Intron gain by tandem genomic duplication: a novel case and a new version of the model

Ming-Yue Ma, Xin-Ran Lan, 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 mechanisms of intron gain because randomly accumulated mutations might erase the evolutionary traces. The mechanisms of intron gain remain unclear due to the presence of very 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. We found that the second intron of the potato RNA-dependent RNA polymerase gene *PGSC0003DMG402000361* originated mainly from a direct duplication of the 3′ side of the upstream intron. The 5′ splicing site of this new intron was recruited from the upstream exonic sequence. In addition to the new intron, a downstream exonic segment of 178 bp also arose from duplication. Most of the splicing signals were inherited directly 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 new version of the tandem genomic duplication model, termed as the partial duplication of the preexisting intron/exon structure. This new version and the widely cited version are not mutually exclusive.

Intron gain by tandem genomic duplication: a novel case and a new version of the model

Ming-Yue Ma, Xin-Ran Lan 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

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 mechanisms of intron gain because randomly accumulated mutations might erase the evolutionary traces. The mechanisms of intron gain remain unclear due to the presence of very 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. We found that the second intron of the potato RNA-dependent RNA polymerase gene *PGSC0003DMG402000361* originated mainly from a direct duplication of the 3′ side of the upstream intron. The 5′ splicing site of this new intron was recruited from the upstream exonic sequence. In addition to the new intron, a downstream exonic segment of 178 bp also arose from duplication. Most of the splicing signals were inherited directly 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 new version of the tandem genomic duplication model, termed as the partial duplication of the preexisting intron/exon structure. This new version and the widely cited version are not mutually exclusive.

INTRODUCTION

Although, spliceosomal introns are the characteristic feature of eukaryotic nuclear genes, their origin and subsequent accumulation during evolution remain obscure. Several models of spliceosomal intron gain have been proposed, including 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. Although intron gains are generally reported at a lower frequency than intron losses, the reported intron gains have been accumulated to a considerable number (Csuros et al. 2011; Fablet et al. 2009; Hooks et al. 2014; Irimia & Roy 2014; Li et al. 2009; Li et al. 2014; Roy & Gilbert 2005; Roy & Penny 2006; Torriani et al. 2011; van der Burgt et al. 2012; Verhelst et al. 2013; Yenerall et al. 2011; Yenerall & Zhou 2012; Zhu & Niu 2013a). Unfortunately, the source sequences of most of these reported intron gains have not been identified. As a consequence, these intron gains provide very limited supporting evidence for the intron gain models. 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). Among the traditional models, intron gain by tandem genomic duplication is not expected to occur rarely, because frequent internal gene duplications are observed (Gao & Lynch 2009). This model was originally put forward by Rogers (1989), suggests that tandem duplication of an exonic segment harboring the AGGT sequence generates two splice sites for the new intron: 5′-GT and 3′-AG. In this model, the new intron comes from the duplication of an exonic sequence and the translated peptide is not altered by the intron gain. An example strictly

consistent with this model is in the vertebrate gene *ATP2A1* (Hellsten et al. 2011). The duplicated region of *ATP2A1* not only has the AGGT signal, but also happen to include a polypyrimidine tract and a branch point. In addition to it, the birth of the intron has been successfully recapitulated in a conserved paralogous gene, *ATP2A2*, by Hellsten et al. (2011). In fission yeasts, multiple tandem duplication of a 24 bp exonic segment containing AGGT occurred in genes *SPOG_01682* and *SOCG_00815*. Comparison of these two genes with their expressed sequence tags indicates an intron across four duplicates in the gene *SPOG_01682* and an intron across two duplicates in the gene *SOCG_00815* (Zhu & Niu 2013b). In these two cases, intronization of the duplicated region possibly alleviated the potential negative effects of the duplications on the translated proteins. In the present study, we found a new intron gained by duplicating a gene segment across an intron-exon boundary in a potato RNA-dependent RNA polymerase (*RdRp*) gene. The *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).

MATERIALS AND METHODS

The genome sequences and annotation files of domesticated potato (*Solanum tuberosum*, PGSC_DM_v3), domesticated tomato (*Solanum lycopersicum*, ITAG2.3), wild tobacco (*Nicotiana benthamiana*, version 0.4.4), and wild tomato (*Solanum pennellii*, spenn_v2.0) were downloaded from Sol Genomics Network (Bombarely et al. 2011), and those of pepper (*Capsicum annuum* L., Zunla-1) were downloaded from the Pepper Genome Database (Qin et al. 2014). The scaffold sequences of Commerson's wild potato (*Solanum commersonii*,

JXZD00000000.1), wild tomato (*Solanum habrochaites*, CBYS000000000.1), and eggplant (*Solanum melongena*, SME_r2.5.1) were downloaded from the NCBI Genome database (http://www.ncbi.nlm.nih.gov/genome/). The SAR files of the whole-genome shotgun (WGS) reads (SRP007439) and the leaf, tuber, and mixed-tissue transcriptomes (SRP022916, SRP005965, SRP040682, and ERP003480) of *S. tuberosum* 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 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 except that the minimum intron length was adjusted to 20 bp for TopHat. The orthologous genes of the *S. tuberosum RdRp* gene *PGSC0003DMG402000361* were identified by using the best reciprocal BLAST hits with a threshold E value of $< 10^{-10}$. In addition, the orthologous relationship between the gene *PGSC0003DMG402000361* and its ortholog in *S. lycopersicum* was confirmed by their synteny using the SynMap (http://genomevolution.org/CoGe/SynMap.pl). The orthologous sequences of the gene *PGSC0003DMG402000361* in *S. commersonii*, *S. habrochaites*, *S. melongena* were manually annotated with references to the annotations in *S. tuberosum*, *S. lycopersicum*, *C. annuum*, and *N. benthamiana*.

