Intron gain by tandem genomic duplication: a novel case in a potato gene encoding RNA-dependent RNA polymerase

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, China

ABSTRACT

The origin and subsequent accumulation of spliceosomal introns are prominent events in the evolution of eukaryotic gene structure. However, the mechanisms underlying intron gain remain unclear because there are few proven cases of recently gained introns. In an *RNA-dependent RNA polymerase* (*RdRp*) gene, we found that a tandem duplication occurred after the divergence of potato and its wild relatives among other *Solanum* plants. The duplicated sequence crosses the intron-exon boundary of the first intron and the second exon. A new intron was detected at this duplicated region, and it includes a small previously exonic segment of the upstream copy of the duplicated sequence and the intronic segment of the downstream copy of the duplicated sequence. The donor site of this new intron was directly obtained from the small previously exonic segment. 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 differed between the intron and exon. In the widely cited model of intron gain by tandem genomic duplication, the duplication of an AGGT-containing exonic segment provides the GT and AG splicing sites for the new intron. Our results illustrate that the tandem duplication model of intron gain should be diverse in terms of obtaining the proper splicing signals.

Subjects Bioinformatics, Computational Biology, Evolutionary Studies, Genomics
Keywords Intron gain, Tandem genomic duplication, Splicing, *RdRp*, *Solanum tuberosum*, *PGSC0003DMG402000361*, Intron/exon structure

Submitted 16 October 2015
Accepted 29 June 2016
Published 26 July 2016

Corresponding author
Deng-Ke Niu, dkniu@bnu.edu.cn

Academic editor
Christophe Dessimoz

Additional Information and Declarations can be found on page 8

DOI **10.7717/peerj.2272**

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, exogenous sequence insertion during double-strand-break repair, group II intron insertion, intron transfer, intronization and introner-like element insertion (*van der Burgt et al., 2012*; *Yenerall & Zhou, 2012*). Comparative analyses of discordant intron positions among conserved homologous

How to cite this article Ma et al. (2016), Intron gain by tandem genomic duplication: a novel case in a potato gene encoding RNA-dependent RNA polymerase. **PeerJ 4:e2272; DOI 10.7717/peerj.2272**

genes have been conducted in diverse eukaryotic lineages. Although dozens of papers have reported intron gains over the last twenty years (*Csuros, Rogozin & Koonin, 2011*; *Fablet et al., 2009*; *Hooks, Delneri & Griffiths-Jones, 2014*; *Irimia & Roy, 2014*; *Li et al., 2009*; *Li et al., 2014*; *Ma et al., 2015a*; *Roy & Gilbert, 2005*; *Roy & Penny, 2006*; *Torriani et al., 2011*; *van der Burgt et al., 2012*; *Verhelst, Van de Peer & Rouze, 2013*; *Yenerall, Krupa & Zhou, 2011*; *Yenerall & Zhou, 2012*; *Zhu & Niu, 2013a*), only a few studies have identified the source sequences of these gained introns (*Collemare et al., 2015*; *Denoeud et al., 2010*; *Hankeln et al., 1997*; *Simmons et al., 2015*; *Torriani et al., 2011*; *van der Burgt et al., 2012*; *Verhelst, Van de Peer & Rouze, 2013*). In most studies (e.g., *Knowles & McLysaght, 2006*; *Zhang, Yang & Niu, 2010*), the source sequences have only been identified for a few introns, whereas tens of intron gains have been reported. Therefore, most of the reported intron gains do not provide supporting evidence for intron gain models. The mechanisms underlying these intron gains might be undetectable because the evolutionary traces have been erased by random mutations. Unexpectedly, the source sequences of most recent intron gains could not be identified. For example, researchers could identify the source sequences of only one intron gain among the seven new introns gained after the recent divergence (two Mya) of *Drosophila persimilis* and *Drosophila pseudoobscura* (*Yenerall, Krupa & Zhou, 2011*). More astonishingly, among the 21 new introns that were gained in certain local populations of *Daphnia pulex*, researchers successfully identified the source sequence of only one intron (*Li et al., 2009*). Because of the lack of identified source sequences, the mechanisms underlying most intron gains are not understood. Therefore, researchers have attempted to draw general conclusions from a small number of cases. Among the traditional models, intron gains by tandem genomic duplication should not occur at a low frequency because internal gene duplications are commonly observed (*Gao & Lynch, 2009*). This model was originally advanced by *Rogers (1989)*, who suggested that the 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, a new intron is derived from the duplication of an exonic sequence, and the translated peptide is not altered by the intron gain. An example that is consistent with this model is the vertebrate gene *ATP2A1* (*Hellsten et al., 2011*). The duplicated region of *ATP2A1* has the AGGT signal and also includes a polypyrimidine tract and a branch point. In addition, the generation of the intron has been experimentally reproduced 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 has been observed in the genes *SPOG_01682* and *SOCG_00815*. A 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 the *Arabidopsis TOUCH3* gene, *Knowles & McLysaght (2006)* observed two tandem internal gene duplications that duplicated an entire preexisting intron along with the exonic sequences on both sides of the intron. However, this finding does not represent the creation of new introns by tandem duplication but rather the multiplication of a preexisting intron by tandem duplication. Segmental duplication containing entire introns has also been observed by *Gao & Lynch (2009)*. In the present paper, we confine our discussion of intron gain to the creation of new introns rather than the propagation of preexisting introns.

