JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 8, 2009
© Mary Ann Liebert, Inc.
Pp. 1071–1083
DOI: 10.1089/cmb.2009.0139

Research Article

Locating Large-Scale Gene Duplication Events through Reconciled Trees: Implications for Identifying Ancient Polyploidy Events in Plants

J.G. BURLEIGH,[1] M.S. BANSAL,[2] A. WEHE,[2] and O. EULENSTEIN[2]

ABSTRACT

Recent analyses of plant genomic data have found extensive evidence of ancient whole genome duplication (or polyploidy) events, but there are many unresolved questions regarding the number and timing of such events in plant evolutionary history. We describe the first exact and efficient algorithm for the Episode Clustering problem, which, given a collection of rooted gene trees and a rooted species tree, seeks the minimum number of locations on the species tree of gene duplication events. Solving this problem allows one to place gene duplication events onto nodes of a given species tree and potentially detect large-scale gene duplication events. We examined the performance of an implementation of our algorithm using 85 plant gene trees that contain genes from a total of 136 plant taxa. We found evidence of large-scale gene duplication events in *Populus*, *Gossypium*, Poaceae, Asteraceae, Brassicaceae, Solanaceae, Fabaceae, and near the root of the eudicot clade that are consistent with previous genomic evidence. However, a lack of phylogenetic signal within the gene trees can produce erroneous evidence of large-scale duplication events, especially near the root of the species tree. Although the results of our algorithm should be interpreted cautiously, they 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.

Key words: algorithms, combinatorial optimization, computational molecular biology.

1. INTRODUCTION

WHOLE GENOME DUPLICATION EVENTS occur in many organisms and are especially widespread in plants (Stebbins, 1950; Grant, 1981; Adams and Wendel, 2005). In flowering plants (angiosperms), whole genome duplication, or polyploidy, may represent 2–4% of all speciation events (Otto and Whitton, 2000), and new genomic evidence suggests that all, or nearly all, angiosperm lineages have experienced at least one whole genome duplication (Cui et al., 2006; Soltis et al., 2009). Furthermore, polyploidy in plants has been linked to high rates of species diversication (De Bodt et al., 2005; Vamosi and Dickinson, 2006; Soltis et al.,

[1]Department of Biology, University of Florida, Gainesville, Florida.
[2]Department of Computer Science, Iowa State University, Ames, Iowa.

2009; Fawcett et al., 2009) and rapid changes in patterns of gene expression (Adams et al., 2003, 2004). To understand the evolutionary implications and effects of polyploidy in plants, it is first necessary to determine when these events occurred. Analyses of genomic data have revealed evidence of cryptic, ancient genome duplications in such lineages as grasses (Vandepoele et al., 2003; Guyot and Keller, 2004; Paterson et al., 2004a; Schlueter et al., 2004; Wang et al., 2005; Yang et al., 2005; Yu et al., 2005), *Arabidopsis* or other Brassicaceae (Vision et al., 2000; Simillion et al., 2002; Blanc et al., 2003; Bowers et al., 2003; Schranz and Mitchell-Olds, 2006), legumes (Shoemaker et al., 1996; Schlueter et al., 2004; Cannon et al., 2006), *Populus* (Sterck et al., 2005), *Gossyplum* (Blanc and Wolfe, 2004; Rong et al., 2004), *Physcomitrella* (Rensing et al., 2007), *Vitis* (Jaillon et al., 2007; Velasco et al., 2007), and the Compositae (Barker et al., 2008). Still, there is no clear consensus on the number of ancient genome duplications or precisely when they happened. In this study, we describe a novel algorithmic approach for mapping large-scale gene duplications, such as whole genome duplications, on a species tree, and we demonstrate its ability to identify ancient whole genome duplications in plant evolutionary history.

The presence of large, duplicated chromosomal segments within a genome provided the first evidence of ancient polyploidy in *Arabidopsis* (Vision et al., 2000; Simillion et al., 2002; Blanc et al., 2003; Bowers et al., 2003), rice (Vandepoele et al., 2003; Guyot and Keller, 2004; Paterson et al., 2004a; Wang et al., 2005; Yu et al., 2005), and *Vitis* (Jaillon et al., 2007; Velasco et al., 2007). We can estimate the timing of these large-scale chromosomal duplications based on the sequence divergence of paralogous genes on duplicated blocks. However, rapid gene loss and gene rearrangements after polyploidy can make it extremely difficult, if not impossible, to detect ancient duplicated chromosomal segments (Lynch and Conery, 2000; Eckhardt, 2001; Wolfe, 2001; Simillion et al., 2002). Without evidence of duplicated blocks of genes, it is still possible to detect ancient polyploidy based on the age distributions of pairs of duplicated (paralogous) genes (Lynch and Conery, 2000; Vision et al., 2000; Vandepoele et al., 2003; Blanc and Wolfe, 2004; Schlueter et al., 2004; Sterck et al., 2005; Cui et al., 2006; Rensing et al., 2007). If gene duplication and loss occurs at a constant rate through time, there should be an exponential decrease in the number of pairs of duplicated genes as they get older, or get more diverged. A whole genome duplication should result in overrepresentation of duplicated gene pairs at the age corresponding to the whole genome duplication. Again, the timing of the inferred duplications can be estimated from the sequence divergence of paralogs based on amino acid or, more commonly, silent (synonymous) substitution rates. Still, it can be difficult to accurately interpret the age distribution graphs, and such analyses can fail to detect known genome duplication events (Blanc and Wolfe, 2004; Paterson et al., 2004b).

