A peer-reviewed version of this preprint was published in PeerJ on 26 July 2016.

Ma M, Lan X, Niu D. 2016. Intron gain by tandem genomic duplication: a novel case in a potato gene encoding RNA-dependent RNA polymerase. PeerJ 4:e2272 https://doi.org/10.7717/peerj.2272

Intron gain by tandem genomic duplication: a novel case and a modification of the traditional model

Ming-Yue Ma, Deng-Ke Niu

Origin and subsequent accumulation of spliceosomal introns are prominent events in the evolution of eukaryotic gene structure. Recently gained introns would be especially useful for the study of the mechanism(s) of intron gain because the evolutionary traces might have not been erased by randomly accumulated mutations. However, the mechanism(s) of intron gain remain unclear due to the presence of a few solid cases. A widely cited model of intron gain is tandem genomic duplication, in which the duplication of an AGGT-containing exonic segment provides the GT and AG splicing sites for the new intron. However, successful recognition and splicing of an intron require many more signals than those at the two splicing sites. We found that the second intron of the potato RNA-dependent RNA polymerase gene *PGSC0003DMG402000361* is absent in the orthologous genes of other Solanaceae plants, and sequence similarity showed that the major part of the new intron is a direct duplication of the 3' side of the upstream intron. In addition to the new intron, a downstream exonic segment of 168bp has also been duplicated. Most of the splicing signals were inherited from the parental intron/exon structure, including a putative branch site, the polypyrimidine tract, the 3' splicing site, two putative exonic splicing enhancers and the GC contents differentiated between the intron and exon. We propose a modified version of the tandem genomic duplication model, termed as the partial duplication of the preexisting intron/exon structure.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1439v1 | CC-BY 4.0 Open Access | rec: 17 Oct 2015, publ: 17 Oct 2015

Intron gain by tandem genomic duplication: a novel case and a modification of the traditional model

Ming-Yue Ma and Deng-Ke Niu*

MOE Key Laboratory for Biodiversity Science and Ecological Engineering and Beijing Key Laboratory of Gene Resource and Molecular Development, College of Life Sciences, Beijing Normal University, Beijing 100875, China

*Corresponding author.

Deng-Ke Niu

No. 19, XinJieKouWai Street, Beijing 100875, China

Email addresses: dengkeniu@hotmail.com; dkniu@bnu.edu.cn

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1439v1 | CC-BY 4.0 Open Access | rec: 17 Oct 2015, publ: 17 Oct 2015

ABSTRACT

Origin and subsequent accumulation of spliceosomal introns are prominent events in the evolution of eukaryotic gene structure. Recently gained introns would be especially useful for the study of the mechanism(s) of intron gain because the evolutionary traces might have not been erased by randomly accumulated mutations. However, the mechanism(s) of intron gain remain unclear due to the presence of a few solid cases. A widely cited model of intron gain is tandem genomic duplication, in which the duplication of an AGGT-containing exonic segment provides the GT and AG splicing sites for the new intron. However, successful recognition and splicing of an intron require many more signals than those at the two splicing sites. We found that the second intron of the potato RNA-dependent RNA polymerase gene *PGSC0003DMG402000361* is absent in the orthologous genes of other Solanaceae plants, and sequence similarity showed that the major part of the new intron is a direct duplication of the 3′ side of the upstream intron. In addition to the new intron, a downstream exonic segment of 168bp has also been duplicated. Most of the splicing signals were inherited from the parental intron/exon structure, including a putative branch site, the polypyrimidine tract, the 3′ splicing site, two putative exonic splicing enhancers and the GC contents differentiated between the intron and exon. We propose a modified version of the tandem genomic duplication model, termed as the partial duplication of the preexisting intron/exon structure.

INTRODUCTION

Although, spliceosomal introns are the characteristic feature of eukaryotic nuclear genes, their origin and subsequent accumulation during evolution remain obscure. There are several purposed models of the spliceosomal introns gain, which include intron transposition, transposon insertion, tandem genomic duplication, insertion of an exogenous sequence during double-strand-break repair, insertion of a group II intron, intron transfer and intronization (Yenerall & Zhou 2012). Comparative analyses of discordant intron positions among conserved homologous genes have been carried out in diverse eukaryotic lineages. Except for a few studies (Fablet et al. 2009; Li et al. 2009; Torriani et al. 2011; van der Burgt et al. 2012; Verhelst et al. 2013), the observed frequency of intron gain has generally been found to be much lower than those of the intron loss, and there is very limited supporting evidence for the intron gain models (Csuros et al. 2011; Hooks et al. 2014; Irimia & Roy 2014; Roy & Gilbert 2005; Roy & Penny 2006; Yenerall et al. 2011; Yenerall & Zhou 2012; Zhu & Niu 2013). Recently, Collemare et al. (2013) claimed that the abundance of introns in extant eukaryotic genomes could not be explained by traditional models of intron gain, but can be possible by a new model, the insertion of introner-like elements (van der Burgt et al. 2012). Here, we investigate a novel case of intron gain and also highlight the importance of tandem genomic duplications in gene evolution.

