Structural dynamics and divergence of the polygalacturonase gene family in land plants

Soon-Jae Kwon[1,2]*, Jae-Han Son[1,2]*, Kyong-Cheul Park[1,2]*, Hae-Young Oh[1,2], Pyeung-Hyeun Kim[1,2], Woo-Hyeon Byeon[1,2] & Nam-Soo Kim[1,2]

[1]Department of Molecular Biosciences, Kangwon National University Chunchon, Korea, 200-701

[2]Institute of Biosciences and Biotechnology, Kangwon National University, Chunchon, Korea 200-701

*These authors contributed equally to this work.

A distinct feature of eukaryotic genomes is the presence of gene families. The polygalacturonase (PG) (EC3.2.1.15) gene family is one of the largest gene families in plants. PG is a pectin-digesting enzyme with a glycoside hydrolase 28 domain. It is involved in numerous plant developmental processes. The evolutionary processes accounting for the functional divergence and the specialized functions of PGs in land plants are unclear. Here, phylogenetic and gene structure analysis of PG genes in algae and land plants revealed that land plant PG genes resulted from differential intron gain and loss, with the latter event predominating. PG genes in land plants contained 15 homologous intron blocks and 13 novel intron blocks. Intron position and phase were not conserved between PGs of algae and land plants but conserved among PG genes of land plants from moss to vascular plants, indicating that the current introns in the PGs in land plants appeared after the split between unicellular algae and multicelluar land plants. These findings demonstrate that the functional divergence and differentiation of PGs in land plants is attributable to intronic loss. Moreover, they underscore the importance of intron gain and loss in genomic adaptation to selective pressure.

Gene families arose from a common ancestor by gene duplication. Purifying selection maintains functional redundancy of the duplicates, while by accumulation of evolutionary neutral or loss-of-function mutations erodes the functional redundancy of paralogs(*1*). Divergence of duplicated pairs is also attributable to gene interruption by introns. Introns are subject to relatively little selective pressure, resulting in rapid changes in the size and sequence of these structures. Nevertheless, a high conservation in intron and exon structures often exists, with intron positions and phase correspondence being noted in orthologs(*2, 3*). Since intron loss or gain is a rare event, comparison of the gene structures among gene families has been used to classify paralogs into subfamilies(*4*). This approach has revealed that, despite the effects of purifying selection on paralogs, PG gene structures within clades of the phylogenetic tree are highly preserved between monocotyledous and dicotyledonous plants(*4*). Here, we investigated the phylogenetic relationships, tandem and segmental duplications, expression, and gene structure dynamics in the whole sets of PGs in *Oryza sativa* (rice), *Arabidopsis thaliana* (flowering plant), *Populus trichocarpa* (poplar), *Physcomitrella patens* (moss), *Chlamydomonas reinhardtii* (green alga), *Phaeodactylum tricornutum* (diatom), and *Aureococcus anaphagefferens* (brown alga). The tissue-specific expression of 13 PGs from other plant species was also investigated to gain insight on the expression of these whole sets of PGs.

Two hundred-eight PGs (Supplementary Table 1) were grouped into six clades (clades A-F), based on their amino acid sequence similarity (Fig. 1). PGs were classified as endo-PGs (A and B), exo-PGs (C and D), or rhamno-PGs (E) as previously described(*5, 6*). Clade F members could not be clearly defined as either endo- or exo-PGs. All algal PGs and 8 of the 11 moss PGs were assigned to clade E. One of the remaining moss PGs was assigned to clade A, while the other was assigned to clade B. One moss PG (MS116593) was out-grouped with clades A-D and F, but not clade E, indicating that endo- and exo-PGs appeared after

multicellular land plants split from unicellular algae about 400 million years ago (Mya)(*7, 8*). Clades C, D, and F did not contain moss PGs, which implies that the PGs in these clades expanded after divergence of vascular plants from non-vascular Bryophytes (e.g., moss) about 200 Mya(*7*). The six clades were further classified into 24 sub-clades. Six of these sub-clades consisted of either dicots (C-I, D-II, D-III, D-IV) or monocots (C-III, D-I). Thus, these PGs might have resulted from lineage-specific expansion after monocotyledonous plants split from their dicotyledonous counterparts about 140-150 Mya(*9, 10*). Three sub-clades (E-X, E-II, A-II) contained PGs from species ranging from moss to dicotyledonous and monocotyledonous plants, implying that ancestral genomes of vascular plants contained at least three PG genes. Expression of the PGs was surveyed by analyzing expressed sequence tags (ESTs) in NCBI, TIGR, or species-specific genome databases (Supplementary Table 2). The PGs exhibited highly redundant expression in the same tissues as shown in previous studies(*11, 12*), and multiple sequence homology with ESTs of different tissues, in agreement with RT-PCR analysis in *A. thaliana(12)*. Among plant PGs, the rhamno-PGs in clade E were the most widely expressed, being matched with ESTs in multiple tissues. In accord with previous findings(*6, 12*), PGs in clade C were prominently expressed in floral organs, intimating that they had undergone functional divergence.

No tandem duplicated PG genes were present in algae and moss, but nine of the moss PG genes were segmentally duplicated (Supplementary Fig. 1). The percent of tandem or segmental duplication among the flowering plant PGs varied, with 45.5% duplication in rice, 46.2% in poplar, and 57% in *A. thaliana*. This rate of segmental duplication and tandem duplication suggests that gene family expansion occurs via whole-genome duplication (WGD)(*12-14*) (Supplementary Fig. 2). WGD and subsequent tandem, local, and regional duplications can distribute gene family members throughout the genome(*1, 15*). WGD has occurred once in rice(*16*) and three times in *A. thaliana* and poplar(*17, 18*). In *A. thaliana*,

many PG genes were derived from large-scale duplication and large numbers of these genes seemed to be lost by gene death(*12*, *19*). Homologous segmental duplication blocks containing PGs were identified using the plant genome duplication database (http://chibba.agtec.uga.edu/duplication/) for rice, *A. thaliana*, and poplar (Supplementary Fig. 2). For moss, the pair wise homology of the PG-flanking genes with those in vascular plants was determined using the CLUSTAL W program. The number of intergenomic PG genes shared between the two dicot plants (i.e., *A. thaliana* vs. poplar) was almost 3-fold greater than the number shared between dicots and monots (i.e., poplar vs. rice or *A. thaliana* vs. rice). More PG genes with segmentally duplicated homologous pairs were shared between dicots than between dicots and monocots, inferring that multiple WGDs and subsequent gene loss occurred after the split between *A. thaliana* and poplar. Five homologous collinear gene blocks were shared among *A. thaliana*, poplar, and rice. While most of the intergenomic segmentally duplicated PG genes were shared by the same phylogenetic clade, 11 sets of intergenomic segmentally duplicated PG genes were differentiated among separate clades. A set of PG genes were present in the intergenomic segmental duplicated gene blocks among moss, *A. thaliana*, poplar, and rice. However, the collinearity was not extended to the *C. reinhardtii* PG, Chlr73470. Chlr73470 was classified into sub-clade E-II, a sub-clade of PGs that exhibited remarkable conservation in gene structure (Fig. 2). Although the structures of PG genes in land plants in this sub-clade were almost identical, the Chrl73470 gene was highly dissimilar in intron/exon structure to these land plant PG genes. One PG from *Aspergillus oryzae* was previously coupled with plant PGs in this sub-clade with a high bootstrap value(*4*), but showed a different gene structure. Therefore, the gene order collinearity and PG gene structures among the sub-clade E-II land plants must have been conserved since their first appearance about 400 Mya(*7*).

Intron position and phase correspondence were analyzed among the PG genes within

and between clades. Seventeen homologous intron blocks were shared among the PG genes of land plants (Supplementary Tables 3 and 4). The PG genes of alga species did not show intron position correspondence with PG genes of land plants. Among the 17 homologous intron blocks, 15 were present in all land plants, with differential intron losses resulting in the current PG genes. One intron block gain (homologous intron block 1) and one loss (homologus intron block 9) occurred after divergence of vascular plants from nonvascular plants. Of the 17 homologous intron blocks, 15 were present within glycosyl hydrolase 28 domain. In land plants, PG genes contained 13 novel introns, likely from intron gain. No intron phase preference was found in these novel introns. The novel introns were more frequently present outside of the domain motifs. BLAST analysis revealed that 8 of the 13 novel introns contained MITE (Miniature Inverted-repeat Transposable Elements). However, MITE-driven intron insertion seemed improbable since MITEs are short and do not have a protein-coding function(*20, 21*).