Examining genomic data from multiple taxa in a phylogenetic context potentially can improve estimates of the timing of large-scale duplication events. For example, in a simple three-taxon case, a phylogenetic tree is constructed with a pair of paralogous genes from one taxon, the best homolog from a second taxon, and a homolog from an outgroup taxon (Bowers et al., 2003; Langkjaer et al., 2003; Vandepoele et al., 2003; Chapman et al., 2004). If the paralogous genes are sister in the resulting phylogeny, they diverged after the most recent common ancestor of the first and second taxa; if a paralog from the first taxa is sister to the sequence from the second taxa, the paralogs diverged prior to the most recent common ancestor of the first and second taxa. This approach has determined the position of large-scale duplications in *Arabidopsis* relative to its divergence with pines, rice, and other eudicots (Bowers et al., 2003) and rice relative to its divergence with pines, *Arabidopsis*, and other monocots (Vandepoele et al., 2003; Chapman et al., 2004).

Guigó et al. (1996) first addressed a more comprehensive phylogenetic approach for detecting large-scale duplication events based on mapping duplications from a collection of rooted, binary gene trees onto a rooted, binary species tree. Page and Cotton (2002) refined this problem 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. (1998), which they proved to be intrinsically difficult. 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. (1979). 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.

For example, in Figure 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]

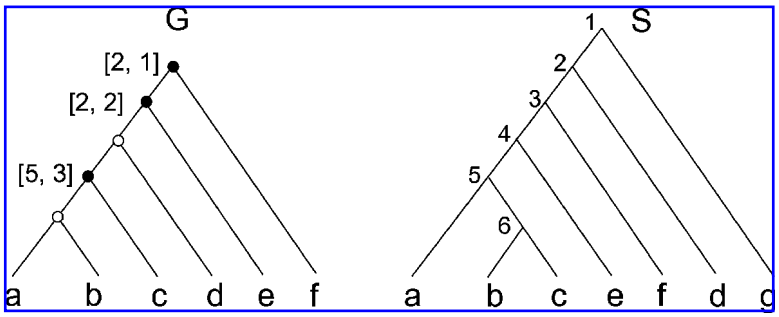

FIG. 1. *G* is a gene tree and *S* is a comparable species tree. The bold nodes in *G* are duplications and their intervals represent their allowed locations in the species tree *S*.

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 Figure 1 can be placed on species nodes 2 and 3. Page and Cotton (2002) observed that the EC problem can be efficiently reduced to the set-cover problem (Garey and Johnson, 1979). They approached the EC problem using a heuristic for the intrinsically difficult set-cover problem. In this article, we present an efficient and exact solution for the EC problem. Our solution is based on the observation that the set-cover formulation of Page and Cotton (2002) only produces instances in which the sets correspond to paths on a tree. We show how this restricted version of the set-cover problem can be solved efficiently using established graph theoretical results.

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.

2.1.1. 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 $\mathrm{Ro}(T)$. 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 $\mathrm{Ro}(T)$ and x. We denote by $x \sim_T y$ 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 $\mathrm{Le}(T)$ 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 $\mathrm{Pa}(x)$ and we call x a *child* of y. The set of all children of y is denoted by $\mathrm{Ch}_T(y)$. 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 $\{x \in V(T): x \leq_T y\}$. 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.

2.1.2. The episode-clustering (EC) problem. The EC problem is to place duplications onto a minimum number of nodes 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 (Page and Holmes, 1998) introduced by Goodman et al. (1979). 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 2.1 (Mapping). *A* leaf-mapping $\mathcal{L}_{G,S}$:Le$(G) \to$ Le(S) *specifies, for each gene g the species from which it was sampled. The extension* $\mathcal{M}_{G,S}: V(G) \to V(S)$ of $\mathcal{L}_{G,S}$ *is the* mapping *defined by* $\mathcal{M}_{G,S}(g) := \text{lca}(\mathcal{L}_{G,S}(\text{Le}(\text{G}_\text{g}))$.

Note, that mathematically speaking a leaf-mapping always exists. However, in the current context, we are only concerned with biologically relevant leaf-mappings.

Definition 2.2 (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 article, $\mathcal{G}$ denotes a collection of input gene trees, S a comparable species tree, and G denotes an arbitrary gene tree in $\mathcal{G}$.

Definition 2.3 (Duplication). *A node* $v \in V(G)$ *is a* (gene) duplication *if* $\mathcal{M}_{G,S}(v) = \mathcal{M}_{G,S}(u)$ *for some* $u \in$ Ch(v), *and we define* $\text{Dup}(G,S) := \{g \in V(G) \mid g \text{ is a duplication}\}$.

Definition 2.4. *For every* $g \in V(G)$, *we define the interval*

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

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* $\bigcup_{g \in \text{Dup}(\mathcal{G},S)} \{I(g)\}$ *in the order* $\leq_S$.
The TIC instance $\bigcup_{g \in \text{Dup}(\mathcal{G},S)} \{I(g)\}$ can be computed in linear time (Zhang, 1997) using efficient lca computation (Bender and Farach-Colton, 2000). 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 $\leq_T$.