MATERIALS AND METHODS

The genome sequences and annotation files of potato (*Solanum tuberosum*) (PGSC_DM_v3), tomato (*Solanum lycopersicum*) (ITAG2.3), and tobacco (*Nicotiana benthamiana*) (version 0.4.4) were downloaded from SGN (Sol Genomics Network) (Bombarely et al. 2011), while Pepper Genome Database (release 2.0) (Qin et al. 2014) was used for the pepper (*Capsicum annuum* L.)

(Zunla-1). The scaffold sequences of eggplant (*Solanum melongena*) were downloaded from NCBI (SME_r2.5.1, http://www.ncbi.nlm.nih.gov/genome/). The SAR files of the potato whole-genome shotgun (WGS) reads (SRP007439) and the leaf, tuber, and mixed-tissue transcriptomes (SRP022916, SRP005965, SRP040682, and ERP003480) were retrieved from the Sequence Read Archive of NCBI (http://www.ncbi.nlm.nih.gov/sra/). We mapped the RNA-Seq reads to the genomes by using TopHat version 2.0.8 (Kim et al. 2013), while BWA (alignment via Burrows-Wheeler transformation, version 0.5.7) (Li & Durbin 2009) was used for the WGS reads. We used default parameters for both programs. The orthologous proteins were identified by using the best reciprocal BLAST hits with a threshold E value of $< 10^{-10}$. In addition, the orthologous relationships were confirmed by using the SynMap (http://genomevolution.org/CoGe/SynMap.pl).

The RNA-dependent RNA polymerase (*RdRp*) genes encode those enzymes which catalyze the replication of RNA from an RNA template. They have been identified in all the major eukaryotic groups and play crucial roles in the regulation of development, maintenance of genome integrity, and defense against the foreign nucleic acids (Willmann et al. 2011; Zong et al. 2009). The *PGSC0003DMG402000361* orthologous sequence in eggplants was manually annotated with references to the annotations of the orthologous genes in potato, tomato, pepper, and tobacco.

By aligning 9,883 groups of orthologous mRNAs among potato, tomato and pepper, we obtained 34,364 conserved introns in potatoes with length > 60 bp. Among these conserved introns, we searched the consensus sequences of the 5′ splicing sites, the branch sites, the polypyrimidine tracts, and the 3′ splicing sites according to Irimia and Roy (2008) and Schwartz, et al. (2008). The information content of these sites was calculated by using the WebLogo 3.4

online (http://weblogo.threeplusone.com/create.cgi) (Crooks et al. 2004). The exonic splicing enhancers of *A. thaliana* were identified by Pertea et al. (2007). We used them as query and found 50 bp exonic sequence downstream of the target intron.

RESULTS AND DISCUSSION

By comparing the orthologous genes of tomatoes (*Solanum lycopersicum*), potatoes (*Solanum tuberosum*) and other Solanaceae plants, we found 11 cases of precise intron loss and six cases of imprecise intron loss (Ma et al. 2015). At the same time, we found the sign of an imprecise intron gain in potato gene, *PGSC0003DMG402000361*. The second intron of this gene is found unique in potatoes (Fig. 1). By analyzing the transcriptomic data of potato, we found 109 RNA-Seq reads that are exclusively mapped to the annotated exon-exon boundary (Supplemental Information 1: Table S1), which confirmed the annotation of this intron. Based on the phylogenetic tree of the species being compared (Fig. 2), there were two possible explanations for the presence/absence of the intron. The first was the gain of a new intron in potatoes, and the second was four parallel intron loss events occurred in the other four species: tomato, eggplant, pepper, and tobacco. According to the principle of parsimony, we concluded that the second intron of the potato gene *PGSC0003DMG402000361* was gained after the divergence of potatoes and tomatoes.

The intron gain was accompanied by a 168 bp insertion in the downstream exon (Fig. 1). For BLAT search (Kent 2002), new intron and the inserted exonic sequence was used as a query sequence against the whole potato genome, and it was found that the combined sequence is a direct duplication of the upstream sequence that co-occurs with a small insertion of an exogenous sequence (10 bp) between the duplicates (Fig. 3A). We were aware of this fact that two nearly identical regions in a reference genome might either be a true duplicate or a false due to an error

in genome assembly. To verify the duplication in the potato gene *PGSC0003DMG402000361*, we found three sources of evidence. Firstly, 62 whole genome shotgun (WGS) reads were exclusively mapped crossing the four boundaries of two duplicates (Supplemental Information 1: Table S2). Secondly, 109 RNA-Seq reads were exclusively mapped crossing the boundary of the two duplicates in mature mRNA, *i.e.,* the position of the new intron (Supplemental Information 1: Table S1). Thirdly, there are ten nucleotide differences between the duplicates (Fig. 3B).