During evolution, gain and loss of introns from three primary gene structures likely generated the current set of land plant PG genes (Fig. 3). Sequential and differential intron losses from a PG gene with a structure similar to MS133517 produced all exo- and endo-PG genes. In vascular plants, rhamno-PG genes are derived from two basic gene structures, MS43415 or MS163808. In contrast, rhamno-PG genes in sub-clade E-II did not undergo changes in gene structure during the evolution of land plants. Based on the present PG gene structures (Supplementary Fig. 3), scenarios for intron gain/loss were deduced (Supplementary Fig. 4). A summary of PG gene structures in land plants (i.e., moss, *A. thaliana*, poplar, and rice) is provided in Supplementary Table 3. Since the intron density among different eukaryotic taxa varies more than three orders magnitude(*22, 23*), frequency of intron loss and gain during evolution is a subject of debate(*22, 24-26*). An extensive study of more than 8,000 orthologs in *A. thaliana* and *O. sativa* revealed that intron losses were

12.6 and 9.8 times greater than intron gains in *A. thaliana* and *O. sativa*, respectively(*27*). Remarkably high intron position conservation was observed among plant, animal, and fungal species over 1.5 billion years(*2*). However, our analysis did not reveal any conservation of intron position or phase between algal PG genes or between algal and land plant PG genes. Intron positions were not conserved among the three PG genes found in *P. tricornutum*, in contrast to the intron position conservation seen among the enslaved nucleomorph-containing algae *Bigelowiella natans*, *C. reinhardtii*, and *A. thaliana(28)*. Algae and bryophytes appeared approximately 1000 and 400 Mya, respectively(*8, 29*). The lack of correspondence in intron position between algae as well as between algae and land plants, taken with the high conservation of intron position among land plants, implies that introns in current PG genes of land plants appeared after multicellular plants split from unicellular algae approximately 400 Mya. Our analysis revealed that intron and exon structures in the PG genes of land plants are highly conserved (gene structures summarized in Supplementary Table 5). Differential intron losses partitioned PG genes into separate clades, which were congruent with phylogenetic classification based on the amino acid sequence similarity. Intron losses dominated intron gains in shaping current PG genes in land plants. Intron gain and loss may be an important genomic adaptation, resulting in genome-specific intron structures during evolution.

Methods

Isolation of polygalacturonase sequences. PG sequences were accessed from species-specific genome databases using either polygalacturonase or glycosyl hydrolase family 28 as a query. Databases were accessed at http://www.tigr.org/tdb/e2k1/osa1/GeneNameSearch.shtml for rice, http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html for poplar, http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html for moss, http://genome.jgi-psf.org/Auran1/Auran1.home.html for *A. anophagefferens*, http://genome.jgi-psf.org/Phatr2/Phatr2.home.html for *P. tricornutum*, and http://genome.jgi-psf.org/Chlre3/Chlre3.home.html for *C. reinhardtii*. The sequences of PG genes from *Arabidopsis* were from Kim et al.(*12*), and sequences of 13 PG genes from other plants were from Park et al(*4*). PG gene sequences from *Cryptomeria japonica* and *Erwinia carotovora* were identified from the NCBI (http://www.ncbi.nlm.nih.gov) using the BLASTP program and the sequence of the glycosyl hydrolase family 28 domain as a query sequence.

Phylogenetic analysis. Multiple alignment of the glycoside hydrolase family 28 domain sequences of the 208 PGs was performed using the MAFFT program (http://align.bmr.kyushu-u.ac.jp/mafft/online/server/), and gaps in the aligned sequences were edited using MEGA4 software (http://megasoftware.net). The phylogenetic tree was constructed with MAFFT using the neighbor joining-JTT method with 100 bootstrap repetitions. The phylogenetic tree was retrieved using TreeView Version 1.6.6 (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

Gene structure analysis. GeneView (http://www.gramene.org/, http://genome.jgi-psf.org) was used to retrieve PG nucleotide sequences and to identify introns and exons. Gene structures of PGs in the same clade were phylogenetically analyzed. Consensus functional domain sequences in the PGs of

each clade were identified using CLUSTAL W (http://www.ebi.ac.uk/clustalw/), and the results were viewed and edited using T-View (http://www.ebi.ac.uk/t-coffee/help.html). Intron phases were manually analyzed from exon information. For intron position analysis, the protein sequences of each clade were analyzed by multiple sequence alignment using CLUSTAL W (http://www.ebi.ac.uk/clustalw/), and the intron locations were manually identified.

References

1. V. E. Prince, F. B. Pickett, *Nat Rev Genet* 3, 827 (2002).
2. I. B. Rogozin, Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, E. V. Koonin, *Curr Biol* 13, 1512 (2003).
3. A. Lecharny, N. Boudet, I. Gy, S. Aubourg, M. Kreis, *J Struct Funct Genomics* 3, 111 (2003).
4. K. C. Park, S. J. Kwon, P. H. Kim, T. Bureau, N. S. Kim, *Genome* 51, 30 (2007).
5. O. Markovic, S. Janecek, *Protein Eng* 14, 615 (2001).
6. K. A. Hadfield, A. B. Bennett, *Plant Physiol* 117, 337 (1998).
7. K. J. Willis, J. C. McElwain, *The evolution of plants*. (Oxford University Press, New York, 2002).
8. S. A. Rensing *et al.*, *Science* 319, 64 (2008).
9. O. Lespinet, Y. I. Wolf, E. V. Koonin, L. Aravind, *Genome Res* 12, 1048 (2002).
10. S. M. Chaw, C. C. Chang, H. L. Chen, W. H. Li, *Journal of Molecular Evolution* 58, 424 (2004).
11. M. Torki, P. Mandaron, R. Mache, D. Falconet, *Gene* 242, 427 (2000).
12. J. Kim, S. H. Shiu, S. Thoma, W. H. Li, S. E. Patterson, *Genome Biol* 7, R87 (2006).
13. K. Horan, J. Lauricha, J. Bailey-Serres, N. Raikhel, T. Girke, *Plant Physiol* 138, 47 (2005).
14. M. C. Nicole *et al.*, *BMC Genomics* 7, 223 (2006).
15. P. S. Soltis, *New Phytologist* 166, 5 (2005).
16. A. H. Paterson, J. E. Bowers, B. A. Chapman, *Proceedings of the National Academy of Sciences of the United States of America* 101, 9903 (2004).
17. J. E. Bowers, B. A. Chapman, J. Rong, A. H. Paterson, *Nature* 422, 433 (2003).

18. G. A. Tuskan *et al.*, *Science* 313, 1596 (2006).

19. H. Kong *et al.*, *The Plant Journal* 50, 873 (2007).

20. S. R. Wessler, T. E. Bureau, S. E. White, *Current Opinion in Genetics & Development* 5, 814 (1995).

21. S. W. Roy, *Genome Biol* 5, 251 (2004).

22. S. W. Roy, W. Gilbert, *Nat. Rev. Genet* 7, 211 (2006).

23. D. C. Jeffares, T. Mourier, D. Penny, *Trends in Genetics* 22, 16 (2006).

24. V. N. Babenko, I. B. Rogozin, S. L. Mekhedov, E. V. Koonin, *Nucleic Acids Research* 32, 3724 (2004).

25. D. G. Knowles, A. McLysaght, *Molecular Biology and Evolution* 23, 1548 (2006).

26. H. Lin, W. Zhu, J. C. Silva, X. Gu, C. R. Buell, *feedback*, (2006).

27. S. W. Roy, D. Penny, *Molecular Biology and Evolution* 24, 171 (2007).

28. P. R. Gilson *et al.*, *Proceedings of the National Academy of Sciences* 103, 9566 (2006).

29. S. S. Merchant *et al.*, *Science* 318, 245 (2007).

Acknowledgements This study was funded by a fellowship grant to S.J.K., J.H.S., K.C.P., and H.Y.O. from the second stage of the BK21 program from the Ministry of Education of Korea.

Figure Legends

Figure 1. Phylogenetic tree of PG from algae, moss, and plants. Ecpeh1 is a PG from *E. carotovora* that served as the out-group. Bootstrap values are indicated at each node. The PGs shown are from moss (MS), *A. thaliana* (AT), poplar (Pop), *O. sativa* (OS), *C. reinhardtii* (Chlr), *P. tricornutum* (Pt), *A. anaphagefferens* (Aa), *Nicotiana tobacuum* (Q05967), *Medicago trunculata* (AJ620946), *Zea mays* (P26216 and AF001000), *Solanum lycorpersicum* (O22313, O22310, Q96487, and P05117), *Brassica napus* (Q42399), *Actiana deleciosa* (L12019), *Turnera sublata* (AY185765), and *Cryptomeria japonica* (BAA06172).