Lemma 2.1. *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* $a = \text{lca}(\bigcup_{[x,y] \in C} x)$ *and* $b = \min(\bigcup_{[x,y] \in C} y)$.

Proof. 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 $u = \text{lca}(\bigcup_{[x,y] \in C'} x)$ and $u' = \min(\bigcup_{[x,y] \in C'} y)$. ■

We first show that $u' \sim_T v'$. 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 $x \sim_T u'$. 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 $v' \sim_T u'$. In the latter case, it follows $v' \leq_T u'$ from $x >_T u'$ and $v' >_T x$.

Now, consider the following two cases:

Case $\mathbf{V} \cap \mathbf{U} \neq \emptyset$. We show that $\bigcap_{I \in C} I$ is an interval in $\leq_T$. From $V \cap U \neq \emptyset$ and $u' \sim_T v'$ 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 $\mathbf{V} \cap \mathbf{U} = \emptyset$. We show that this case is not possible. Consider the two possible cases for $u' \sim_T v'$:

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. ■

Lemma 2.2. *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($\mathcal{I}$).*

Proof. 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*($\mathcal{I}$). ■

Theorem 2.1. *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 $F{:}\mathcal{C} \rightarrow V(T)$ that maps $f(C)$ to some element in $\bigcap_{I \in C} I$. Note, f is well defined by Lemma 2.1. Then, the node set $f(\mathcal{C})$ is a minimum interval cover of $\mathcal{I}$.*

Theorem 2.1. 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 2.2, $\mathcal{C}' := \bigcup_{v \in V} \{cover(v)\}$ 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. ■

The following two results are well known (Monma and Wei, 1985; Golumbic, 2004).

Lemma 2.3. *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 2.4. *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 2.2. *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 $l = |\mathrm{Le}(T)|$.*

Proof. Theorem 2.1 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 $a = \text{lca}(\bigcup_{[x,y]\in C} x)$ 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 2.4. Also, the node a for the lca computation) or b can be computed in $O(n)$ time (e.g., Bender and Farach-Colton, 2000) 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 $O(n^2 + nm + l)$. ■

Corollary 2.1. *Let $\mathcal{G}$ be a collection of gene trees and S a comparable species tree, where $k = \Sigma_{G\in\mathcal{G}}|\text{Le}(G)|$ and $l = |\text{Le}(S)|$. 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 $\mathcal{G}$.*

Proof. The EC problem for the instance $(\mathcal{G}, S)$ is the TIC problem for the instance $\mathcal{I} = \bigcup_{g\in\text{Dup}(\mathcal{G},S)} I(g)$. 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 (Zhang, 1997). 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 2.1. In summary the EC problem can be solved in time $O(k^2 + km + l)$. ■

2.3. Plant study

2.3.1. Data set. We evaluated the performance and utility of our algorithm using a set of plant gene family trees made from alignments obtained from the Phytome, an online comparative genomics database (Hartmann et al., 2006). For our analysis, we selected the masked amino acid alignments from all gene families that contain sequences from at least 100 of the 136 total taxa represented in Phytome. The gene trees were inferred from maximum likelihood (ML) phylogenetic analyses on the 13 gene family alignments using RAxML-VI-HPC version 2.2.3 (Stamatakis et al., 2005). The ML gene trees were first rooted using mid-point rooting as implemented in PHYLIP Retree (Felsenstein, 2005). We then searched for alternate rootings that implied fewer gene duplications needed to reconcile the gene trees with the species tree. If such a rooting existed, we re-rooted the gene trees to minimize the overall 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 pruned the single-taxon clade by removing all but a single leaf. We used a species tree based on the APG II classification (APG II, 2000; Soltis et al., 2000).

2.3.2. Inferring gene duplications events. We used our algorithm to infer the minimum number of duplication locations for the set of ML gene trees on the species tree. Our implementation of the algorithm provides a solution for the minimum number of duplication locations and 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. Also, in order to examine how the phylogenetic signal in the gene trees effects the performance of our algorithm, we created 1000 gene tree data sets by randomly permuting the leaf labels from each of the gene trees in our original data set. This experiment will provide an expectation of the results of our analysis if there was no phylogenetic signal in the gene trees, or if the gene tree topologies were essentially random.

3. RESULTS

3.1. Data set

We identified 85 gene families in Phytome that each contained sequences from at least 100 of the 136 total taxa. The gene trees contained a minimum of 160 and a maximum of 933 leaves (average = 347; median = 290). On average, each taxon is represented in 68.3 of the gene trees (median = 78).

3.2. Plant duplication analysis

Our algorithm identified gene duplication events on a minimum of 119 internal nodes in the species tree. Some nodes have evidence of many duplications while others have evidence of very few duplications (Figs. 2 and 3). For example, while 51 nodes have evidence of $\leq$10 duplications, four nodes have evidence of $\geq$1000 duplications (Fig. 2). Also, there is evidence of duplications involving five or fewer of the gene trees in 40 nodes, but there is evidence of duplications involving 75 or more gene trees on five nodes. Since we are most interested in identifying large-scale duplications, we focus on the nodes with duplications involving at least half ($\geq$43) of the gene trees. There are 25 such nodes (Table 1 and Fig. 4). Though these nodes occur throughout the plant species tree, they are especially abundant near the root of the species tree (Fig. 3). However, they also are associated with several major groups of angiosperms including Poaceae, Solanaceae, Asteraceae, Brassicaceae, and some model systems for plant genomics such as *Populus* and *Gossypium* (Table 1 and Fig. 4).