Close examination of the coding region confirmed that the duplication and the insertion did not cause any frame-shifts. Furthermore, we tested whether *PGSC0003DMG402000361* is still a functional gene, or a pseudogene, by surveying its nonsynonymous and synonymous substitution rates, d_N and d_S, respectively. Using the phylogenetic tree of *PGSC0003DMG402000361* and its orthologous genes in tomato, pepper, and tobacco, we performed a likelihood-ratio test (LRT) to compare two models. The first model was the null hypothesis that the gene was undergoing neutral evolution, in which case the d_N/d_S value of *PGSC0003DMG402000361* would be equal to one. In the alternative model, the estimated value of d_N/d_S would be < 1 (Yang 2007). The d_N/d_S was 0.3101; the LRT statistic, $2\Delta\ell$ (twice the log likelihood difference between the two compared models), was 74.7; and the χ^2 test supported the hypothesis that the *PGSC0003DMG402000361* gene was subject to purifying selection ($P < 10^{-16}$).

According to Logsdon et al. (1998), strong evidence of intron gain must satisfy the two conditions. The first one is a clear phylogeny to provide support for the intron gain, while the second is an identified source element of the gained intron. Given the clear phylogeny and the identity of the source sequence, we consider the second intron of the potato gene *PGSC0003DMG402000361* to be a well-supported case of a newly gained intron.

125 The present case of intron gain is somewhat different from the tandem genomic duplication
126 model of intron gain that was originally put forward by Rogers (1989). In that model, tandem
127 duplication of an exonic segment harboring the AGGT sequence generates two splice sites for
128 the new intron: 5′-GT and 3′-AG, and the new intron comes from the duplication of exonic
129 sequence. It is now well known that the two splice sites do not contain sufficient information to
130 unequivocally determine the exon-intron boundaries (Lim & Burge 2001). Accurate recognition
131 and efficient splicing of an intron also requires a polypyrimidine tract, an adenine nucleotide at
132 the branch site, and many other *cis*-acting regulatory motifs (Schwartz et al. 2009; Spies et al.
133 2009; Wang & Burge 2008; Wang et al. 2004). Duplication of exonic segments might happen to
134 generate a combination of the required signals and also produce a functional intron (Hellsten et
135 al. 2011), but it is unlikely that preexisting polypyrimidine tracts and other splicing signals are
136 commonly present in coding sequences. Additionally, introns are often remarkably richer in AU
137 than exons (Amit et al. 2012), and this difference has been demonstrated to be a requirement for
138 efficient splicing (Carle-Urioste et al. 1997; Luehrsen & Walbot 1994). The phenomena of direct
139 introns gain from duplicated exonic segments is particularly unlikely in plants, and it is due to
140 the striking difference of base content between exons and introns. In the present case of intron
141 gain, the duplication includes the 3′ side sequence of an intron and the 5′ side of the downstream
142 exon (Fig. 3A). The 3′ splicing site signal (CAG), the polypyrimidine tract
143 (TCTTCCAATGCCT), and the putative branch site (TTTAC) of this novel intron was inherited
144 from the parental intron (Fig. 3B, 3C). Moreover, the two overlapped putative exonic splicing
145 enhancers of the 3′ flanking exon, TCAGCT and CAGCTC, and the GC contents differentiated
146 between the intron and exon (36% *vs.* 46%) were also inherited from the parental copy. The 5′
147 splicing signal of the novel intron, GTAAG, was provided by the exogenous sequence of 10 bp.

Utilization of some of the active splicing signals of the parental intron is apparently a more efficient method of intron gain. Therefore, we propose a modified version of the tandem genomic duplication model, termed as partial duplication of a preexisting intron and the flanking exon. Segmental duplication containing entire introns would be more likely to increase the gene intron number and also has been observed previously (Gao & Lynch 2009). In the present paper, we confine our discussion to the creation of new introns rather than the propagation of preexisting introns.

The modified version of the tandem genomic duplication model of intron gain could also be termed as imprecise intron gain, which highlights the co-occurring insertion of the coding sequence. Generally, the researchers seek intron gains in highly conserved orthologous genes. Thus, only introns flanking conserved exonic sequences are likely to be identified as a new one. Due to this methodology, the frequency of intron gain by segmental duplication might have been underestimated previously. To be consistent with this idea, a study that specifically explored intron gains by segmental duplications revealed tens of new introns in humans, mice, and *Arabidopsis thaliana* (Gao & Lynch 2009). This result is in stark contrast to the comparative studies of their highly conserved orthologous genes, which found very few or no intron gains at all (Coulombe-Huntington & Majewski 2007; Fawcett et al. 2012; Roy et al. 2003; Yang et al. 2013). Considering the high frequency of internal gene duplications, which is 0.001–0.013 duplications/gene per million years (Gao & Lynch 2009), it can be stated that intron gain by segmental duplication may be an important force shaping the eukaryotic gene structure. With the increasing number of very closely related genomes (*i.e.*, diverged within ten million years) to be sequenced, we expect to find more intron gains by segmental duplication in the near future.