Figure 2. PG gene structure and collinearity in selected organisms. (A) Structures of PG genes in sub-clade E-II and (B) gene order collinearity in rice, poplar, Arabidopsis, and moss. Orthologous genes are shown in the same color, and levels of sequence identity between genes in the collinear block are indicated with colored lines. The first and last nucleotides of the genes are shown in (B).

Figure 3. Gene structure dynamics PG genes in land plants. Endo- and exo-PGs are shown in (A), while rhamo-PGs are shown in (B, C). Clade designations are indicated by the color of the gene name, in accordance with the color classification in Fig. 1. Number of nucleotide is provided on the exons (thick bars), and intron phases (Phase 0-2) are provided on the introns (thin lines). Phase 0 introns are present between codons. Phase 1 introns are present between the first and second nucleotides in a codon. Phase 2 introns are present between the second and third codon nucleotide.

Supplementary information

Supplementary Figure S1. Segmentally duplicated chromosomal regions containing PG genes in *P. patens* (moss).

Supplementary Figure S2. Chromosomal distributions of the PG genes.

Supplementary Figure S3. PG gene structures.

Supplementary Figure S4. Intron losses and gains in PG genes of each clade.

Supplementary Table 1. Polypeptide sequences of the PGs

Supplementary Table 2. Expression analysis by EST matching of the PGs

Supplementary Table 3. Homologous introns in PGs in land plants

Supplementary Table 4. Novel introns in PG genes of land plants

Supplementary Table 5. Overalls of the gene structures of the land plant PGs

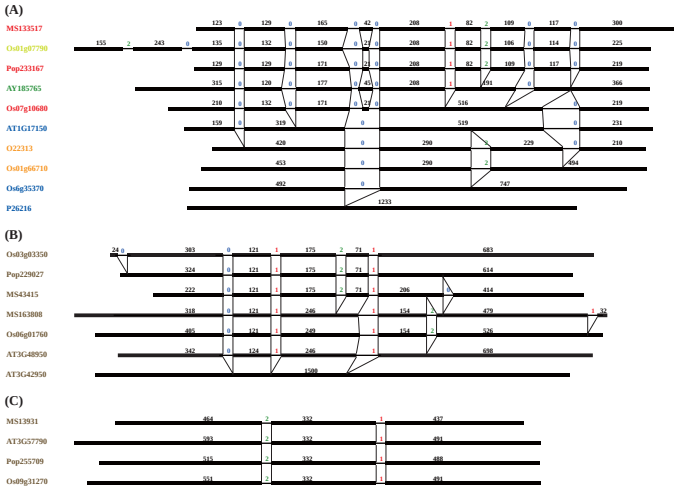


Figure 3.

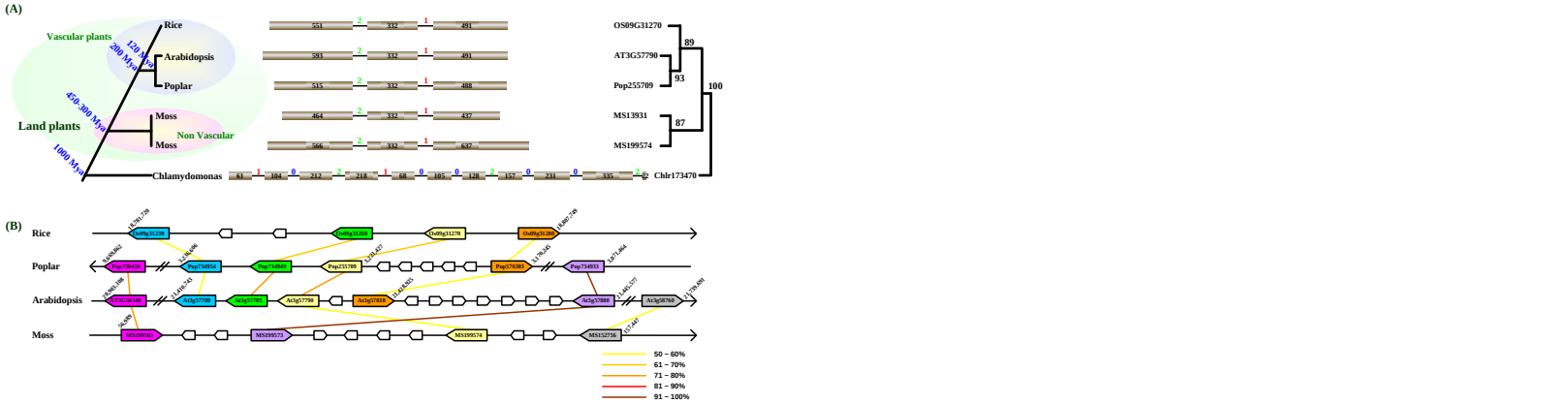


Figure 2.