By comparing the orthologous genes of *Solanum lycopersicum*, *Solanum tuberosum*, and other Solanaceae plants, we found 11 cases of precise intron loss and six cases of imprecise intron loss (*Ma et al., 2015b*). Moreover, we found indications of an intron gain in one of the potato RNA-dependent RNA polymerase (*RdRp*) genes, *PGSC0003DMG402000361* (Fig. 1). The *RdRp* genes encode enzymes that catalyze the replication of RNA from an RNA template, and these genes 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 foreign nucleic acids (*Willmann et al., 2011*; *Zong et al., 2009*). In this study, we confirmed that the new intron was created by the duplication of a gene segment crossing one intron-exon boundary. The 5′ donor site of the new intron was activated by a cryptic donor site that previously occurred in the exonic region, whereas other splicing signals were inherited from the preexisting intron/exon structure.

MATERIALS AND METHODS

The genome sequences and annotation files of domesticated potato *S. tuberosum* (PGSC_DM_v3), domesticated tomato *S. lycopersicum* (ITAG2.3), wild tobacco *Nicotiana benthamiana* (version 1.0.1), and wild tomato *Solanum pennellii* (spenn_v2.0) were downloaded from the Sol Genomics Network (*Bombarely et al., 2011*), and those for hot 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), another 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 scaffold sequences and annotation files of *Mimulus guttatus* (version 2.0) were downloaded from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html). The following files were retrieved from the Sequence Read Archive of the NCBI (http://www.ncbi.nlm.nih.gov/sra/): 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*; the transcriptomes (SRP015739 and SRP018993) of *S. lycopersicum*; the transcriptome (SRP067562) of *S. pennellii*; the transcriptome (SRP019256) of *C. annuum*; and the transcriptome (SRP018508) of *N. benthamiana*. We mapped the RNA-Seq reads to the genomes using TopHat version 2.0.8 (*Kim et al., 2013*), whereas BWA (alignment via Burrows-Wheeler transformation, version 0.5.7) (*Li & Durbin, 2009*) was used for the WGS reads. We used the default parameters for both programs, although the minimum intron length was adjusted to 20 bp for TopHat.

The orthologous genes of *PGSC0003DMG402000361* were identified using the best reciprocal BLAST hits with a threshold E value of $< 10^{-10}$. In addition, the orthologous relationship between the genes in *S. tuberosum* and *S. lycopersicum* was confirmed by their synteny using SynMap (http://genomevolution.org/CoGe/SynMap.pl). Using the RNA-Seq data, we examined the available annotations of the *RdRp* genes in *S. lycopersicum*, *S. pennellii*, *N. benthamiana*, and *C. annuum*. The annotations in *S. pennellii* and

```
Stub  QRDHYDPRPSTFRDR--ASTRGISEQLLALN1IVGDASDSPTSAPRIPSPPMSPVTTSFQ
Slyc  QKCHYDPSPSKFRDR--ASTRGISEQLLALN0----------------------------
Smel  QRDHYDPKPSEFRDR--ASTRGISEQLLALS0----------------------------
Cann  QRDHYARRSSEFRLRNSASTRGISEQLLALS0----------------------------
Nben  QRESCDPMPSEYRNR--AGIQGISEQLLALS0----------------------------

Stub  RDHYDPRPSTFRDRASTRGISEQLLALSKLEFRKFFLILNYIGR1RKVEDVIMLHDVGDI
Slyc  ----------------------------KLEFRKFFLILNYIGR1RKVEDVITLHDVGDI
Smel  ----------------------------TLEFRKFFLILNYIGR1RKVEDVIMLHDIGDI
Cann  ----------------------------KLEFRKFFLILNYIGR1RKLEDVIMLHDVGDI
Nben  ----------------------------DVEFRKLFLILHYIGR1RKLEDVIMLHDVGEI
```

