1*, and *C2* gene families. These intronless sub-family genes were most enriched in the biological process “transcriptional regulation”, which function drought responsive pathways (Liu *et al.*, 2018), suggesting that they play significant roles in responding to drought

and salt stresses. On the other hand, intronless sub-family genes in other gene families were enriched in GO terms like "protein phosphorylation", "negative regulation of catalytic activity", and "transmembrane transport", which may suggest involvement in different biological processes, like epigenetic processes and signal transduction.

Stress-related genes have been shown to have contributed to the terrestrial adaptation of green plants subjected to many kind stress factors such as brighter sunlight (UV radiation), lack of efficient support against gravity, dehydration, and high carbon dioxide levels in the atmosphere (Fang *et al.*, 2017, Morozov and Solovyev, 2019). Since relatively few genes have been characterized in detail, it requires additional experiments to specifically determine the functions of intronless sub-family genes in stress response and other developmental processes.

Materials and Methods

Sequence Data

The genome sequences, as well as their annotations, were downloaded from public databases such as Phytozome (version 12.1, http://www.phytozome.net/) (Goodstein *et al.*, 2012) and the NCBI (National Center for Biotechnology Information https://www.ncbi.nlm.nih.gov). We selected seven species to represent land plants, namely *Arabidopsis thaliana* (eudicots) (TAIR10), *Oryza sativa* (monocots) (v7.0), *Amborella trichopoda* (an early diverging angiosperm) (v1.0), *Ginkgo biloba* (gymnosperm, seed plant) (v1.0), *Selaginella moellendorffii* (a vascular plant) (v1.0), and *Physcomitrella patens* (a moss) (v3.3), and one alga, *Chlamydomonas reinhardtii* (green algae) (v5.5). The protein kinase data in the seven species were obtained from the latest iTAK release (17.09, http://itak.feilab.net/cgi-bin/itak/db_home.cgi) (Zheng *et al.*, 2016).

Prediction and classification of intronless genes

Intronless genes of the seven species were obtained by surveying the General Feature Format Version 3 (GFF3) files from Phytozome v. 12.1. All candidate intronless genes were further submitted to Pfam (http://pfam.xfam.org/) to verify their gene structure and the presence of

domains (Finn *et al.*, 2014). To improve the accuracy of intronless genes identification, all the genes were submitted to the online software tool SMART ((http://smart.embl-heidelberg.de/) for further confirmation . A Hidden Markov Model (HMM) was used to ascribe genes to gene families using HMMER v. 3.0 with an E-value cut-off of <1.0 (Liu *et al.*, SR, 1998). We considered five intronless genes in each gene family as a reasonable cut-off for further study while putative gene families with fewer than five intronless genes were discarded. Finally, genes for which there was high confidence for being 'true' genes were assigned to each gene family in the seven species.

Multiple sequence alignment, phylogenetic analysis and exon-intron structure analysis

The full-length protein sequences of each gene family were aligned by MEGA 7 using the MUSCLE program (Kumar *et al.*, 2008). Phylogenetic trees were constructed using maximum likelihood (ML) as implemented in the FastTree software (Chen *et al.*, 2018b).

The online Gene Structure Display Server (GSDS: http://gsds.cbi.pku.edu.cn) was used to generate the exon-intron structures of all candidate genes by DNA and cDNA sequences (Guo *et al.*, 2007). The software TBtools (version 0.58) was exploited to show exon-intron structures for each gene family based on GFF3 files (Chen *et al.*, 2018a).

Detection of duplication types

BLASTP (*E-value* < $1e-10$) was used to search for potential homologs amongst all protein sequences for each gene family in the seven species (Mahram and Herbordt, 2010). To investigate gene synteny, blast hits and gene locations were used as inputs for MCScanX (Multiple Collinearity Scan toolkit), using default settings (Wang *et al.*, 2013). Genes and duplicates were ascribed to one of five different duplication categories: singleton, tandem, proximal, Whole Genome Duplication (WGD) or segmental, and dispersed (Wang *et al.*, 2013). Transposed duplicates of intronless genes in the intron-poor sub-families were searched for in the Plant Duplicate Gene Database (http://pdgd.njau.edu.cn:8080) (Qiao *et al.*, 2019).

Calculation of *Ka* and *Ks* values

The nucleotide sequences of CDSs in each gene family were aligned based on the protein sequence with ClustalW in MEGA 7.0, using default settings (Larkin *et al.*, 2007). Nonsynonymous substitutions (*Ka*) and synonymous substitutions (*Ks*) and nonsynonymous to synonymous substitution ratios (*Ka/Ks*) were estimated in each gene family by the use of Perl scripts (Zhong *et al.*, 2018).

Estimation of the GC content

The GC content was calculated by counting G and C bases as percentages of the total number of bases in sequences with intronless, intron-poor, and intron-rich genes, in each gene family. To this end, the online software tool EMBOSS (http://www.bioinformatics.nl/cgi-bin/emboss/geecee) was used (Rice *et al.*, 2000, Singh *et al.*, 2016).

Chromosomal location analysis

The chromosomal locations of intronless sub-family genes in *Arabidopsis* and rice were retrieved from annotation information as available in the genome database Phytozome (version 12.1, http://www.phytozome.net/). The intronless sub-family genes in the two species were mapped to chromosomes with MapChart (Zhu *et al.*, 2018).

RNA-seq data and quantification

Single- and/or paired-end RNA-seq reads were downloaded from NCBI SRA (https://www.ncbi.nlm.nih.gov/sra) (Team, 2011). RNA-seq samples of *Arabidopsis* and rice were obtained from different tissues and developmental stages under drought and salt stress conditions (Table S7). The data were filtered using Trim_galore, a high throughput sequence quality confine analysis tool (Brown *et al.*, 2017, Wang *et al.*, 2018). Then, the filtered reads were mapped to the *Arabidopsis* and rice reference genomes using HISAT2 (Yang-Ming *et al.*, 2016). The reads of

each gene were counted by Subread-featureCounts (Liao *et al.*, 2013), using default parameters. The differentially expressed genes (DEGs) were then identified using the edgeR package (Nikolayeva, 2014). A false discovery rate (FDR) ≤0.01 and an absolute value of the | logFC | ≥1.5 were used as thresholds to evaluate the significance of gene expression differences. We compared expressions of all control to all treatments. We only investigated DEGs between intronless genes and intron-rich genes in families that contain intron-poor sub-families.

Gene ontology (GO) analysis

To determine the possible functions of intronless, intron-poor and intron-rich genes in each gene family, gene ontology (GO) annotations for *Arabidopsis* and rice were retrieved from the Gene Ontology Consortium (http://www.geneontology.org/) Website (Consortium, 2004).

