When gene marriages don’t work out: divorce by subfunctionalization

Brian P. Cusack and Kenneth H. Wolfe

Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland

We describe how a bifunctional gene, encoding two proteins by alternative splicing, arose when the chloroplast gene *RPL32* integrated into an intron of the nuclear gene *SODcp* in an ancestor of mangrove and poplar trees. Mangrove retains the alternatively spliced chimeric gene, but in poplar it underwent duplication and complete subfunctionalization, through complementary structural degeneration, to re-form separate *RPL32* and *SODcp* genes.

Gene duplication and subfunctionalization

Subfunctionalization – the partitioning of different subsets of the functions of an ancestral gene among daughter copies after gene duplication – provides an attractive explanation for why so many duplicated genes exist in eukaryotes, without requiring each duplication event to have conferred a selective advantage [1]. For many duplicated genes, however, it has been difficult to pinpoint different subfunctions of the ancestral gene that were partitioned among the daughter genes. Often, our knowledge of the functions of the ancestral gene is so limited that we might not be able to recognize subfunctionalization even if it has occurred. Most of the examples of subfunctionalization reported to date involve changes in gene expression profiles [1–4], and there are only a few reports of duplicate gene pairs that have undergone subfunctionalization by means of substantial changes in gene structure relative to their common ancestor [5–9]. Here, we report an example of a structural subfunctionalization event where the ancestral functions being partitioned among the daughter genes can be readily identified and are clearly distinct. Unlike previously reported examples of subfunctionalization of alternatively spliced genes [5,6], the two subfunctions being partitioned here have nothing in common except their subcellular localization in chloroplasts; one is a redox enzyme and the other is a structural component of a ribosome. Moreover, our example illustrates the reversibility of gene fusion by subsequent fission through the duplication-degeneration-complementation (DDC) mechanism [1], with both processes being observed in the short lifetime of a single gene.

Formation of the *SODcp-RPL32* chimeric gene

The gene for chloroplast ribosomal protein L32 (*RPL32*) is located in the chloroplast genome of most flowering plants, but is not present in the chloroplast genomes of two poplar species (*Populus trichocarpa* and *P. alba*; [10–12]). Loss of *RPL32* from chloroplast DNA occurred after *Populus* (order Malpighiales) diverged from other members of the Eurosid I clade such as cucumber (order Cucurbitales) and legumes (order Fabales). We identified database EST (expressed sequence tag) sequences from a copy of *RPL32* that has become relocated to the nuclear genome in poplar. The coding sequence of *RPL32* in this transcript is fused in-frame downstream of a sequence resembling chloroplast Cu–Zn superoxide dismutase (SOD). Further comparisons with ESTs and genomic sequence data from *P. trichocarpa* [11,13] and *Bruguiera gymnorrhiza* [14] (Burma mangrove, also in the order Malpighiales) enabled us to reconstruct the events that occurred subsequent to the transfer of the gene to the nucleus (Figure 1).

Plants have several isozymes of Cu–Zn SOD, which is an enzyme functioning in redox balance. Some of these isozymes are cytosolic and some are imported into chloroplasts by means of an N-terminal transit peptide [15]. In the legume *Medicago truncatula* the chloroplast isozyme is encoded by a single nuclear gene (*SODcp*) with eight exons (Figure 2). In an ancestor of poplar and mangrove, the *RPL32* sequence from the chloroplast genome was transferred to the nuclear genome, where it became inserted into the last intron (intron 7) of *SODcp*. The newly formed chimeric *SODcp-RPL32* gene was alternatively spliced, producing one transcript identical in structure to the original *SODcp* mRNA, and one in which exons 1–7 were spliced onto a novel exon (exon X) corresponding almost exactly to the whole *RPL32* coding region, instead of onto the last exon (exon 8) of *SODcp*. This alternatively spliced gene still exists in mangrove, in which we identified ESTs corresponding to two types of transcript: one coding for SOD (transcript B, 219 amino acids), and the other encoding a chimeric protein with residues 1–211 of SOD fused to residues 2–54 of RPL32 (transcript A; Figure 2 and see Figure S1 in the supplementary material online). We confirmed that alternative splicing occurs in mangrove by sequencing a genomic PCR product that contains exons 7, X and 8 (Figure 2) and perfectly matches the sequences of ESTs of the two types of transcript.

Disintegration of the chimeric gene in the poplar lineage

In poplar, after its divergence from mangrove, the chimeric *SODcp-RPL32* gene was duplicated twice. The first duplication (node A on the phylogenetic tree in Figure 2) resulted in subfunctionalization of the chimeric gene, producing daughter genes that encode either RPL32 (*Poplar1* gene) or SOD, but not both. The SOD-encoding daughter

Corresponding author: Wolfe, K.H. (khwolfe@tcd.ie).

