Ancestral genome organization as a diagnosis tool for phylogenomics

Eric Tannier, Adelme Bazin, Adrian Davin, Laurent Guéguen, Sèverine Bérard, ***Cedric Chauve***

INRIA and Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive
Laboratory of Bioinformatics Analyses for Genomics and Metabolism, CEA
RIKEN Center for Advanced Intelligence Project (AIP)
Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive
ISEM, Université de Montpellier, CNRS, IRD, EPHE
Department of Mathematics, Simon Fraser University

Alphy 2019

Overview

Background: a phylogenomics pipeline to reconstruct ancestral gene orders

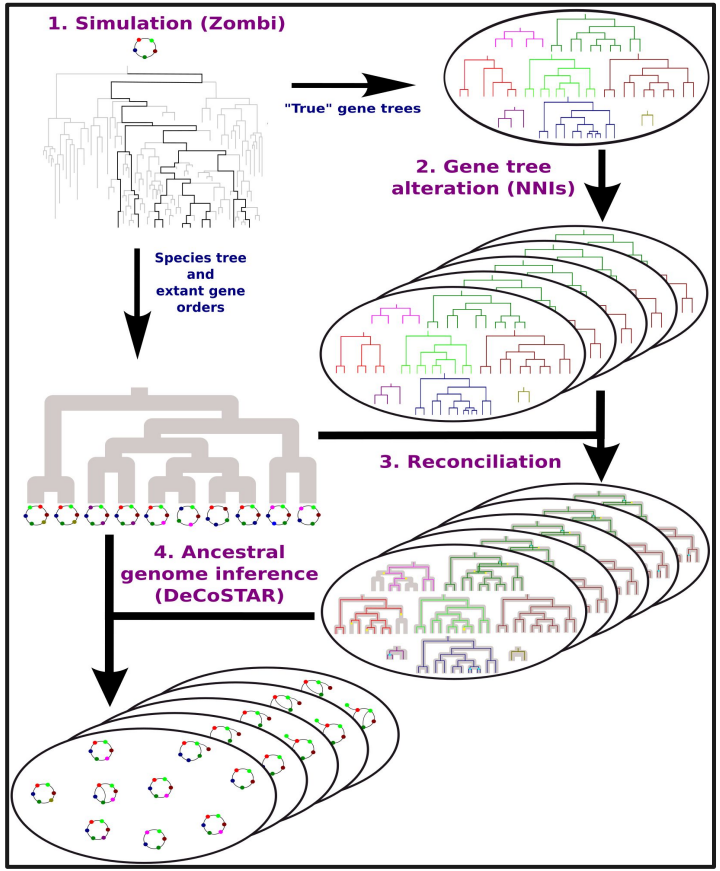
- **Input:** Assembled genomes, with annotated genes, a given species tree
- **Step 1:** Cluster genes into **gene families** (OrthoMCL, …)
- **Step 2:** Compute/sample **gene trees** for each gene family (MrBayes, RAxML, IQ-TREE, …)
- **Step 3:** **Reconcile gene trees** with the species tree (ecceTERA, ALE, …)
- **Step 4:** Reconstruct **ancestral gene adjacencies** (DeCoStar, …) …
- **Step 5**: Clear **syntenic conflicts** among inferred ancestral gene adjacencies

Syntenic conflict: an ancestral gene belongs to more than two inferred ancestral gene adjacencies. This can not happen in a valid gene order.

Question: Is-there some useful signal in the syntenic conflicts ?

Hypothesis: The extent of syntenic conflict is correlated to the extent of noise/errors in the reconstructed gene trees: syntenic conflicts can help to detect errors in gene trees.

