Supplementary material for the article

“***TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees***”.

Marina Marcet-Houben and Toni Gabaldón.

Performance of treeKO in simulated phylomes

To evaluate the performance of treeKO under controlled conditions we ran evolutionary simulations using the sequence YER147C as a seed (this is the only seed in the T12a phylome whose gene tree is fully congruent to the species tree). This sequence was made to evolve along the species tree using Rose (1). Mutation frequencies used were computed with TreePuzzle (2), assuming a 16 rate gamma distribution. For each position in the alignment we took the category and associated relative rate that contributed the most to the likelihood. The remaining parameters for the simulation were the ones described in (3).

Four different sets of 1000 simulations each were ran with the following conditions: A) probability of insertions and deletions (indel) was set to 0.0007 and no gene duplications and losses were allowed; B) as A, but using an indel probability of 0.003 and the mutation frequencies were multiplied by two; C) gene duplication and losses were simulated by making the sequences evolve on a simulated gene tree (see below). Indel probabilities were set to 0.0007. D) as C, but using an indel probability of 0.003 and the mutation frequencies were multiplied by two. Additionally, two datasets of 1000 randomly chosen trees from the T12a phylome were chosen for comparison.

Simulated trees in C and D were generated by ETE (4), as follows: Each simulated gene tree started at the root of the canonical species tree, and then it was made evolve along it until the last species diverged. At any point either a speciation or a duplication event occurred, at rates that were determined by lineage-specific ratios inferred from the observed frequencies in the T12a phylome. When a duplication happened at a given point in the evolution, then all the species that had not been speciated (or lost) yet were affected by the duplication.

Simulated sequences were then aligned and a tree was reconstructed using the same approach as the one used in the phylome (see main text). Thus each set of 1000 simulations constitutes a “simulated” phylome, to which alternative topologies can be evaluated as described in the main text. In all cases tested, the distances to the known species topology used to perform the simulations were lower than those to alternative topologies. Some examples are shown in supplementary table 2 (below).

Supplementary table legends:

Supplementary table 1.- Results for the t-test when comparing the distance distribution of the original T12a topology (5) and each one of the six alternative topologies that result from swapping pairs of branches of the post-whole genome duplication species. Distance distributions were calculated under six different conditions as explained in the table. Bold numbers represent distance distributions of alternative topologies that are not significantly different (p-value > 0,05) to the distribution of distances of each alternative topology to the original topology.

Supplementary table 2.- Results for the t-test comparing sets of 1000 simulated trees. Simulations were run under four different conditions. Two of the conditions contained trees without duplications, they each had a different indel probability (0,0007 and 0,003) and in the second case the mutation frequency was multiplied by two. The two other simulated sets contained trees with duplications and were generated under the same indel probability and mutation frequency as the previous sets. Additionally, two sets of 1000 trees randomly selected from the phylome were also included to avoid any sample size effect in the comparison. Bold numbers represent distance distributions of alternative topologies that are not significantly different to the distribution of distances of each alternative topology to the original topology.

Supplementary figures legends:

Supplementary figure 1.- Set of topologies obtained by swapping branches of the post-Whole genome duplication species.

Supplementary figure 2.- Relationship between the logarithms of the number of pruned trees calculated by treeKO (x-axis) and by TOPD (y-axis). Trees were taken from the T12a phylome (5). Note the change of scale between the x-axis and the y-axis showing that the number of trees used in TOPD is several times larger than the set predicted by treeKO.

Supplementary figure 3.- Relationship between the number of duplications (x-axis) found in a tree and the resulting number of pruned trees (y-axis). All the trees in the T12a phylome (5) was used for this analysis. Data seems to follow a linear distribution that can be represented by the equation: number_duplications = number_pruned_trees + 1. This indicated that most duplications in this phylome are nested.

Literature cited

1. Stoye, J., Evers, D., and Meyer, F. (1998). Rose: generating sequence families. Bioinformatics , 14 (2), 157–163.

2. Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics , 18 (3), 502–504.

3. Talavera, G. and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol , 56 (4), 564–577.

4. Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010). Ete: a python environment for tree exploration. BMC Bioinformatics , 11 , 24.

5. Marcet-Houben, M. and Gabaldón, T. (2009). The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. PLoS One , 4 (2), e4357.