Available online 5 April 2007.

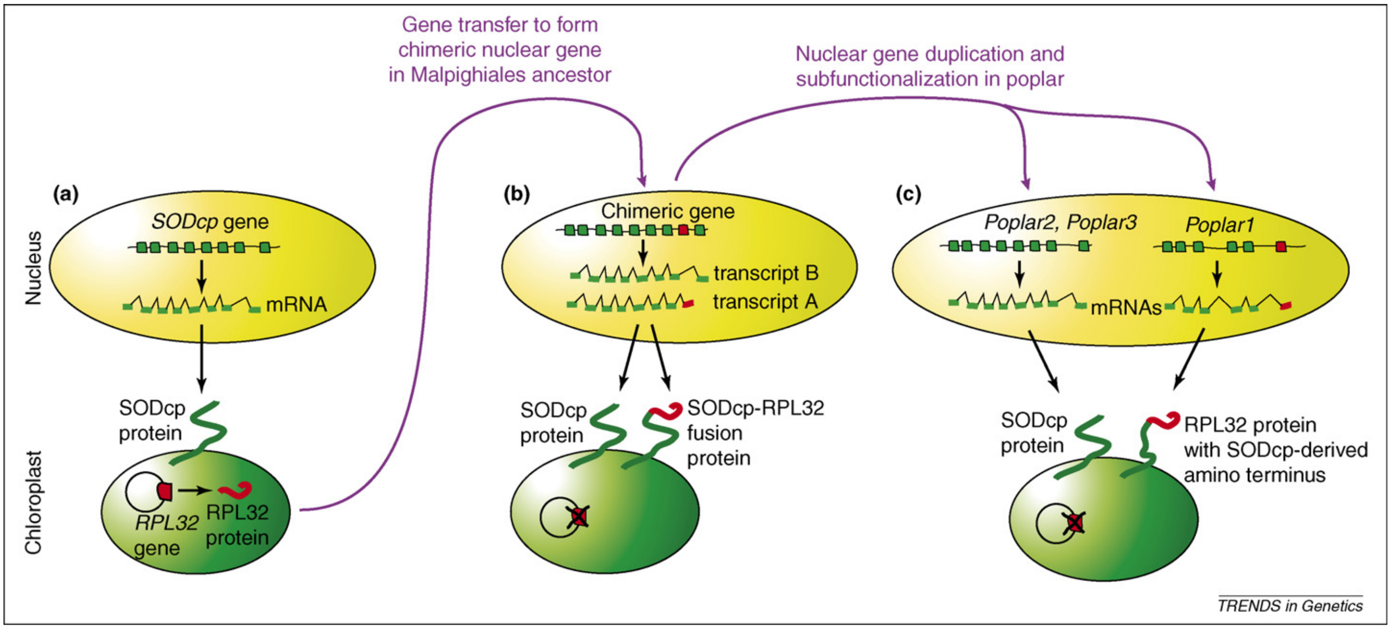
Gene transfer to form chimeric nuclear gene in Malpighiales ancestor
Nuclear gene duplication and subfunctionalization in poplar
(a)
(b)
(c)
Nucleus
Chloroplast
SODcp gene
mRNA
SODcp protein
RPL32 gene
RPL32 protein
Chimeric gene
transcript B
transcript A
SODcp protein
SODcp-RPL32 fusion protein
Poplar2, Poplar3
Poplar1
mRNAs
SODcp protein
RPL32 protein with SODcp-derived amino terminus
TRENDS in Genetics

Figure 1. History of the *SODcp-RPL32* chimeric gene. **(a)** The ancestral configuration that exists in most flowering plants (e.g. *Medicago*), with SODcp (chloroplast Cu–Zn superoxide dismutase) encoded by a nuclear gene and RPL32 (chloroplast ribosomal protein L32) encoded by the chloroplast genome. The SODcp protein is imported into chloroplasts by means of the transit peptide at its amino terminus. **(b)** The configuration inferred to have existed in a Malpighiales ancestor. The nuclear gene is alternatively spliced, as currently seen in mangrove. The chloroplast gene has been lost, as currently observed in poplar and postulated for mangrove. **(c)** The current configuration in poplar.

later became duplicated a second time (node B) to produce two genes (*Poplar2* and *Poplar3*) that have virtually identical structures. EST analysis shows that all three poplar genes are transcribed and none of them is alternatively spliced. The *Poplar2* and *Poplar3* genes have lost exon X and encode proteins that can be aligned along their whole length to *Medicago* SOD. Reciprocally, the RPL32-encoding copy (*Poplar1*) has retained exon X but has lost exons 4, 7 and 8. Exon 4 of *Poplar1* is a pseudo-exon containing a frameshift mutation and is skipped in all nine database ESTs we identified from the gene. There are also deletions in exons 1 and 2 of *Poplar1* relative to *Poplar2*, *Poplar3* and the

