Locating Multiple Gene Duplications Through Reconciled Trees

J. G. Burleigh[1], M. S. Bansal[2], A. Wehe[2], O. Eulenstein[2]

[1] National Evolutionary Synthesis Center, Durham, NC, USA
jgb12@duke.edu
[2] Department of Computer Science, Iowa State University, Ames, IA, USA
{bansal, awehe, oeulenst}@cs.iastate.edu

Abstract. We introduce the first exact and efficient algorithm for Guigó et al.'s problem that, given a collection of rooted, binary gene trees and a rooted, binary species tree, determines a minimum number of locations for gene duplication events from the gene trees on the species tree. We examined the performance of our algorithm using a set of 85 gene trees that contain genes from a total of 136 plant taxa. There was evidence of large-scale gene duplication events in *Populus, Gossypium*, Poaceae, Asteraceae, Brassicaceae, Solanaceae, Fabaceae, and near the root of the eudicot clade. However, error in gene trees can produce erroneous evidence of large-scale duplication events, especially near the root of the species tree. Our algorithm can provide hypotheses for precise locations of large-scale gene duplication events with data from relatively few gene trees and can complement other genomic approaches to provide a more comprehensive view of ancient large-scale gene duplication events.

1 Introduction

Polyploidy is a major component of plant genome evolution [27, 14]. Analyses of genomic data from numerous plants such as grasses [16, 24], *Arabidopsis* or Brassicaceae [30, 26, 3], poplar [28], cotton [4], *Physcomitrella* [25], and *Vitis* [10] have revealed evidence of ancient genome duplications. Yet the number of ancient genome duplications and their precise location in the evolutionary history of plants is still unclear. We describe the first exact polynomial time algorithm for Guigó et al.'s problem [15] that maps large-scale gene duplications, such as polyploidy, on a species tree, and we demonstrate its ability to identify and place ancient polyploidy events in plants.

The presence of large, duplicated chromosomal segments within a genome provided the first evidence of ancient polyploidy (e.g. [30, 26, 3, 5, 16, 24, 10]). These duplications can be dated based on the sequence divergence between paralogous genes on duplicated blocks. However, rapid gene loss and gene rearrangements after a polyploidy event can make it difficult or impossible to detect ancient duplicated chromosomal segments [20, 26], and few plant taxa have adequate gene mapping data. It is also possible to detect ancient polyploidy based solely on the age distributions of pairs of duplicated (paralogous) genes (e.g. [20, 30, 4, 28, 8, 25]). The date of the inferred duplications is estimated from amino acid or, more commonly, silent (synonymous) substitution rates, using molecular clock assumptions. Examining genomic data from multiple taxa in a

comparative phylogentic context has the potential to improve estimates of the timing of large-scale duplication events (e.g. [5, 7]). In the simplest approach, a phylogenetic tree is constructed with a pair of paralogous genes from one taxon, and the best homolog from a second taxon and from an outgroup taxon [5, 7]. This allows one to date the duplication from the first taxon relative to the divergence with the second taxon. Yet placing a duplication event relative to a single taxonomic divergence is not very specific.

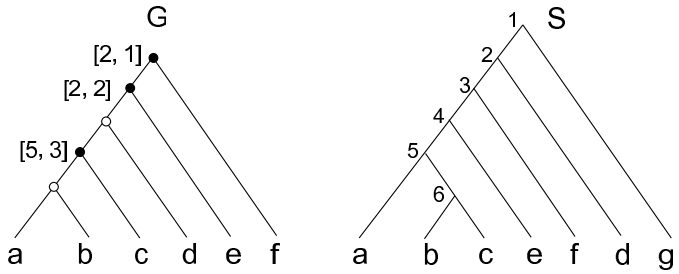

Fig. 1. A gene tree G and a comparable species tree S is depicted. The bold nodes in G are duplications and their intervals represent their allowed locations in the species tree S.

Guigó et al. [15] first addressed a more comprehensive phylogenetic approach that maps duplication events from a collection of rooted, binary gene trees onto a rooted, binary species tree. Later on, Page and Cotton [22] refined this problem and used it to examine gene duplication events in vertebrates. We refer to the refined problem as the Episode Clustering problem. An alternative version of this problem was introduced by Fellows et al., which they proved to be intrinsically difficult [9]. Hence, we direct the focus of this work to the Episode Clustering problem. This problem determines duplication events using the Gene Duplication Model from Goodman et al. [13]. Each duplication can be placed on any species on a path between the two (not necessarily distinct) most recent species that could have contained the duplication and its parent respectively. In case the parent does not exist, the path runs between the most recent species for the duplication and the root of the species tree. An example is depicted in Fig. 1. The duplications in gene tree G are represented by the three bold nodes. Associated with each bold node is its path represented by an interval. For example, the interval $[5, 3]$ represents the path $5, 4, 3$ in the species trees S. Let g denote the node corresponding to the interval $[5, 3]$. Species 5 is the most recent species that could have contained g and the parent of species 3, i.e. 2, is the most recent species that could have contained the parent of g. The *Episode Clustering (EC)* problem is, given a collection of gene trees and a species tree, find a minimum number of locations in the species tree where all duplications in the gene trees can be placed. For example, all three duplications in Fig. 1 can be placed on species nodes 2 and 3. Page and Cotton [22] observed that the EC problem can be efficiently reduced to the set-cover problem [11]. They approach the EC problem using a heuristic for the intrinsically difficult set-cover problem. In this paper we present an efficient and exact solution for the EC problem,

which is based on established graph theoretical results. Note, that the gene duplications and the paths where duplications can be placed are computable in linear time using efficient least common ancestor computations [2, 31].