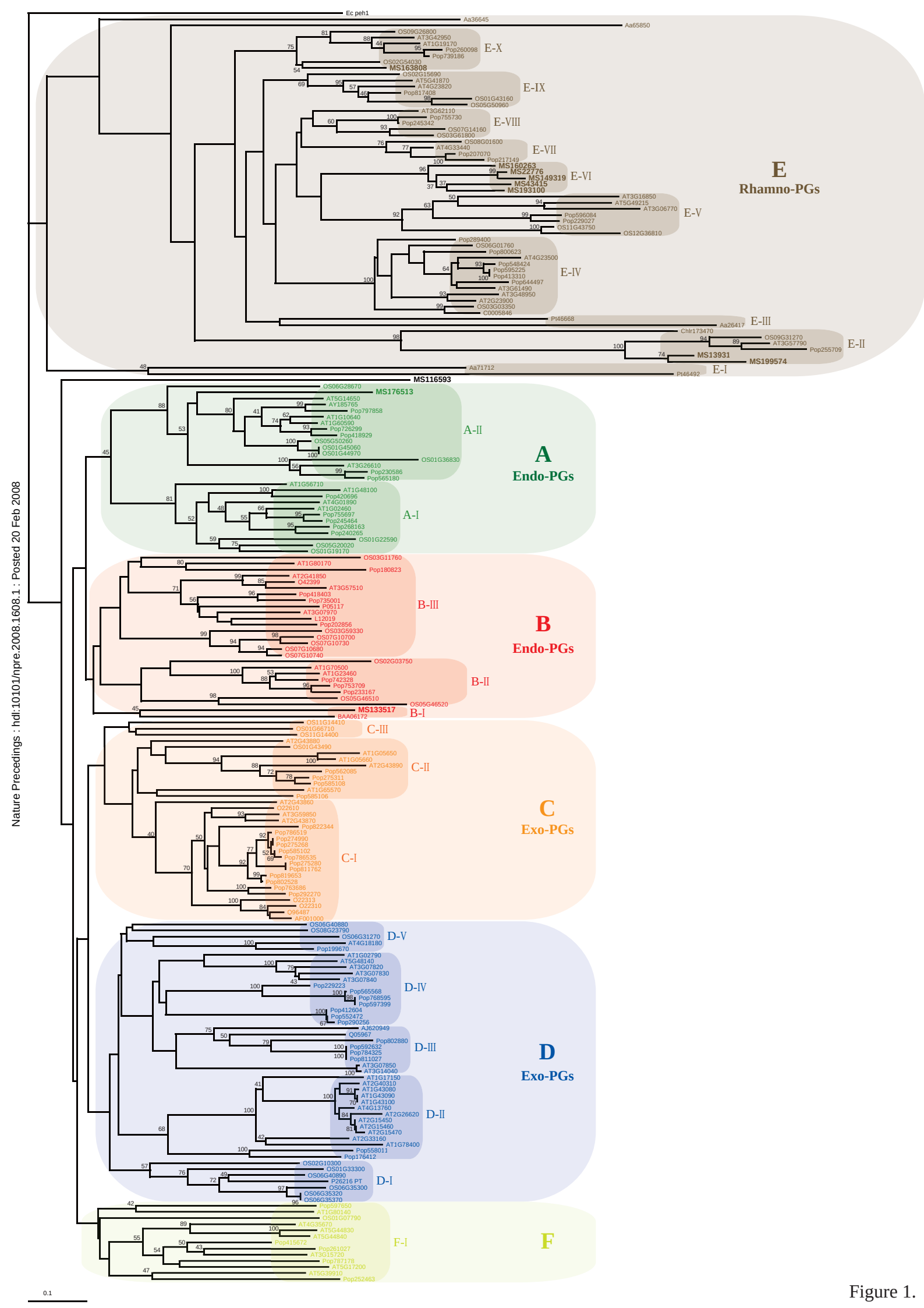
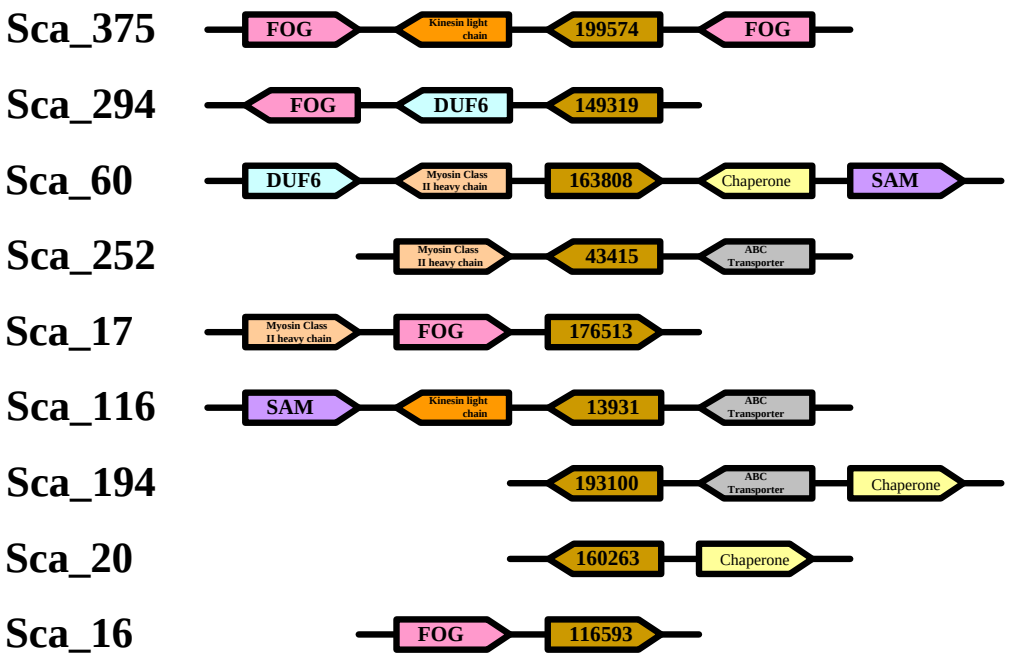
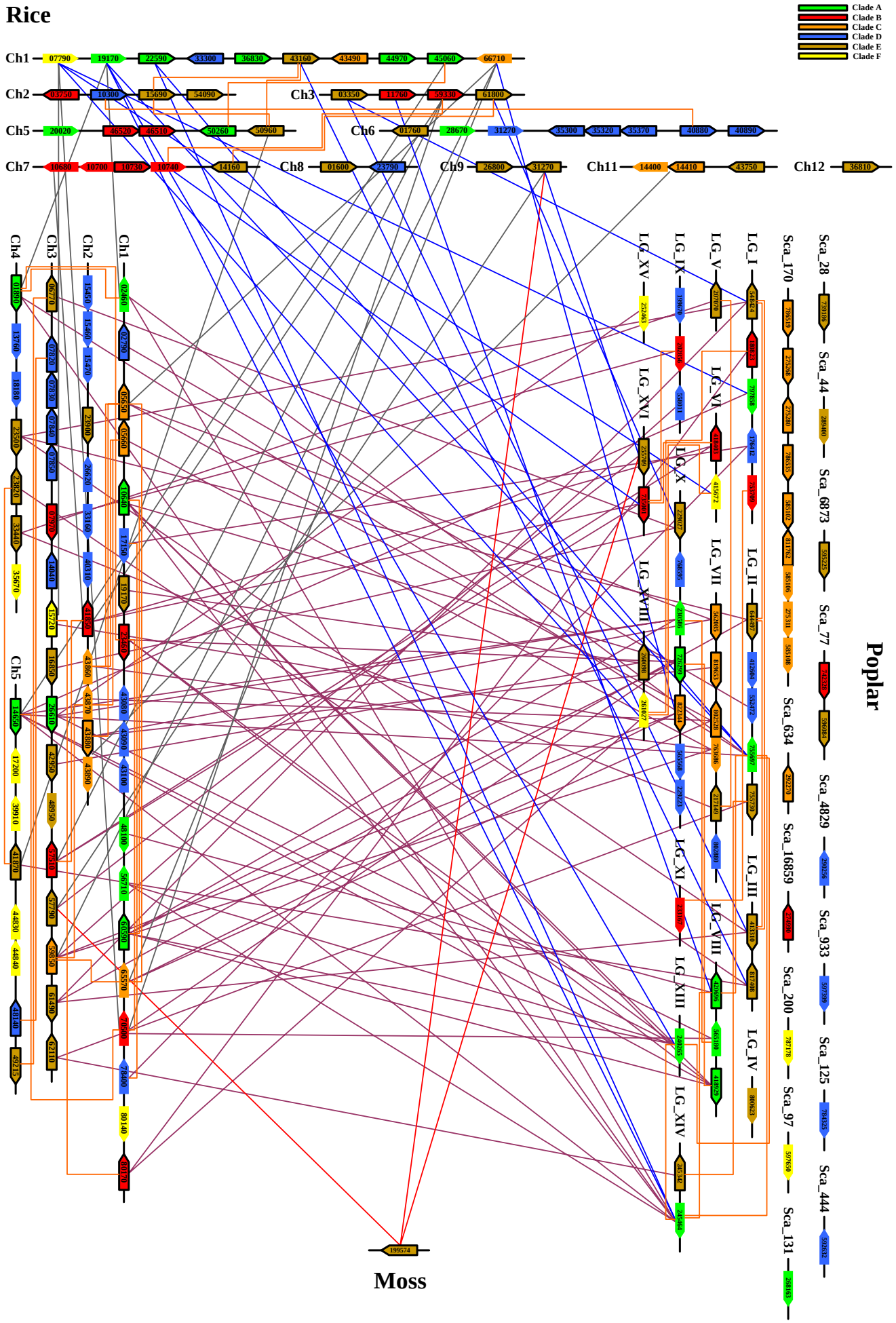

Ec peh1
Aa36645
Aa65850
OS09G26800
AT3G42950
AT1G19170
Pop260098
Pop739186
E-X
OS02G54030
MS163808
OS02G15690
AT5G41870
AT4G23820
Pop817408
OS01G43160
OS05G50960
E-IX
AT3G62110
Pop755730
Pop245342
E-VIII
OS07G14160
OS03G61800
OS08G01600
AT4G33440
Pop207070
E-VII
Pop217149
MS160263
MS22776
MS149319
MS43415
MS193100
E-VI
E
Rhamno-PGs
AT3G16850
AT5G49215
AT3G06770
E-V
Pop596084
Pop229027
OS11G43750
OS12G36810
Pop289400
OS06G01760
Pop800623
AT4G23500
Pop548424
Pop595225
Pop413310
Pop644497
AT3G61490
AT3G48950
E-IV
AT2G23900
OS03G03350
C0005846
Pt46668
Aa26417
E-III
Chlr173470
OS09G31270
AT3G57790
Pop255709
E-II
MS13931
MS199574
Aa71712
Pt46492
E-I
MS116593
OS06G28670
MS176513
AT5G14650
AY185765
Pop797858
AT1G10640
AT1G60590
Pop726299
Pop418929
OS05G50260
OS01G45060
OS01G44970
OS01G36830
A-II
AT3G26610
Pop230586
Pop565180
A
Endo-PGs
AT1G56710
AT1G48100
Pop420696
AT4G01890
AT1G02460
Pop755697
Pop245464
Pop268163
Pop240265
OS01G22590
A-I
OS05G20020
OS01G19170
OS03G11760
AT1G80170
Pop180823
AT2G41850
O42399
AT3G57510
Pop418403
Pop735001
P05117
AT3G07970
L12019
Pop202856
OS03G59330
OS07G10700
OS07G10730
OS07G10680
OS07G10740
B-III
B
Endo-PGs
OS02G03750
AT1G70500
AT1G23460
Pop742328
Pop753709
Pop233167
OS05G46510
OS05G46520
B-II
MS133517
BAA06172
B-I
OS11G14410
OS01G66710
OS11G14400
C-III
AT2G43880
OS01G43490
AT1G05650
AT1G05660
AT2G43890
C-II
Pop562085
Pop275311
Pop585108
AT1G65570
Pop585106
AT2G43860
O22610
AT3G59850
AT2G43870
Pop822344
Pop786519
Pop274990
Pop275268
Pop585102
Pop786535
Pop275280
Pop811762
Pop819653
Pop802528
Pop763686
Pop292270
O22313
O22310
O96487
AF001000
C-I
C
Exo-PGs
OS06G40880
OS08G23790
OS06G31270
AT4G18180
Pop199670
D-V
AT1G02790
AT5G48140
AT3G07820
AT3G07830
AT3G07840
Pop229223
Pop565568
Pop768595
Pop597399
Pop412604
Pop552472
Pop290256
D-IV
AJ620949
Q05967
Pop802880
Pop592632
Pop784325
Pop811027
AT3G07850
AT3G14040
D-III
AT1G17150
AT2G40310
AT1G43080
AT1G43090
AT1G43100
AT4G13760
AT2G26620
AT2G15450
AT2G15460
AT2G15470
AT2G33160
AT1G78400
D-II
Pop558011
Pop176412
OS02G10300
OS01G33300
OS06G40890
P26216 PT
OS06G35300
OS06G35320
OS06G35370
D-I
D
Exo-PGs
Pop807650
AT1G80140
OS01G07790
AT4G35670
AT5G44830
AT5G44840
Pop415672
Pop261027
AT3G15720
Pop787178
AT5G17200
AT5G39910
Pop252463
F-I
F
0.1
Nature Precedings : hdl:10101/npre.2008.1608.1 : Posted 20 Feb 2008