We found that the intron gain was involved in a duplication using BLAT search (Kent 2002) and then identified the exact duplicated sequences using the programs REPuter (Kurtz et al. 2001) and Tandem Repeats Finder (Benson 1999).

By aligning 9,883 groups of orthologous mRNAs among *S. tuberosum*, *S. lycopersicum*, and *C. annuum*, we found all the introns conserved among these three species. After filtering them with a length of > 60 bp in *S. tuberosum*, 34,364 groups of conserved introns were retained.

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). Sequence logos were generated using the WebLogo 3.4 online (http://weblogo.threeplusone.com/create.cgi) (Crooks et al. 2004) from multiple alignments of the 34,364 conserved introns in potatoes. The exonic splicing enhancers (ESEs) of *Arabidopsis thaliana* were identified by Pertea et al. (2007). We used them as query and searched 50 bp exonic sequences upstream and downstream of the target intron.

The phylogenetic tree of the gene *PGSC0003DMG402000361* and its orthologs was constructed using MEGA 6.0 by employing the Neighbor-Joining method (Tamura et al. 2013). The tree topology is consistent with the species tree constructed by Särkinen et al. (2013). The schematic diagram of gene structures was drawn using the program GSDraw (Wang et al. 2013).

RESULTS AND DISCUSSION

By comparing the orthologous genes of *S. lycopersicum*, *S. 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 intron gain in the *S. tuberosum* gene, *PGSC0003DMG402000361* (Fig. 1). According to the potato genome version PGSC_DM_v3, this gene has eight introns and nine exons. By comparing the annotations of other Solanaceae genomes, we manually annotated 16 exons in the orthologous gene in *S. commersonii* (Fig. 2). The orthologous genes in *S. lycopersicum*, *S. habrochaites*, *S. pennellii*, *S. melongena*, *C. annuum*, and *N. benthamiana* have 15, 15, 16, 15, 18, and 17 exons, respectively. The second introns of *S. tuberosum* and *S. commersonii* are absent from other Solanaceae genomes. Meanwhile, the third exons of these two species have sequences similar to the upstream ones as

well as the second exons of other Solanaceae species (Fig. 2). By analyzing the transcriptomic data of *S. tuberosum*, we found 106 RNA-Seq reads that are exclusively mapped to the annotated exon-exon boundary (Supplemental Information 1: Table S1; Supplemental Information 2: Fig. S1), which confirmed the annotation of this intron.

Based on the phylogenetic tree constructed using the gene *PGSC0003DMG402000361* and its orthologs (Fig. 2), there were two possible explanations for the presence/absence of the intron. The first was the gain of a new intron in the common ancestor of *S. tuberosum* and *S. commersonii*, and the second was four intron loss events independently occurred in the other four evolutionary branches: tomatoes (including *S. lycopersicum*, *S. pennellii*, and *S. habrochaites*), *S. melongena*, *C. annuum*, and *N. benthamiana*. According to the principle of parsimony, we concluded that the second intron of the gene *PGSC0003DMG402000361* was gained after the divergence of potatoes (*S. tuberosum* and *S. commersonii*) from other *Solanum* plants, but prior to the divergence between *S. tuberosum* and *S. commersonii*.

The new intron and the inserted exonic sequence (Fig. 1) was used as a query sequence against the whole genome of *S. tuberosum*. We found that this insertion is a tandem genomic duplication (Fig. 3A). The major part of the new intron and inserted exon region is a direct duplicate of the upstream intron-exon structure (Fig. 3B). Meanwhile, 10 nucleotides at the 5′ end of the new intron was recruited from the upstream exon (Fig. 3A). We were aware of the fact that two nearly identical regions in a reference genome might either be a true duplication or a false due to an error in genome assembly. To verify the duplication, we found three sources of evidence in *S. tuberosum*. Firstly, 53 WGS reads were exclusively mapped crossing the three boundaries of two duplicates (Supplemental Information 1: Table S2; Supplemental Information 2: Fig. S2-S4). Secondly, 106 RNA-Seq reads were exclusively mapped crossing the exon

147 boundary of the mature mRNA (Supplemental Information 1: Table S1; Supplemental
148 Information 2: Fig. S1). The exon boundary sequence would not exist in mature mRNA if the
149 duplication did not happen. Thirdly, there are ten nucleotides different between the duplicates
150 (Fig. 3B).