2 Methods

2.1 Basic Definitions, Notation, and Preliminaries

In this section we first introduce basic definitions and notation that we will be dealing with and then define preliminaries required for this work.

Basic Definitions and Notation A *tree* T is a connected graph with no cycles, consisting of a node set $V(T)$ and an edge set $E(T)$. T is *rooted* if it has exactly one distinguished node called the *root* which we denote by $\mathtt{Ro}(T)$. Let T be a rooted tree. We define $\leq_T$ to be the partial order on $V(T)$ where $x \leq_T y$ if y is a node on the path between $\mathtt{Ro}(T)$ and x. We denote by $x \backsim_T y$ that x, y are related by $\leq_T$, and by $<_T$ the strict counterpart of the relation $\leq_T$. The set of minima under $\leq_T$ is denoted by $\mathtt{Le}(T)$ and its elements are called *leaves*. If $x \leq_T y$ and $\{x, y\} \in E(T)$, then we call y the *parent* of x denoted by $\mathtt{Pa}(x)$ and we call x a *child* of y. The set of all children of y is denoted by $\mathtt{Ch}_T(y)$. The *least common ancestor (lca)* of a non-empty subset $L \subseteq V(T)$ denoted as $\mathrm{lca}(L)$, is the unique smallest upper bound of L under $\leq_T$. A subtree of T rooted at node $y \in V(T)$, denoted by T_y, is the tree induced by $\{x \in V(T) \colon x \leq_T y\}$. T is called (fully) binary if every node has either zero or two children.

The *interval* for $a \leq_T b$ is defined as $[a, b] = \{x \in V(T) \mid a \leq_T x \leq_T b\}$. Let $\mathcal{I}$ be a collection of intervals in $\leq_T$. The *node cover* of a node $v \in V(T)$ is defined as $cover(v) := \{I \in \mathcal{I} \mid v \in I\}$ and the *node cover* of a node set $V \subseteq V(T)$ is defined as $cover(V) = \bigcup_{v \in V} cover(v)$. A set $V \subseteq V(T)$ is called a *cover* of $\mathcal{I}$, if $cover(V) = \mathcal{I}$. If V is a cover of minimum cardinality, we call V a *minimum cover* of $\mathcal{I}$.

The *intersection graph* of a collection of intervals $\mathcal{I}$, denoted $int(\mathcal{I})$, is the graph $(\mathcal{I}, E)$ where $\{I, I'\} \in E$ precisely if $I \cap I' \neq \emptyset$. Let $G = (V, E)$ be a graph, then $V(G) = V$ and $E(G) = E$. A *clique* in G is a set $C \subseteq V$ which induces a completely connected subgraph in G. A *clique cover* of a G is a set of cliques $\mathcal{C}$ in G such that $\bigcup_{C \in \mathcal{C}} C = V$. A *minimum clique cover* is a clique cover of minimum size.

Problem 1 *Tree Interval Cover (TIC)*
Instance: *A collection of intervals $\mathcal{I}$ in the order $\leq_T$.*
Find: *A minimum cover of $\mathcal{I}$.*

The Episode Clustering problem is a special case of the TIC problem.

The Episode-Clustering (EC) Problem The EC problem is to place duplications onto a minimum number of species in a species tree, where each duplication is associated with an interval in the species tree describing the locations where that duplication can be placed. The definition of duplication and its associated interval are based on the Gene

Duplication (GD) model [23] introduced by Goodman et al. [13]. Here we only provide definitions necessary to state the EC problem.

The GD model is based on a gene and species tree from which gene duplications and their associated intervals can be derived. A *species tree* is a tree that depicts the evolutionary relationships of a set of species. Given a gene family for a set of species, a *gene tree* is a tree that depicts the evolutionary relationships among the sequences encoding only that gene family in the given species. Thus the nodes in a gene tree represent genes. To compare a gene tree G with a species tree S a mapping from each gene $g \in V(G)$ to the most recent species in S that could have contained g is required.