CONCLUSIONS

In the potato gene *PGSC0003DMG402000361*, we found a novel intron originated from tandem duplication. The duplicate includes the 3′ side sequence of an intron and the 5′ side of the downstream exon. Most splicing signals which include, a putative branch site, the polypyrimidine tract, the 3′ splicing site, two putative exonic splicing enhancers and the GC contents differentiated between the intron and exon were inherited from the parental intron/exon structure. By contrast, the widely cited model of intron gain is tandem duplication of an exonic segment containing AGGT, which would create the GT and AG splicing sites. The case of intron gain which we observed, requires a modified version of the tandem genomic duplication model: partial duplication of the preexisting intron/exon structure. As we see, this modified version is more consistent with the mechanisms of intron recognition and splicing (Schwartz et al. 2009; Spies et al. 2009; Wang & Burge 2008; Wang et al. 2004).

ACKNOWLEDGEMENT

We are thankful to Sidra Aslam for her help in the improvement of English language of this paper.

REFERENCES

Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Lev-Maor G, Burstein D, Schwartz S, Postolsky B, Pupko T, and Ast G. 2012. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Reports* 1:543-556. DOI 10.1016/j.celrep.2012.03.013

Bombarely A, Menda N, Tecle IY, Buels RM, Strickler S, Fischer-York T, Pujar A, Leto J, Gosselin J, and Mueller LA. 2011. The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Research* 39:D1149-D1155. DOI 10.1093/Nar/Gkq866

196 Carle-Urioste JC, Brendel V, and Walbot V. 1997. A combinatorial role for exon, intron and
197 splice site sequences in splicing in maize. *The Plant Journal* 11:1253-1263. DOI
198 10.1046/j.1365-313X.1997.11061253.x
199 Collemare J, van der Burgt A, and de Wit PJGM. 2013. At the origin of spliceosomal introns: Is
200 multiplication of introner-like elements the main mechanism of intron gain in fungi?
201 *Communicative & Integrative Biology* 6:e23147. DOI 10.4161/cib.23147
202 Coulombe-Huntington J, and Majewski J. 2007. Characterization of intron loss events in
203 mammals. *Genome Research* 17:23-32. DOI 10.1101/gr.5703406
204 Crooks GE, Hon G, Chandonia JM, and Brenner SE. 2004. WebLogo: a sequence logo generator.
205 *Genome Research* 14:1188-1190. DOI 10.1101/gr.849004
206 Csuros M, Rogozin IB, and Koonin EV. 2011. A detailed history of intron-rich eukaryotic
207 ancestors inferred from a global survey of 100 complete genomes. *PLoS Computational*
208 *Biology* 7:e1002150. DOI 10.1371/journal.pcbi.1002150
209 Fablet M, Bueno M, Potrzebowski L, and Kaessmann H. 2009. Evolutionary origin and functions
210 of retrogene introns. *Molecular Biology and Evolution* 26:2147-2156. DOI
211 10.1093/molbev/msp125
212 Fawcett JA, Rouzé P, and Van de Peer Y. 2012. Higher intron loss rate in *Arabidopsis thaliana*
213 than *A. lyrata* is consistent with stronger selection for a smaller genome. *Molecular*
214 *Biology and Evolution* 29:849-859. DOI 10.1093/molbev/msr254
215 Gao X, and Lynch M. 2009. Ubiquitous internal gene duplication and intron creation in
216 eukaryotes. *Proceedings of the National Academy of Sciences of the United States of*
217 *America* 49:20818-20823. DOI 10.1073/pnas.0911093106
218 Hellsten U, Aspden JL, Rio DC, and Rokhsar DS. 2011. A segmental genomic duplication
219 generates a functional intron. *Nature Communications* 2:454. DOI 10.1038/ncomms1461
220 Hooks KB, Delneri D, and Griffiths-Jones S. 2014. Intron Evolution in Saccharomycetaceae.
221 *Genome Biology and Evolution* 6:2543-2556. DOI 10.1093/Gbe/Evu196
222 Irimia M, and Roy SW. 2008. Evolutionary convergence on highly-conserved 3' intron structures
223 in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS*
224 *Genetics* 4:e1000148. DOI 10.1371/journal.pgen.1000148
225 Irimia M, and Roy SW. 2014. Origin of spliceosomal introns and alternative splicing. *Cold*
226 *Spring Harbor Perspectives in Biology* 6:a016071. DOI 10.1101/cshperspect.a016071
227 Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Research* 12:656-664. DOI
228 10.1101/gr.229202
229 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL. 2013. TopHat2: accurate
230 alignment of transcriptomes in the presence of insertions, deletions and gene fusions.
231 *Genome Biology* 14:R36. DOI 10.1186/Gb-2013-14-4-R36
232 Li H, and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
233 transform. *Bioinformatics* 25:1754-1760. DOI 10.1093/bioinformatics/btp324
234 Li W, Tucker AE, Sung W, Thomas WK, and Lynch M. 2009. Extensive, recent intron gains in
235 *Daphnia* populations. *Science* 326:1260-1262. DOI 10.1126/science.1179302

Lim LP, and Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences of the United States of America* 98:11193-11198. DOI 10.1073/pnas.201407298

Logsdon Jr JM, Stoltzfus A, and Doolittle WF. 1998. Molecular evolution: Recent cases of spliceosomal intron gain? *Current Biology* 8:R560-R563. DOI 10.1016/S0960-9822(07)00361-2

Luehrsen K, and Walbot V. 1994. Addition of A- and U-rich sequence increases the splicing efficiency of a deleted form of a maize intron. *Plant Molecular Biology* 24:449-463. DOI 10.1007/BF00024113

Ma M-Y, Zhu T, Li X-N, Lan X-R, Liu H-Y, Yang Y-F, and Niu D-K. 2015. Imprecise intron losses are less frequent than precise intron losses but are not rare in plants. *Biology Direct* 10:24.

