Annex 1: Gene tree reconciliations: detailed procedure

1. The reconciliation of gene and species tree requires both trees to be rooted and highly depends on where the root are placed. Usual approaches for rooting based on balancing the tree topology (e.g. using midpoint rooting) can be inappropriate in case of certain lineages evolving at different rates. In our context, we favoured rooting criteria that already consider the gene tree together with the species tree, in order to minimize the impact of the rooting on the subsequently inferred scenario, i.e. to find a root that is parsimonious in terms of implied DTL events. This was achieved by minimizing a score that combines two criteria: the first one aims at maximizing the size of subtrees with unicopy sequences while the second one tries to maximize the accuracy of taxonomic affectations for nodes (Bigot et al. 2013). Because TPMS roots poorly trees in absence of duplication and presence of many transfers, we used alternative method in unicopy gene trees: we searched for horizontal tranfers using Prunier to then re-root them consistently to the parsimonious transfer scenario.

2. Nodes with unicity conflict, i.e. where a species is represented in both children nodes, were listed in each tree, using the 'Unicity' algorithm from TPMS (Bigot et al. 2013). To decide if this unicity conflict originated from duplication or rather additive transfer events, we computed the duplication-induced loss rate (DLR) as the ratio of the number of loss necessary to complete a duplication scenario to the number of ancestral species nodes below the ancestor to which the duplication would map. Manual examination of trees led to choose an arbitrary threshold of DLR = 0.2 above which the duplication hypothesis seems too costly and the additive transfer hypothesis is preferred. Use of an alternate criterion was tried with the duplication consistency score (DC score) (Vilella et al. 2009), but showed poorer discrimination of duplication or transfer-like patterns, certainly because it was first implemented to assess duplication confidence in a vertebrate genome evolution model not considering transfer events (Vilella et al. 2009). Note that this decision is not definitive, as further detection of transfer events can lead to resolution of a certain amount of unicity conflict, and therefore, abolition of upper duplication events. Unicopy subsets of leaves were sampled in the full gene tree so that they gather the maximal set of species without putting together paralogs, i.e. gathering only orthologs or co-orthologs (sensu (Kristensen et al. 2011)). All possible combination of orthologs were explored, with each leaf being potentially represented in several leaf combination sets. The corresponding subtrees (same topology as the full gene tree, but restricted to an unicopy leaf set) where then searched for transfer events

3. Transfer were searched using Prunier software (version 2.0, fast algorithm, forward search depth = 2, branch support threshold = 0.9) that detects supported topological incongruence with the reference tree of species (Abby et al. 2010).

4. Unicopy subtrees were mapped back on the full gene tree, associating each node of the unicopy subtree to one or several collapsed nodes of the full gene tree, and transfer events found by Prunier were annotated on a node of the full gene tree. Doing so with every unicopy subtree, a count is made by node of the number of times a node of the full gene tree is tested for transfer (i.e. covered by an unicopy subtree submitted to Prunier ananlysis, Prunier coverage) and of the number of times is is detected as a particular transfer scenario. The latter count gives a support for each

transfer scenario. The total Prunier coverage minus the sum of supports for all proposed transfer scenario gives the support for an absence of transfer at this node (support for a speciation or duplication, depending on the further completion of the reconciliation).

5. To choose between several events proposed at he same node, we used two criteria: first, inclusion in the largest block event, and second, best coverage by Prunier replicate tests.

(a) At this step, a preliminary search for block of co-transferred genes was performed (see section 5). Several conflicting transfer scenarios could be proposed at the same node, especially when a sparse sample of species in the tested unicopy subtree gave poor context to orientate the transfer. This often results in inferring a scenario of transfer from species A to species B and another scenario of transfer from B to A. Gene families locally co-evolve (for instance during ponctual events of co-transfer), but the remaining majority of their history should be independent, giving potentially different patterns of losses in respective unicopy subtrees of neighbouring genes. Although it can be hard to decide on a scenario of a single transfer event, a series of neighbouring genes with compatible scenarios gives a better confidence in the shared scenario. Therefore, when several events are considered, the one participating in the longest block event is chosen.

(b) If no choice can be done considering block events (a majority of transfer events involve only one gene), the event supported by the most replicates of Prunier is retained. This can notably retain an absence of transfer, when a majority of Prunier tests did not infer any transfer event at the node (often by explaining the potential phylogenetic conflict by transfers in another part of the tree).