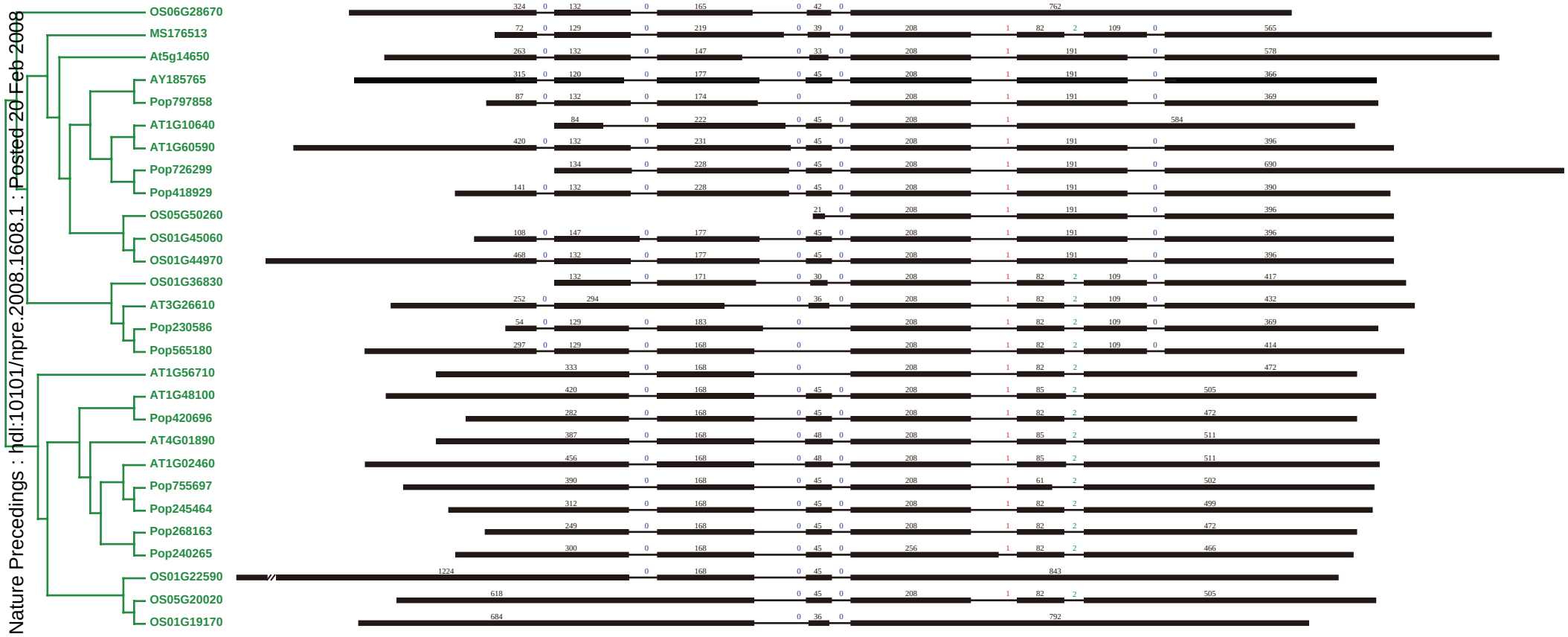
Figure 1.



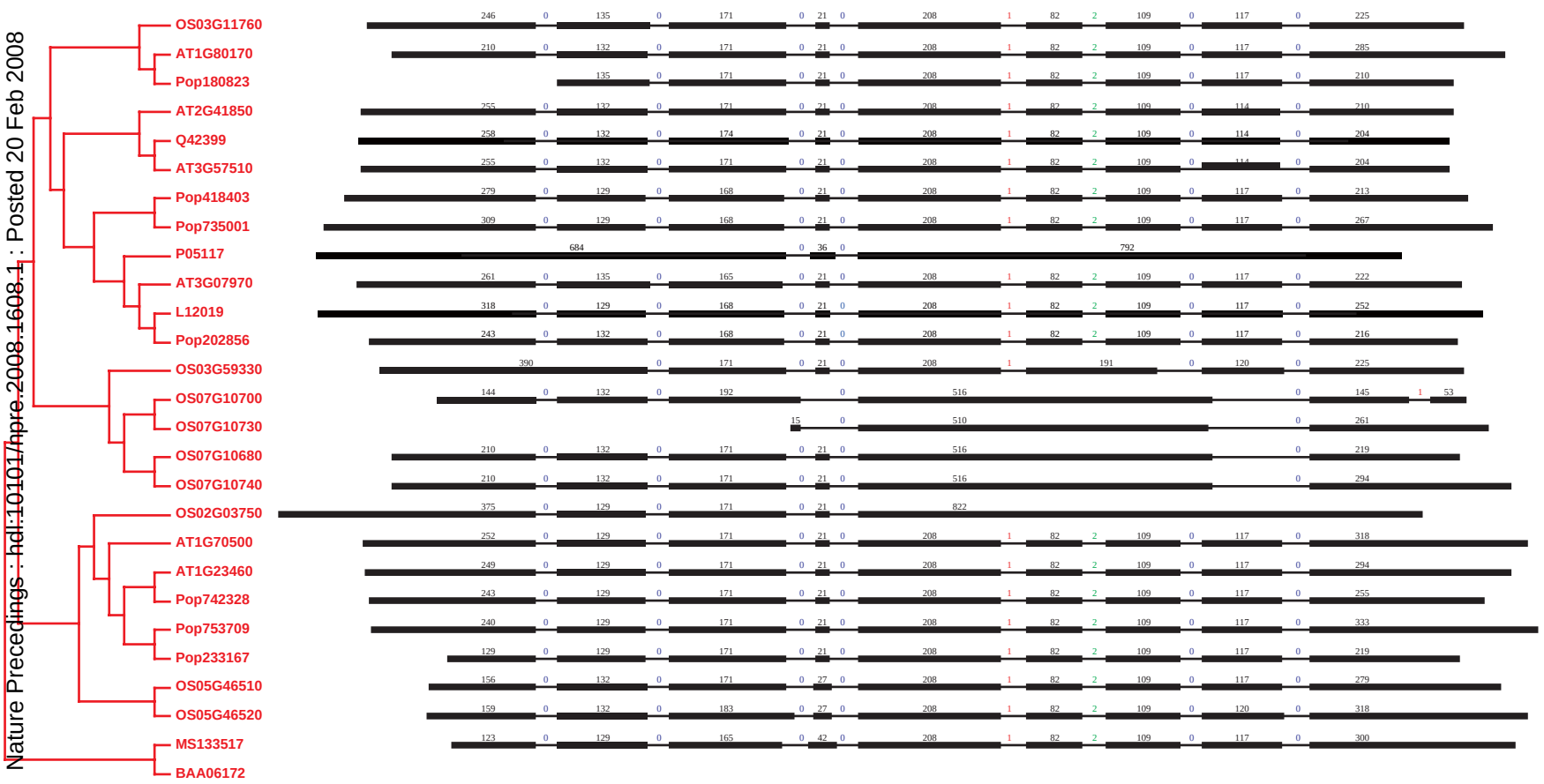
Supplementary Figure S1. Segmentally duplicated chromosomal regions containing PG genes in *P. patens* (moss). The orthologous genes were designated by same colors.



Supplementary Figure S2. Chromosomal distributions of the PG genes. PG genes matching ESTs are encircled and non-matching PG genes are not encircled. Tandemly duplicated PG genes are clustered, and segmentally duplicated PGs within a genome are connected by yellow lines. Orthologous PG genes in segmentally duplicated regions between genomes are connected by blue (between rice and poplar), gray (rice and *A. thaliana*), purple (*A. thaliana* and poplar), and brown lines (moss and vascular plants). The locations of poplar PG genes that are not assigned to chromosomes are indicated with a scaffold.