151 Close examination of the coding region confirmed that the duplication did not cause any
152 frame-shifts. Furthermore, using the phylogenetic tree of *PGSC0003DMG402000361* and its
153 orthologous genes in tomato, pepper, and tobacco, we performed a likelihood-ratio test (LRT) to
154 compare two hypotheses. The null hypothesis is that the gene is actually a pseudogene and so
155 was undergoing neutral evolution, in which case the d_N/d_S value of *PGSC0003DMG402000361*
156 would be equal to one. In the alternative hypothesis, the gene is still functional and under
157 purifying selection, in which the estimated value of d_N/d_S would be < 1 (Yang 2007). The d_N/d_S
158 that we observed was 0.3101; the LRT statistic, $2\Delta\ell$ (twice the log likelihood difference between
159 the two compared models), was 74.7; and the χ^2 test supported the second model ($P < 10^{-16}$).
160 Although this result indicates that this protein-coding gene is still functional after the duplication,
161 we do not think that producing functional proteins is a prerequisite in the identification of a
162 sequence as a new intron. An intron is defined by its being spliced out during the maturation of
163 any RNA molecules, including both protein-coding mRNAs and noncoding RNAs. In recent
164 years, numerous sequences have been found to be spliced out of long noncoding RNAs, and been
165 described as introns without any debate (Derrien et al. 2012; Guttman et al. 2009; Jayakodi et al.
166 2015; Kapusta & Feschotte 2014).

167 According to Logsdon et al. (1998), strong evidence of intron gain must satisfy the two
168 conditions. The first one is a clear phylogeny to provide support for the intron gain, while the
169 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.

The present case of intron gain is somewhat different from the tandem genomic duplication model of intron gain that was originally put forward by Rogers (1989). In that model, tandem duplication of an exonic segment harboring the AGGT sequence generates two splice sites for the new intron: 5′-GT and 3′-AG, and the new intron comes from the duplication of exonic sequence. It is now well known that the two splice sites do not contain sufficient information to unequivocally determine the exon-intron boundaries (Lim & Burge 2001). Accurate recognition and efficient splicing of an intron also requires a polypyrimidine tract, an adenine nucleotide at the branch site, and many other *cis*-acting regulatory motifs (Schwartz et al. 2009; Spies et al. 2009; Wang & Burge 2008; Wang et al. 2004). In addition, introns are often remarkably richer in AU than exons (Amit et al. 2012), and this difference has been demonstrated to be a requirement for efficient splicing (Carle-Urioste et al. 1997; Luehrsen & Walbot 1994). At the first glance, it seems unlikely for a coding segment to have a full set of the splicing signals. Contrary to this expectation, intronization of coding regions has been observed in several different organisms including both animals and plants (Irimia et al. 2008; Kang et al. 2012; Szczesniak et al. 2011; Zhan et al. 2014; Zhu et al. 2009). These observations indicate that it is possible for coding sequences to contain cryptic splice signals. Furthermore, an experimentally duplicated coding segment of the vertebrate gene, *ATP2A2*, has been shown to be successfully spliced out of the mature mRNA (Hellsten et al. 2011). Therefore, a full set of the splicing signals require for active splicing is present in the coding sequence of the gene *ATP2A2*. Although a full set of the splicing signals could preexist in coding sequences, we believe that utilization of the active splicing signals of the parental intron/exon structure is a more efficient method of intron gain. In

193 the potato gene *PGSC0003DMG402000361*, the duplication includes the 3′ side sequence of an
194 intron and the 5′ side of the downstream exon (Fig. 3A). The 3′ splicing site signal (CAG), the
195 polypyrimidine tract (TCTTCCAATGCCT), and the putative branch site (TTTAC) of this novel
196 intron was inherited from the parental intron (Fig. 3B, 3C). Moreover, the two overlapped
197 putative ESEs of the 3′ flanking exon, TCAGCT and CAGCTC, and the GC contents
198 differentiated between the intron and exon (36% *vs.* 46%) were also inherited from the parental
199 copy. The 5′ splicing signal of the novel intron, GTAAG, was activated from a cryptic splice site
200 which was recruited from the upstream exon. One putative 5′ ESE, GAGGAA, has been
201 identified in the 5′ flanking exon of this new intron. Before the duplication event, the signal
202 GAGGAA was 73 bp far from its downstream intron. It was more likely a cryptic ESE than an
203 active one. The duplication event made it close to an intron and so ready to act as an ESE.
204 Therefore, we propose a new version of the tandem genomic duplication model, termed as partial
205 duplication of a preexisting intron/exon structure. Apparently, the traditional version of the
206 tandem genomic duplication model and this new version is not mutually exclusive. Each of them
207 might account for some cases of intron gain in evolution. Segmental duplication containing
208 entire introns would be more likely to increase the intron number of genes and also has been
209 observed previously (Gao & Lynch 2009). In the present paper, we confine our discussion to the
210 creation of new introns rather than the propagation of preexisting introns.
211 The new version of the tandem genomic duplication model also highlights the co-occurring
212 insertion of coding sequence with an intron gain. Generally, the researchers seek intron gains in
213 highly conserved orthologous genes. Thus, only introns flanking conserved exonic sequences are
214 likely to be identified as a new one. Due to this methodology, the frequency of intron gain by
215 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 *A. 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 gene *PGSC0003DMG402000361* of last common ancestor of domesticated potato *S. tuberosum* and wild potato *S. commersonii*, a tandem duplication event created a novel intron. 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 ESEs 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 new version of the tandem genomic duplication model: partial duplication of the preexisting intron/exon structure. This version is a supplement to the widely cited version of the tandem genomic duplication model (Rogers 1989; Yenerall & Zhou 2012).