3.3. Random leaves analysis

The 1000 analyses using gene trees with randomly permuted leaf labels found evidence for gene duplication events on only 21–35 (average 27.9) 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. 4) of the species tree as well as the root nodes of the eudicots (nodes L, M, N, and R in Fig. 4), but there were generally few genes in the other nodes of the species tree (Table 1 and Fig. 4).

4. DISCUSSION

4.1. Gene and genome duplications in plants

The analysis of gene duplications in plants provides a plausible hypothesis for the history of gene duplication events in plants. Our results first emphasize the ubiquity of gene duplications throughout the evolutionary history of plants. Although we examined only 85 gene trees with incomplete sampling, we found evidence of gene duplications on at least 88% of the internal nodes, with most nodes having few duplications (Figs. 2 and 3). The randomly permuted gene tree replicates resulted in duplications on far fewer nodes, suggesting that the phylogenetic signal in the gene trees greatly increases estimated minimum number of locations for gene duplication events. These results emphasize the role of both large and small-scale duplications in generating plant gene family diversity (Lynch and Conery, 2000).

Our analyses also provide a hypothesis for the placement of large-scale gene duplications in plants that is generally consistent with estimates based on other data and methods (Bowers et al., 2003; Cui et al., 2006;

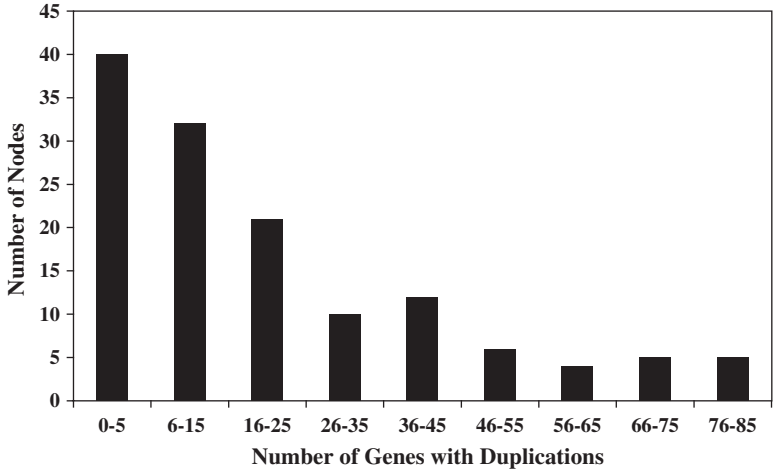

FIG. 2. Distribution of the number of gene trees with duplications across internal nodes. For example, 40 internal nodes have 0–5 gene trees with duplications.

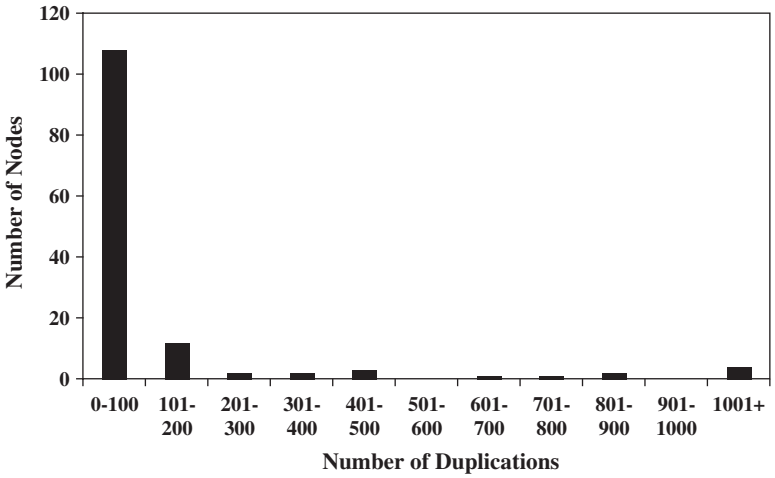

FIG. 3. Distribution of the number of gene duplications across internal nodes. For example, 108 internal nodes have 0–100 gene duplications.