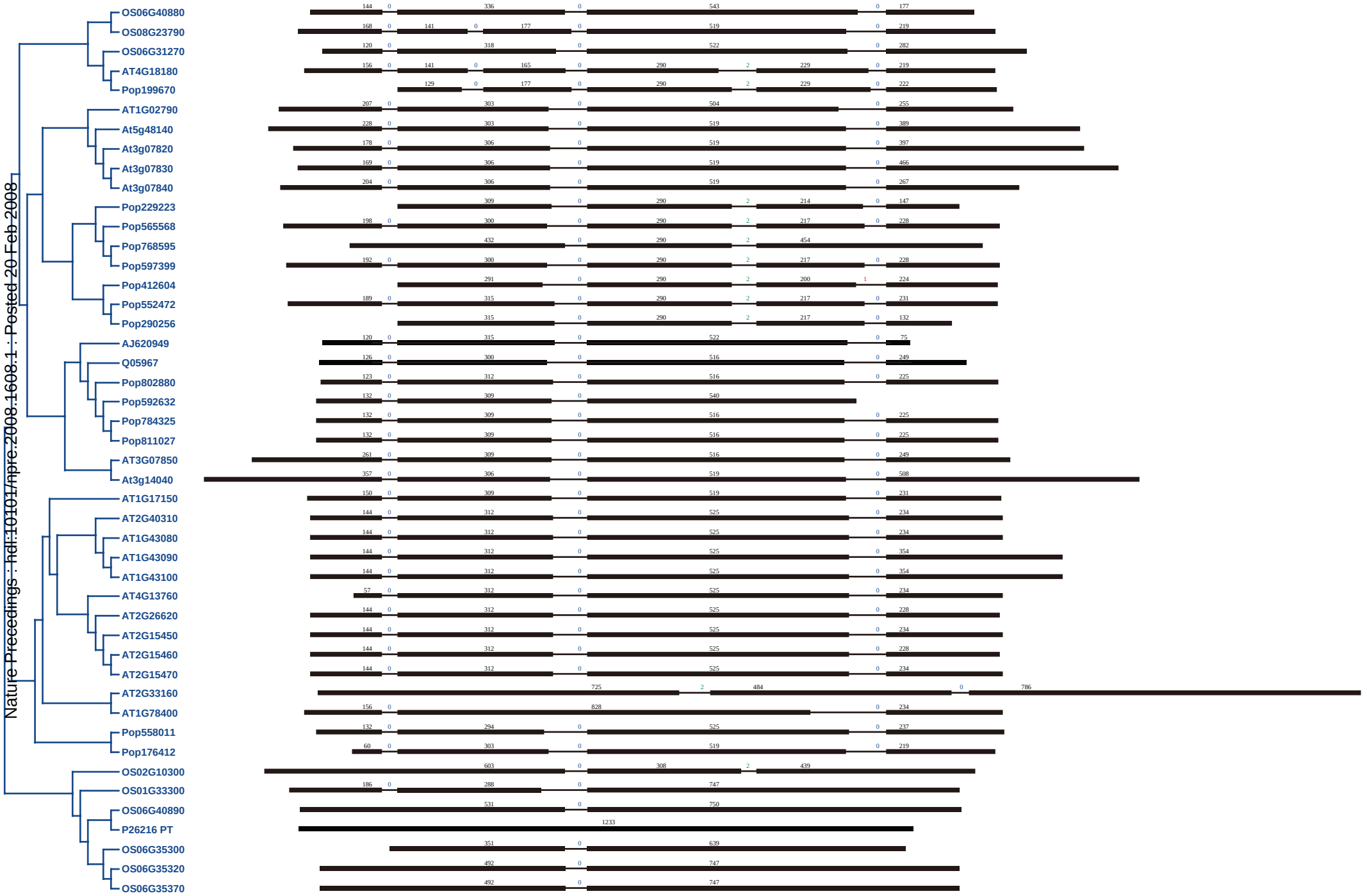


Supplementary Figure S3-1. PG gene structures in clade A. The thick bars are exons and thin lines between them are introns. The numbers on the exons are the number of nucleotides and the numbers on the introns are their phases (0-2). Phase 0 introns are present between codons. Phase 1 introns are present between the first and second nucleotides in a codon, and phase 2 introns are present between the second and third nucleotide.

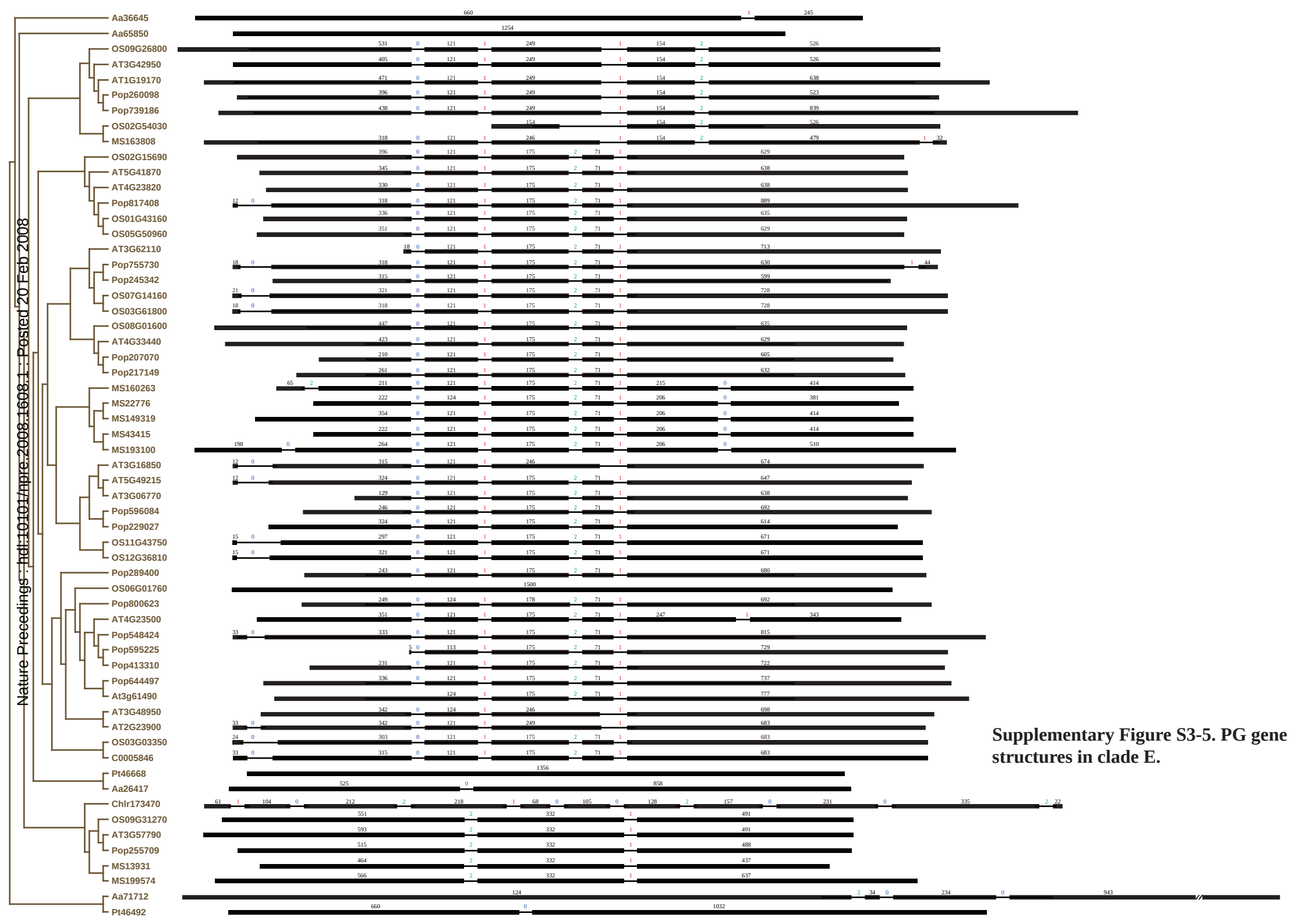


Supplementary Figure S3-2. PG gene structures in clade B.