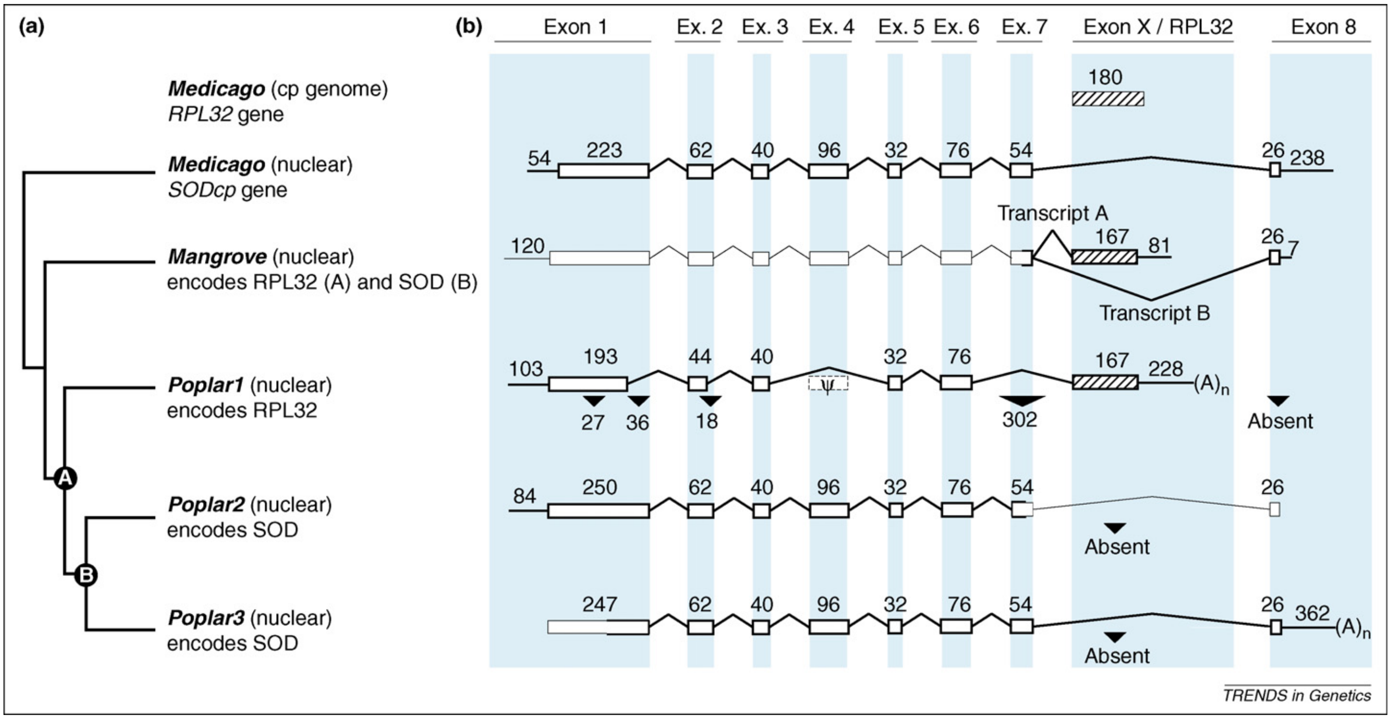

Figure 2. Organization of *SODcp*, *RPL32* and chimeric genes. **(a)** The branching order of the nuclear genes, based on their pairwise *dS* (synonymous nucleotide substitution) values. Nodes A and B represent gene duplications in poplar. Node B corresponds to a large segmental or whole-genome duplication [11,13] in poplar, because many of the genes neighboring *Poplar2* have homologs neighboring *Poplar3*. In **(b)**, boxes represent exons, horizontal lines represent untranslated regions, and the lengths (bp) of some exons are shown. Introns are not drawn to scale. Unfilled boxes show *SODcp*-related exons and hatched boxes show *RPL32*-related exons. Triangles indicate sequences deleted in the poplar genes (with deletion lengths where known), and ψ indicates the pseudo-exon 4 in *Poplar1*. Thick outlines to boxes denote the parts of gene structures that were verified directly by comparing genomic and cDNA or EST sequences from the same species. Thin box outlines in poplar show parts of genes for which only genomic sequence is available, and in mangrove show regions where only EST data are available. The intron-exon structure of the 5′ part of the mangrove gene is assumed to be the same as in other species. Sources of sequence data are listed in Table S1 in the supplementary material online. $(A)_n$, poly(A) tail; cp, chloroplast.