Pertea M, Mount SM, and Salzberg SL. 2007. A computational survey of candidate exonic splicing enhancer motifs in the model plant *Arabidopsis thaliana*. *BMC Bioinformatics* 8:159. DOI 10.1186/1471-2105-8-159

Qin C, Yu C, Shen Y, Fang X, Chen L, Min J, Cheng J, Zhao S, Xu M, Luo Y, Yang Y, Wu Z, Mao L, Wu H, Ling-Hu C, Zhou H, Lin H, González-Morales S, Trejo-Saavedra DL, Tian H, Tang X, Zhao M, Huang Z, Zhou A, Yao X, Cui J, Li W, Chen Z, Feng Y, Niu Y, Bi S, Yang X, Li W, Cai H, Luo X, Montes-Hernández S, Leyva-González MA, Xiong Z, He X, Bai L, Tan S, Tang X, Liu D, Liu J, Zhang S, Chen M, Zhang L, Zhang L, Zhang Y, Liao W, Zhang Y, Wang M, Lv X, Wen B, Liu H, Luan H, Zhang Y, Yang S, Wang X, Xu J, Li X, Li S, Wang J, Palloix A, Bosland PW, Li Y, Krogh A, Rivera-Bustamante RF, Herrera-Estrella L, Yin Y, Yu J, Hu K, and Zhang Z. 2014. Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proceedings of the National Academy of Sciences of the United States of America* 111:5135-5140. DOI 10.1073/pnas.1400975111

Rogers JH. 1989. How were introns inserted into nuclear genes. *Trends in Genetics* 5:213-216. DOI 10.1016/0168-9525(89)90084-X

Roy SW, Fedorov A, and Gilbert W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proceedings of the National Academy of Sciences of the United States of America* 100:7158-7162. DOI 10.1073/pnas.1232297100

Roy SW, and Gilbert W. 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. DOI 10.1073/pnas.0500383102

Roy SW, and Penny D. 2006. Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. *Molecular Biology and Evolution* 23:2259-2262. DOI 10.1093/molbev/msl098

Sarkinen T, Bohs L, Olmstead RG, and Knapp S. 2013. A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. *BMC Evolutionary Biology* 13:214. DOI 10.1186/1471-2148-13-214

276 Schwartz S, Gal-Mark N, Kfir N, Oren R, Kim E, and Ast G. 2009. *Alu* exonization events reveal
277 features required for precise recognition of exons by the splicing machinery. *PLoS*
278 *Computational Biology* 5:e1000300. DOI 10.1371/journal.pcbi.1000300
279 Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, and Ast G. 2008. Large-scale comparative
280 analysis of splicing signals and their corresponding splicing factors in eukaryotes.
281 *Genome Research* 18:88-103. DOI 10.1101/Gr.6818908
282 Spies N, Nielsen CB, Padgett RA, and Burge CB. 2009. Biased chromatin signatures around
283 polyadenylation sites and exons. *Molecular Cell* 36:245-254. DOI
284 10.1016/j.molcel.2009.10.008
285 Torriani SFF, Stukenbrock EH, Brunner PC, McDonald BA, and Croll D. 2011. Evidence for
286 extensive recent intron transposition in closely related fungi. *Current Biology* 21:2017-
287 2022. DOI 10.1016/j.cub.2011.10.041
288 van der Burgt A, Severing E, de Wit Pierre JGM, and Collemare J. 2012. Birth of new
289 spliceosomal introns in fungi by multiplication of intron-like elements. *Current Biology*
290 22:1260-1265. DOI 10.1016/j.cub.2012.05.011
291 Verhelst B, Van de Peer Y, and Rouze P. 2013. The complex intron landscape and massive
292 intron invasion in a picoeukaryote provides insights into intron evolution. *Genome*
293 *Biology and Evolution* 5:2393-2401. DOI 10.1093/Gbe/Evt189
294 Wang ZF, and Burge CB. 2008. Splicing regulation: From a parts list of regulatory elements to
295 an integrated splicing code. *RNA* 14:802-813. DOI 10.1261/rna.876308
296 Wang ZF, Rolish ME, Yeo G, Tung V, Mawson M, and Burge CB. 2004. Systematic
297 identification and analysis of exonic splicing silencers. *Cell* 119:831-845. DOI
298 10.1016/j.cell.2004.11.010
299 Willmann MR, Endres MW, Cook RT, and Gregory BD. 2011. The functions of RNA-dependent
300 RNA polymerases in *Arabidopsis*. *The Arabidopsis book/American Society of Plant*
301 *Biologists* 9. DOI 10.1199/tab.0146
302 Yang YF, Zhu T, and Niu DK. 2013. Association of intron loss with high mutation rate in
303 *Arabidopsis*: implications for genome size evolution. *Genome Biology and Evolution*
304 5:723-733. DOI 10.1093/gbe/evt043
305 Yang ZH. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology*
306 *and Evolution* 24:1586-1591. DOI 10.1093/molbev/msm088
307 Yenerall P, Krupa B, and Zhou L. 2011. Mechanisms of intron gain and loss in *Drosophila*.
308 *BMC Evolutionary Biology* 11:364. DOI 10.1186/1471-2148-11-364
309 Yenerall P, and Zhou L. 2012. Identifying the mechanisms of intron gain: progress and trends.
310 *Biology Direct* 7:29. DOI 10.1186/1745-6150-7-29
311 Zhu T, and Niu DK. 2013. Frequency of intron loss correlates with processed pseudogene
312 abundance: a novel strategy to test the reverse transcriptase model of intron loss. *BMC*
313 *Biology* 11:23. DOI 10.1186/1741-7007-11-23
314 Zong J, Yao X, Yin JY, Zhang DB, and Ma H. 2009. Evolution of the RNA-dependent RNA
315 polymerase (RdRP) genes: Duplications and possible losses before and after the