Definition 1 (Mapping). *A* leaf-mapping $\mathcal{L}_{G,S}\colon \mathsf{Le}(G) \to \mathsf{Le}(S)$ *specifies, for each gene g the species from which it was sampled. The extension $\mathcal{M}_{G,S}\colon V(G) \to V(S)$ of $\mathcal{L}_{G,S}$ is the* mapping *defined by $\mathcal{M}_{G,S}(g) = \mathrm{lca}(\mathcal{L}_{G,S}(\mathsf{Le}(G_g))$.*

Definition 2 (Comparability). *The trees G and S are* comparable *if there exists a leaf-mapping $\mathcal{L}_{G,S}$. A set of gene trees $\mathcal{G}$ and S are* comparable *if each gene tree in $\mathcal{G}$ is comparable with S.*

Throughout the remainder of this paper, $\mathcal{G}$ denotes a collection of input gene trees, S a comparable species tree, and G denotes an arbitrary gene tree in $\mathcal{G}$.

Definition 3 (Duplication). *A node $v \in V(G)$ is a* (gene) duplication *if $\mathcal{M}_{G,S}(v) = \mathcal{M}_{G,S}(u)$ for some $u \in \mathsf{Ch}(v)$ and we define $\mathsf{Dup}(G,S) = \{g \in V(G) \mid g \text{ is a duplication}\}$.*

Definition 4. *For every $g \in V(G)$ we define the interval*

$$I(g) = \begin{cases} [\mathcal{M}(g), \mathsf{Ro}(S)], & \text{if } g = \mathsf{Ro}(G), \\ [\mathcal{M}(g), \mathcal{M}(g)], & \text{if } \mathcal{M}(g) = \mathcal{M}(\mathsf{Pa}(g)), \\ [\mathcal{M}(g), \mathcal{M}(\mathsf{Pa}(g))] - \{\mathcal{M}(\mathsf{Pa}(g))\}, & \text{otherwise.} \end{cases} \quad (1)$$

Problem 2 *Episode Clustering (EC)*
Instance: *A collection of gene trees $\mathcal{G}$ and a comparable species tree S.*
Find: *A solution to the TIC instance $\bigcup_{g \in \mathsf{Dup}(\mathcal{G},S)} \{I(g)\}$ in the order $\leq_S$.*

The TIC instance $\bigcup_{g \in \mathsf{Dup}(\mathcal{G},S)} \{I(g)\}$ can be computed in linear time [31] using efficient lca computation (e.g. [2]). To solve the EC problem we give an efficient solution for the TIC problem in the following section.

2.2 Solving the TIC Problem

Let $\mathcal{I}$ be a collection of intervals in the order $\leq_T$. In the interest of brevity, proofs for Lemmas 1 and 2, Theorems 1 and 2, and Corollary 1 appear in the Appendix.

Lemma 1. *Let C be a clique in the intersection graph $int(\mathcal{I})$. Then, $\bigcap_{I \in C} I$ is an interval in the order $\leq_T$. In particular $\bigcap_{I \in C} I = [a, b]$ where $a = \mathrm{lca}(\bigcup_{[x,y] \in C} x)$ and $b = \min(\bigcup_{[x,y] \in C} y)$.*

Lemma 2. *Let $\mathcal{I}$ be a collection of intervals over $\leq_T$ and $V \subseteq V(T)$ covers $\mathcal{I}$. Then, $\mathcal{C} := \bigcup_{v \in V} \{cover(v)\}$ forms a clique cover of the intersection graph $int(\mathcal{I})$.*

Theorem 1. *Let $\mathcal{I}$ be a collection of intervals over $\leq_T$, and $\mathcal{C}$ be a minimum clique cover of the intersection graph $int(\mathcal{I})$. Define the function $f\colon \mathcal{C} \to V(T)$ that maps $f(C)$ to some element in $\bigcap_{I \in C} I$. Note, f is well defined by Lemma 1. Then, the node set $f(\mathcal{C})$ is a minimum interval cover of $\mathcal{I}$.*

The following two results are well known (see [21], and [12]).

Lemma 3. *If G is the intersection graph of a family of paths on a tree, then G is triangulated.*

Every interval in $\leq_T$ is equivalent to a path on T. Thus, the intersection graph $int(\mathcal{I})$ is triangulated.

Lemma 4. *Given a triangulated graph G with n nodes and m edges, a minimum clique cover for G can be computed in $O(n + m)$ time.*

Theorem 2. *Given a collection of intervals $\mathcal{I}$ in $\leq_T$ that are presented through paths on the tree T. Then, the TIC problem can be solved in $O(n^2 + nm + l)$ where $n = |V(int(\mathcal{I}))|$, $m = |E(int(\mathcal{I}))|$ and $l = |\operatorname{Le}(T)|$.*

Corollary 1. *Let $\mathcal{G}$ be a collection of gene trees and S a comparable species tree, where $k = \Sigma_{G \in \mathcal{G}} |\operatorname{Le}(G)|$ and $l = |\operatorname{Le}(S)|$. Then, the EC problem for the instance $\mathcal{G}$ and S can be solved in $O(k^2 + km + l)$ time, where m is the number of intersecting intervals that are associated with the duplications in the collection of gene trees $\mathcal{G}$.*