Statistical analysis

A student's *t*-test ($P<0.01$ and $P<0.05$) was performed, where appropriate, using the IBM SPSS Statistics v25 software (SPSS, Inc., USA) (Dunn, 2013).

Data availability statement

All relevant data are included in the manuscript and its supporting materials.

Acknowledgements

This work was supported by the open funds of the State Key Laboratory of Crop Genetics and Germplasm Enhancement (ZW201813), the Priority Academic Program Development of Jiangsu Higher Education Institutions, and supported by the Bioinformatics Center of Nanjing Agricultural University. YVdP acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 833522).

We thank Dr. Wei Qu, Dr. Jin-Song Xiong, and Dr. Yan Zhong for bioinformatics assistance.

Authors' contributions

HL and ZMC designed and initiated this study. HL and HML carried out the bioinformatics analyses. HL wrote the manuscript. KKZ, YVdP and ZMC critically revised the manuscript. All authors read and approved the final manuscript.

Conflicts of interests

The authors declare that they have no competing interests.

Supporting information legends

Table S1 The gene ID of the intronless genes in *Arabidopsis* and rice. A total of 6308 and 10314 intronless genes were identified and assigned according to species.

Table S2 List of 149 gene families contain at least five intronless genes in *Arabidopsis* genome.

Table S3 List of the number of intron-poor and intronless genes in intron-poor sub-families of 33 gene families in the species of *C. reinhardtii*, *P. patens*, *S. moellendorffii*, *G. biloba*, and *A. trichopoda*, *A. thaliana*, and *O. sativa*.

Table S4 List of the number of intronless genes in intron-poor sub-families of 33 gene families in the species of *C. reinhardtii*, *P. patens*, *S. moellendorffii*, *G. biloba*, and *A. trichopoda*, *A. thaliana*, and *O. sativa*.

Table S5 List of the number of orthologs and paralogs in intronless genes in each gene family and their average *Ks* values.

Table S6 List of the *Ka/Ks* values in gene pairs among 33 gene families of

intron-poor sub-families and intron-rich sub-families in the seven species.

Table S7 List of the RNA-seq studies in *A. thaliana*, and *O. sativa.*

Table S8 List of the intronless DEGs of *A. thaliana*, and *O. sativa.*

Figure S1 A: Distribution of the identified intronless genes in 33 genes families on 5 chromosomes of *Arabidopsis*. **B:** Distribution of the identified intronless genes in 33 genes families on 12 chromosomes of rice.

Figure S2 The percentage of intronless duplicated genes and the all duplicates of different duplication types among the seven species.

Figure S3 The relative abundance of GO functions of intronless, intron-poor and intron-rich genes in each gene family in *Arabidopsis* and their detail classifications. The (m) represents molecular function. The (c) represents cellular component. The (b) represents biological process.

Figure S4 The relative abundance of GO functions of intronless, intron-poor and intron-rich genes in each gene family in rice and their detail classifications. The (m) represents molecular function. The (c) represents cellular component. The (b) represents biological process.

References

Baier, T., Wichmann, J., Kruse, O. and Lauersen, K.J. (2018) Intron-containing algal transgenes mediate efficient recombinant gene expression in the green microalga Chlamydomonas reinhardtii. *Nucleic acids research*, **46**, 6909-6919.

Brown, J., Pirrung, M. and Mccue, L.A. (2017) FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics*, **33**.

Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V. and Kondrashov, F.A. (2002) Selection for short introns in highly expressed genes. *Nature Genetics*, **31**, 415.

Chen, C., Xia, R., Chen, H. and He, Y. (2018a) TBtools, a Toolkit for Biologists integrating various HTS-data handling tools with a user-friendly interface. *bioRxiv*, 289660.

Chen, F., Zhang, L., Lin, Z. and Cheng, Z.M.M. (2018b) Identification of a novel fused gene family implicates convergent evolution in eukaryotic calcium signaling. *Bmc Genomics*, **19**, 306.

Clark, J.W. and Pcj, D. (2017) Constraining the timing of whole genome duplication in plant evolutionary history. *Proc Biol Sci*, **284**, 20170912.

Consortium, G.O. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, **32**, D258-D261.

Coulombe-Huntington, J. and Majewski, J. (2007) Characterization of intron loss events in mammals. *Genome Research*, **17**, 23-32.

Da, L., Jean-Luc, Janecek and Stefan (2013) Gene make-up: rapid and massive intron gains after horizontal transfer;of a bacterial alpha-amylase gene to Basidiomycetes. *BMC Evolutionary Biology,13,1(2013-02-13)*, **13**, 40-40.

De Bodt, S., Maere, S. and Van de Peer, Y. (2005) Genome duplication and the origin of angiosperms. *Trends in ecology & evolution*, **20**, 591-597.

Dunn, P. (2013) SPSS survival manual: a step by step guide to data analysis using IBM SPSS. *Australian & New Zealand Journal of Public Health*, **37**, 597-598.

Fang, H., Huangfu, L., Chen, R., Li, P., Xu, S., Zhang, E., Cao, W., Li, L., Yao, Y. and Liang, G. (2017) Ancestor of land plants acquired the DNA-3-methyladenine glycosylase (MAG) gene from bacteria through horizontal gene transfer. *Scientific Reports*, **7**, 9324.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L. and Mistry, J. (2014) Pfam: the protein families database. *Nucleic Acids Research*, **42**, 222-230.

Foster, C.S., Sauquet, H., Van der Merwe, M., McPherson, H., Rossetto, M. and Ho, S.Y. (2016) Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Systematic Biology*, **66**, 338-351.

Gilbert, W. (1978) Why genes in pieces? *Nature*, **271**, 501-501.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U. and Putnam, N. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, **40**, D1178-D1186.

Guo, A.Y., Zhu, Q.H., Chen, X. and Luo, J.C. (2007) GSDS: a gene structure display server. *Hereditas*, **29**, 1023-1026.

Jurica, M.S. and Moore, M.J. (2003) Pre-mRNA Splicing : Awash in a Sea of Proteins. *Molecular Cell*, **12**, 5-14.

Kousuke, H., Cheng, Z., Lehti-Shiu, M.D., Kazuo, S. and Shin-Han, S. (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiology*, **148**, 993-1003.

Kumar, S., Nei, M., Dudley, J. and Tamura, K. (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics*, **9**, 299-306.

Lang, D., Ullrich, K.K., Murat, F., Fuchs, J., Jenkins, J., Haas, F.B., Piednoel, M., Gundlach, H., Van Bel, M. and Meyberg, R. (2018) The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant Journal*, **93**, 515-533.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., Mcgettigan, P.A., Mcwilliam, H.,

Valentin, F., Wallace, I.M., Wilm, A. and Lopez, R. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947-2948.