316 divergence of major eukaryotic groups. *Gene* 447:29-39. DOI
317 10.1016/j.gene.2009.07.004

318 Figures

319 **Figure 1. Alignments of coding sequences showing the intron gain and a flanking insertion**
320 **in the potato gene *PGSC0003DMG402000361*.**

321 The presence and absence of the intron are represented by 1 and 0, respectively. The orthologous
322 genes used as references are *Solyc12g008410.1* in tomato, *Capana09g000243* in pepper, and
323 *NbS00003153g0003* in tobacco. The orthologous region in eggplants was manually identified by
324 the best reciprocal program, BLAST, and manually annotated.

325

326 **Figure 2. Phylogenetic tree used to identify the intron gain in *Solanum tuberosum*.**

327 The tree was adapted from Särkinen et al. (2013) and is not scaled according to substitution rates.
328 The presence and absence of the intron are represented by + and –, respectively.

329

330 **Figure 3. The intron gain by tandem genomic duplication in the potato gene**
331 ***PGSC0003DMG402000361*.**

332 **(A)** A schematic diagram showing the creation of a new intron by partial duplication of the
333 parental intron (marked in blue line) and the insertion of a short exogenous sequence (marked in
334 red line). **(B)** Alignment of the two copies of the duplication and the inserted exogenous
335 sequence (marked in red). The splicing sites, the putative branch site, the polypyrimidine tract,
336 and two overlapping putative exonic splicing enhancers (ESE; TCAGCT and CAGCTC) are
337 underlined. Sites differing between the two copies are indicated with bold blue letters. **(C)** The

consensus sequences of the introns conserved among potatoes, tomatoes and peppers. These sequences were used to recognize the splicing signals for the new intron.

Figure 1(on next page)

Alignments of coding sequences showing the intron gain and a flanking insertion in the potato gene *PGSC0003DMG402000361*

The presence and absence of the intron are represented by 1 and 0, respectively. The orthologous genes used as references are *Solyc12g008410.1* in tomato, *Capana09g000243* in pepper, and *NbS00003153g0003* in tobacco. The orthologous regions in eggplants were manually identified by the best reciprocal program, BLAST, and manually annotated.