6. Transferred subtrees were pruned from the full gene tree to yield a forest of subtrees free of transfer events. Search of unicity conflict was performed again (cf. step 2) on each tree of this forest. Nodes bearing unicity conflict were explored in a post-order traversal. Under each conflicting node, groups of leaves representing monophyletic group of species and present in multiple copies were searched for phylogenetic incongruence with the species tree using the algorithm for detection of taxonomic incongruence from (Bigot et al. 2013). When taxonomic incongruence was found, the source of the conflict was identified by iteratively testing to prune a clade within the subtree and evaluating how much conflict was removed. Once the ‘alien’ clade was identified, a transfer was inferred and the transferred subtree was pruned, potentially resolving the unicity conflict. After each transfer detection, search of unicity conflict was redone on the pruned tree. When this iterative search reached the root of the tree and no more transfer could be inferred, remaining uncity conflict was explained by annotating duplications at conflicting nodes. On one hand, this procedure allowed us to avoid over-annotating duplications in cases of likely events of additive transfers, where the topological incongruence was not enough supported to be previously detected by Prunier. On the other hand, applying this transfer detection method, which disregards gene tree branch supports, as stand-alone HGT detection procedure over the whole gene tree could be hazardous, given the possible errors in the estimation of the gene tree topology; applying it only in case of unicity conflict is thus more conservative, as it considers the signature of additive transfers.

7. Ancestral content inference was done using asymmetric Wagner parsimony using Count software (Csűrös 2008)

(a) The gene trees were first atomised into a forest of orthologous subtrees by systematically pruning the recipient subtree under a transfer or duplication event (choosing arbitrarily the duplication 'recipient' as the smaller subtree). This was done by exploring the node events in a post-order traversal. Each leaf (gene) set of these subtrees were considered as orthologous subfamilies for which the gain/loss history and ancestral presence states were estimated independently.

(b) Xenologous replacement consist of the entry of a transferred gene in the genome followed by coincidental loss of the resident gene. This can be instantaneous in case of gene conversion by homologous recombination. This kind of transfer does not result in a new gene copy and if it was given by a close relative, the replacing gene is expected to be quite similar in function to the native gene it was substituted to. Therefore, in order to keep subfamilies as (functionally) coherent groups of leaves rather than true orthologous groups, subtrees containing only speciations and *transfers that do not disturb the monophyly of the represented species* were kept as unique subfamilies punctuated of allelic replacements.

(c) Having previously detected transfer and duplication events, inferring the ancestral states and events should be straightforward, by inferring a gain at the last common ancestor (LCA) of the represented species, and completing the history with losses (Dollo parsimony). However, subfamilies with patchy distribution of genes in the species tree suggests conspicuous transfer events (not in topological conflict with the species tree) rather than gain at an ancient LCA followed by several losses. Asymmetric Wagner parsimony corrects for this bias by inferring additional transfers that are annotated on the gene tree. Since these transfers do not induce phylogenetic or unicity conflict, the genes in the subfamily are still considered orthologous, though they do not match exactly the definition of a group of leaves only related by speciation events. Again, this allow to keep subfamilies as coherent clusters of genes with probably similar functions – a goal for which orthology is often a proxy.

(d) Then, gain/loss histories of all subfamilies are integrated to the global family history. It implies to correct the ancestor to which some gain have been mapped. Indeed, the ancestor in which happened a duplication is the LCA of the merged set of species represented in paralogous children subtrees. This global species set can be larger than each individual orthologous species set, because of independent parallel losses. The ancestor of each orthologous may then be inferred lower in the species tree that it should, what is corrected.

(e) Global integration of gain/loss/duplication/transfer reception/emission/allelic replacement history was done across all families to describe the history of the whole genomes.

References:

Abby SS, Tannier E, Gouy M, Daubin V. 2010. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. BMC Bioinformatics. 11:324–324. doi: 10.1186/1471-2105-11-324.

Bigot T, Daubin V, Lassalle F, Perrière G. 2013. TPMS: a set of utilities for querying collections of gene trees. BMC Bioinformatics. 14:109. doi: 10.1186/1471-2105-14-109.

Csűrös M. 2008. Ancestral Reconstruction by Asymmetric Wagner Parsimony over Continuous Characters and Squared Parsimony over Distributions. In: Comparative Genomics. Nelson, CE & Vialette, S, editors. Lecture Notes in Computer Science Springer Berlin Heidelberg pp. 72–86. http://link.springer.com/chapter/10.1007/978-3-540-87989-3_6 (Accessed March 7, 2013).

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. 2011. Computational Methods for Gene Orthology Inference. Brief. Bioinform. 12:379–391. doi: 10.1093/bib/bbr030.

Vilella AJ et al. 2009. EnsemblCompara GeneTrees: Complete, Duplication-Aware Phylogenetic Trees in Vertebrates. Genome Res. 19:327–335. doi: 10.1101/gr.073585.107.