Liao, Y., Smyth, G.K. and Shi, W. (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, **41**, e108.

Liu, H., Zhong, Y., Guo, C., Wang, X.L., Xiong, J., Cheng, Q. and Cheng, Z.M. (2017) Genome-wide analysis and evolution of the bZIP transcription factor gene family in six Fragaria species. *Plant Systematics & Evolution*, 1-13.

Liu, J., Chen, N., Chen, F., Cai, B., Santo, S.D., Tornielli, G.B., Pezzotti, M. and Cheng, Z.M. (2014) Genome-wide analysis and expression profile of the bZIP transcription factor gene family in grapevine (Vitis vinifera). *Bmc Genomics*, **15**, 281-281.

Liu, X., Ting, w., Ezra, B., Kezia, B., Mingming, D., Yaqi, Z., Sen, Y., Yanling, C., Shudan, X. and Yiqun, W. Comprehensive analysis of NAC transcription factors and their expression during fruit spine development in cucumber (Cucumis sativus L.). *Horticulture Research*, **5**, 31-.

Liu, Z., Qin, J., Tian, X., Xu, S., Wang, Y., Li, H., Wang, X., Peng, H., Yao, Y. and Hu, Z. (2018) Global profiling of alternative splicing landscape responsive to drought, heat and their combination in wheat (Triticum aestivum L.). *Plant Biotechnology Journal*, **16**, 714-726.

Mahram, A. and Herbordt, M.C. (2010) Fast and accurate NCBI BLASTP:acceleration with multiphase FPGA-based prefiltering. In *International Conference on Supercomputing, 2010, Tsukuba, Ibaraki, Japan, June*, pp. 73-82.

McDonell, L. and Drouin, G. (2012) The abundance of processed pseudogenes derived from glycolytic genes is correlated with their expression level. *Genome*, **55**, 147-151.

Min, X.J., Powell, B., Braessler, J., Meinken, J., Yu, F. and Sablok, G. (2015) Genome-wide cataloging and analysis of alternatively spliced genes in cereal crops. *Bmc Genomics*, **16**, 1-13.

Mizoi, J., Shinozaki, K. and Yamaguchi-Shinozaki, K. (2012) AP2/ERF family transcription

factors in plant abiotic stress responses. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, **1819**, 86-96.

Morozov, S. and Solovyev, A. (2019) Emergence of intronless evolutionary forms of stress response genes: possible relation to terrestrial adaptation of green plants. *Frontiers in plant science*, **10**, 83.

Morris, J.L., Puttick, M.N., Clark, J.W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Yang, Z., Schneider, H. and Donoghue, P.C. (2018) The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences*, **115**, E2274-E2283.

Nikolayeva, O. (2014) edgeR for Differential RNA-seq and ChIP-seq Analysis: An Application t. *Methods Mol Biol*, **1150**, 45-79.

Nilsen, T.W. (2003) The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, **25**, 1147-1149.

Pierrehumbert, R.T., Abbott, D.S., Voigt, A. and Koll, D.D.B. (2011) Climate of the Neoproterozoic. *Annual Review of Earth and Planetary Sciences*, **39**, 417-460.

Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L. and Carter, D.R.F. (2011) Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA*, **17**, 792-798.

Prave, A.R., Condon, D.J., Hoffmann, K.H., Tapster, S. and Fallick, A.E. (2016) Duration and nature of the end-Cryogenian (Marinoan) glaciation. *Geology*, **44**, 631-634.

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S. and Paterson, A.H. (2019) Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biology*, **20**, 38-38.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E.A. and Kamisugi, Y. (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64-69.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276-277.

Riechmann, J.L. and Meyerowitz, E.M. (1998) The AP2/EREBP family of plant transcription factors. *Biological Chemistry*, **379**, 633-646.

Rogozin, I.B. (2012) Origin and evolution of spliceosomal introns. *Biology Direct*, **7**, 11-11.

Roy, S.W. and Penny, D. (2007) Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of O. sativa and A. thaliana. *Molecular Biology & Evolution*, **24**, 171-181.

Ruhfel, B.R., Gitzendanner, M.A., Soltis, P.S., Soltis, D.E. and Burleigh, J.G. (2014) From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology*, **14**, 23-23.

Ruprecht, C., Lohaus, R., Vanneste, K., Mutwil, M., Nikoloski, Z., Van de Peer, Y. and Persson, S. (2017) Revisiting ancestral polyploidy in plants. *Science Advances*, **3**, e1603195.

Sakharkar, M.K., Yu, L., Chaturvedi, I. and Peng, L. (2007) A Tale of Intronless Genes in Eukaryotic Genomes. In *Biomedical and Pharmaceutical Engineering, 2006. ICBPE 2006. International Conference on*.

Scott William, R. and Walter, G. (2005) Rates of intron loss and gain: implications for early eukaryotic evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 5773-5778.

Shigyo, M., Hasebe, M. and Ito, M. (2006) Molecular evolution of the AP2 subfamily. *Gene*, **366**, 256-265.

Singh, R., Ming, R. and Yu, Q. (2016) Comparative Analysis of GC Content Variations in Plant Genomes. *Tropical Plant Biology*, **9**, 1-14.

SR, E. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755.

Team, G.E. (2011) Closure of the NCBI SRA and implications for the long-term future of genomics data storage. *Genome Biology*, **12**, 402.

Tutar, Y. (2012) Pseudogenes. *Comparative and functional genomics*, **2012**.

Verhelst, B., Peer, Y.V.D. and Rouzé, P. (2013) The Complex Intron Landscape and Massive

Intron Invasion in a Picoeukaryote Provides Insights into Intron Evolution. *Genome Biology and Evolution,5,12(2013-11-22)*, **5**, 2393-2401.

Wang, H., Chang, X., Lin, J., Chang, Y., Chen, J.-C., Reid, M.S. and Jiang, C.-Z. (2018) Transcriptome profiling reveals regulatory mechanisms underlying corolla senescence in petunia. *Horticulture research*, **5**, 1-13.

Wang, Y., Li, J. and Paterson, A.H. (2013) MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. *Bioinformatics*, **29**, 1458-1460.

Yang-Ming, S.I., Xing, Y.Q. and Cai, L. (2016) Differential splicing event analysis of liver tumor-educated blood platelets RNA-seq data with Hisat2 and MISO. *Journal of Inner Mongolia University of Science & Technology*.