2.3 Plant Gene Analysis

We tested our algorithm using a set of plant gene family trees made from alignments obtained from Phytome, an online comparative genomics database for plants [18]. We selected the masked amino acid alignments from all 85 gene families in Phytome that contain sequences from at least 100 of the 136 total taxa. The gene trees were inferred with maximum likelihood (ML) phylogenetic analyses using RAxML-VI-HPC version 2.2.3. The ML analyses used the JTT amino acid substitution model [19] with the PROTMIX option for modeling rate variation among sites. The ML gene trees were first rooted using mid-point rooting. However, if any alternate rootings of the gene trees decreased the minimum number of gene duplications needed to reconcile the gene trees with the species tree, we chose a rooting that minimizes the number of duplications. Finally, since it is difficult to distinguish allelic variants of a single gene from paralogs, if a gene tree had any clades that contain only sequences from a single taxon, we removed all but a single leaf from the clade. We used a species tree based on currently accepted plant phylogenetic hypotheses (e.g. [1]).

Inferring Gene Duplications Events. We used our EC algorithm to infer the minimum number of duplication locations for the set of ML gene trees on the specified species tree. Our algorithm provides a solution for the minimum number of duplication locations that also includes the total number of duplications at each node, the number of duplication episodes at each node, and the number of genes with duplications at each node. In order to examine the performance of our algorithm in the absence of phylogenetic signal, we also performed 10 replicates our analysis after randomly permuting the leaf labels from each of the gene trees. This experiment will provide an expectation of the results of our algorithm if there was no phylogenetic signal in the gene trees, or if the gene trees were essentially random.

3 Results

Plant Duplication Analysis. We found that gene duplication events involving at least one of the 85 gene trees occur on a minimum of 119 of the 135 internal nodes. While some nodes show evidence of many duplications, others have evidence of very few duplications. For example, 51 nodes have evidence of ≤ 10 duplications, and 4 nodes have evidence of ≥ 1000 duplications. Since we are most interested in identifying large-scale duplications, we focus on the 25 nodes with duplications involving at least half (≥ 43) of the gene trees (Table 1 and Fig. 2). These are especially abundant among the root nodes (Fig. 2). However, they are also common at the base of major clades including Poaceae, Solanaceae, Asteraceae, Brassicaceae, as well as *Populus* and *Gossypium* (Table 1 and Fig. 2). Each analysis of the 85 gene data set took approximately 15 minutes on a Macintosh Power PC laptop computer with a 1.5 GHz G4 processor and Mac OSX 10.4 operating system.

Random Leaves Analysis. The 10 analyses using gene trees with randomly permuted leaf labels found evidence for gene duplication events on only between 25 and 33 (ave. 28.3) internal nodes. In all replicates there was evidence for gene duplications involving many if not all genes in the root nodes (A-C, F-I in Fig. 2) of the species tree as well as the root nodes of the eudicots (nodes L, M, N, and R in Fig. 2), but generally few genes in the other nodes of the species tree (Table 1 and Fig. 2).

4 Discussion

Gene and Genome Duplications in Plants. Our analyses first emphasize the ubiquity of gene duplications throughout the evolutionary history of plants. While we examined only 85 gene families with incomplete sampling, there is evidence of gene duplications on nearly 90% of the internal nodes. Our analyses also provide a hypothesis for the history of large-scale gene duplications in plants that is generally consistent with previous hypotheses (e.g., [8]). Our focus on the 25 nodes with evidence of duplications in at least half of the gene families identified many previously hypothesized ancient polyploidy events. These include events at the base of the Poaceae (node J [16, 24]), Brassicaceae (node T [30, 26]), and Asteraceae (node Q [8]), within Solanaceae (nodes O and P [8]) and Fabaceae (node W [6]), and in *Populus* (nodes X and Y [28]) and *Gossypium*

(node V [4] Fig. 2). In some cases, our analyses provide more precise hypotheses of the phylogenetic location of these duplications because of our higher taxon sampling. For example, while there has been evidence of a large-scale gene duplication common to many grasses (e.g. [29, 24]), our analysis places it between the divergence of *Ananas* and the Poaceae (node J, Fig. 2). There is little evidence for large-scale duplications at the root nodes (nodes A-C, F-I Fig. 2), and at most of the early eudicot nodes (nodes L-N, R-S; Fig. 2); yet, these also are the nodes where large numbers of duplications map in our analysis of the randomly permuted gene trees (Table 1; Fig. 2). When mapping duplications from a single gene tree to a species tree, error in the gene trees erroneously places duplications towards the root of the species tree [17]. Our results suggest that erroneously placed genes in gene trees also provide erroneous evidence of large-scale duplications at the root nodes. Thus, we advise interpreting evidence of large-scale duplications near the root of a tree with great caution. If we disregard the potentially erroneous events at the root nodes, our analysis provides an overall picture of ancient polyploidy in angiosperms that is largely consistent with the recent data from the *Vitis* genome [10]. We hypothesize that the two genome duplications in *Arabidopsis* since its common ancestor with *Vitis* occurred at the base of the Brassicaceae (node T; Fig. 2) and at the base of the eurosid I + eurosid II clade (node R; Fig. 2). The ancestral hexiploidization of the *Vitis* and *Arbaidopsis* genomes occured at nodes L and/or M (Fig. 2), after the divergence of eudicots and monocots.