239

240 ACKNOWLEDGEMENT

241 We are thankful to the anonymous referees for their useful comments and Sidra Aslam for her

242 help in the improvement of English language of this paper.

243

244 REFERENCES

245 Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Lev-Maor G, Burstein D,
246 Schwartz S, Postolsky B, Pupko T, and Ast G. 2012. Differential GC content between
247 exons and introns establishes distinct strategies of splice-site recognition. *Cell Reports*
248 1:543-556. DOI 10.1016/j.celrep.2012.03.013
249 Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids*
250 *Research* 27:573-580. DOI 10.1093/nar/27.2.573
251 Bombarely A, Menda N, Tecle IY, Buels RM, Strickler S, Fischer-York T, Pujar A, Leto J,
252 Gosselin J, and Mueller LA. 2011. The Sol Genomics Network (solgenomics.net):
253 growing tomatoes using Perl. *Nucleic Acids Research* 39:D1149-D1155. DOI
254 10.1093/Nar/Gkq866
255 Carle-Urioste JC, Brendel V, and Walbot V. 1997. A combinatorial role for exon, intron and
256 splice site sequences in splicing in maize. *The Plant Journal* 11:1253-1263. DOI
257 10.1046/j.1365-313X.1997.11061253.x
258 Collemare J, van der Burgt A, and de Wit PJGM. 2013. At the origin of spliceosomal introns: Is
259 multiplication of introner-like elements the main mechanism of intron gain in fungi?
260 *Communicative & Integrative Biology* 6:e23147. DOI 10.4161/cib.23147
261 Coulombe-Huntington J, and Majewski J. 2007. Characterization of intron loss events in
262 mammals. *Genome Research* 17:23-32. DOI 10.1101/gr.5703406
263 Crooks GE, Hon G, Chandonia JM, and Brenner SE. 2004. WebLogo: a sequence logo
264 generator. *Genome Research* 14:1188-1190. DOI 10.1101/gr.849004
265 Csuros M, Rogozin IB, and Koonin EV. 2011. A detailed history of intron-rich eukaryotic
266 ancestors inferred from a global survey of 100 complete genomes. *PLoS Computational*
267 *Biology* 7:e1002150. DOI 10.1371/journal.pcbi.1002150
268 Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel
269 A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P,
270 Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR,
271 Hubbard TJ, Notredame C, Harrow J, and Guigó R. 2012. The GENCODE v7 catalog of
272 human long noncoding RNAs: Analysis of their gene structure, evolution, and
273 expression. *Genome Research* 22:1775-1789. DOI 10.1101/gr.132159.111
274 Fablet M, Bueno M, Potrzebowski L, and Kaessmann H. 2009. Evolutionary origin and functions
275 of retrogene introns. *Molecular Biology and Evolution* 26:2147-2156. DOI
276 10.1093/molbev/msp125