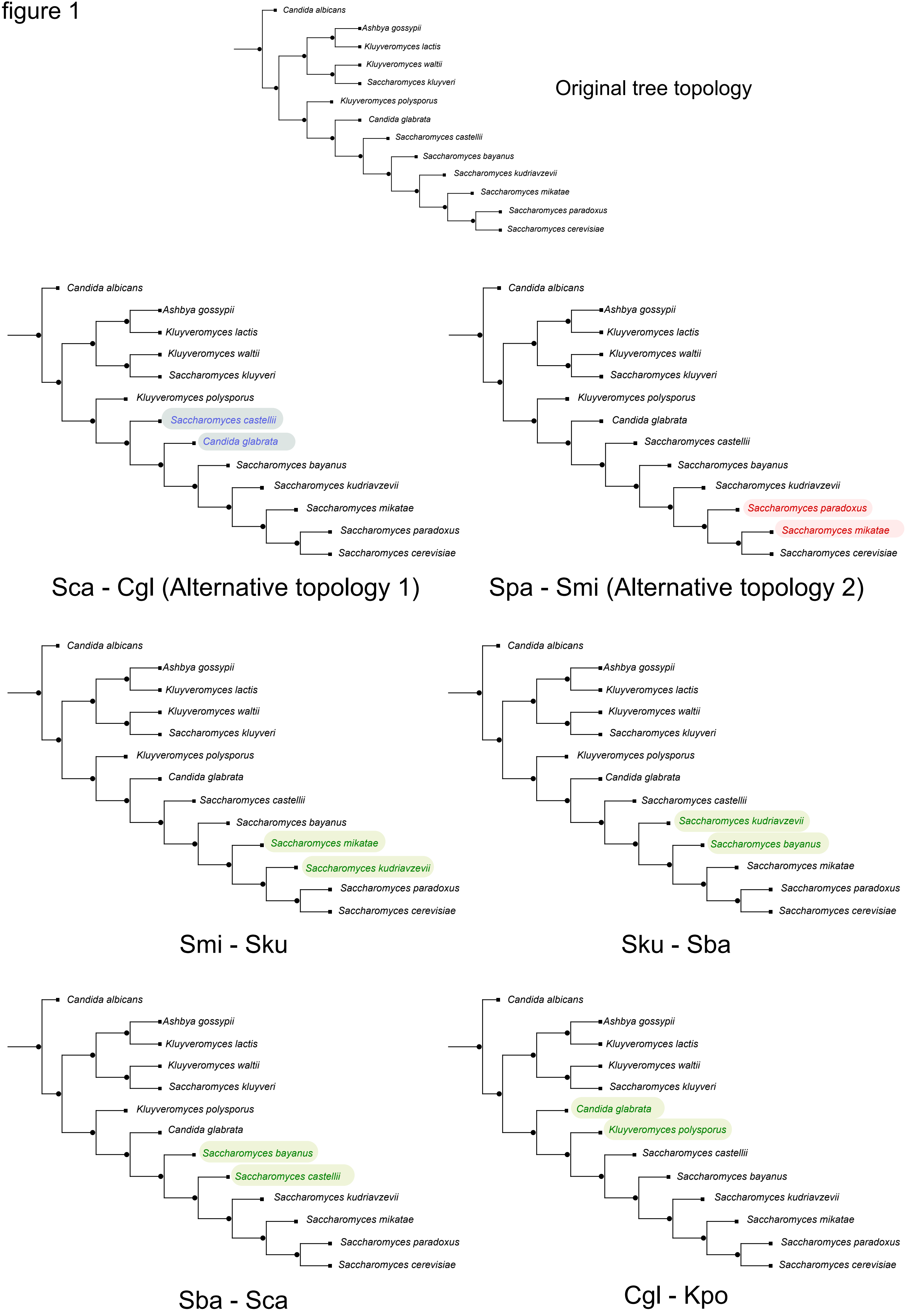
Supplementary table 1

Conditions	Alternative topologies					
	Spa – Smi	Smi – Sku	Sku – Sba	Sba – Sca	Sca – Cgl	Cgl – Kpo
Speciation distance +Midpoint rooting	2,20E-016	2,20E-016	2,20E-016	2,20E-016	0,0955	1,17E-005
Speciation distance +minimal duplications rooting	2,20E-016	2,20E-016	5,77E-016	2,20E-016	0,08554	0,001328
Speciation distance +Midpoint rooting +collapse branches with a support <0.5	2,20E-016	2,20E-016	2,20E-016	2,20E-016	0,236	6,82E-005
Reconciliation distance (duplication distance)	0,01538	0,1521	0,06113	0,1197	0,9353	0,4085
Reconciliation distance (duplication +loss distance)	0,00217	0,0367	0,01652	4,34E-005	0,8736	0,1916

