Supplementary material for: Detecting and Locating Whole Genome Duplications on a phylogeny: a probabilistic approach

1 Whole genome triplications

Whole genome triplications (WGT) are modeled similarly to WGD, with each gene lineage being instantaneously triplicated and 1 or 2 or 3 copies being retained with probabilities q_1, q_2 and q_3. In other words, out of the 3 gene copies, a maximum of 2 can be lost immediately after the triplication. This model requires 2 parameters (because $q_1+q_2+q_3=1$), instead of a single retention rate q. Under the assumption that the 2 extra copies are each retained independently with probability q, then the model simplifies to $q_3 = q^2$, $q_2 = 2q(1-q)$ and $q_1 = (1-q)^2$, using a single parameter.

1.1 Gene count probabilities

To determine the probability that a lineage entering the WGT is doomed with no descendants below the WGT, we consider the possible number of copies retained immediately after the WGT and get the following recursion formula

$$d(u_{\text{WGTbefore}}) = q_3 d(u_{\text{WGTafter}})^3 + q_2 d(u_{\text{WGTafter}})^2 + q_1 d(u_{\text{WGTafter}}).$$

Based on our model for WGTs, the transition probability from i genes just before the WGT to j genes immediately after is

$$P_{\text{WGT}}(j|i) = \sum_{\substack{k_1,\, k_2,\, k_3 \,\geq\, 0,\, k_1 \,+\, k_2 \,+\, k_3 \,=\, i \\ k_1 \,+\, 2k_2 \,+\, 3k_3 \,=\, j}} \frac{i!}{k_1!\, k_2!\, k_3!}\, q_1^{k_1} q_2^{k_2} q_3^{k_3}$$

where k_1, k_2, k_3 are the number of original genes ending up with 1, 2, or 3 retained copies after the WGT. The linear constraints on k_1, k_2 and k_3 imply that $k_1 = 2i - j + k_3$ and $k_2 = j - i - 2k_3$, and the non-negativity of each k_i implies that $\max\{0, j-2i\} \leq k_3 \leq (j-i)/2$. We get that $P_{\text{WGT}}(j|i) = 0$ if $j < i$ or $j > 3i$, and otherwise

$$P_{\text{WGT}}(j|i) = \sum_{k=\max\{0,j-2i\}}^{\lfloor \frac{j-i}{2} \rfloor} \frac{i!}{(2i-j+k)!\,(j-i-2k)!\,k!}\, q_1^{2i-j+k} q_2^{j-i-2k} q_3^{k}\,.$$

These transition probabilities can be computed fast recursively using

$$P_{\text{WGT}}(j|i+1) = q_1 P_{\text{WGT}}(j-1|i) + q_2 P_{\text{WGT}}(j-2|i) + q_3 P_{\text{WGT}}(j-3|i)\,,$$

which is derived by considering the fate of the first and remaining i lineages entering the WGT. Similarly, the survival transition probabilities w^* along the WGT edge into $u = u_{\text{WGDafter}}$ are obtained by considering the fate of the last surviving lineage. Conditional on $i = 1$ lineage entering the WGT we get:

$$\begin{aligned}
w^*_u(1|1) &= q_1(1-d(u)) + 2q_2 d(u)(1-d(u)) + 3q_3 d(u)^2(1-d(u)) \\
w^*_u(2|1) &= q_2(1-d(u))^2 + 3q_3 d(u)(1-d(u))^2 \\
w^*_u(3|1) &= q_3(1-d(u))^3 \text{ and} \\
w^*_u(i|1) &= 0 \text{ for } i = 0 \text{ or } i \geq 4.
\end{aligned}$$

The remaining values are then obtained recursively with

$$w^*_u(j|i) = w^*_u(1|1)w^*_u(j-1|i-1) + w^*_u(2|1)w^*_u(j-2|i-1) + w^*_u(3|1)w^*_u(j-3|i-1).$$