Figure 1 Alignment of protein sequences close to the intron variation site. The presence and absence of the intron are represented by 1 and 0, respectively. The genes shown in this figure are *PGSC0003DMG402000361* in *S. tuberosum*, *Solyc12g008410.1* in *S. lycopersicum*, *Capana09g000243* in *C. annuum*, and *Niben101Scf04189g00002* in *N. benthamiana*. The orthologous region in eggplant was manually identified by the reciprocal best BLAST hits and manually annotated. Abbreviations: Stub, *S. tuberosum*; Slyc, *S. lycopersicum*; Smel, *S. melongena*; Cann, *C. annuum*; and Nben, *N. benthamiana*.

N. benthamiana have been confirmed, and those in *S. lycopersicum* and *C. annuum* have been revised (Data S1). The orthologous sequences in *S. commersonii*, *S. habrochaites*, and *S. melongena* were manually annotated with references to the annotations in *S. lycopersicum*, *C. annuum*, and *N. benthamiana*. The annotation files are provided in Data S1.

According to the annotation files of the domesticated potato genome, PGSC_DM_v3, the 3′ end of the gene *PGSC0003DMG402000361* overlaps with the 5′ end of the downstream gene *PGSC0003DMG401000361* (Fig. S1). The orthologous genomic regions of *C. annuum*, *M. guttatus*, *N. benthamiana*, *S. lycopersicum*, and *S. pennellii* present a long gene sequence rather than two overlapping genes. We examined the annotation of this overlapping region using the RNA-Seq data of *S. tuberosum* and found a paired read (Read ID: 127022 in SRR866275) that crosses the overlapping region. It appears that a long transcript similar to the orthologs in other species occurs in *S. tuberosum*. In addition, the 3′ end of the coding sequence of the gene *PGSC0003DMG402000361*, GTAATCTGA, has been annotated as the beginning of the ninth intron of the longer transcript (Fig. S1). In total, we found 37 RNA-Seq reads that support the removal of the ninth intron from certain mature mRNA molecules and four RNA-Seq reads supporting the presence of the small segment GTAATCTGA in other mature mRNA molecules. In addition, we found transcription termination signals at the ends of both transcripts using POLYAH (*Salamov & Solovyev, 1997*). It appears that the potato *RdRp* gene *PGSC0003DMG402000361* undergoes alternative cleavage and polyadenylation during transcription, which produces two isoforms with different lengths. We named the shorter one *PGSC0003DMG402000361.S* and the longer one *PGSC0003DMG402000361.L*. The newly identified intron is spliced from the identical region of these two transcripts; therefore, potential annotation errors in either transcript do not affect the validity of the identification of the new intron. For convenience, we present *PGSC0003DMG402000361.L* in this paper.

We found that the intron gain involved duplication using a 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*).

We searched the 5′ splicing sites, the branch sites, the polypyrimidine tracts, and the 3′ splicing sites according to *Irimia & Roy (2008)* and *Schwartz et al. (2008)*. The exonic splicing enhancers (ESEs) of *Arabidopsis thaliana* were identified by *Pertea, Mount & Salzberg (2007)* and used as the query in a search of the 50 bp exonic sequences upstream and downstream of the target intron.

The phylogenetic tree was constructed using MEGA 6.0 by employing the maximum likelihood method with the Tamura-Nei substitution model and uniform rates (*Tamura et al., 2013*). The number of bootstrap replications was 1,000. The schematic diagram of the gene structures was constructed using the program GSDraw (*Wang et al., 2013*).