Supplementary table 2

Conditions			Alternative topologies	
			Spa – Smi	Sca – Cgl
Speciation distance	Without duplications	indel probability: 0,0007	2,20E-016	2,20E-016
		indel probability: 0,003; mutation frequency x 2	2,20E-016	2,20E-016
	With duplications	indel probability: 0,0007	2,20E-016	2,20E-016
		indel probability: 0,003; mutation frequency x 2	2,20E-016	2,20E-016
	Phylome data 1 (1000 random trees)		0,468	2,20E-016
	Phylome data 2 (1000 random trees)		0,7558	6,71E-014
Reconciliation (duplications)	Without duplications	indel probability: 0,0007	2,20E-016	2,20E-016
		indel probability: 0,003; mutation frequency x 2	2,20E-016	2,20E-016
	With duplications	indel probability: 0,0007	0,0288	6,60E-012
		indel probability: 0,003; mutation frequency x 2	1,06E-010	2,20E-016
	Phylome data 1 (1000 random trees)		1	0,28
	Phylome data 2 (1000 random trees)		0,979	0,342
	Without duplications	indel probability: 0,0007	2,20E-016	2,20E-016
		indel probability: 0,003; mutation frequency x 2	2,20E-016	2,20E-016
	With duplications	indel probability: 0,0007	0,00836	9,86E-011
		indel probability: 0,003; mutation frequency x 2	1,58E-009	2,20E-016
	Phylome data 1 (1000 random trees)		0,987	0,179
	Phylome data 2 (1000 random trees)		0,956	0,216

Supplementary figure 1



Original tree topology

Sca - Cgl (Alternative topology 1)

Spa - Smi (Alternative topology 2)