Yuan, N., P, F.C.S., Tianqi, Z., Ru, Y., A, D.D., W, H.S.Y. and Bojian, Z. (2019) Accounting for Uncertainty in the Evolutionary Timescale of Green Plants Through Clock-Partitioning and Fossil Calibration Strategies. *Systematic Biology*, 1.

Zhang, Z., Harrison, P.M., Liu, Y. and Gerstein, M. (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Research*, **13**, 2541.

Zheng, Y., Jiao, C., Sun, H., Rosli, H.G., Pombo, M.A., Zhang, P. and Banf, M. (2016) iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors,Transcriptional Regulators, and Protein Kinases. *Molecular Plant*, **9**, 1667-1670.

Zhong, Y., Zhang, X. and Cheng, Z.M. (2018) Lineage-specific duplications of NBS-LRR genes occurring before the divergence of six Fragaria species. *Bmc Genomics*, **19**, 128.

Zhu, K., Fei, C., Liu, J., Chen, X., Hewezi, T. and Cheng, Z.M. (2016) Evolution of an intron-poor cluster of the CIPK gene family and expression in response to drought stress in soybean. *Scientific Reports*, **6**, 28225.

Zhu, K., Wang, X., Liu, J., Tang, J., Cheng, Q., Chen, J.G. and Cheng, Z.M. (2018) The grapevine kinome: annotation, classification and expression patterns in developmental processes and stress responses. *Hortic Res*, **5**, 19.

Zou, M., Guo, B. and He, S. (2011) The Roles and Evolutionary Patterns of Intronless Genes in Deuterostomes. *Comparative & Functional Genomics*, **2011**, 680673.

Zwaenepoel, A. and Van de Peer, Y. (2019) Inference of Ancient Whole-Genome Duplications and the Evolution of Gene Duplication and Loss Rates. *Molecular biology and evolution*, **36**, 1384-1404.

Figure/Table legends

Figure 1 Schematic representation of a gene family (CBL-interacting protein kinases CIPKs (A)) with intronless and intron-poor (3 or less introns per gene) genes. Number of genes and intronless genes (B) in the seven species representative used in the current study (C).

Figure 2 Number of gene families with intron-poor gene sub-families present in *C. reinhardtii*, *P. patens*, *S. moellendorffii*, and *G. biloba*.

Figure 3 The number of intronless genes in intron-poor gene sub-families first appeared in the species of *C. reinhardtii*, *P. patens*, *S. moellendorffii*, and *G. biloba.* The circle size represents the number of gene families.

Figure 4 A: The percentage of intronless duplicated genes in each gene family of different duplication types among the seven species. B: The *Ks* ranges of intronless duplicated genes in WGD, segmental, and tandem duplication type in the 33 gene families among the seven species.

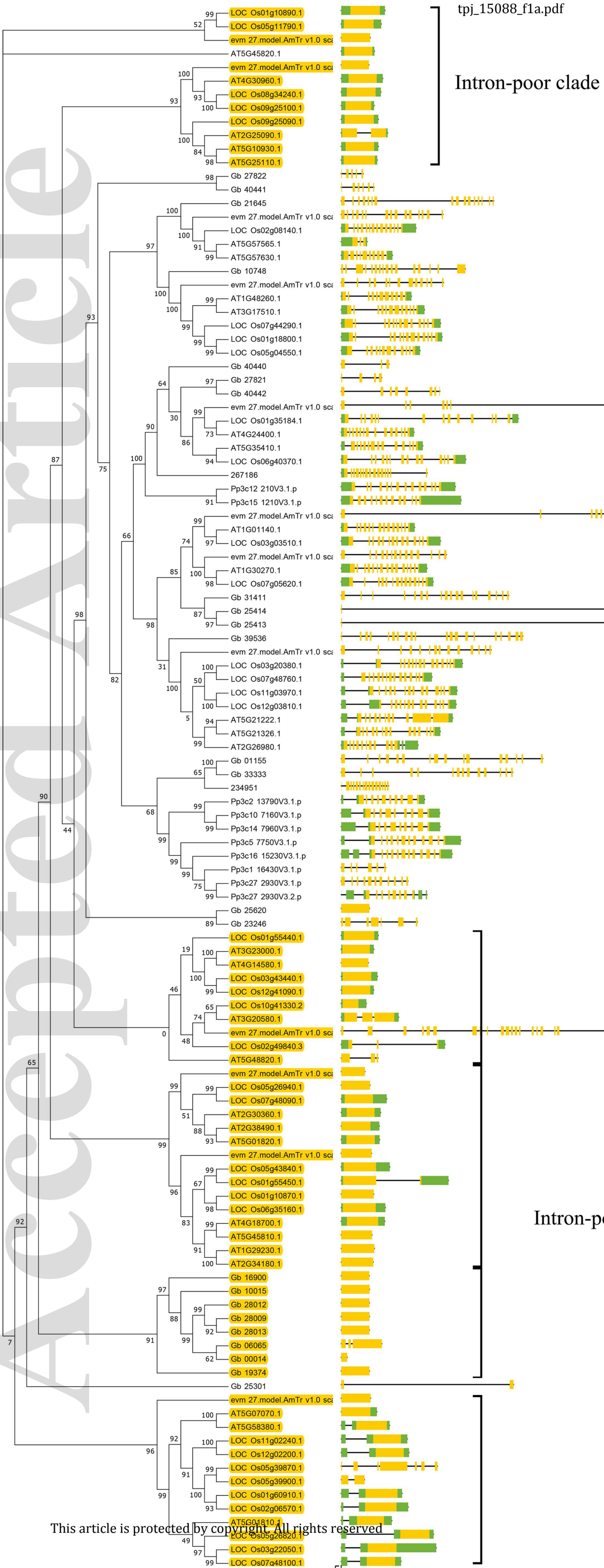
Figure 5 A: *Ks* range of intronless, intron-poor and intron-rich genes in the seven species investigated in this study. B: The *Ka/Ks* ratios of intronless,

intron-poor and intron-rich genes. See text for details.

Figure 6 GC-content (%) for intronless, intron-poor, and intron-rich genes in 33 gene families for seven representative plant species.

Figure 7 A: The ratios of DEGs in intronless, intron-poor and intron-rich genes under drought treatments in *Arabidopsis* and rice. B: The ratios of DEGs in intronless, intron-poor and intron-rich genes under salt treatments in *Arabidopsis* and rice.