RESULTS AND DISCUSSION

Among the cluster of orthologous genes for *RdRp*, the members of *S. tuberosum* and *S. commersonii* have 20 exons and those of the other Solanaceae species have 19 exons. A comparison of the annotations clearly showed that the second introns of *S. tuberosum* and *S. commersonii* are absent from the other Solanaceae genomes (Fig. 2). By analyzing the transcriptomic data of *S. tuberosum*, we found 106 RNA-Seq reads that were exclusively mapped to the annotated exon-exon boundary (Table S1; Fig. S2), which confirmed the annotation of this intron. Based on the phylogenetic tree constructed using the gene *PGSC0003DMG402000361.L* and its orthologs (Fig. 2), there were two possible explanations for the presence/absence of the intron: the gain of a new intron in the common ancestor of *S. tuberosum* and *S. commersonii*, or four independent intron loss events in the other four evolutionary branches (*S. lycopersicum–S. habrochaites–S. pennellii*; *S. melongena*; *C. annuum*; and *N. benthamiana*). According to the principle of parsimony, we concluded that the second intron of the gene *PGSC0003DMG402000361.L* was gained after the divergence of potato (*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 were used as a query sequence to search against the entire genome of *S. tuberosum*. We found that this insertion is a tandem genomic duplication (Fig. 2). The major part of the new intron and inserted exon region was a direct duplicate of the upstream intron-exon structure (Fig. 2). In addition, 10 nucleotides at the 5′ end of the new intron were recruited from the upstream exon (Fig. 2). Because two nearly identical regions in a reference genome might represent a true duplication or a false result caused by errors in genome assembly, we verified the duplication by examining the following three sources of evidence in *S. tuberosum*. First, 53 WGS reads were exclusively mapped crossing the three boundaries of two duplicates (Figs. S3–S5; Table S2). Second, 106 RNA-Seq reads were exclusively mapped crossing the mature mRNA exon boundary (Fig. S2; Table S1), which would not be observed in mature mRNA if the duplication had not occurred. Third, ten nucleotides were different between the duplicates (Fig. 3).

An intron is spliced out during the maturation of any RNA molecule, including protein-coding mRNAs and noncoding RNAs. In recent years, numerous spliced out sequences have been identified as originating in long noncoding RNAs and conclusively described as introns (*Derrien et al., 2012*; *Guttman et al., 2009*; *Jayakodi et al., 2015*;

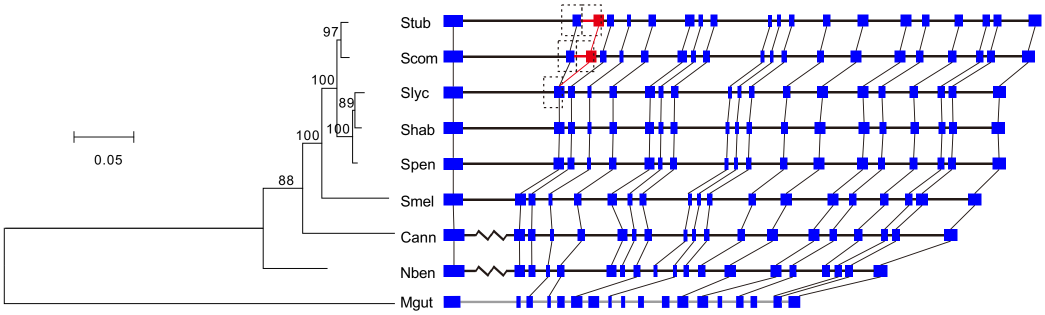

Figure 2 Identification of the intron gain in potatoes. The phylogenetic tree was constructed using the coding sequences of the gene *PGSC0003DMG402000361.L* and its orthologs *Solyc12g008410.1* in *S. lycopersicum*, *Sopen12g003370.2* in *S. pennellii*, *Capana09g000243* in *C. annuum*, *Niben101Scf04189g00002* in *N. benthamiana*, and *Migut.K00531* in *M. guttatus* as well as the orthologous regions manually annotated in *S. commersonii*, *S. habrochaites*, and *S. melongena*. Numbers above the branches indicate the percentage of bootstrap support after 1,000 replicates. In the schematic diagram of the gene structures, the presented sequences start from the initiation codon ATG, the boxes represent exons, and the horizontal lines represent introns. Because of space limitations, the extraordinarily long introns are not scaled according to their lengths, and they are represented by broken lines. To avoid crowding together the slashed lines, the introns of *M. guttatus* have been scaled up by a factor of two. The new intron/exon structure is marked in red. Abbreviations: Stub, *S. tuberosum*; Scom, *S. commersonii*; Slyc, *S. lycopersicum*; Shab, *S. habrochaites*; Spen, *S. pennellii*; Smel, *S. melongena*; Cann, *C. annuum*; Nben, *N. benthamiana*; Mgut, *M. guttatus*.

```
Downstream  attagtttaggcttaatgaagaacttgttcaattttttattggttgcgatctcttcttc
Upstream    attagtttaggcttaatgaagaacttgttcaaatttttattggttgcgatctcttcttc

Downstream  c-tttttttgcatatttacaactctacatgtaaactatgttgctcggactctcaaaaact
Upstream    cttttttttgcatatttacaactctacatgtaaactatgttgctcggactctcaaaaact

Downstream  gttgaacccgtgttggattctccaaaatgcactacttttggagtattcgatacacacttt
Upstream    gttgaacccgtgttggattctccaaaatgcactacttttggagtattcgatacacacttt

Downstream  tgaagagtccgaacaacacaacatgt-aatgtactcagacctttcaagaattctagttta
Upstream    tgaagagtctgaacaacacaacatgtaaacatactcagacctttcaaaaattctagttta

Downstream  ccaatgatggtcttccaatgcctgcagACATAGTAGGTGATGCATCTGATTCTCCTACAT
Upstream    ccaatgatggtcttccaatgcctacagACATAGTAGGTGATGCATCTGATTCTCCTACAT

Downstream  CAGCTCCACGAATACCATCACCTCCAATGAGTCCAGTGACAACTAGCTTTCAAAGAGATC
Upstream    CAGCTCCACGAATACCATCACCTCCAATGAGTCCAGTGACAACTAGCTTTCAAAGAGATC

Downstream  ATTACGATCCTAGGCCATCTACATTCAGAGACAGGGCCAGCACACGAGGAATAAGTGAGC
Upstream    ATTACGATCCTAGGCCATCTACATTCAGAGACAGGGCCAGCACACGAGGAATAAGTGAAC

Downstream  AGTTACTGGCACTCAGTAAGCTTGA
Upstream    AGTTGCTGGCACTCAgtaagcttga
```