1.2 Tree probability at a WGT

Along the edge of a WGT event, there are 3 different possible observed reconciled subtrees T_k with $k = 1, 2$ or 3 leaves. Their associated probability terms are still $g(\nu, u_{\text{WGTafter}}, T_k) = f(T, T_k, R)\ h(u_{\text{WGTafter}}, k)$ with

$$h(u_{\text{WGTafter}}, k) = \begin{cases} q_3 & \text{if } k = 3, \\ q_2 + 3q_3 d(u_{\text{WGTafter}}) & \text{if } k = 2, \\ q_1 + 2q_2 d(u_{\text{WGTafter}}) + 3q_3 d(u_{\text{WGTafter}})^2 & \text{if } k = 1. \end{cases}$$

to account for the fact that more than k gene copies may be retained after the WGT, but then doomed later on. The factor $f(T, T_k, R)$ accounts for topological symmetries at internal nodes in the gene tree. At a WGT, the birth-death process does not apply, so equation (13c) in Rasmussen and Kellis (2011) simplifies to $f(T, T', R) = N_2(T, T', R)$ (using their notations). Their rationale still applies to calculate the N_2 labeling factor, but we need to expand it to polytomies. Indeed, when all 3 copies are retained $T' = T_3$ is the 3-tip tree with a polytomy. In this case, $N_2(T, T_3, R) = 1$, 3 or 6, corresponding to 1, 2, or 3 distinct “color labels” at the 3 tips of T_3. Two tips in T' have the same color label if the gene subtrees that they subtend have the same topology, when their leaves are labeled by the species in which they belong.

2 Reconciliation method

2.1 Simulation results with unknown WGD location

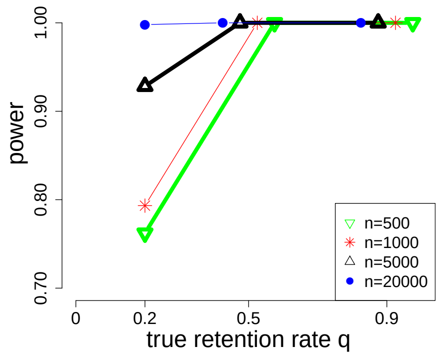

Figure 1: Estimated power of the reconciliation method on a 4-taxon tree from n gene families when 2 possible locations for the WGD are considered.

2.2 Most parsimonious reconciliation at the WGD

We focus here on an edge (v, u) in the species tree, where a WGD is hypothesized and on which nodes $u_{\text{WGDbefore}}$ and u_{WGDafter} have been added to model the WGD. We provide here an algorithm to refine the reconciliation of a gene subtree T' whose root ν has been reconciled to v and tips reconciled to u, to map each internal node in T' to either $u_{\text{WGDbefore}}$, u_{WGDafter}, or u. To do so, we seek to maximize the number of duplications at the WGD and then minimize the number of losses at the WGD. A post-order traversal of T' (algorithm 1) is first used from the tips to the root to maximize the number of WGD duplications on each subtree, when constrained to originate before the WGD. This is followed by a pre-order tree traversal of T' (algorithm 2) to determine the reconciliation of each internal node. An example of most parsimonious reconciliation is shown in figure 2. Note that this method cannot handle more than one WGD along each branch of the species tree, and that it is specific to WGDs, not whole genome triplications.

Algorithm 1: WGDbestPosition

```
Input: tree $T'$ rooted at node $\nu$
Output: Bestposition: most parsimonious reconciliation of $\nu$ under the constraint that its
        parent is reconciled at $v$ or $u_{\text{WGDbefore}}$ (i.e. before the WGD).
        Ndup: most parsimonious number of duplications at the WGD in $T'$.
if $\nu$ is a leaf of $T'$ then
    Bestposition $= u$
    Ndup=0
else
    $T'_1 =$ subtree of $T'$ rooted at the left child of $\nu$
    $T'_2 =$ subtree of $T'$ rooted at the right child of $\nu$
    $\text{Ndup}_i =$ Ndup from WGDbestPosition$(T'_i)$, $i = 1, 2$
    if $Ndup_1 + Ndup_2 > 1$ then
        Bestposition $= u_{\text{WGDbefore}}$
        Ndup $= \text{Ndup}_1 + \text{Ndup}_2$
    else
        Bestposition $= u_{\text{WGDafter}}$
        Ndup $= 1$
return Bestposition, Ndup
```