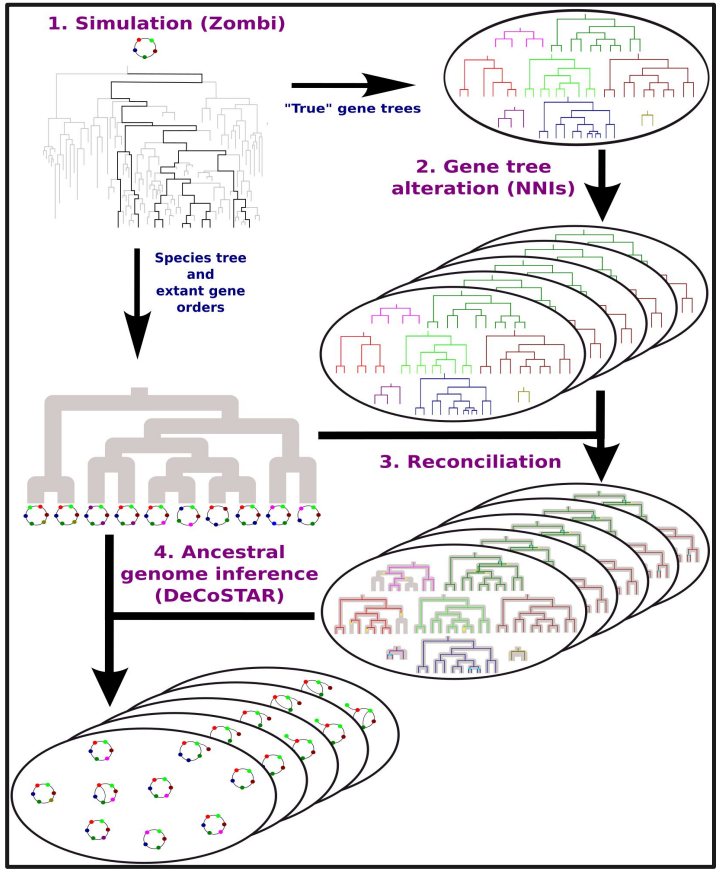
Methods: generating simulated noisy gene trees



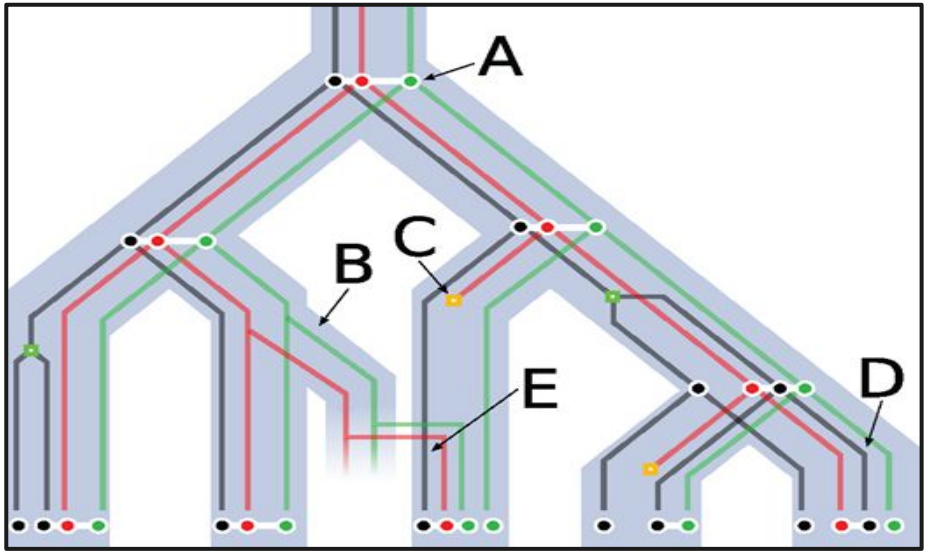
1. **Zombi (Davin *et al*, 2018):**
   Species tree (**ST,** 26 sampled species out of 151).
   Root genome: circular genome of 1000 genes, with no in-paralog.
   Gene families (**GF**) / gene tree (**GT**) evolution:
   Dataset 1: Duplication-Loss (**DL**);
   Dataset 2: Duplication-Loss-Transfer (**DLT**).
   Gene order evolution: random genome rearrangements (inversions, …)

2. **Introducing errors in the true GTs:**
   Random Nearest-neighbour Interchange **(NNI)**, whose number is chosen at random following a Poisson distribution of parameter **λ**.