Figure 3 Splicing signals of the new intron in the potato gene *PGSC0003DMG402000361*. 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′ splicing signal, GTAAG, was activated by the duplication event. This duplication was also found in the orthologous region of the wild potato *S. commersonii*. In addition to this duplication, we 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*; therefore, it is not described here in detail. Sites that differed between the two copies are indicated in green letters.

Kapusta & Feschotte, 2014). Therefore, the production of functional proteins by spliced RNA molecules should not be considered as a prerequisite for identifying a sequence spliced out of RNA molecules as a new intron. In this study, our search for evidence of intron gains is limited to whether the intron sequence occurred in the potato genome and whether the intron sequence has been removed from the mature mRNA. Although the WGS and RNA-Seq reads could demonstrate that the duplication is real and the intron is spliced, the best method of validating this assumption would be to perform a PCR assay for the genomic DNA and RT-PCR on RNA. This methodology underlies the objectives of our study.

According to *Logsdon, Stoltzfus & Doolittle (1998)*, strong evidence of intron gain must satisfy two conditions. The first is a clear phylogeny that provides support for the intron gain, and the second is an identified source element for the gained intron. Because of the clear phylogeny and the identity of the source sequence, we consider the second intron of the potato gene *PGSC0003DMG402000361.L* to be a well-supported case of a newly gained intron.

According to the tandem genomic duplication model originally proposed by *Rogers (1989)*, tandem duplication of an exonic segment harboring the AGGT sequence generates two splice sites for the new intron, 5′-GT and 3′-AG, and a new intron is derived from the duplication of the exonic sequence. However, 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 A and U compared with exons (*Amit et al., 2012*), and this difference is considered a requirement for efficient splicing (*Carle-Urioste, Brendel & Walbot, 1997*; *Luehrsen & Walbot, 1994*). At first glance, it appears unlikely that a coding segment will have a full set of splicing signals. However, the intronization of coding regions has been observed in several different organisms, including animals and plants (*Irimia et al., 2008*; *Kang et al., 2012*; *Szczesniak et al., 2011*; *Zhan et al., 2014*; *Zhu, Zhang & Long, 2009*). These observations indicate that it is possible for certain coding sequences to contain a full set of cryptic splice signals. Furthermore, an experimental duplication of a coding segment of the vertebrate gene *ATP2A2*, which harbors the AGGT sequence, has been shown to generate the new intron (*Hellsten et al., 2011*). Therefore, a full set of the splicing signals required for active splicing is present in the coding sequence of the gene *ATP2A2*. Although a full set of the splicing signals may have been contained in the coding sequences, we believe that utilization of the active splicing signals of the parental intron/exon structure represent a more frequent method of intron gain. In the potato gene *PGSC0003DMG402000361.L*, the duplication includes the 3′ side sequence of an intron and the 5′ side of the downstream exon (Fig. 2). The 3′ splicing site signal (CAG), the polypyrimidine tract (TCTTCCAATGCCT), and the putative branch site (TTTAC) of this novel intron were inherited from the parental intron (Fig. 3). Moreover, the two overlapped putative ESEs of the 3′ flanking exon, TCAGCT and CAGCTC, and the GC contents that were different between the intron and exon (36% *vs.* 46%) inherited from the parental copy (Fig. 3). The 5′ splicing signal of the novel intron GTAAG was activated from a cryptic splice site that was recruited from the upstream exon. This case of intron gain indicates that the tandem duplication model should not be narrowly considered according to its original proposal 27 years ago (*Rogers, 1989*).