Algorithm Performance and Limitations. The results of analysis of plant gene trees also suggest some weaknesses in our approach and directions for future research. First, though our analysis uses only 85 gene trees, we find evidence of duplications on nearly all of the internal nodes. With more gene trees, there will doubtlessly be evidence for duplications on every node of the tree. In this case, an algorithm that seeks to find the minimum number of nodes with duplications will cease to be informative. It may be more informative to find the duplication mappings that minimize the overall number of duplication episodes. The randomized leaf analysis also suggests that gene tree error can produce evidence of apparently anomalous large-scale gene duplication events. Unfortunately, some error is likely inherent in any gene tree inference. Even if the unrooted gene tree topology is correct, it is extremely difficult to determine the correct rooting when there is a history of duplications. It may be useful to develop methods for mapping large-gene duplication events that can account for possible error in the gene trees, either by utilizing unresolved or unrooted gene trees or by allowing small changes in the topology of the gene trees if they will lead to better solutions.

5 Conclusion

We introduce a new exact algorithm that solves a biological problem: how can we reconstruct the history of gene duplications across a phylogeny in a way that minimizes the locations of the duplications. By placing large-scale duplication events in such a phylogenetic context, we can help specify the precise location and timing of the duplications. Unlike other methods, our approach does not require gene map data and does not rely on molecular clock assumptions. Furthermore, it can be used with relatively few