Algorithm 2: WGDresetReconciliation

```
Input: tree $T'$ rooted at $\nu$, reconciliation $R$ mapping $\nu$ to either $v$ or $u$ and all other nodes
       to $u$, and Bestposition for each node from algorithm 1
Output: updated reconciliation $R$
if $\nu$ is a leaf of $T'$ then
    reconciliation $R[\nu]$ is unchanged at $u$
else
    if $Bestposition[\nu] = u_{WGDafter}$ then
        $R[\nu] = u_{\text{WGDafter}}$
        label $\nu$ as a WGD duplication
        stop (all descendants of $\nu$ have reconciliation $R$ unchanged at $u$)
    if $Bestposition[\nu] = u_{WGDbefore}$ then
        $R[\nu] = u_{\text{WGDbefore}}$
        $T'_1 =$ subtree of $T'$ rooted at the left child of $\nu$
        $T'_2 =$ subtree of $T'$ rooted at the right child of $\nu$
        WGDresetReconciliation$(T'_1)$
        WGDresetReconciliation$(T'_2)$
```

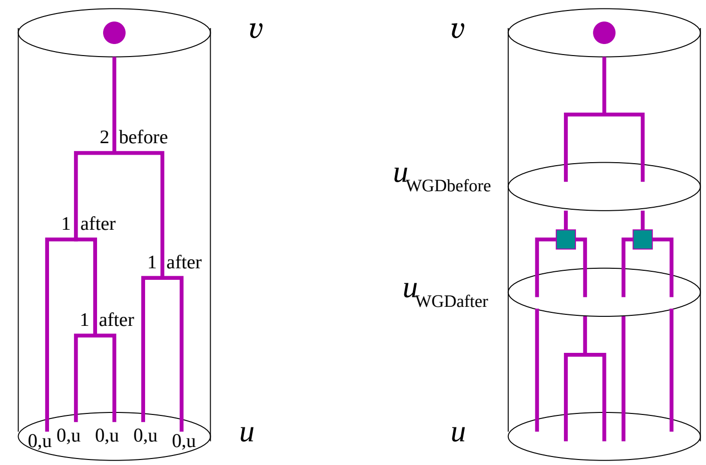

Figure 2: Example of a gene subtree T' within a branch of the species tree with a WGD (left) and its most parsimonious reconciliation at the WGD (right). Node numbers and annotations indicate the "Ndup" and "Bestposition" values calculated by algorithm 1. Blue squares indicate duplications from the WGD.

2.3 Changes to SPIMAP

In addition to the changes for the presence of WGDs and a geometric prior distribution for the number of genes at the root, we implemented two conditional probability calculations. By default, probabilities are conditional on the data being observed (non-extinct families). Users also have the option to condition on filtering families having at least one gene in each subtree from the root.

In addition, we implemented the non-molecular-clock LOCAL sampler of Larget and Simon (1999), called SubtreeSlides in SPIMAP. It modifies the tree only in a small neighborhood of a randomly chosen internal branch, leaving the remainder of the tree unchanged. We used a tuning parameter value of $\lambda = 0.2$ in order to propose new branch lengths ℓ^* within a maximum of about 10% of the initial branch length ℓ according to $\ell^* = \ell \exp(\lambda(U - 0.5))$, where U is random, uniform in $(0, 1)$. Users can therefore choose between three tree proposals: a subtree pruning and regrafting (SPR), a nearest neighbor interchange (NNI) (both already present in Rasmussen and Kellis, 2011), and the new SubtreeSlides sampler. At each iteration, two proposals are performed, with hill climbing optimization. The first proposes a new tree with a possibly new topology, using one of the SPR, NNI or SubtreeSlides operation. The second proposal evaluates changes in branch lengths. With probability 0.8, we use the branch length optimization of the branch likelihood term proposed by Rasmussen and Kellis (2011, section Rapid Tree Search). With probability 0.2, small random changes are proposed to the length of b branches, where b is the total number of branches in the gene tree and branches are chosen at random. A new length is proposed within 10% of the current length using tuning parameter $\lambda = 0.2$ as above.

References

Larget B and Simon D. 1999. Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol Biol Evol* **16**: 750–759.

Rasmussen M and Kellis M. 2011. A bayesian approach for fast and accurate gene tree reconstruction. *Mol Biol Evol* **28**: 273–290.