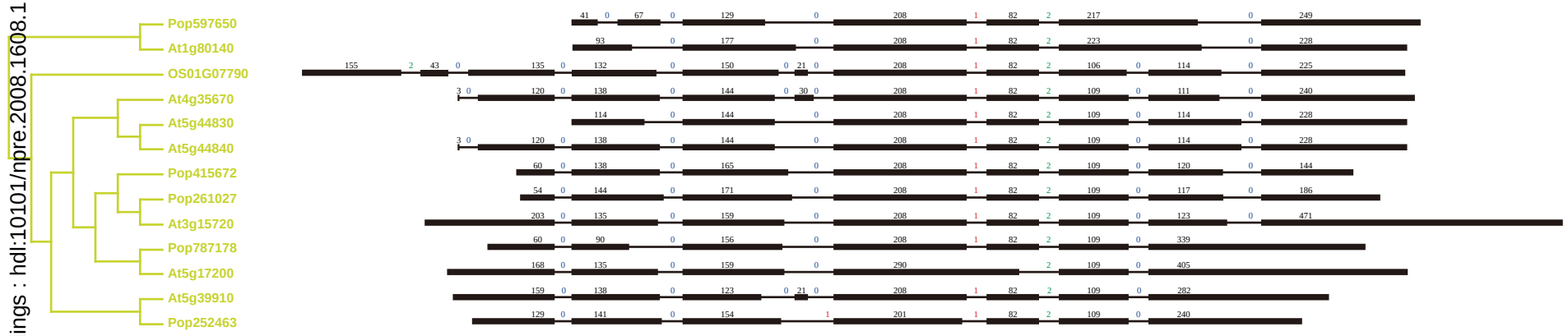
Supplementary Figure S3-3. PG gene structures in clade C.



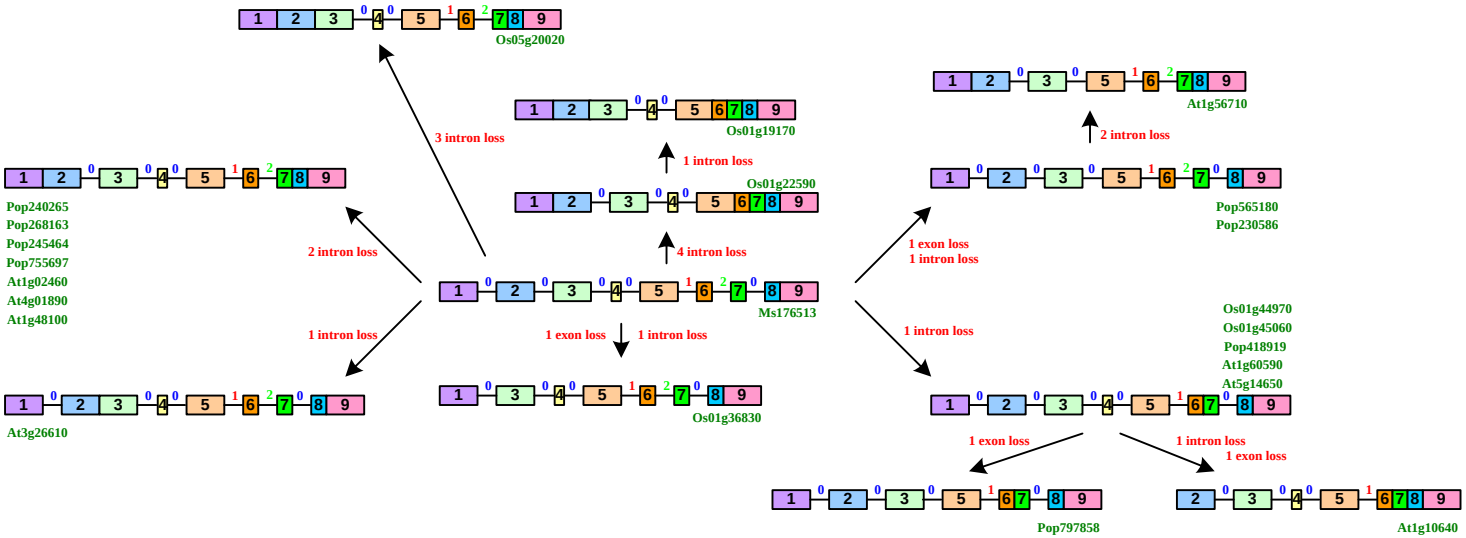
Supplementary Figure S3-4. PG gene structures in clade D.



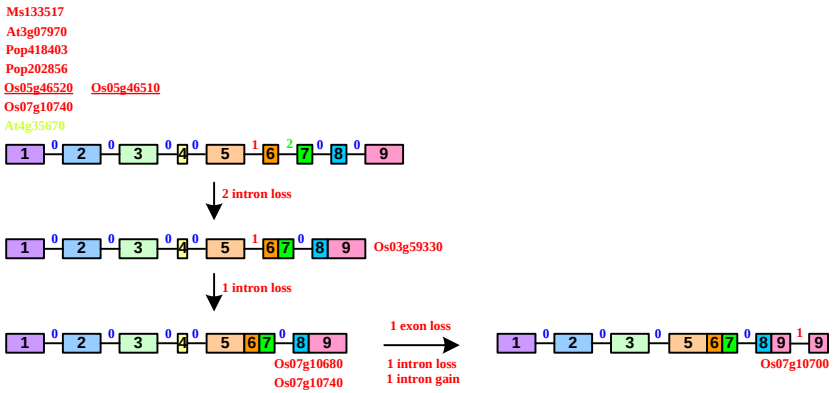
Supplementary Figure S3-5. PG gene structures in clade E.



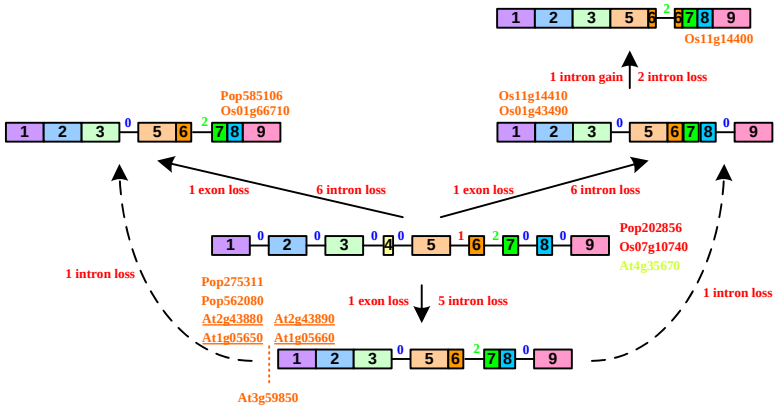
Supplementary Figure S3-6. PG gene structures in clade F.



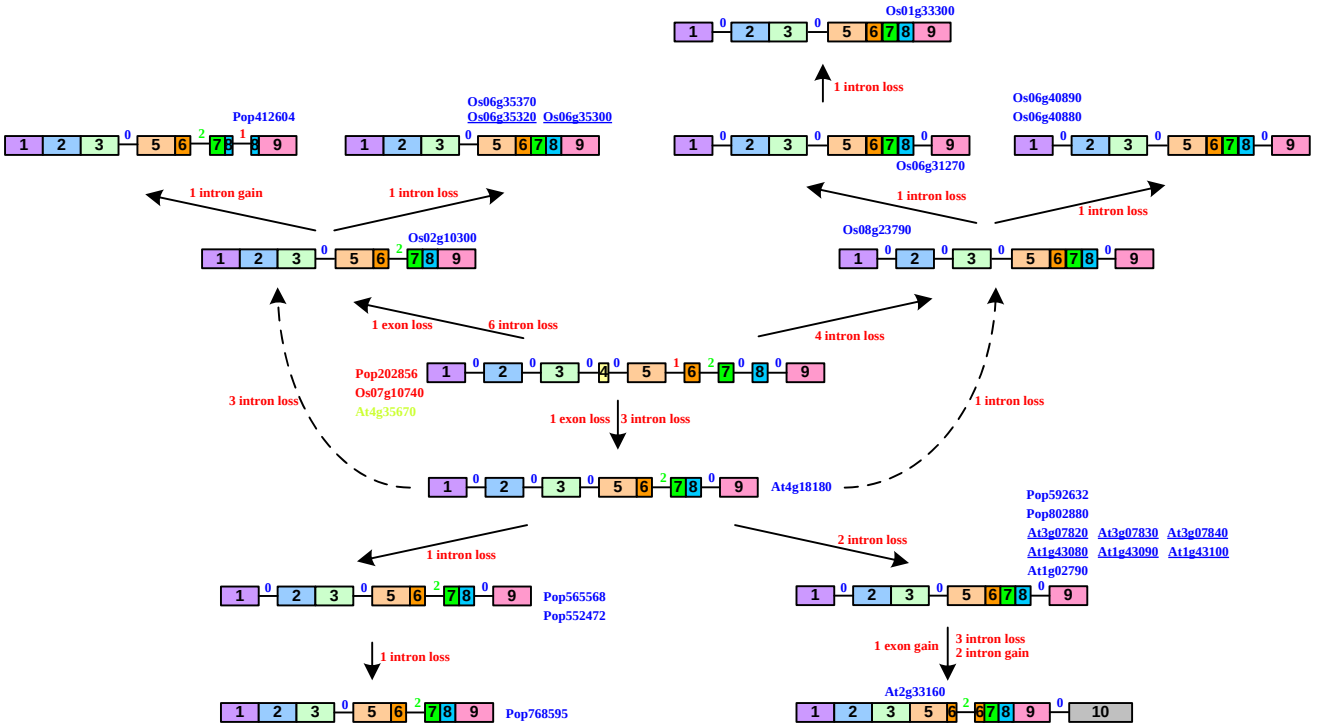
Supplementary Figure S4-1. Intron losses and gains in PG genes in clade A.
Seventeen intron losses and four exon gains have occurred in genes that are structurally similar to MS176513.