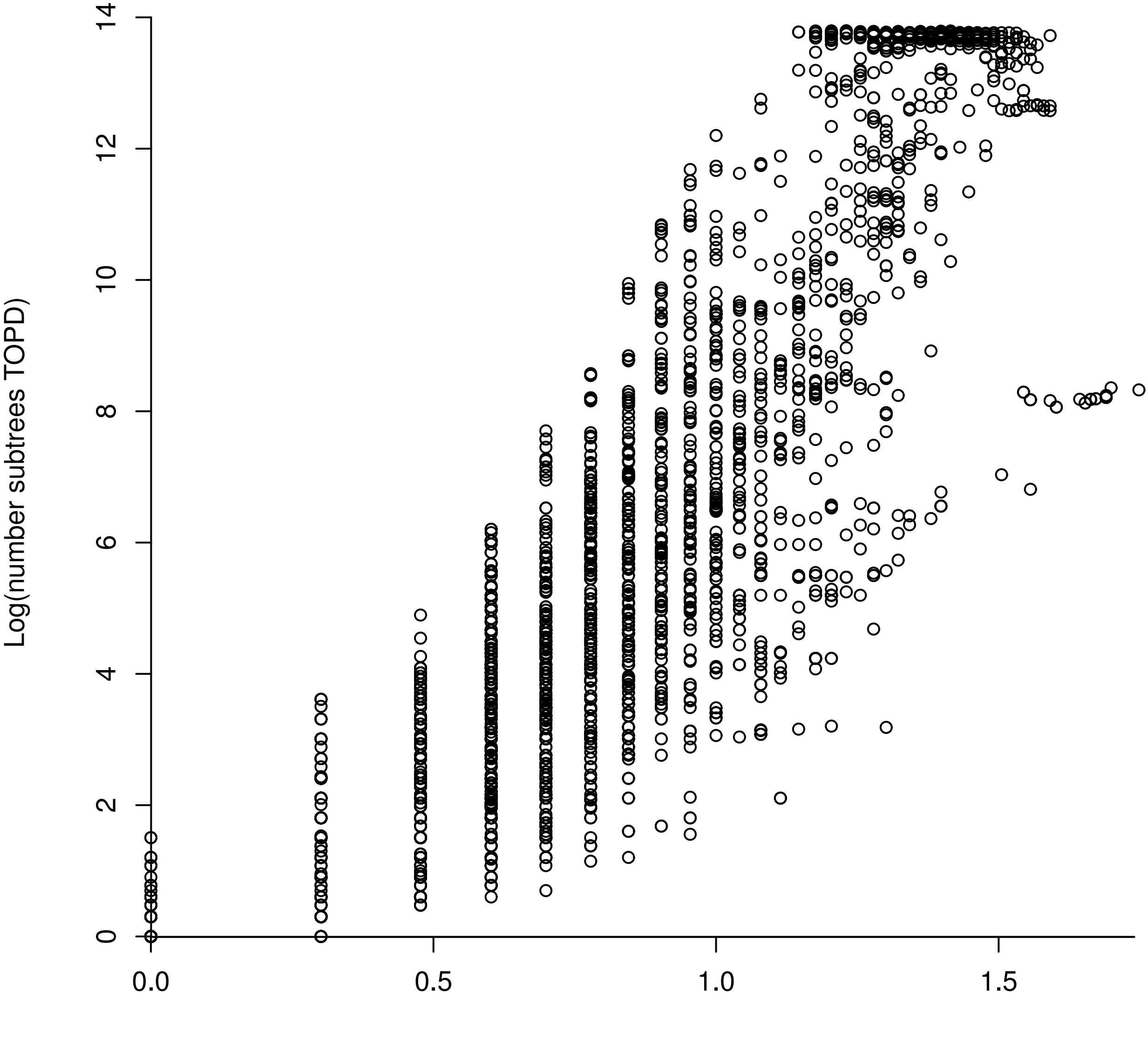
Smi - Sku

Sku - Sba

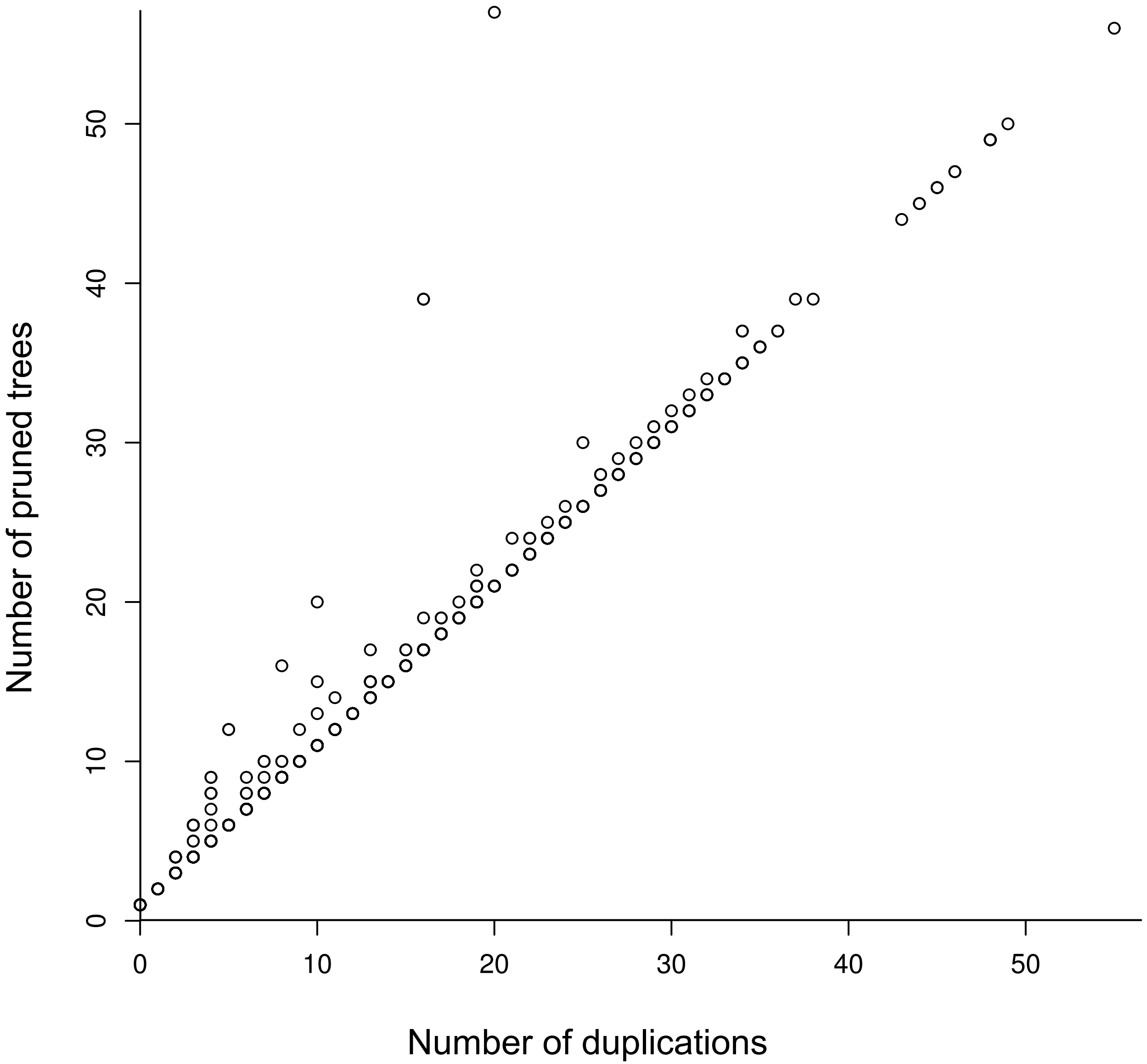
Sba - Sca

Cgl - Kpo

Supplementary figure 2



Supplementary figure 3


Number of pruned trees
Number of duplications