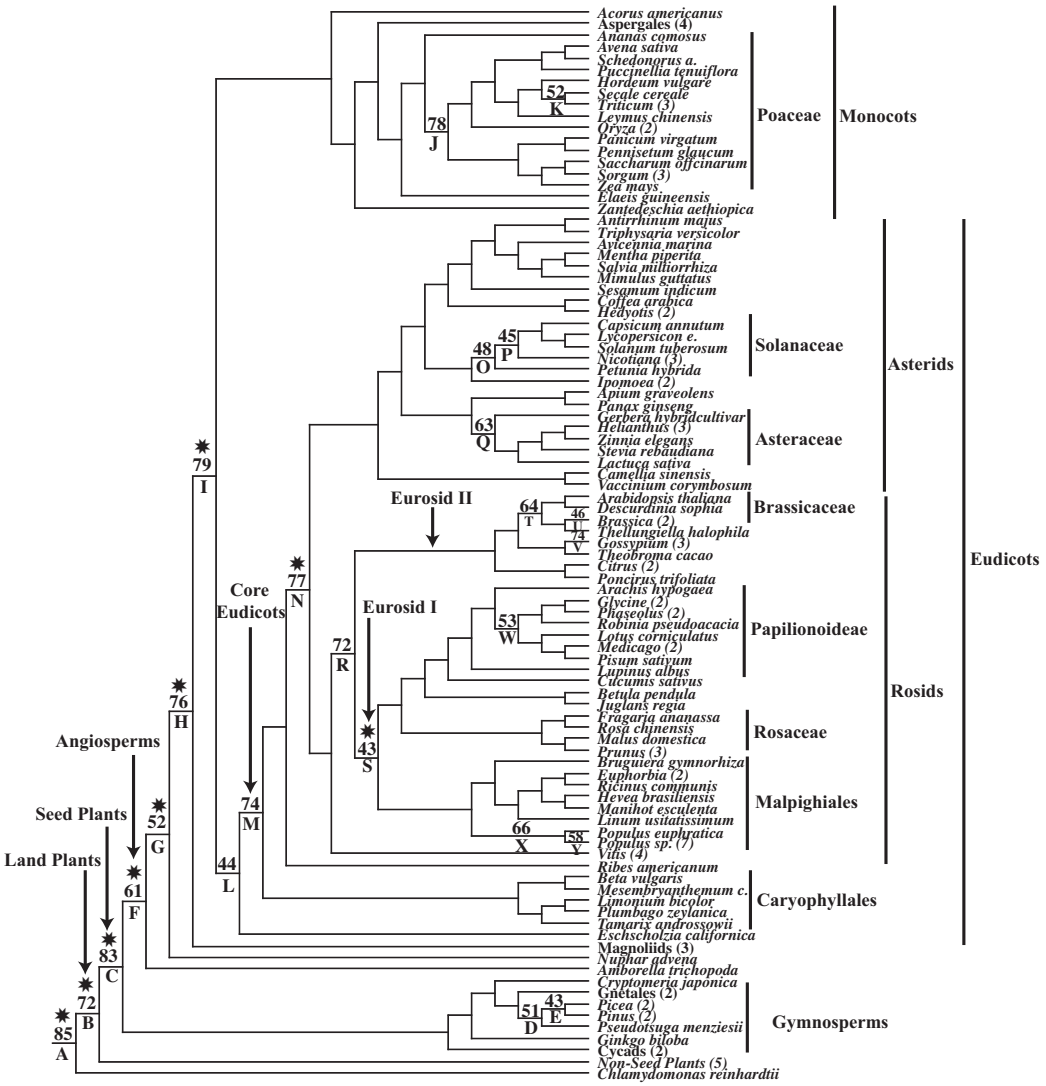

Fig. 2. Species tree with potential locations of large-scale gene duplication events. The species tree used in the analysis contains 136 taxa, and in some cases, multiple (usually congeneric) species in a clade were combined into a single taxon for this figure. In these cases, the total number of species in the combined group is written in parentheses beside the leaf name. The internal nodes with duplications from $\geq$ 43 of the 85 gene trees have letters under the branch leading to the node, and the number of gene trees with duplications on top of the branch. Stars on top of the branch denote nodes where the analyses using gene trees with randomly permuted leaf labels identified gene duplications from as many gene trees as the analysis with ML gene trees. In other words, the estimated number of duplicated genes at the nodes with stars may be greatly influenced by, if not totally due to, error in the gene trees.

Table 1. Internal nodes in the species tree with duplications from at least 43 gene trees. The letter in the Node column denoted the location of the node on the species tree figure (Fig. 2). Dup. Genes shows the number of genes (out of 85) with duplications located at the specified node, and Random Dup. Genes shows the number of duplicated genes in the 10 replicates that used the gene trees with randomly permuted leaf labels. Taxa are the taxa in the clade descending from the specified node.

Node	Dup. Genes	Random Dup. Genes	Taxa
A	85	85	All Taxa
B	72	84-85	Land Plants
C	83	84-85	Seed Plants
D	51	0	Pinaceae
E	43	0	Pinus, Abies
F	61	45-65	Angiosperms
G	52	49-58	Angiosperms except Amborella
H	76	79-83	Magnoliids + Monocots + Eudicots
I	79	85	Monocots + Eudicots
J	78	0-20	Poaceae
K	52	0	Secale + Triticum
L	44	26-36	Eudicots
M	74	55-69	Core Eudicots
N	77	84-85	Rosids + Asterids
O	48	0	Solanaceae
P	45	0	within Solanaceae
Q	63	0	Asteraceae
R	72	62-68	Eurosid I + Eurosid II
S	43	28-44	Eurosid I
T	64	0-5	Brassicaceae
U	46	0	Brassica
V	74	0	Gossypium
W	53	0-18	within Fabaceae
X	66	0-8	Populus
Y	58	0	within Populus

gene family trees. However, error in the gene trees, and possibly the species tree, can confound the results from our approach, creating evidence for apparently anomalous large-scale duplication events. Thus, our approach may be most effective as a complement to other methods for detecting large-scale duplications from genomic data of one or few taxa.

References

1. A. P. G. (APG II). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG II. *Bot. J. Linn. Soc.*, 141:399–436, 2000.
2. M. A. Bender and M. Farach-Colton. The LCA problem revisited. In *LATIN*, pages 88–94, 2000.
3. G. Blanc, K. Hokamp, and K. H. Wolfe. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.*, 13:137–144, 2003.