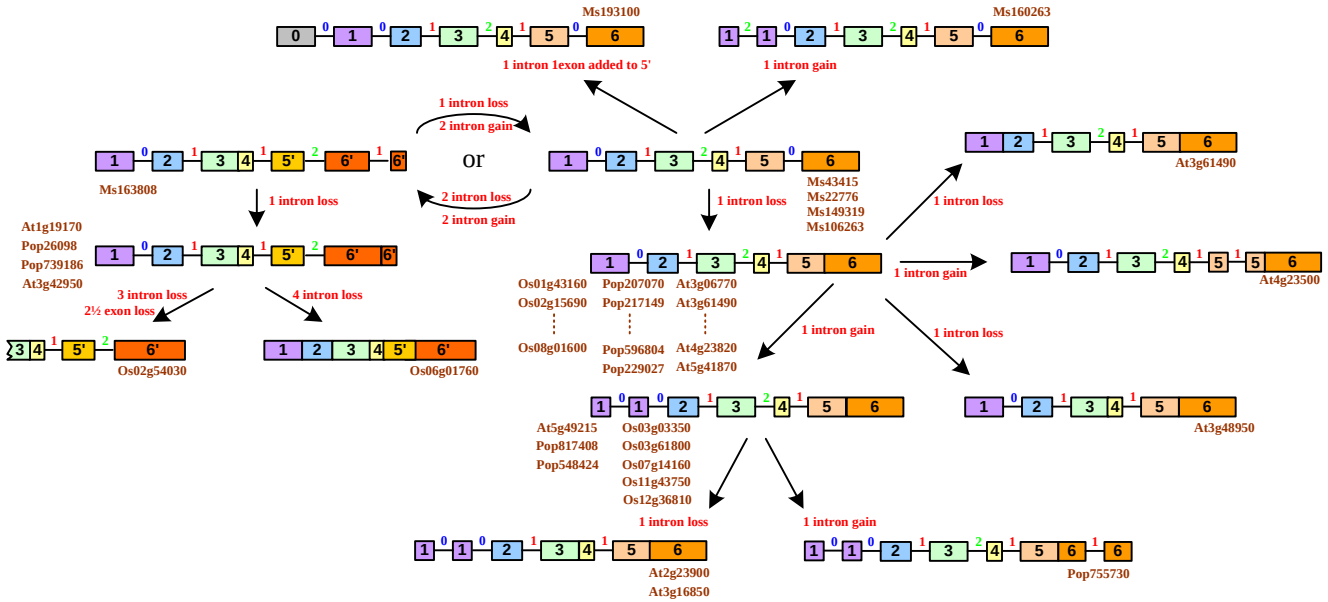
Supplementary Figure S4-2. Intron losses and gains in PG genes in clade B.
Four intron losses, one intron gain, and one exon loss had occurred.



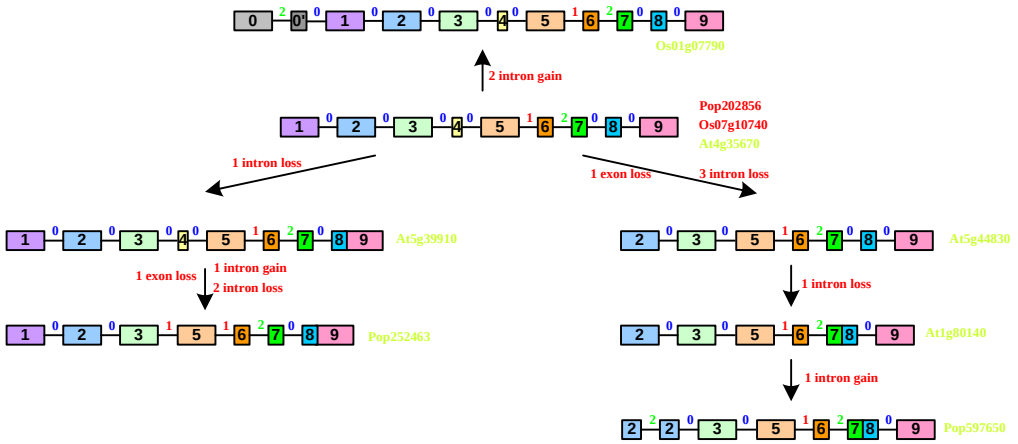
Supplementary Figure S4-3. Intron losses and gains in PG genes in clade C. Nine to nineteen intron losses had occurred. Since rice does not harbor a gene structurally related to At2g43890, the Os01g43490 gene likely resulted from as many as six intron losses and one exon gain in Os07g10740. However, if a rice PG gene structurally similar to At2g43890 was lost by gene death, then Os01g43490 could have been produced from this lost gene by a single intron loss. The same logic can applies to Os06g66710 from the extant rice gene structure as the At2g43890. Three exon losses and one intron gain could be counted.



Supplementary Figure S4-4. Intron losses and gains in PG genes in clade D.
Eighteen to twenty-five intron losses occurred, depending on whether a rice PG gene structurally similar to At4g18180 existed. These intron losses were accompanied by three intron gains and two exon losses.



Supplementary Figure S4-5. Intron losses and gains in PG genes in clade E.
Two different PG gene structures exist in moss (one in MS163808 and one in MS43415, MS22776, MS49319, and MS10263), making it unclear which represents the primary structure. Derivation of each structure from the other would require either one intron loss and two intron gains or two intron losses and two intron gains. Thus, 13-14 intron losses, 6 intron gains, 1 exon gain and 2.5 exon losses occurred. The 5' deletion in Os02g54030 resulted in 2.5 exon losses and 2 intron losses.



Supplementary Figure S4-6. Intron losses and gains in PG genes in clade F.
Seven intron losses, four intron gains, and two exon losses occurred.

Supplement Table 3. Homologous introns in PGs in land plants.

Intron Numb	Clade	Phase	Moss	Arabidopsis	Poplar	Rice	Within domain
1	E	0	–	+	+	+	No
2	E	0	+	+	+	+	Yes
3	E	2	+	+	+	+	Yes
4	E	1	+	+	+	+	Yes
5	E	2	+	+	+	+	Yes
6	E	1	+	+	+	+	Yes
7	E	1	+	+	+	+	Yes
8	E	2	+	+	+	+	Yes
9	E	0	+	–	–	–	Yes
10	AB D F	0	+	+	+	+	No
11	AB D F	0	+	+	+	+	Yes
12	AB F	0	+	+	+	+	Yes
13	ABCD F	0	+	+	+	+	Yes
14	AB F	1	+	+	+	+	Yes
15	ABCD F	2	+	+	+	+	Yes
16	AB F	0	+	+	+	+	Yes
17	BCD F	0	+	+	+	+	Yes

Supplement Table 4. Novel introns in PGs in land plants.

Intron numbers	Clade	Phase	Gene ID	Within Domain	Related with Transposon
1	E	1	Ms163808	No	MITE
2	E	1	Pop755730	No	MITE
3	E	2	Ms160263	No	MITE
4	E	0	Ms193100	No	No
5	E	1	At4g23500	Yes	No
6	F	0	Os01g07790	No	MITE
7	F	2	Os01g07790	No	No
8	F	0	Pop597650	Yes	No
9	F	1	Pop252463	Yes	MITE
10	B	1	Os07g10700	No	No
11	D	0	At2g33160	No	MITE
12	D	2	At2g33160	Yes	MITE
13	D	1	Pop412604	Yes	MITE

Supplement Table 5. Overalls of the gene structures of the land plant PGs

No. of Genes	186
No. of Introns	788
Aver. no. intron per genes	4.2
Intron loss	77~87
Intron gain	15
Exon loss	14.5
No. of homologous gene blocks	17
No. of novel genes	13