Soltis et al., 2009). It is difficult to determine an objective cutoff to distinguish evidence of ancient whole genome duplications polyploidy from merely evidence of a large number of gene duplications. However, the 25 nodes with evidence of duplications in at least half of the gene families include the locations of many accepted ancient plant whole genome duplications. These include an event at the base of the Poaceae (node J in Fig. 4) (Guyot and Keller, 2004; Paterson et al., 2004a; Schlueter et al., 2004; Wang et al., 2005),

TABLE 1. INTERNAL NODES IN THE SPECIES TREE WITH DUPLICATIONS FROM AT LEAST 43 GENE TREES

Node	*Dup. trees*	*Random dup. trees*	*Taxa*
A	85	84.9 (84–85)	All taxa
B	72	84.9 (83–85)	Land plants
C	83	85 (85–85)	Seed plants
D	51	1.3 (0–18)	Pinaceae
E	43	1.4 (0–17)	*Pinus, Abies*
F	61	57.8 (45–67)	Angiosperms
G	52	54.2 (41–64)	Angiosperms except *Amborella*
H	76	81.4 (74–85)	Magnoliids + Monocots + Eudicots
I	79	85 (85–85)	Monocots + Eudicots
J	78	7.7 (0–27)	Poaceae
K	52	0.6 (0–15)	*Secale* + *Triticum*
L	44	32.7 (20–46)	Eudicots
M	74	64.9 (49–76)	Core Eudicots
N	77	84.7 (82–85)	Rosids + Asterids
O	48	0.5 (0–14)	Solanaceae
P	45	0.7 (0–16)	Within Solanaceae
Q	63	0.2 (0–6)	Asteraceae
R	72	67.4 (56–77)	Eurosid I + Eurosid II
S	43	35.2 (21–56)	Eurosid I
T	64	1.2 (0–11)	Brassicaceae
U	46	0.1 (0–7)	*Brassica*
V	74	0.7 (0–19)	*Gossypium*
W	53	0.9 (0–22)	Within Fabaceae
X	66	1.7 (0–19)	*Populus*
Y	58	1.3 (0–16)	Within *Populus*

The letter in the "Node" column denotes the location of the node on the species tree figure (Fig. 4). "Dup. Trees" shows the number of gene trees (out of 85) with duplications located at the specified node, and "Random Dup. Trees" shows the average number (and range) of duplicated gene trees in the 1000 replicates that used the gene trees with randomly permuted leaf labels. "Taxa" are the taxa in the clade descending from the specified node.

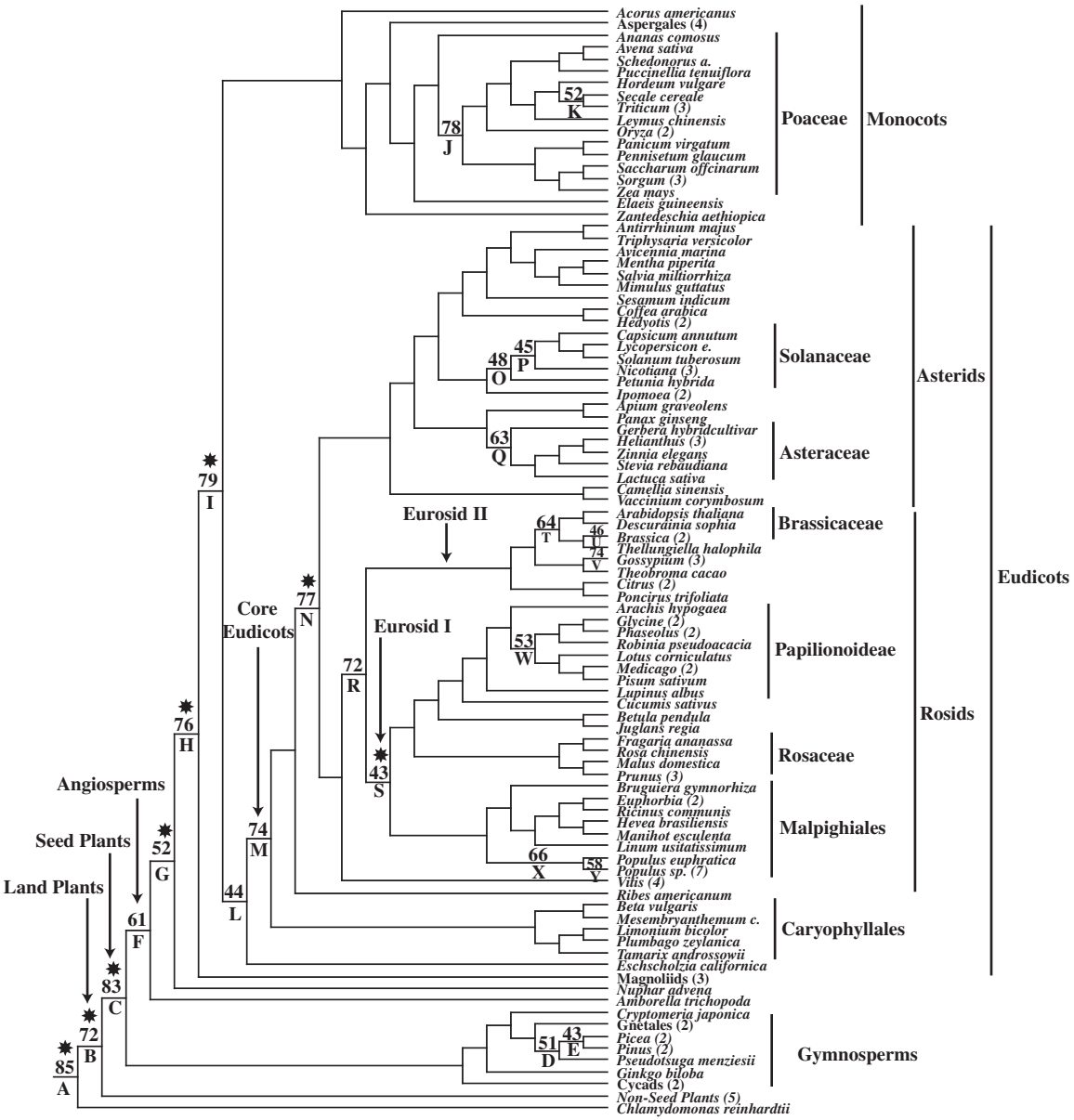

FIG. 4. Species tree with potential locations of large-scale gene duplication events. The species tree used in the analysis contains 136 taxa, and in some cases, multiple (usually congeneric) species in a clade were combined into a single taxon for this figure. In these cases, the total number of species 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, on average, 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.