Methods: ancestral adjacencies

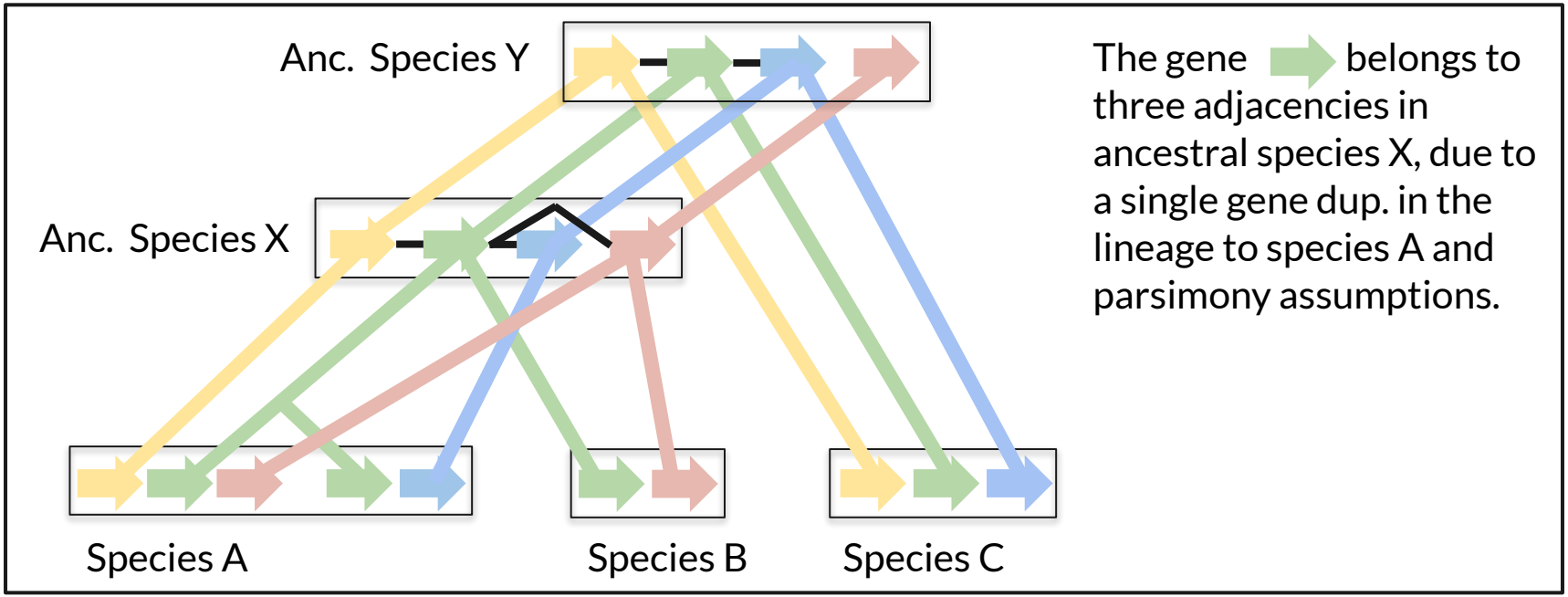


3. **ecceTERA (Jacox *et al*, 2016):** Parsimonious GT/ST reconciliation.
4. **DeCoSTAR (Duchemin *et al*, 2017):** Parsimonious reconstruction of ancestral gene adjacencies, **one at a time.**



Syntenic conflict

Syntenic conflict: An ancestral gene belongs to more than two gene adjacencies. This can be observed in pipelines where ancestral gene adjacencies are inferred individually, independently of other potential ancestral gene adjacencies