4. G. Blanc and K. H. Wolfe. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, 16:1093–1101, 2004.
5. J. E. Bowers, B. A. Chapman, J. Rong, and A. H. Paterson. Unravelling angiosperm genome eolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422:433–438, 2003.
6. S. B. Cannon *et al.* Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc. Natl. Acad. Sci.*, 103:14959–14964, 2006.
7. B. A. Chapman, J. E. Bowers, S. R. Schulze, and A. H. Paterson. A comparative phylogenetic approach for dating whole genome duplication events. *Bioinformatics*, 20:180–185, 2004.
8. L. Cui *et al.* Widespread genome duplications throughout the history of flowering plants. *Genome Res.*, 16:738–749, 2006.
9. M. Fellows, M. Hallet, and U. Stege. On the multiple gene duplication problem. In *ISAAC*, pages 347–356, 1998.
10. F.-I. P. C. for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 2007.
11. M. R. Garey and D. S. Johnson. *Computers and Intractability: A guide to the theory of NP-completeness*. W. H. Freeman, New York, 1979.
12. M. R. Golumbic. *Algorithmic Graph Theory and Perfect Graphs*, volume 57 of *Annals of Discrete Mathematics*. Academic Press, 2nd edition, 2004.
13. M. Goodman, J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28:132–163, 1979.
14. V. Grant. *Plant speciation*, volume 2nd Edition. Columbia University Press, 1981.
15. R. Guigó, I. Muchnik, and T. F. Smith. Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution*, 6(2):189–213, 1996.
16. Guyot and Keller. Ancestral genome duplication in rice. *Genome*, 47:610–614, 2004.
17. M. Hahn. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.*, 8:R141, 2007.
18. S. Hartmann, D. Lu, J. Phillips, and T. J. Vision. Phytome: A platform for plant comparative genomics. *Nucleic Acids Research*, 34:D724–D730, 2006.
19. D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comp. Appl. Biosci.*, 8:25–282, 1992.
20. M. Lynch and J. S. Conery. The evolutionary fate and consequence of duplicate genes. *Science*, 290:1151–1155, 2000.
21. C. L. Monma and V. K. Wei. Intersection graphs of paths in a tree. *Journal of Combinatorial Theory*, 41:141 – 181, 1985.
22. R. D. M. Page and J. A. Cotton. Vertebrate phylogenomics: reconciled trees and gene duplications. In *Pacific Symposium on Biocomputing*, pages 536–547, 2002.
23. R. D. M. Page and E. C. Holmes. *Molecular evolution: a phylogenetic approach*. Blackwell Science, 1998.
24. A. H. Paterson, J. E. Bowers, and B. A. Chapman. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci.*, 101:9903–9908, 2004.
25. S. A. Rensing, J. Ick, J. A. Fawcett, D. Lang, A. Zimmer, Y. Van de Peer, and R. Reski. An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol. Biol.*, 7:130, 2007.
26. C. Simillion, K. Vandepoele, M. C. E. Van Montagu, M. Zabeau, and Y. Van de Peer. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.*, 99:13627–32, 2002.
27. G. Stebbins. Variation and evolution in plants. *Columbia Univ. Press*, 1950.
28. L. Sterck, S. Rombauts, S. Jansson, F. Sterky, P. Rouzé, , and Y. Van de Peer. EST data suggest that poplar is an ancient polyploidy. *New Phytologist*, 167:165–170, 2005.

29. K. Vandepoele, C. Simillion, and Y. van de Peer. Evidence that rice and other cereals are ancient aneuploids. *Plant Cell*, 15:2192–2202, 2003.
30. T. J. Vision, D. G. Brown, and S. Tanksley. The origins of genome duplications in *Arabidopsis*. *Science*, 290:2114–2117, 2000.
31. L. Zhang. On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *Journal of Computational Biology*, 4(2):177–187, 1997.

A Appendix

Proof (Lemma 1). The proof is by induction on $|C|$. Clearly, the result holds for $|C| \leq 1$. Now, assume that $|C| \geq 2$ and that the result holds for all cliques with fewer nodes. Let $V = [v, v']$ be an interval in C. Then, for $C' = C - \{V\}$ it holds by the inductive assumption that $\bigcap_{I \in C'} I$ is an interval, say $U = [u, u']$ where $u = \mathrm{lca}(\bigcup_{[x,y] \in C'} x)$ and $u' = \min(\bigcup_{[x,y] \in C'} y)$.

We first show that $u' \backsim_T v'$. Any interval $W \in C'$ intersects with V since $V, W \in C$, and thus there exists $x \in V \cap W$ where $x \leq v'$. The interval W also contains the interval U and especially the element u', since $U = \bigcap_{I \in C'} I$. Since $x, u' \in W$ it follows $x \backsim_T u'$. Thus either $x \leq_T u'$ or $x >_T u'$. In the first case x is a lower bound on u' and a lower bound on v', since $x \leq_T v'$. Thus $v' \backsim_T u'$. In the latter case it follows $v' \leq_T u'$ from $x >_T u'$ and $v' \geq_T x$.

Now, consider the following two cases:

Case $V \cap U \neq \emptyset$ We show that $\bigcap_{I \in C} I$ is an interval in $\leq_T$. From $V \cap U \neq \emptyset$ and $u' \backsim_T v'$ follows that $V \cap U = [\mathrm{lca}(u, v), \min(u', v')]$. With our hypothesis $u = \mathrm{lca}(\bigcup_{[x,y] \in C'} x)$ and $u' = \min(\bigcup_{[x,y] \in C'} y)$, the desired statement follows.

Case $V \cap U = \emptyset$ We show that this case is not possible. Consider the two possible cases for $u' \backsim_T v'$:

- **Case** $u' \leq_T v'$ Thus $[u', v']$ is an interval, and $[u', v'] \cap V$ is an interval with the minimum element $v'' := \mathrm{lca}(u', v)$. With $U \cap V = \emptyset$ follows that $u' < v''$ and further that $v'' \notin U$. We show that v'' is an element in every $W \in C'$ and thus $v'' \in U$, a contradiction. Consider any $W \in C'$, then $u' \in W$, and there exists $x \in W \cap V$, since $W, V \in C$. With $u' < v''_T$ we follow that $w \leq u' <_T v'' \leq_T x \leq_T w'$ and further $v'' \in W$ as desired.
- **Case** $v' <_T u'$ Thus $[v', u']$ is an interval. We show that v' is an element in every $W \in C'$ and thus $v' \in U$, a contradiction to $V \cap U = \emptyset$. Consider any $W \in C'$ we have $u' \in W$, and there exists $x \in V \cap W$ where $x \leq_T v'$. Therefore we have $w \leq_T x \leq_T v' < u'' \leq_T u' \leq_T w'$ from which follows that $v' \in W$ as desired.