Brassicaceae (node T in Fig. 4) (Vision et al., 2000; Simillion et al., 2002; Bowers et al., 2003), and Asteraceae (node Q in Fig. 4) (Cui et al., 2006; Barker et al., 2008), within Solanaceae (nodes O and P in Fig. 4) (Cui et al., 2006) and Fabaceae (node W in Fig. 4) (Schlueter et al., 2004; Cannon et al., 2006), and in *Populus* (nodes X and Y in Fig. 4) (Sterck et al., 2005), and *Gossypium* (node V in Fig. 4) (Blanc and Wolfe, 2004; Rong et al., 2004). Since our analyses include many taxa, in some cases, they provide more precise hypotheses of the phylogenetic location of these duplications than previous methods. For example, while previous studies found evidence of a large-scale gene duplication that is common to several grass

taxa (Vandepoele et al., 2003; Paterson et al., 2004a), our analysis suggests it occurred between the divergence of *Ananas* and the Poaceae (node J) (Fig. 4).

There is little previous evidence for large-scale duplications at all the root nodes (nodes A–C, F–I; four), and at most of the early eudicot nodes (nodes L–N, R–S; four). Yet these also are the nodes where large numbers of duplications map in our analysis of the randomly permuted gene trees (Table 1 and Fig. 4). 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 (Hahn, 2007). Our analyses suggest that the lack of phylogenetic signal within the gene trees also will provide evidence of large-scale duplications at the root nodes. Thus, we advise interpreting such evidence with great caution. Several studies have suggested a whole genome duplication at the base of the angiosperms (node F in Fig. 4) (Bowers et al., 2003; Cui et al., 2006; Jaillon et al., 2007). Our analysis found evidence of duplications in slightly more gene trees than we would expect in the absence of phylogenetic signal (Table 1). Thus, our data suggests it is possible that a whole genome duplication did precede the origin of angiosperms. However, using our approach on this data set, it may be difficult to distinguish evidence for polyploidy at the root of the angiosperms from the effects of gene tree error.

We also found several questionable placements of large-scale duplications closer to the tips of the species tree. For example, although there is evidence of a single ancient polyploidy in *Populus* (Sterck et al., 2005), our analyses suggest two large-scale duplication events in *Populus* (nodes X and Y in Fig. 4). Also, our analysis revealed evidence of a large number of duplications in Pinaceae (nodes D and E in Fig. 4); however, there is very little evidence of polyploidy in gymnosperms (Khoshoo, 1959), and a recent analysis of genomic sequences from *Pinus* found no evidence of ancient whole genome duplications (Cui et al., 2006). Hybridization among some of the *Populus* or Pinaceae taxa, incomplete lineage sorting, and error in the species tree could all lead to overestimates of gene duplication events.

Overall, if we disregard the potentially erroneous large-scale duplication 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 (Jaillon et al., 2007; Velasco et al., 2007). We hypothesize that the two genome duplications in *Arabidopsis* since its common ancestor with *Vitis* occurred at the base of the Brassicaceae (node T in Fig. 4) and at the base of the eurosid I + eurosid II clade (node R in Fig. 4). The ancestral hexiploidization of the *Vitis* and *Arbaidopsis* genomes occurred at nodes L and/or M (Fig. 4), after the divergence of eudicots and monocots.

4.2. Algorithm performance and limitations

The plant gene analysis demonstrates that our algorithm can be useful for identifying large-scale duplications from relatively few gene trees. Still, our experiments also suggest some weaknesses in our approach and potentially informative directions for future research. First, though our analysis used only 85 gene trees, we found evidence of duplications on nearly all of the internal nodes. With enough gene trees, there will certainly be evidence for duplications on every node of the tree. In this case, all possible gene duplication mappings will be equally optimal, and our algorithm would not be useful.