277 Fawcett JA, Rouzé P, and Van de Peer Y. 2012. Higher intron loss rate in *Arabidopsis thaliana*
278 than *A. lyrata* is consistent with stronger selection for a smaller genome. *Molecular*
279 *Biology and Evolution* 29:849-859. DOI 10.1093/molbev/msr254
280 Gao X, and Lynch M. 2009. Ubiquitous internal gene duplication and intron creation in
281 eukaryotes. *Proceedings of the National Academy of Sciences of the United States of*
282 *America* 49:20818-20823. DOI 10.1073/pnas.0911093106
283 Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW,
284 Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE,
285 Kellis M, Regev A, Rinn JL, and Lander ES. 2009. Chromatin signature reveals over a
286 thousand highly conserved large non-coding RNAs in mammals. *Nature* 458:223-227.
287 DOI 10.1038/nature07672
288 Hellsten U, Aspden JL, Rio DC, and Rokhsar DS. 2011. A segmental genomic duplication
289 generates a functional intron. *Nature Communications* 2:454. DOI 10.1038/ncomms1461
290 Hooks KB, Delneri D, and Griffiths-Jones S. 2014. Intron Evolution in Saccharomycetaceae.
291 *Genome Biology and Evolution* 6:2543-2556. DOI 10.1093/Gbe/Evu196
292 Irimia M, and Roy SW. 2008. Evolutionary convergence on highly-conserved 3' intron structures
293 in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS*
294 *Genetics* 4:e1000148. DOI 10.1371/journal.pgen.1000148
295 Irimia M, and Roy SW. 2014. Origin of spliceosomal introns and alternative splicing. *Cold*
296 *Spring Harbor Perspectives in Biology* 6:a016071. DOI 10.1101/cshperspect.a016071
297 Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, and Roy SW. 2008. Origin of
298 introns by 'intronization' of exonic sequences. *Trends in Genetics* 24:378-381. DOI
299 10.1016/j.tig.2008.05.007
300 Jayakodi M, Jung JW, Park D, Ahn Y-J, Lee S-C, Shin S-Y, Shin C, Yang T-J, and Kwon HW.
301 2015. Genome-wide characterization of long intergenic non-coding RNAs (lincRNAs)
302 provides new insight into viral diseases in honey bees *Apis cerana* and *Apis mellifera*.
303 *BMC Genomics* 16:S1. DOI 10.1186/s12864-015-1868-7
304 Kang LF, Zhu ZL, Zhao Q, Chen LY, and Zhang Z. 2012. Newly evolved introns in human
305 retrogenes provide novel insights into their evolutionary roles. *BMC Evolutionary*
306 *Biology* 12:128. DOI 10.1186/1471-2148-12-128
307 Kapusta A, and Feschotte C. 2014. Volatile evolution of long noncoding RNA repertoires:
308 mechanisms and biological implications. *Trends in Genetics* 30:439-452. DOI
309 10.1016/j.tig.2014.08.004
310 Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Research* 12:656-664. DOI
311 10.1101/gr.229202
312 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL. 2013. TopHat2: accurate
313 alignment of transcriptomes in the presence of insertions, deletions and gene fusions.
314 *Genome Biology* 14:R36. DOI 10.1186/Gb-2013-14-4-R36
315 Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, and Giegerich R. 2001.
316 REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids*
317 *Research* 29:4633-4642. DOI 10.1093/nar/29.22.4633
318 Li H, and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
319 transform. *Bioinformatics* 25:1754-1760. DOI 10.1093/bioinformatics/btp324
320 Li W, Tucker AE, Sung W, Thomas WK, and Lynch M. 2009. Extensive, recent intron gains in
321 *Daphnia* populations. *Science* 326:1260-1262. DOI 10.1126/science.1179302

Li WL, Kuzoff R, Wong CK, Tucker A, and Lynch M. 2014. Characterization of newly gained introns in *Daphnia* populations. *Genome Biology and Evolution* 6:2218-2234. DOI 10.1093/Gbe/Evu174

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. DOI 10.1186/s13062-015-0056-7

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

Schwartz S, Gal-Mark N, Kfir N, Oren R, Kim E, and Ast G. 2009. *Alu* exonization events reveal features required for precise recognition of exons by the splicing machinery. *PLoS Computational Biology* 5:e1000300. DOI 10.1371/journal.pcbi.1000300

368 Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, and Ast G. 2008. Large-scale comparative
369 analysis of splicing signals and their corresponding splicing factors in eukaryotes.
370 *Genome Research* 18:88-103. DOI 10.1101/Gr.6818908
371 Spies N, Nielsen CB, Padgett RA, and Burge CB. 2009. Biased chromatin signatures around
372 polyadenylation sites and exons. *Molecular Cell* 36:245-254. DOI
373 10.1016/j.molcel.2009.10.008
374 Szczesniak MW, Ciomborowska J, Nowak W, Rogozin IB, and Makalowska I. 2011. Primate
375 and rodent specific intron gains and the origin of retrogenes with splice variants.
376 *Molecular Biology and Evolution* 28:33-37. DOI 10.1093/molbev/msq260
377 Tamura K, Stecher G, Peterson D, Filipski A, and Kumar S. 2013. MEGA6: Molecular
378 evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* 30:2725-
379 2729. DOI 10.1093/molbev/mst197
380 Torriani SFF, Stukenbrock EH, Brunner PC, McDonald BA, and Croll D. 2011. Evidence for
381 extensive recent intron transposition in closely related fungi. *Current Biology* 21:2017-
382 2022. DOI 10.1016/j.cub.2011.10.041
383 van der Burgt A, Severing E, de Wit Pierre JGM, and Collemare J. 2012. Birth of new
384 spliceosomal introns in fungi by multiplication of introner-like elements. *Current Biology*
385 22:1260-1265. DOI 10.1016/j.cub.2012.05.011
386 Verhelst B, Van de Peer Y, and Rouze P. 2013. The complex intron landscape and massive
387 intron invasion in a picoeukaryote provides insights into intron evolution. *Genome*
388 *Biology and Evolution* 5:2393-2401. DOI 10.1093/Gbe/Evt189
389 Wang Y, You FM, Lazo GR, Luo MC, Thilmony R, Gordon S, Kianian SF, and Gu YQ. 2013.
390 PIECE: a database for plant gene structure comparison and evolution. *Nucleic Acids*
391 *Research* 41:D1159-D1166. DOI 10.1093/nar/gks1109
392 Wang ZF, and Burge CB. 2008. Splicing regulation: From a parts list of regulatory elements to
393 an integrated splicing code. *RNA* 14:802-813. DOI 10.1261/rna.876308
394 Wang ZF, Rolish ME, Yeo G, Tung V, Mawson M, and Burge CB. 2004. Systematic
395 identification and analysis of exonic splicing silencers. *Cell* 119:831-845. DOI
396 10.1016/j.cell.2004.11.010
397 Willmann MR, Endres MW, Cook RT, and Gregory BD. 2011. The functions of RNA-dependent
398 RNA polymerases in *Arabidopsis*. *The Arabidopsis book/American Society of Plant*
399 *Biologists* 9. DOI 10.1199/tab.0146
400 Yang YF, Zhu T, and Niu DK. 2013. Association of intron loss with high mutation rate in
401 *Arabidopsis*: implications for genome size evolution. *Genome Biology and Evolution*
402 5:723-733. DOI 10.1093/gbe/evt043
403 Yang ZH. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology*
404 *and Evolution* 24:1586-1591. DOI 10.1093/molbev/msm088
405 Yenerall P, Krupa B, and Zhou L. 2011. Mechanisms of intron gain and loss in *Drosophila*.
406 *BMC Evolutionary Biology* 11:364. DOI 10.1186/1471-2148-11-364
407 Yenerall P, and Zhou L. 2012. Identifying the mechanisms of intron gain: progress and trends.
408 *Biology Direct* 7:29. DOI 10.1186/1745-6150-7-29
409 Zhan LL, Meng QH, Chen R, Yue Y, and Jin YF. 2014. Origin and evolution of a new retained
410 intron on the vulcan gene in Drosophila melanogaster subgroup species. *Genome* 57:567-
411 572. DOI 10.1139/gen-2014-0132