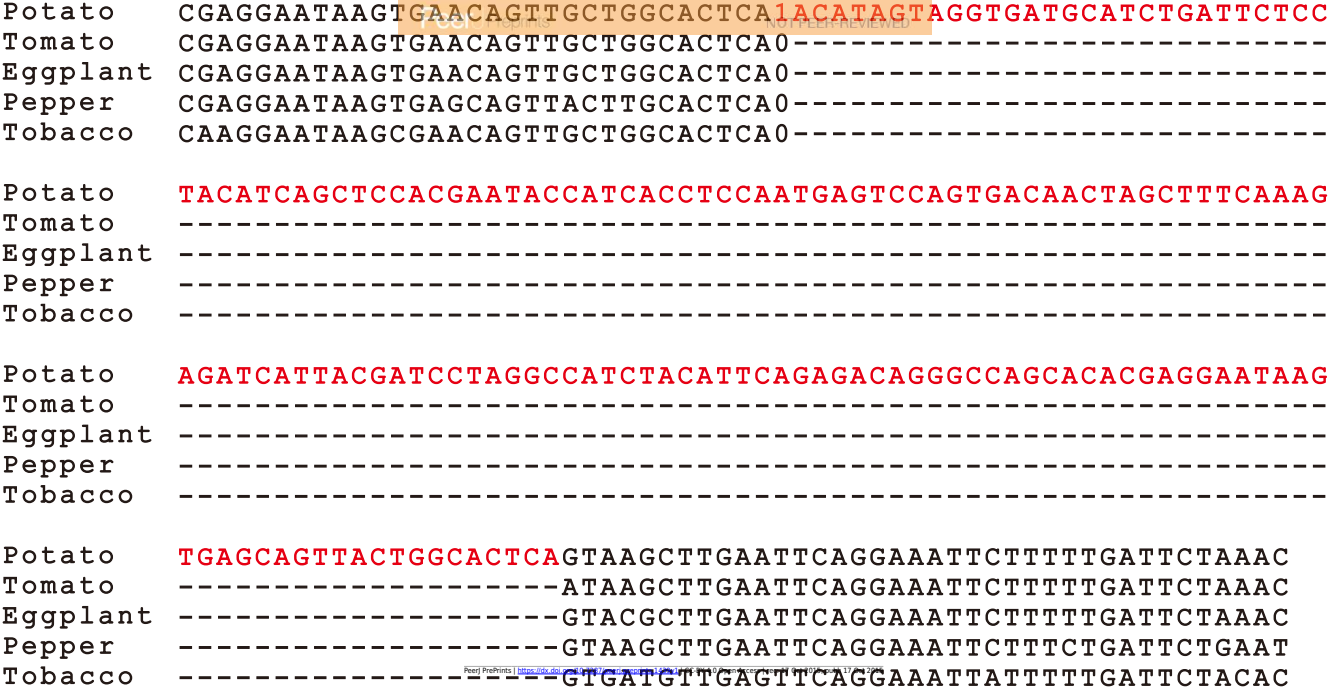

Potato CGAGGAATAAGTGAACAGTTGCTGGCACTCA1ACATAGTAGGTGATGCATCTGATTCTCC
Tomato CGAGGAATAAGTGAACAGTTGCTGGCACTCA0----------
Eggplant CGAGGAATAAGTGAACAGTTGCTGGCACTCA0----------
Pepper CGAGGAATAAGTGAGCAGTTACTTGCACTCA0----------
Tobacco CAAGGAATAAGCGAACAGTTGCTGGCACTCA0----------
Potato TACATCAGCTCCACGAATACCATCACCTCCAATGAGTCCAGTGACAACTAGCTTTCAAAG
Tomato ----------
Eggplant ----------
Pepper ----------
Tobacco ----------
Potato AGATCATTACGATCCTAGGCCATCTACATTCAGAGACAGGGCCAGCACACGAGGAATAAG
Tomato ----------
Eggplant ----------
Pepper ----------
Tobacco ----------
Potato TGAGCAGTTACTGGCACTCAGTAAGCTTGAATTCAGGAAATTCTTTTTGATTCTAAAC
Tomato ----------ATAAGCTTGAATTCAGGAAATTCTTTTTGATTCTAAAC
Eggplant ----------GTACGCTTGAATTCAGGAAATTCTTTTTGATTCTAAAC
Pepper ----------GTAAGCTTGAATTCAGGAAATTCTTTCTGATTCTGAAT
Tobacco ----------GTGATGTTGAGTTCAGGAAATTATTTTTGATTCTACAC

Figure 2 (on next page)

Phylogenetic tree used to identify the intron gain in ***Solanum tuberosum***

The tree was adapted from Särkinen et al. (2013) and is not scaled according to substitution rates. The presence and absence of the intron are represented by + and - , respectively.