A similar, but alternate approach is to identify the duplication mappings that minimize the overall number of duplication episodes (Page and Cotton, 2002; Bansal and Eulenstein, 2008). In fact, the EC problem was first proposed as a heuristic for finding the minimum number of gene duplication episodes (Guigó et al., 1996; Page and Cotton, 2002). Bansal and Eulenstein (2008) later introduced an efficient algorithm to find the minimum number of gene duplication episodes and demonstrated that the solution to EC problem does not necessarily correspond to the minimum number of episodes. However, minimizing the number of gene duplication episodes does not necessarily accurately reflect the actual process of gene duplication. Minimizing the number of duplication episodes suggests that gene duplication events are rare, but they involve many genes. However, gene duplication appear to be relatively common, and nearly all involve only a single gene or very few genes, with the much more rare large-scale or whole-genome duplication events. Thus, assuming an episode model of gene duplication may provide a very misleading interpretation of genome evolution. Also, the minimum number of duplication episodes at a node in the species tree is the maximum number of episodes implied by any single gene tree at that node. In other words, a single gene tree can (and often does) have huge effects on the minimum number of episodes, and many small trees will have no effect on the number of gene duplication episodes, no matter where there duplications are mapped. If we want to find the location of whole-genome duplications that by definition affect all genes, we need to use a criterion

that considers all genes, not just a few of the largest genes. A more accurate approach to identify whole genome duplications might be to find the mapping that maximizes the size of the largest gene duplication episodes without necessarily minimizing the number of episodes. However, ideally we would identify whole genome duplications based on empirically derived models of gene family evolution.

The plant gene family analysis also strongly suggests that gene tree error can produce evidence of apparently anomalous large-scale gene duplication events. Unfortunately, a certain level of error is likely inherent in any gene tree inference. Our approach can take advantage of large-scale genomic data, such as EST sequences, that are being produced from many plant taxa. Yet the fragmentary and incomplete nature of some of these data often produces gappy alignments, which in turn may complicate phylogenetic analyses. Furthermore, even if the gene tree topology is correct, it is extremely difficult to identify the correct rooting when there is a history of duplication events. Midpoint rooting relies on molecular clock assumptions, but the molecular clock is rejected in over 70% of the Phytome gene families (Hartmann et al., 2006). We choose a rooting that minimizes the overall number of duplications, but using this criterion, there often are multiple optimal rootings. It may be useful to develop methods for mapping large-gene duplications that can account for possible error in the gene trees, either by utilizing unresolved or possibly unrooted gene trees or by allowing small changes in the topology of the gene trees if they will lead to better solutions. It will also be useful to more fully examine the effects of error on mapping of large-scale gene duplications in order to better diagnose potentially erroneous results.

5. CONCLUSION

Our study uses a cross-disciplinary approach to examine the problem of identifying ancient large-scale duplications across the plant tree of life. We introduce a new exact algorithm that attempts to solve the following 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 gene duplication events in such a phylogenetic context, we support biologists in their efforts to specify the precise location and timing of duplication events. We suggest that our approach can be an informative complement to other large-scale genomic approaches. Unlike other methods, our approach does not need gene map data for many species and does not rely on molecular clock assumptions to place gene duplications. Furthermore, it can be useful with relatively little genomic data, and thus can include many taxa. Although our approach assumes a very simple model of gene duplication, our analysis of plant gene trees produces a credible hypothesis for the placement of ancient whole genome duplications. 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. When these methods detect evidence of a whole genome duplication, our algorithm can help place this duplication in a large-scale phylogenetic context.

REFERENCES

Adams, K.L., and Wendel, J.F. 2005. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* 8, 135–141.

Adams, K.L., Cronn, R., Percifield, R., et al. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. USA* 100, 4649–4654.

Adams, K.L., Percifield, R., and Wendel, J.F. 2004. Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* 168, 2217–2226.

APG II. (Angiosperm Phylogeny Group) 2000. 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.

Bansal, M.S., and Eulenstein, O. 2008. The multiple gene duplication problem revisited. *Bioinformatics* 24, i132–i138.

Barker, M.S., Kane, N.C., Matvienko, M., et al. 2008. Multiple paleopolyploidizations during the evolution of the compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* 25, 2445–2455.

Bender, M.A., and Farach-Colton, M. 2000. The LCA problem revisited. *Lect. Notes Comput. Sci.* 1776, 88–94.

Blanc, G., and Wolfe, K.H. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, 16, 1093–1101.

Blanc, G., Hokamp, K., and Wolfe, K.H. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res.* 13, 137–144.

Bowers, J.E., Chapman, B.A., Rong, J., et al. 2003. Unravelling angiosperm genome eolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438.

Cannon, S.B., Sterck, L., Rombauts, S., et al. 2006. Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc. Natl. Acad. Sci. USA* 103, 14959–14964.

Chapman, B.A., Bowers, J.E., Schulze, S.R., et al. 2004. A comparative phylogenetic approach for dating whole genome duplication events. *Bioinformatics* 20, 180–185.

Cui, L., Wall, P.K., Leebens-Mack, J.H., et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16, 738–749.

De Bodt, S., Maere, S., and Van de Peer, Y. 2005. Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.* 20, 591–597.

Eckhardt, N. 2001. A sense of self: the role of DNA sequence elimination in allopolyploidization. *Plant Cell* 13, 1699–1704.

Fawcett, J.A., Maere, S., and Van de Peer, Y. 2009. Plants with double genomes might have had a better chance to survive the cretaceous-tertiary extinction event. *Proc. Natl. Acad. Sci. USA* 106, 5737–5742.

Fellows, M.R., Hallett, M.T., and Stege, U. 1998. On the multiple gene duplication problem. *Lect. Notes Comput. Sci.* 1533, 347–356.

Felsenstein, J. 2005. PHYLIP *(Phylogeny Inference Package). Version 3.6.* University of Washington, Seattle, WA.

Garey, M.R., and Johnson, D.S. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman, New York.