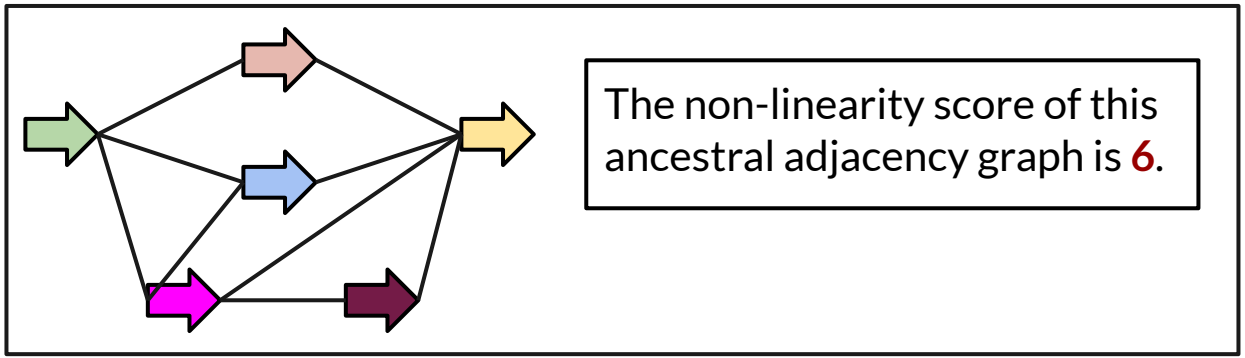


Alphy 2019

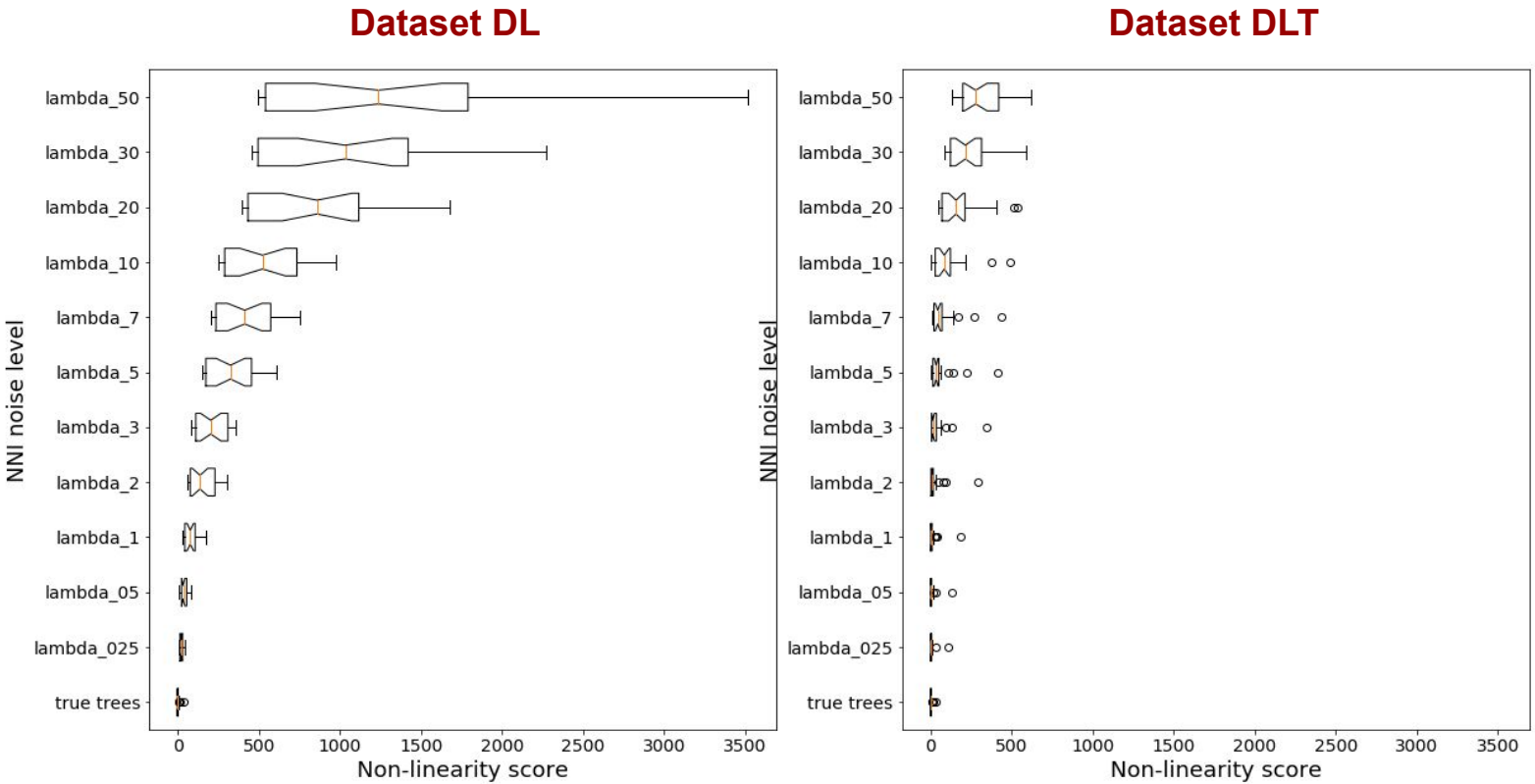
Methods: measuring the extent of syntenic conflict

For each ancestral species, the gene adjacencies define an **ancestral adjacency graph** where vertices represent ancestral genes and edges represent ancestral gene adjacencies.
To measure the extent of syntenic conflict we use the **non-linearity score**, defined for each ancestral species as:

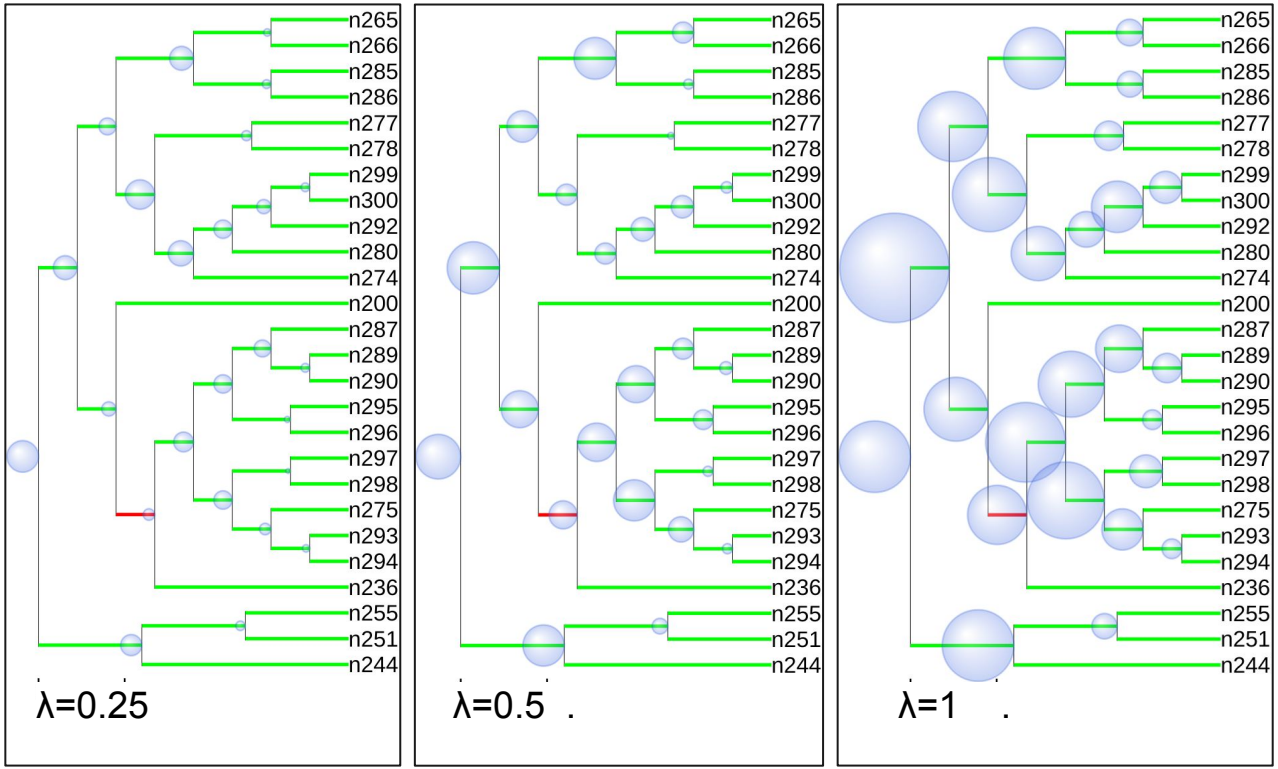
Sum over all vertices (genes) *g* of |*deg(g)-2*|