Zhu T, and Niu DK. 2013a. Frequency of intron loss correlates with processed pseudogene abundance: a novel strategy to test the reverse transcriptase model of intron loss. *BMC Biology* 11:23. DOI 10.1186/1741-7007-11-23

Zhu T, and Niu DK. 2013b. Mechanisms of intron loss and gain in the fission yeast *Schizosaccharomyces*. *PLoS ONE* 8:e61683. DOI 10.1371/journal.pone.0061683

Zhu ZL, Zhang Y, and Long MY. 2009. Extensive structural renovation of retrogenes in the evolution of the *Populus* genome. *Plant Physiology* 151:1943-1951. DOI 10.1104/pp.109.142984

Zong J, Yao X, Yin JY, Zhang DB, and Ma H. 2009. Evolution of the RNA-dependent RNA polymerase (RdRP) genes: Duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene* 447:29-39. DOI 10.1016/j.gene.2009.07.004

Figures

Figure 1. Alignments indicating an intron gain and a flanking insertion of coding sequence in the potato gene *PGSC0003DMG402000361*.

The orthologous genes used as references are *Solyc12g008410.1* in *S. lycopersicum*, *Capana09g000243* in *C. annuum*, and *NbS00003153g0003* in *N. benthamiana*. The orthologous region in eggplants was manually identified by the best reciprocal program, BLAST, and manually annotated. Only aligned sequences close to the intron variation are shown here. Abbreviations: Stub: *S. tuberosum*; Slyc: *S. lycopersicum*; Smel: *S. melongena*; Cann: *C. annuum*; Nben: *Nicotiana benthamiana*.

Figure 2. Identification of the intron gain in potatoes.

The phylogenetic tree was constructed using the coding sequences of the gene *PGSC0003DMG402000361* and its orthologs: *Solyc12g008410.1* in *S. lycopersicum*, *Sopen12g003370* in *S. pennellii*, *Capana09g000243* in *C. annuum*, and *NbS00003153g0003* in *N. benthamiana*, and the orthologous regions manually annotated in *S. commersonii*, *S. habrochaites*, and *S. melongena*. The tree is not scaled according to substitution rates. As the untranslated regions have not been annotated in *S. commersonii*, *S. lycopersicum*, *S. habrochaites*, or *C. annuum*, the presented sequences start from the initiation codon ATG. In the schematic diagram of gene structures, boxes represent exons and horizontal lines represent introns. Due to the limited space, two extraordinarily long introns are not scaled according to their lengths. They are represented by broken lines. The new intron/exon structure is marked in red color. Abbreviations: Stub: *S. tuberosum*; Scom: *S. commersonii*; Slyc: *S. lycopersicum*; Shab: *S. habrochaites*; Spen: *S. pennellii*; Smel: *S. melongena*; Cann: *C. annuum*; Nben: *Nicotiana benthamiana*.

Figure 3. An intron gained by tandem genomic duplication within 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 recruitment of a 10 bp exonic segment (marked in red line). **(B)** Alignment of the two copies of the duplication. The splicing sites, the putative branch site, the polypyrimidine tract, and putative exonic splicing enhancers (TCAGCT, CAGCTC and

GAGGAA) are underlined. A cryptic 5′ exonic splicing enhancer, GAGGAA, and a cryptic 5′ splicing signal, GTAAG, was activated by the duplication event. This duplication was also found in the orthologous region of the wild potato *S. commersonii*. Besides this duplication, we also detected another 83 bp tandem genomic duplication within the first intron of the gene *PGSC0003DMG402000361*, but not in the orthologous region of *S. commersonii*. The second duplication did not change the intron/exon structure of the gene *PGSC0003DMG402000361*. So it is not described here in detail. Sites differing between the two copies are indicated with green 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 indicating an intron gain and a flanking insertion of coding sequence in the potato gene *PGSC0003DMG402000361.*