CONCLUSIONS

In the last common ancestor of domesticated potato *S. tuberosum* and wild potato *S. commersonii*, a tandem duplication event in the gene *PGSC0003DMG402000361.L* 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 that included a putative branch site, a polypyrimidine tract, a 3′ splicing site, two putative ESEs, and GC contents that were differentiated between the intron and exon inherited from the parental intron/exon structure. However, the widely cited model of intron gain includes the tandem duplication of an exonic segment containing AGGT, which would create the GT and AG splicing sites. The case of intron gain observed here illustrates that the tandem duplication model of intron gain should be diversified so that the proper splicing signals can be obtained.

ACKNOWLEDGEMENTS

We are thankful to the anonymous referees for their useful comments and Sidra Aslam for her help improving the English in this paper.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Natural Science Foundation of China (grant numbers 31421063 and 31371283) and the Fundamental Research Funds for the Central Universities. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China: 31421063 and 31371283.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Ming-Yue Ma analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Xin-Ran Lan analyzed the data, reviewed drafts of the paper.
- Deng-Ke Niu conceived and designed the experiments, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Data Deposition

The following information was supplied regarding data availability:

The raw data has been supplied as Supplemental Dataset Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.2272#supplemental-information.

REFERENCES

Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Lev-Maor G, Burstein D, Schwartz S, Postolsky B, Pupko T, Ast G. 2012. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Reports* **1**(**5**)**:**543–556 DOI 10.1016/j.celrep.2012.03.013.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**(**2**)**:**573–580 DOI 10.1093/nar/27.2.573.

Bombarely A, Menda N, Tecle IY, Buels RM, Strickler S, Fischer-York T, Pujar A, Leto J, Gosselin J, Mueller LA. 2011. The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Research* **39**(**Suppl 1**)**:**D1149–D1155 DOI 10.1093/nar/gkq866.

Carle-Urioste JC, Brendel V, Walbot V. 1997. A combinatorial role for exon, intron and splice site sequences in splicing in maize. *The Plant Journal* **11**(**6**)**:**1253–1263 DOI 10.1046/j.1365-313X.1997.11061253.x.

Collemare J, Beenen HG, Crous PW, de Wit PJGM, van der Burgt A. 2015. Novel intronerlike elements in fungi are involved in parallel gains of spliceosomal introns. *PLoS ONE* **10**(**6**)**:** e129302 DOI 10.1371/journal.pone.0129302.

Csuros M, Rogozin IB, Koonin EV. 2011. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Computational Biology* **7**(**9**)**:** e1002150 DOI 10.1371/journal.pcbi.1002150.

Denoeud F, Henriet S, Mungpakdee S, Aury J-M, Da Silva C, Brinkmann H, Mikhaleva J, Olsen LC, Jubin C, Canestro C, Bouquet J-M, Danks G, Poulain J, Campsteijn C, Adamski M, Cross I, Yadetie F, Muffato M, Louis A, Butcher S, Tsagkogeorga G, Konrad A, Singh S, Jensen MF, Cong EH, Eikeseth-Otteraa H, Noel B, Anthouard V, Porcel BM, Kachouri-Lafond R, Nishino A, Ugolini M, Chourrout P, Nishida H, Aasland R, Huzurbazar S, Westhof E, Delsuc F, Lehrach H, Reinhardt R, Weissenbach J, Roy SW, Artiguenave F, Postlethwait JH, Manak JR, Thompson EM, Jaillon O, Du Pasquier L, Boudinot P, Liberles DA, Volff J-N, Philippe H, Lenhard B, Crollius HR, Wincker P, Chourrout D. 2010. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330**(**6009**)**:**1381–1385 DOI 10.1126/science.1194167.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research* **22**(**9**)**:**1775–1789 DOI 10.1101/gr.132159.111.

Fablet M, Bueno M, Potrzebowski L, Kaessmann H. 2009. Evolutionary origin and functions of retrogene introns. *Molecular Biology and Evolution* **26**(**9**)**:**2147–2156 DOI 10.1093/molbev/msp125.

Gao X, Lynch M. 2009. Ubiquitous internal gene duplication and intron creation in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **106**(**49**)**:**20818–20823 DOI 10.1073/pnas.0911093106.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458(7235):**223–227 DOI 10.1038/nature07672.