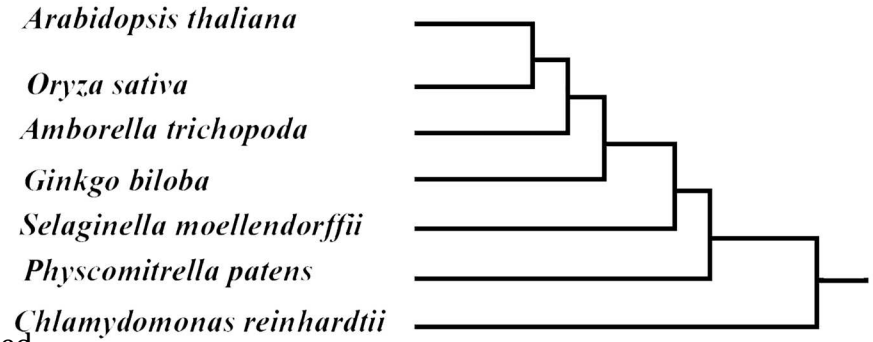
tpj_15088_f1a.pdf

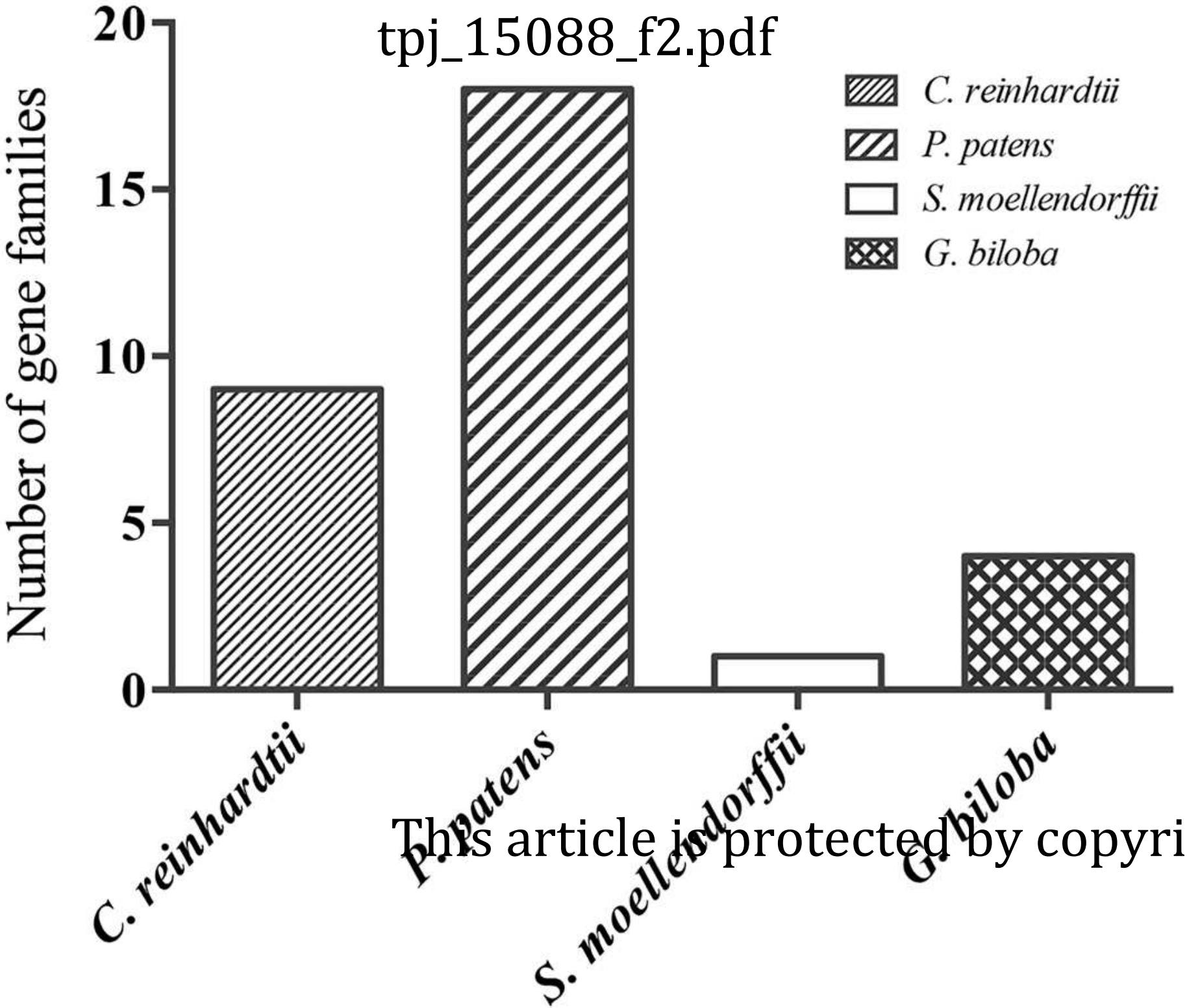


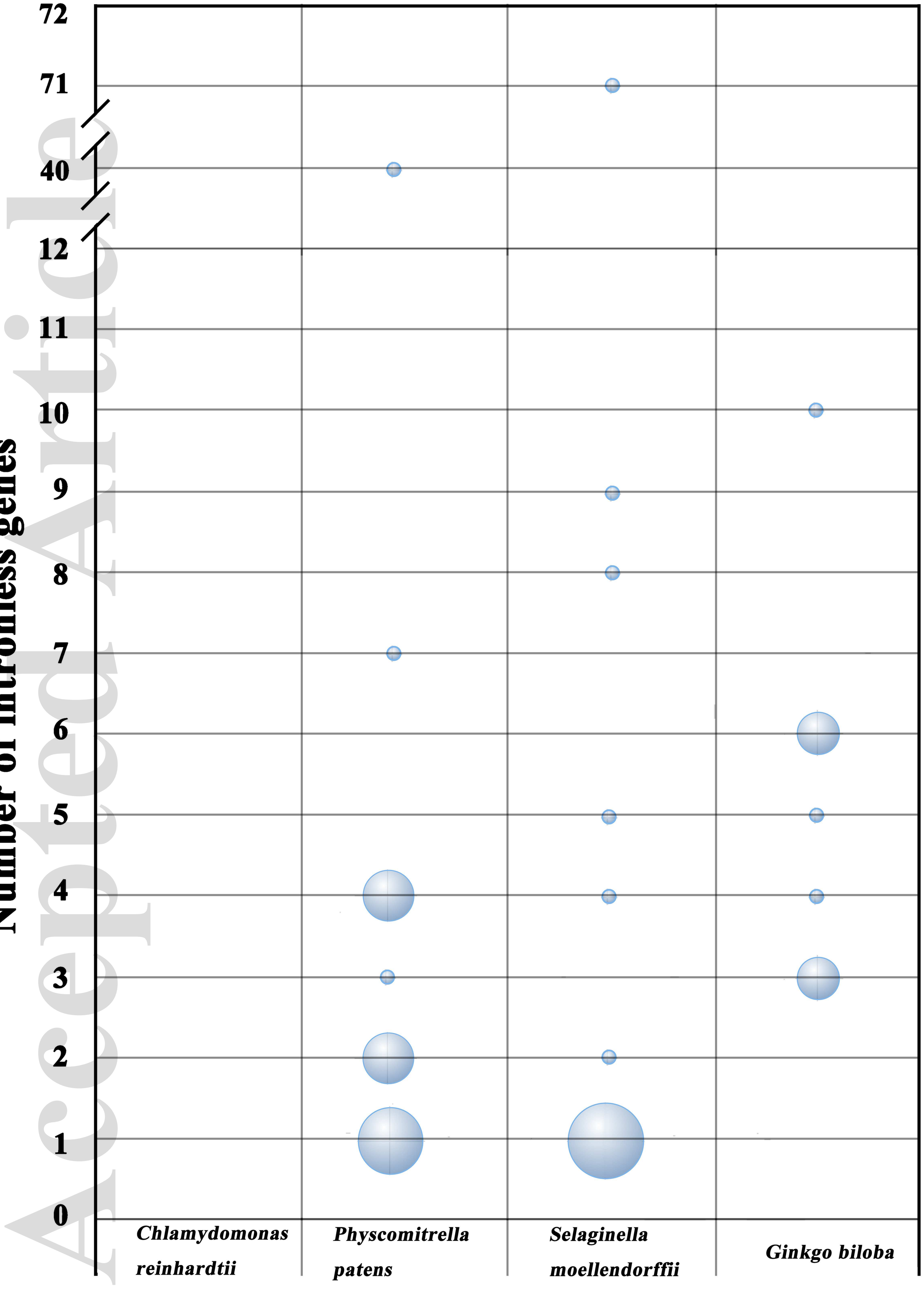
B

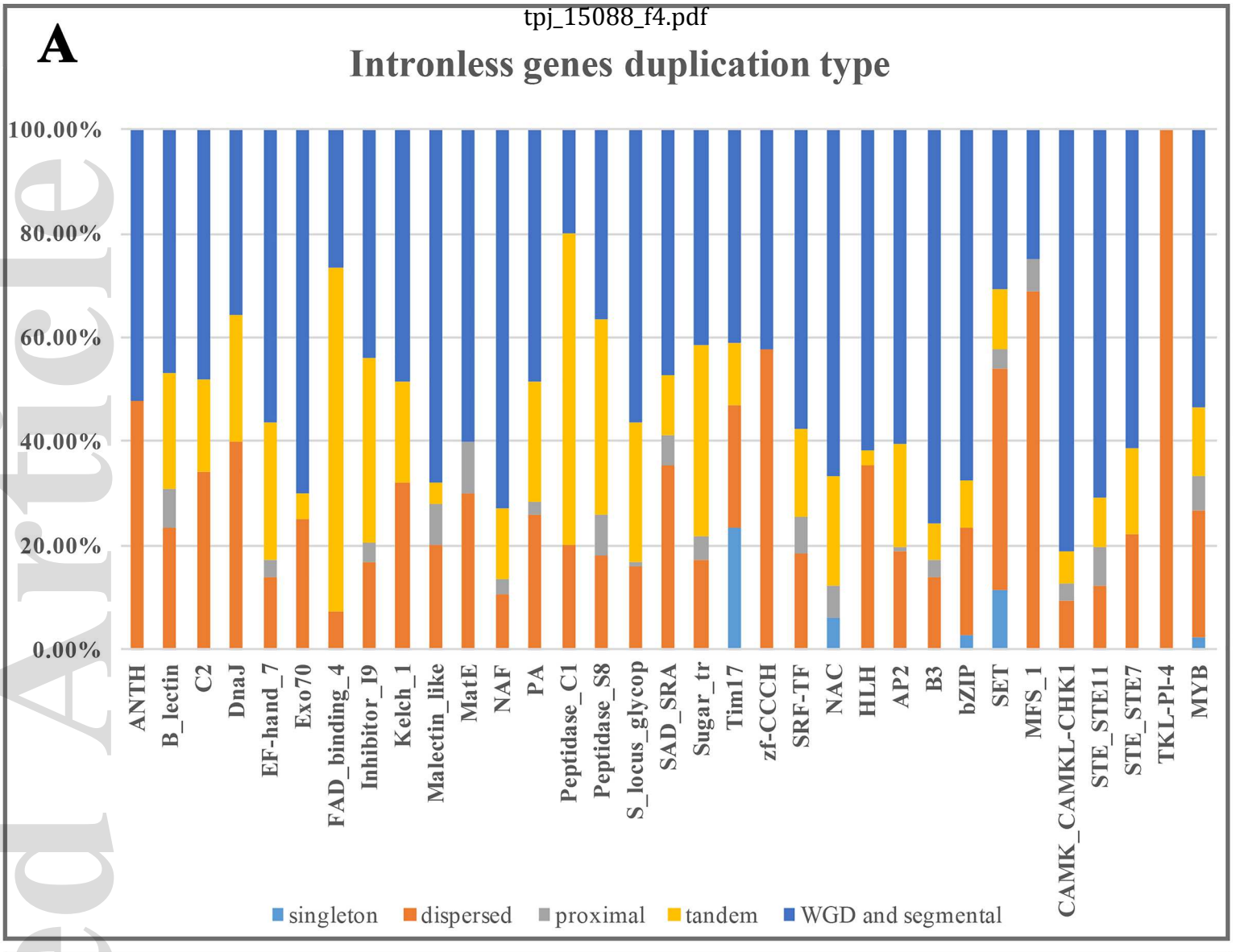
Number of Genes	Number of Intronless Genes	Percentage
27416	6308	23.01%
42189	10320	24.46%
26846	5659	21.08%
41840	10894	26.04%
22273	3484	15.64%
32926	11298	34.31%
17744	1073	6.05%