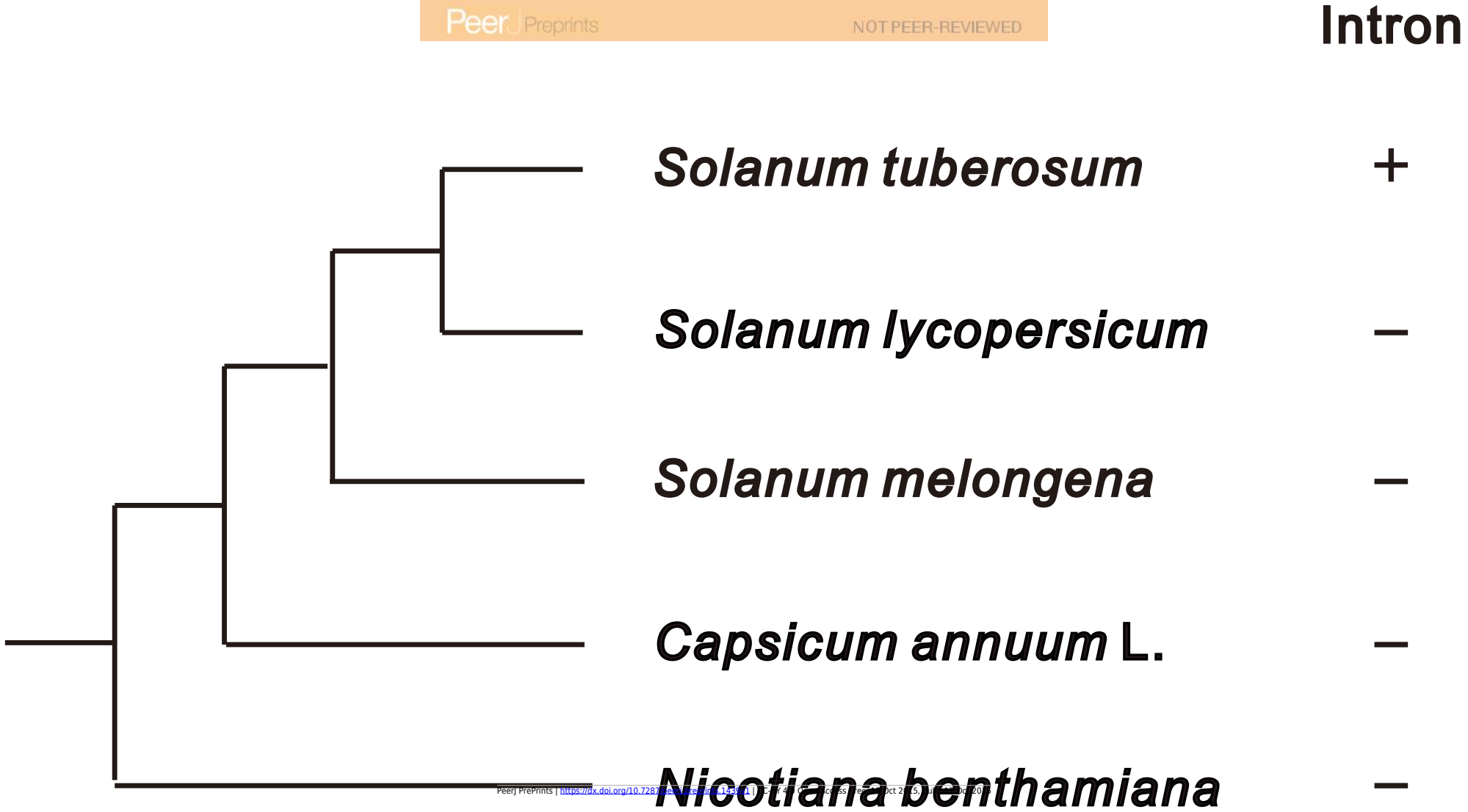
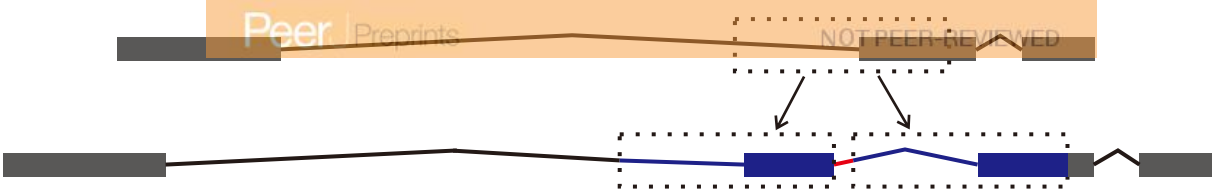


Figure 3(on next page)

The intron gain by tandemgenomic duplication in the potato gene ***PGSC0003DMG402000361***

(A) A schematic diagram showing the creation of a new intron by partial duplication of the parental intron (marked in blue line) and the insertion of a short exogenous sequence (marked in red line). **(B)** Alignment of the two copies of the duplication and the inserted exogenous sequence (marked in red). The splicing sites, the putative branch site, the polypyrimidine tract, and two overlapping putative exonic splicing enhancers (ESE; TCAGCT and CAGCTC) are underlined. Sites differing between the two copies are indicated with bold blue letters. **(C)** The consensus sequences of the introns conserved among potatoes, tomatoes and peppers. These sequences were used to recognize the splicing signals for the new intron.

A



B

```
Downstream gtaagcttgaattagtttaggcttaatgaagaacttgttcaatttttttattggttgcga
Upstream   ----------attagtttaggcttaatgaagaacttgttcaaatttttttattggttgcga

Downstream tctcttcttcc-ttttttttgcatatttacaactctacatgtaaactatgttgctcggact
Upstream   tctcttcttcctttttttttgcatatttacaactctacatgtaaactatgttgctcggact

Downstream ctcaaaaactgttgaacccgtgttggattctccaaaatgcactacttttggagtattcga
Upstream   ctcaaaaactgttgaacccgtgttggattctccaaaatgcactacttttggagtattcga

Downstream tacacacttttgaagagtccgaacaacacaacatgt-aatgtactcagacctttcaagaa
Upstream   tacacacttttgaagagtctgaacaacacaacatgtaaacatactcagacctttcaaaaa

Downstream ttctagtttaccaatgatggtcttccaatgcctgcagACATAGTAGGTGATGCATCTGAT
Upstream   ttctagtttaccaatgatggtcttccaatgcctacagACATAGTAGGTGATGCATCTGAT

Downstream TCTCCTACATCAGCTCCACGAATACCATCACCTCCAATGAGTCCAGTGACAACTAGCTTT
Upstream   TCTCCTACATCAGCTCCACGAATACCATCACCTCCAATGAGTCCAGTGACAACTAGCTTT

Downstream CAAAGAGATCATTACGATCCTAGGCCATCTACATTCAGAGACAGGGCCAGCACACGAGGA
Upstream   CAAAGAGATCATTACGATCCTAGGCCATCTACATTCAGAGACAGGGCCAGCACACGAGGA

Downstream ATAAGTGAGCAGTTACTGGCACTCA
Upstream   ATAAGTGAACAGTTGCTGGCACTCA
```

C