SODcp genes of other plant species (see Figure S1 in the supplementary material online). The *Poplar1* gene still has a continuous open reading frame between the former *SODcp* start codon and the *RPL32* stop codon, and the N-terminus of its protein product is strongly predicted to be a chloroplast transit peptide [16]. However, the protein encoded by *Poplar1* cannot be a functional SOD enzyme because it lacks many residues normally conserved in SOD proteins, including all six active site residues (four are deleted and two are substituted; see Figure S1 in the supplementary material online). In addition to the deletions, the remaining SOD-derived parts of the Poplar1 protein also show deconstrained sequence evolution: in exons 1–6 there is only 60% amino acid sequence identity between Poplar1 and mangrove, lower than for Poplar2 or Poplar3 versus mangrove (both 77% identity). Analysis of nonsynonymous (*dN*) and synonymous (*dS*) nucleotide substitutions shows that the *SODcp*-derived exons of *Poplar1* have been evolving almost free of selective constraint ($dN/dS = 0.9$; see Figure S2 in the supplementary material online). These exons have lost the requirement to specify a functional SOD and instead are constrained only to provide a working transit peptide for the RPL32 protein.

Concluding remarks

The marriage of *RPL32* to *SODcp* and their subsequent divorce in the poplar lineage provides an unusually graphic example of the partitioning of multiple functions of an ancestral gene among daughter genes formed by duplication. This partitioning process can be categorized as subfunctionalization because the structural changes in the poplar genes indicate unambiguously that, after the duplication at node A (Figure 2), a complementary loss of subfunctions of the ancestral chimeric gene occurred in its two daughters. The losses of exon X (encoding the RPL32 subfunction) in the *Poplar2* and *Poplar3* lineage, and of exons 4, 7 and 8 (encoding the SOD subfunction) in *Poplar1*, were caused by degenerative mutations that are likely to have been selectively neutral because in each case the subfunction lost by one gene copy was maintained by the other. Consequently, the gene pair was preserved in the genome by subfunctionalization as envisaged by Lynch and Force [1,17].

Acknowledgements

We thank M.S. Islam for *B. gymnorrhiza* DNA, G. Butler for PCR amplification, and M. Woolfit and M. Sémon for comments. This study was supported by Science Foundation Ireland.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2007.03.010.

References

1 Force, A. *et al.* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545

2 Lynch, M. (2004) Gene duplication and evolution. In *Evolution: From Molecules to Ecosystems* (Moya, A. and Font, E., eds), pp. 33–47, Oxford University Press

3 Cresko, W.A. *et al.* (2003) Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. *Dev. Dyn.* 228, 480–489

4 Huminiecki, L. and Wolfe, K.H. (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res.* 14, 1870–1879

5 Altschmied, J. *et al.* (2002) Subfunctionalization of duplicate *mitf* genes associated with differential degeneration of alternative exons in fish. *Genetics* 161, 259–267

6 Yu, W.P. *et al.* (2003) Duplication, degeneration and subfunctionalization of the nested *synapsin-Timp* genes in *Fugu*. *Trends Genet.* 19, 180–183

7 de Souza, F.S. *et al.* (2005) Subfunctionalization of expression and peptide domains following the ancient duplication of the proopiomelanocortin gene in teleost fishes. *Mol. Biol. Evol.* 22, 2417–2427

8 Korneev, S. and O'Shea, M. (2002) Evolution of nitric oxide synthase regulatory genes by DNA inversion. *Mol. Biol. Evol.* 19, 1228–1233

9 Wang, W. *et al.* (2004) Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. *Nat. Genet.* 36, 523–527

10 Steane, D.A. (2005) Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum, *Eucalyptus globulus* (Myrtaceae). *DNA Res.* 12, 215–220

11 Tuskan, G.A. *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604

12 Okumura, S. *et al.* (2006) Transformation of poplar (*Populus alba*) plastids and expression of foreign proteins in tree chloroplasts. *Transgenic Res.* 15, 637–646

13 Sterck, L. *et al.* (2005) EST data suggest that poplar is an ancient polyploid. *New Phytol.* 167, 165–170

14 Miyama, M. *et al.* (2006) Sequencing and analysis of 14,842 expressed sequence tags of burma mangrove, *Bruguiera gymnorrhiza*. *Plant Sci.* 171, 234–241

15 Schinkel, H. *et al.* (2001) A small family of novel CuZn-superoxide dismutases with high isoelectric points in hybrid aspen. *Planta* 213, 272–279

16 Emanuelsson, O. *et al.* (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016

17 Lynch, M. and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459–473

doi:10.1016/j.tig.2007.03.010