Hankeln T, Friedl H, Ebersberger I, Martin J, Schmidt ER. 1997. A variable intron distribution in globin genes of *Chironomus*: evidence for recent intron gain. *Gene* **205(1–2):**151–160 DOI 10.1016/S0378-1119(97)00518-0.

Hellsten U, Aspden JL, Rio DC, Rokhsar DS. 2011. A segmental genomic duplication generates a functional intron. *Nature Communications* **2:**454 DOI 10.1038/ncomms1461.

Hooks KB, Delneri D, Griffiths-Jones S. 2014. Intron evolution in saccharomycetaceae. *Genome Biology and Evolution* **6(9):**2543–2556 DOI 10.1093/gbe/evu196.

Irimia M, Roy SW. 2008. Evolutionary convergence on highly-conserved 3′ intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS Genetics* **4(8):** e1000148 DOI 10.1371/journal.pgen.1000148.

Irimia M, Roy SW. 2014. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harbor Perspectives in Biology* **6(6):**a016071 DOI 10.1101/cshperspect.a016071.

Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, Roy SW. 2008. Origin of introns by 'intronization' of exonic sequences. *Trends in Genetics* **24(8):**378–381 DOI 10.1016/j.tig.2008.05.007.

Jayakodi M, Jung JW, Park D, Ahn Y-J, Lee S-C, Shin S-Y, Shin C, Yang T-J, Kwon HW. 2015. Genome-wide characterization of long intergenic non-coding RNAs (lincRNAs) provides new insight into viral diseases in honey bees *Apis cerana* and *Apis mellifera*. *BMC Genomics* **16(1):**680 DOI 10.1186/s12864-015-1868-7.

Kang L-F, Zhu Z-L, Zhao Q, Chen L-Y, Zhang Z. 2012. Newly evolved introns in human retrogenes provide novel insights into their evolutionary roles. *BMC Evolutionary Biology* **12(1):**128 DOI 10.1186/1471-2148-12-128.

Kapusta A, Feschotte C. 2014. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends in Genetics* **30(10):**439–452 DOI 10.1016/j.tig.2014.08.004.

Kent WJ. 2002. BLAT–the BLAST-like alignment tool. *Genome Research* **12(4):**656–664 DOI 10.1101/gr.229202.

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

Knowles DG, McLysaght A. 2006. High rate of recent intron gain and loss in simultaneously duplicated *Arabidopsis* genes. *Molecular Biology and Evolution* **23(8):**1548–1557 DOI 10.1093/molbev/msl017.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research* **29(22):**4633–4642 DOI 10.1093/nar/29.22.4633.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25(14):**1754–1760 DOI 10.1093/bioinformatics/btp324.

Li W, Tucker AE, Sung W, Thomas WK, Lynch M. 2009. Extensive, recent intron gains in *Daphnia* populations. *Science* **326(5957):**1260–1262 DOI 10.1126/science.1179302.

Li WL, Kuzoff R, Wong CK, Tucker A, Lynch M. 2014. Characterization of newly gained introns in *Daphnia* populations. *Genome Biology and Evolution* **6(9):**2218–2234 DOI 10.1093/gbe/evu174.

Lim LP, 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(20):**11193–11198 DOI 10.1073/pnas.201407298.

Logsdon JM Jr, Stoltzfus A, Doolittle WF. 1998. Molecular evolution: recent cases of spliceosomal intron gain? *Current Biology* **8(16):**R560–R563 DOI 10.1016/S0960-9822(07)00361-2.

Luehrsen KR, 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(3):**449–463 DOI 10.1007/BF00024113.

Ma M-Y, Che X-R, Porceddu A, Niu D-K. 2015a. Evaluation of the mechanisms of intron loss and gain in the social amoebae Dictyostelium. *BMC Evolutionary Biology* **15(1):**286 DOI 10.1186/s12862-015-0567-y.

Ma M-Y, Zhu T, Li X-N, Lan X-R, Liu H-Y, Yang Y-F, Niu D-K. 2015b. Imprecise intron losses are less frequent than precise intron losses but are not rare in plants. *Biology Direct* **10(1):**24 DOI 10.1186/s13062-015-0056-7.

Pertea M, Mount SM, Salzberg SL. 2007. A computational survey of candidate exonic splicing enhancer motifs in the model plant *Arabidopsis thaliana. BMC Bioinformatics* **8(1):**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, 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(14):**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, 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(16):**5773–5778 DOI 10.1073/pnas.0500383102.

Roy SW, Penny D. 2006. Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. *Molecular Biology and Evolution* **23(12):**2259–2262 DOI 10.1093/molbev/msl098.