The orthologous genes used as references are *Solyc12g008410.1* in *S. lycopersicum*, *Capana09g000243* in *C. annuum*, and *NbS00003153g0003* in *N. benthamiana*. The orthologous region in eggplants was manually identified by the best reciprocal program, BLAST, and manually annotated. Only aligned sequences close to the intron variation are shown here. Abbreviations: Stub: *S. tuberosum*; Slyc: *S. lycopersicum*; Smel: *S. melongena*; Cann: *C. annuum*; Nben: *Nicotiana benthamiana*.

```
Stub   GAGACAGGGCCAGCACACGAGGAATAAGTGAACAGTTGCTGGCACTCAgtaagcttgaat
Slyc   GAGACAGGGCCAGCACACGAGGAATAAGTGAACAGTTGCTGGCACTCA------------
Smel   GAGACAGGGCCAGCACACGAGGAATAAGTGAACAGTTGCTGGCACTCA------------
Cann   GAAACAGTGCCAGCACACGAGGAATAAGTGAGCAGTTACTTGCACTCA------------
Nben   GAAACAGGGCTGGCATACAAGGAATAAGCGAACAGTTGCTGGCACTCA------------

Stub   tagtttaggcttaatgaagaacttgttcaatttttttattggttgcgatctcttcttcct
Slyc   ------------------------------------------------------------
Smel   ------------------------------------------------------------
Cann   ------------------------------------------------------------
Nben   ------------------------------------------------------------

Stub   ttttttgcatatttacaactctacatgtaaactatgttgctcggactctcaaaaactgtt
Slyc   ------------------------------------------------------------
Smel   ------------------------------------------------------------
Cann   ------------------------------------------------------------
Nben   ------------------------------------------------------------

Stub   gaacccgtgttggattctccaaaatgcactacttttggagtattcgatacacacttttga
Slyc   ------------------------------------------------------------
Smel   ------------------------------------------------------------
Cann   ------------------------------------------------------------
Nben   ------------------------------------------------------------

Stub   agagtccgaacaacacaacatgtaatgtactcagacctttcaagaattctagtttaccaa
Slyc   ------------------------------------------------------------
Smel   ------------------------------------------------------------
Cann   ------------------------------------------------------------
Nben   ------------------------------------------------------------

Stub   tgatggtcttccaatgcctgcagACATAGTAGGTGATGCATCTGATTCTCCTACATCAGC
Slyc   ------------------------------------------------------------
Smel   ------------------------------------------------------------
Cann   ------------------------------------------------------------
Nben   ------------------------------------------------------------

Stub   TCCACGAATACCATCACCTCCAATGAGTCCAGTGACAACTAGCTTTCAAAGAGATCATTA
Slyc   ------------------------------------------------------------
Smel   ------------------------------------------------------------
Cann   ------------------------------------------------------------
Nben   ------------------------------------------------------------

Stub   CGATCCTAGGCCATCTACATTCAGAGACAGGGCCAGCACACGAGGAATAAGTGAGCAGTT
Slyc   ------------------------------------------------------------
Smel   ------------------------------------------------------------
Cann   ------------------------------------------------------------
Nben   ------------------------------------------------------------

Stub   ACTGGCACTCAGTAAGCTTGAATTCAGGAAATTCTTTTTGATTCTAAACTACATTGGGAG
Slyc   ------------ATAAGCTTGAATTCAGGAAATTCTTTTTGATTCTAAACTACATTGGGAG
Smel   ------------GTACGCTTGAATTCAGGAAATTCTTTTTGATTCTAAACTACATTGGGAG
Cann   ------------ATAAGCTTGAATTCAGGAAATTCTTTCTGATTCTGAATTACATAGGGAG
Nben   ------------GTGATGTTGAGTTCAGGAAATTATTTTTGATTCTACACTACATTGGAAG
```

Figure 2(on next page)

Identification of the intron gain in potatoes.

The phylogenetic tree was constructed using the coding sequences of the gene *PGSC0003DMG402000361* and its orthologs: *Solyc12g008410.1* in *S. lycopersicum*, *Sopen12g003370* in *S. pennellii*, *Capana09g000243* in *C. annuum*, and *NbS00003153g0003* in *N. benthamiana*, and the orthologous regions manually annotated in *S. commersonii*, *S. habrochaites*, and *S. melongena*. The tree is not scaled according to substitution rates. As the untranslated regions have not been annotated in *S. commersonii*, *S. lycopersicum*, *S. habrochaites*, or *C. annuum*, the presented sequences start from the initiation codon ATG. In the schematic diagram of gene structures, boxes represent exons and horizontal lines represent introns. Due to the limited space, two extraordinarily long introns are not scaled according to their lengths. They are represented by broken lines. The new intron/exon structure is marked in red color. Abbreviations: Stub: *S. tuberosum*; Scom: *S. commersonii*; Slyc: *S. lycopersicum*; Shab: *S. habrochaites*; Spen: *S. pennellii*; Smel: *S. melongena*; Cann: *C. annuum*; Nben: *Nicotiana benthamiana*.