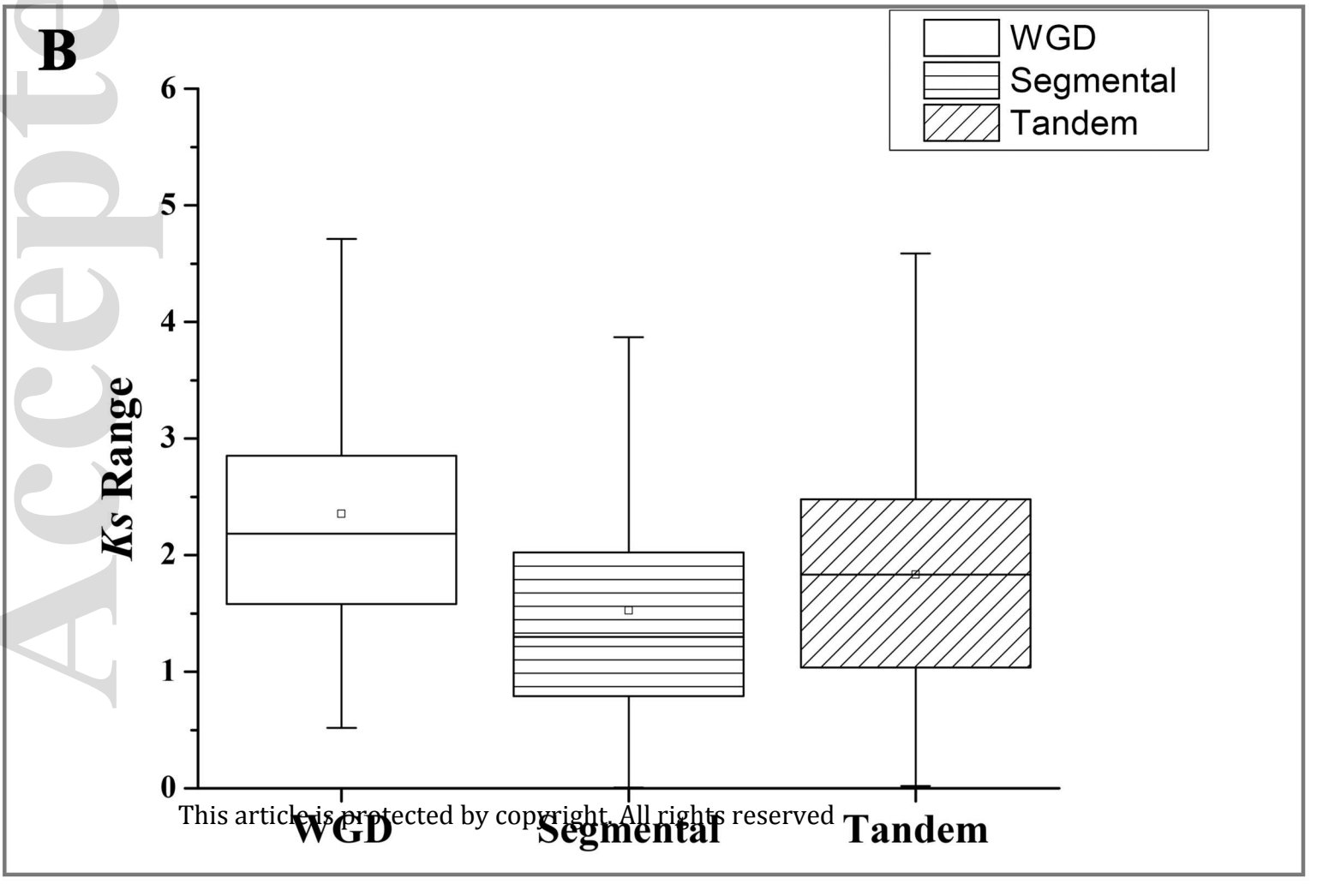
C

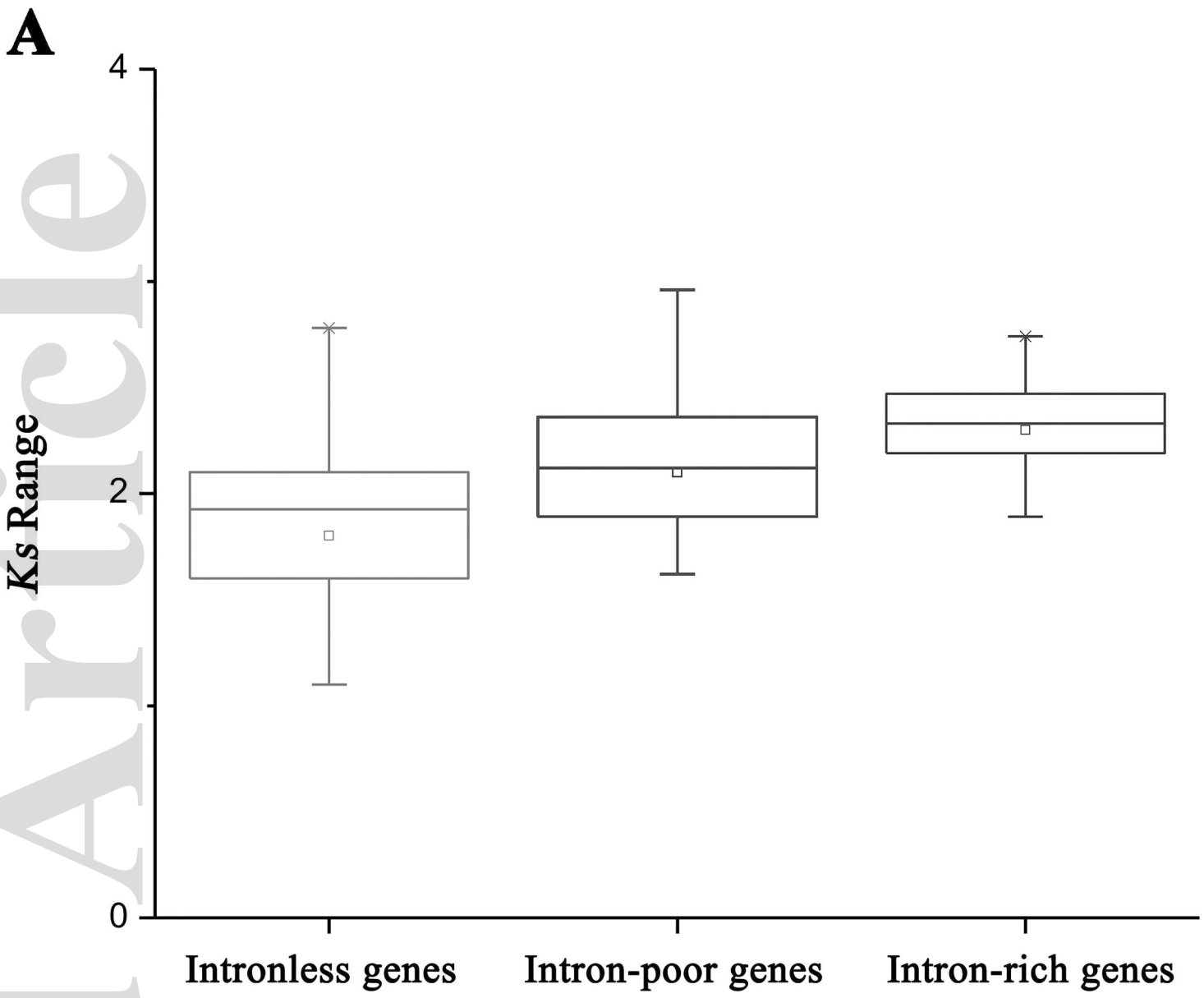


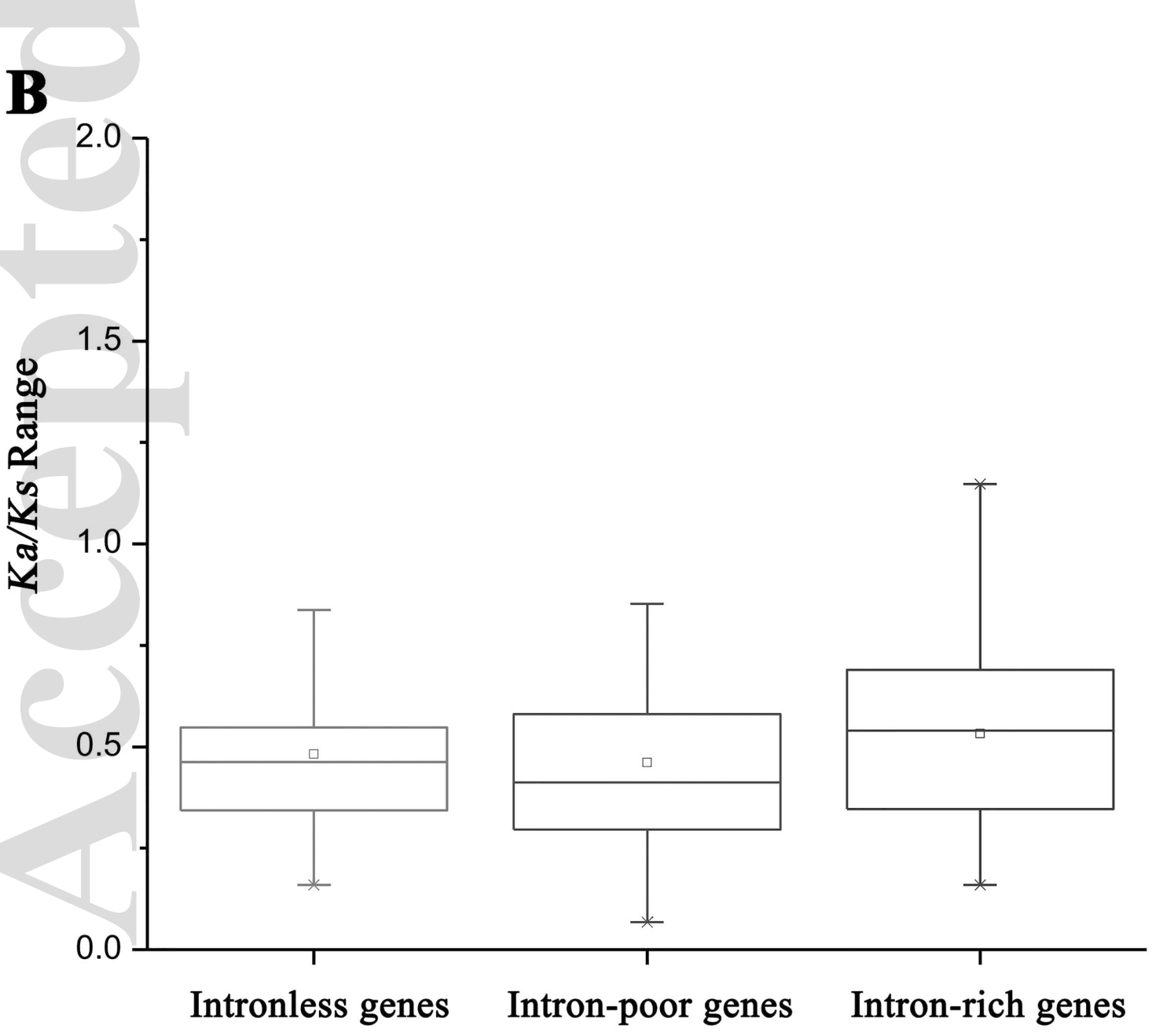