Salamov AA, Solovyev VV. 1997. Recognition of 3'-processing sites of human mRNA precursors. *CABIOS* **13(1):**23–28 DOI 10.1093/bioinformatics/13.1.23.

Schwartz S, Gal-Mark N, Kfir N, Oren R, Kim E, Ast G. 2009. *Alu* exonization events reveal features required for precise recognition of exons by the splicing machinery. *PLoS Computational Biology* **5(3):**e1000300 DOI 10.1371/journal.pcbi.1000300.

Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G. 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Research* **18(1):**88–103 DOI 10.1101/gr.6818908.

Simmons MP, Bachy C, Sudek S, van Baren MJ, Sudek L, Ares M Jr, Worden AZ. 2015. Intron invasions trace algal speciation and reveal nearly identical arctic and antarctic micromonas populations. *Molecular Biology and Evolution* **32(9):**2219–2235 DOI 10.1093/molbev/msv122.

Spies N, Nielsen CB, Padgett RA, Burge CB. 2009. Biased chromatin signatures around polyadenylation sites and exons. *Molecular Cell* **36(2):**245–254 DOI 10.1016/j.molcel.2009.10.008.

Szczesniak MW, Ciomborowska J, Nowak W, Rogozin IB, Makalowska I. 2011. Primate and rodent specific intron gains and the origin of retrogenes with splice variants. *Molecular Biology and Evolution* **28(1):**33–37 DOI 10.1093/molbev/msq260.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* **30(12):**2725–2729 DOI 10.1093/molbev/mst197.

Torriani SFF, Stukenbrock EH, Brunner PC, McDonald BA, Croll D. 2011. Evidence for extensive recent intron transposition in closely related fungi. *Current Biology* **21(23):**2017–2022 DOI 10.1016/j.cub.2011.10.041.

van der Burgt A, Severing E, de Wit PJGM, Collemare J. 2012. Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. *Current Biology* **22(13):**1260–1265 DOI 10.1016/j.cub.2012.05.011.

Verhelst B, Van de Peer Y, Rouze P. 2013. The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. *Genome Biology and Evolution* **5(12):**2393–2401 DOI 10.1093/gbe/evt189.

Wang Y, You FM, Lazo GR, Luo MC, Thilmony R, Gordon S, Kianian SF, Gu YQ. 2013. PIECE: a database for plant gene structure comparison and evolution. *Nucleic Acids Research* **41(D1):** D1159–D1166 DOI 10.1093/nar/gks1109.

Wang ZF, Burge CB. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14(5):**802–813 DOI 10.1261/rna.876308.

Wang ZF, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* **119(6):**831–845 DOI 10.1016/j.cell.2004.11.010.

Willmann MR, Endres MW, Cook RT, Gregory BD. 2011. The functions of RNA-dependent RNA polymerases in *Arabidopsis*. *The Arabidopsis Book* **9:**e146 DOI 10.1199/tab.0146.

Yenerall P, Krupa B, Zhou L. 2011. Mechanisms of intron gain and loss in *Drosophila*. *BMC Evolutionary Biology* **11(1):**364 DOI 10.1186/1471-2148-11-364.

Yenerall P, Zhou L. 2012. Identifying the mechanisms of intron gain: progress and trends. *Biology Direct* **7(1):**29 DOI 10.1186/1745-6150-7-29.

Zhan LL, Meng QH, Chen R, Yue Y, Jin YF. 2014. Origin and evolution of a new retained intron on the vulcan gene in Drosophila melanogaster subgroup species. *Genome* **57(10):**567–572 DOI 10.1139/gen-2014-0132.

Zhang L-Y, Yang Y-F, Niu D-K. 2010. Evaluation of models of the mechanisms underlying intron loss and gain in *Aspergillus* fungi. *Journal of Molecular Evolution* **71(5–6):**364–373 DOI 10.1007/s00239-010-9391-6.

Zhu T, Niu D-K. 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(1):**23 DOI 10.1186/1741-7007-11-23.

Zhu T, Niu D-K. 2013b. Mechanisms of intron loss and gain in the fission yeast *Schizosaccharomyces*. *PLoS ONE* **8(4):**e61683 DOI 10.1371/journal.pone.0061683.

Zhu ZL, Zhang Y, Long MY. 2009. Extensive structural renovation of retrogenes in the evolution of the *Populus* genome. *Plant Physiology* **151(4):**1943–1951 DOI 10.1104/pp.109.142984.

Zong J, Yao X, Yin JY, Zhang DB, 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(1):**29–39 DOI 10.1016/j.gene.2009.07.004.