□

Proof (Lemma 2). We first show that $cover(v)$ forms a clique in the intersection graph $int(\mathcal{I})$ for any $v \in V$. Let U, V be distinct intervals in $cover(v)$, then $v \in (U \cap V)$. Thus $\{U, V\} \in E(int(\mathcal{I}))$ and it follows that $int(\mathcal{I})$ is a clique.

From the proven statement above follows that $\mathcal{C}$ is a collection of cliques in $int(\mathcal{I})$. To show that $\mathcal{C}$ covers $int(\mathcal{I})$ consider an interval $I \in V(int(\mathcal{I}))$. Since V covers $\mathcal{I}$, there exists an element $v \in V$ such that $I \in cover(v)$. We have shown that $cover(v)$ is a clique in $\mathcal{C}$. Hence, $\mathcal{C}$ covers $int(\mathcal{I})$. □

Proof (Theorem 1). We first show that $f(\mathcal{C})$ is an interval cover of $\mathcal{I}$, and then we show the minimality of the interval cover $f(\mathcal{C})$.

$f(\mathcal{C})$ **is an interval cover for** $\mathcal{I}$**:** Let $I \in \mathcal{I}$. Since $\mathcal{C}$ is a clique cover of $int(\mathcal{I})$, there exists a clique $C \in \mathcal{C}$ where $I \in C$. Thus $f(C)$ is an element in I and therefore covers I. Hence, every interval $I \in \mathcal{I}$ is covered by $f(\mathcal{C})$.

$f(\mathcal{C})$ **is a minimum interval cover for** $\mathcal{I}$**:** We first prove that $|f(\mathcal{C})| = |\mathcal{C}|$ by showing that f is injective. Suppose that there exist distinct cliques $C, C' \in \mathcal{C}$ such that $f(C) = f(C')$. Then, $f(C) \in I$ for every interval $I \in (C \cup C')$. Therefore, $C \cup C'$ forms a clique in $int(\mathcal{I})$, and $\mathcal{C}' = \mathcal{C} - \{C, C'\} \cup \{C \cup C'\}$ is a clique cover of $int(\mathcal{I})$ where $|\mathcal{C}'| < |\mathcal{C}|$. Hence, $\mathcal{C}$ is not a minimum clique cover of $int(\mathcal{I})$, a contradiction.

Now, suppose for the purpose of a contradiction that there exists an interval cover $V \subseteq V(T)$ such that $|V| < |f(\mathcal{C})|$. By Lemma 2, $\mathcal{C}' := \bigcup_{v \in V} \{cover(v)\}$ is a clique cover and $|\mathcal{C}'| \leq |V| < |f(\mathcal{C})| = |\mathcal{C}|$. Hence, $\mathcal{C}$ is not a minimum clique cover, a contradiction.

□

Proof (Theorem 2). Theorem 1 states that the TIC problem for an instance $\mathcal{I}$ can be solved by finding a minimum clique cover $\mathcal{C}$ in the intersection graph $int(\mathcal{I})$ and then constructing an interval cover by selecting for every clique $C \in \mathcal{C}$ a node $v \in [a, b]$ where $a = \mathrm{lca}(\bigcup_{[x,y] \in C} x)$ and $b = \min_{[x,y] \in C} y$.

The intersection graph $int(\mathcal{I})$ can be constructed naively through a tree traversal of T in time $O(n^2 + l)$. A minimum clique cover $\mathcal{C}$ of $int(\mathcal{I})$ can be found in $O(n + m)$ by Lemma 4. Also naively the node a (using [2] for the lca computation) or b can be computed in $O(n)$ time for each clique in $\mathcal{C}$. This results in $O(nm)$ time to construct an interval cover from $\mathcal{C}$. In summary the TIC problem can be solved in time $O(n^2 + nm + l)$. □

Proof (Corollary 1). The EC problem for the instance $(\mathcal{G}, S)$ is the TIC problem for the instance $\mathcal{I} = \bigcup_{g \in \mathrm{Dup}(\mathcal{G},S)} I(g)$. Therefore, the overall time to solve the EC problem is the time to compute the instance $\mathcal{I}$ in addition to the running time to solve the TIC problem for the instance $\mathcal{I}$.

After $O(l)$ preprocessing time, the mapping $\mathcal{M}$ for all gene trees in $\mathcal{G}$ can be computed in $O(k)$ time [31] (using [2]). Traversing all trees $G \in \mathcal{G}$ the gene duplications and their intervals can computed in $O(k)$ time. Hence, the desired TIC problem instance can be computed in $O(k + l)$ time. The TIC problem for the $O(k)$ intervals over $\leq_S$ can be solved in time $O(k^2 + km + l)$ by Theorem 1. In summary the EC problem can be solved in time $O(k^2 + km + l)$. □