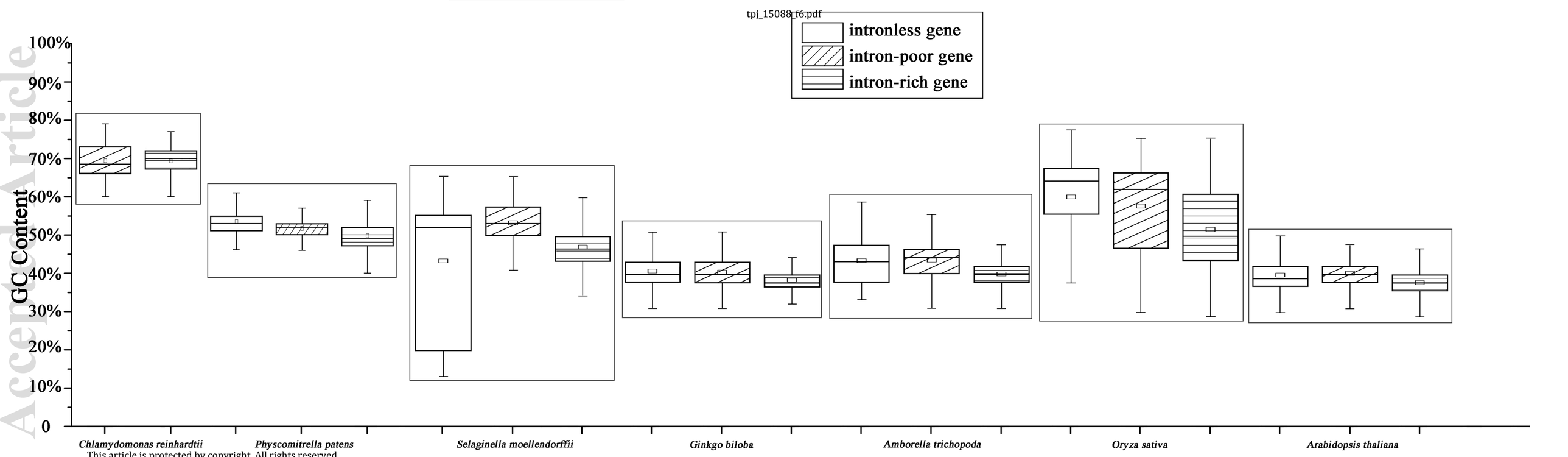








tpj_15088_f6.pdf



tpj_15088_f7.pdf