Golumbic, M.C. (2004). *Algorithmic Graph Theory and Perfect Graphs*. *Volume 57 of Annals of Discrete Mathematics*, 2nd ed. Academic Press, New York.

Goodman, M., Czelusniak, J., Moore, G.W., et al. 1979. Fitting the gene lineage into its species lineage. a parsimony strategy illustrated by cladograms constructed from globin sequences. *System. Zool.* 28, 132–163.

Grant, V. 1981. *Plant Speciation*, 2nd ed. Columbia University Press, New York.

Guigó, R., Muchnik, I., and Smith, T.F. 1996. Reconstruction of ancient molecular phylogeny. *Mol. Phylogenet. Evol.* 6, 189–213.

Guyot, R., and Keller, B. 2004. Ancestral genome duplication in rice. *Genome* 47, 610–614.

Hahn, M. 2007. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.* 8, R141.

Hartmann, S., Lu, D., Phillips, J., et al. 2006. Phytome: a platform for plant comparative genomics. *Nucleic Acids Res.* 34, D724–D730.

Jaillon, O., Aury, J.-M., Noel, B., et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467.

Khoshoo, T.N. 1959. Polyploidy in gymnosperms. *Evolution* 13, 24–39.

Langkjaer, R.B., Cliften, P.F., Johnson, M., et al. 2003. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. *Nature* 421, 848–852.

Lynch, M., and Conery, J.S. 2000. The evolutionary fate and consequence of duplicate genes. *Science* 290, 1151–1155.

Monma, C.L., and Wei, V.K. 1985. Intersection graphs of paths in a tree. *J. Combin. Theory* 41, 141–181.

Otto, S.P., and Whitton, J. 2000. Polyploid incidence and evolution. *Annu. Rev. Genet.* 34, 401–437.

Page, R.D.M., and Cotton, J.A. 2002. Vertebrate phylogenomics: reconciled trees and gene duplications. *Pacif. Symp. Biocomput.* 536–547.

Page, R.D.M., and Holmes, E.C. 1998. *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, New York.

Paterson, A.H., Bowers, J.E., and Chapman, B.A. 2004a. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. USA* 101, 9903–9908.

Paterson, A.H., Chapman, B.A., Peterson, D.G., et al. 2004b. Comparative genome analysis of monocots and dicots, toward characterization of angiosperm diversity. *Curr. Opin. Biotechnol.* 15, 120–125.

Rensing, S.A., Ick, J., Fawcett, J.A., et al. 2007. An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol. Biol.* 7, 130.

Rong, J., Abbey, C., Bowers, J.E., et al. 2004. A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). *Genetics* 166, 389–417.

Schlueter, J.A., Dixon, P., Granger, C., et al. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47, 868–876.

Schranz, M.E., and Mitchell-Olds, T. 2006. Independent ancient polyploidy events in sister families Brassicaceae and Cleomaceae. *Plant Cell* 18, 1152–1165.

Shoemaker, R.C., Polzin, K., Labate, J., et al. 1996. Genome duplication in soybean (*Glycine* subgenus *soja*). *Genetics* 144, 329–338.

Simillion, C., Vandepoele, K., Van Montagu, M.C. E., et al. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* 99, 13627–1632.

Soltis, D.E., Soltis, P.S., Chase, M.W., et al. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Bot. J. Linn. Soc.* 133, 381–341.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., et al. 2009. Polyploidy and angiosperm diversification. *Am. J. Botany* 96, 336–348.

Stamatakis, A., Ludwig, T., and Meier, H. 2005. Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21, 456–463.

Stebbins, G.L. 1950. *Variation and Evolution in Plants.* Columbia University Press, New York.

Sterck, L., Rombauts, S., Jansson, S., et al. 2005. EST data suggest that poplar is an ancient polyploidy. *New Phytol.* 167, 165–170.

Vamosi, J.C., and Dickinson, T.A. 2006. Polyploidy and diversification: a phylogenetic investigation in Rosaceae. *Int. J. Plant Sci.* 167, 349–358.

Vandepoele, K., Simillion, C., and van de Peer, Y. 2003. Evidence that rice and other cereals are ancient aneuploids. *Plant Cell* 15, 2192–2202.

Velasco, R., Zharkikh, A., Troggio, M., et al. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* 2, e1326.

Vision, T.J., Brown, D.G., and Tanksley, S.D. 2000. The origins of genome duplications in *Arabidopsis*. *Science* 290, 2114–2117.

Wang, X., Shi, X., Hao, B., et al. 2005. Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol.* 165, 937–946.

Wolfe, K.H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2, 333–341.

Yang, Z., Xu, G., Guo, X., et al. 2005. Two rounds of polyploidy in rice genome. *J. Zhejiang Univ.* 6B, 87–90.

Yu, J., Wang, J., Lin, W., et al. 2005. The genomes of *Oryza sativa*: a history of duplication. *PLoS Biol.* 3, 266–281.

Zhang, L. 1997. On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *J. Comput. Biol.* 4, 177–187.

Address correspondence to:
Dr. Oliver Eulenstein
Department of Computer Science
Iowa State University
212 Atanasoff Hall
Ames, IA 50011

E-mail: oeulenst@cs.iastate.edu