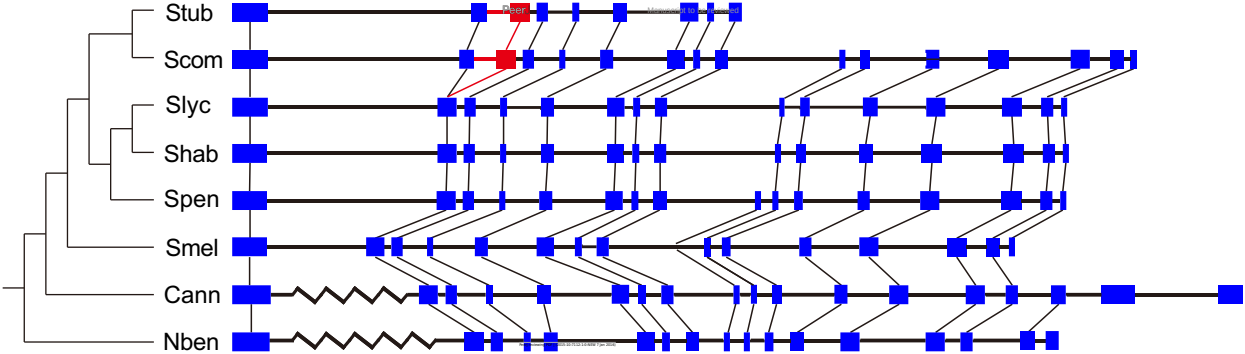
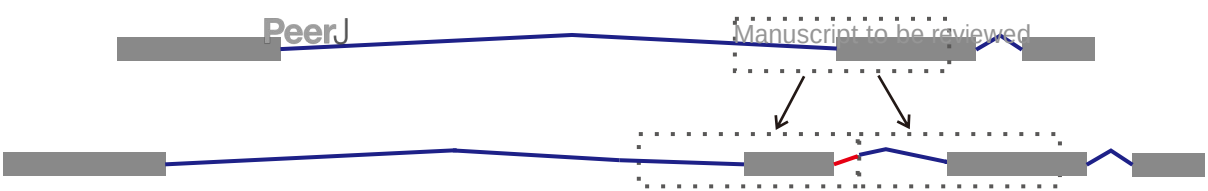

Stub
Scom
Slyc
Shab
Spen
Smel
Cann
Nben

Figure 3(on next page)

An intron gained by tandem genomic duplication within 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 recruitment of a 10 bp exonic segment (marked in red line). **(B)** Alignment of the two copies of the duplication. The splicing sites, the putative branch site, the polypyrimidine tract, and putative exonic splicing enhancers (TCAGCT, CAGCTC and GAGGAA) are underlined. A cryptic 5′ exonic splicing enhancer, GAGGAA, and a cryptic 5′ splicing signal, GTAAG, was activated by the duplication event. This duplication was also found in the orthologous region of the wild potato *S. commersonii*. Besides this duplication, we also detected another 83 bp tandem genomic duplication within the first intron of the gene *PGSC0003DMG402000361*, but not in the orthologous region of *S. commersonii*. The second duplication did not change the intron/exon structure of the gene *PGSC0003DMG402000361*. So it is not described here in detail. Sites differing between the two copies are indicated with green 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 attagtttaggcttaatgaagaacttgttcaattttttattggttgcgatctcttcttc
Upstream   attagtttaggcttaatgaagaacttgttcaaattttttattggttgcgatctcttcttc

Downstream c-ttttttgcatatttacaactctacatgtaaactatgttgctcggactctcaaaaact
Upstream   cttttttttgcatatttacaactctacatgtaaactatgttgctcggactctcaaaaact

Downstream gttgaacccgtgttggattctccaaaatgcactacttttggagtattcgatacacacttt
Upstream   gttgaacccgtgttggattctccaaaatgcactacttttggagtattcgatacacacttt

Downstream tgaagagtccgaacaacacaacatgt-aatgtactcagacctttcaagaattctagtttta
Upstream   tgaagagtctgaacaacacaacatgtaaacatactcagacctttcaaaaattctagtttta

Downstream ccaatgatggtcttccaatgcctgcagACATAGTAGGTGATGCATCTGATTCTCCTACAT
Upstream   ccaatgatggtcttccaatgcctacagACATAGTAGGTGATGCATCTGATTCTCCTACAT

Downstream CAGCTCCACGAATACCATCACCTCCAATGAGTCCAGTGACAACTAGCTTTCAAAGAGATC
Upstream   CAGCTCCACGAATACCATCACCTCCAATGAGTCCAGTGACAACTAGCTTTCAAAGAGATC

Downstream ATTACGATCCTAGGCCATCTACATTCAGAGACAGGGCCAGCACACGAGGAATAAGTGAGC
Upstream   ATTACGATCCTAGGCCATCTACATTCAGAGACAGGGCCAGCACACGAGGAATAAGTGAAC

Downstream AGTTACTGGCACTCAGTAAGCTTGA
Upstream   AGTTGCTGGCACTCAgtaagcttga
```

C