The non-linearity score correlates with RF distance



Results: Dataset DL

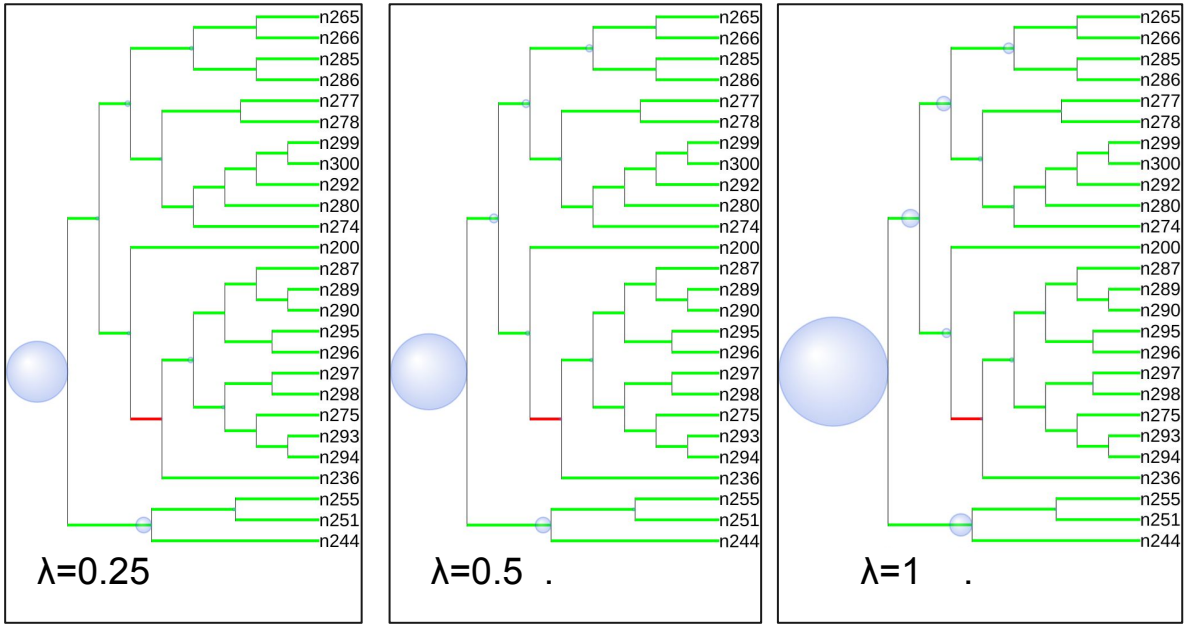


We observe a trend that the non-linearity score increases with the noise in GTs.

Generally, the level of syntenic conflict increases in more ancient species.

Even with a low level of noise, syntenic conflict occurs in recent ancestral species (cherries).

Results: Dataset DLT

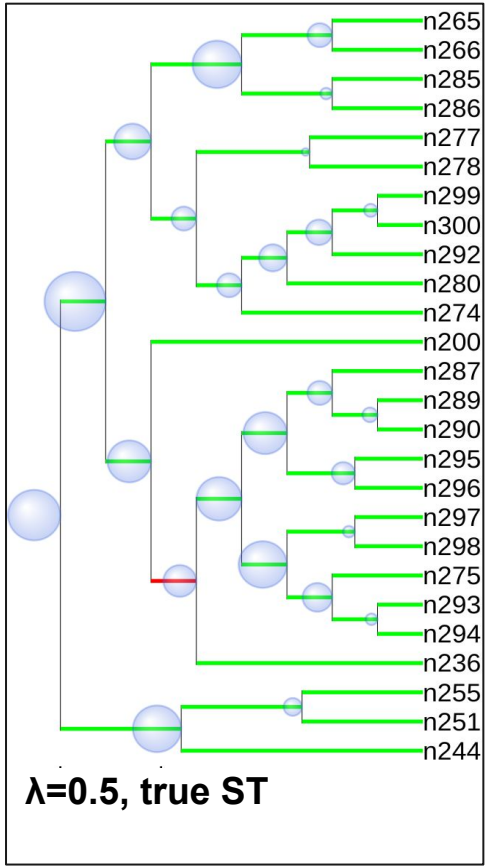


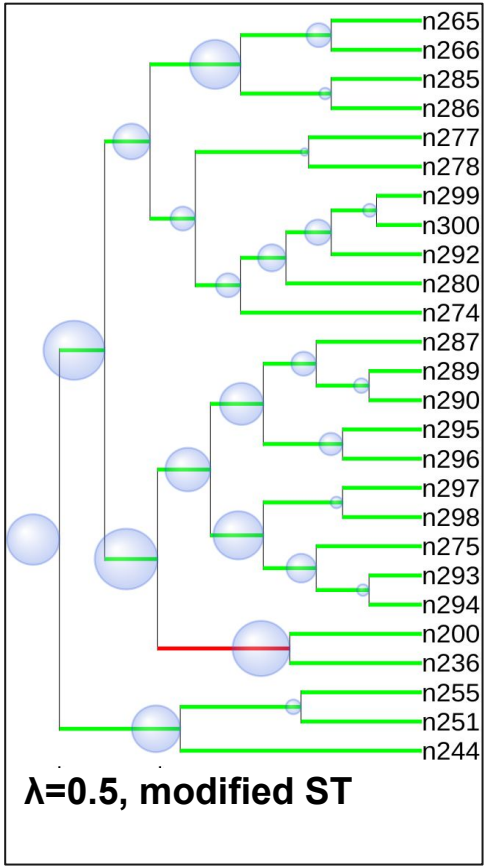
The trend of a non-linearity score increasing with the noise in GTs is present too, but it is much milder, but at the level of the root species.

Results: Adding noise in the species tree

Experiment:
We modify ST by doing a single NNI (red branch).

Observation (Dataset DL):
The non-linearity score increases around the branch impacted by the NNI, suggesting the non-linearity score can also capture the signal of errors in the ST.





Conclusion

With our (limited) simulation-based experiments, we observe a clear trend that the non-linearity score correlates with the extent of noise in the gene trees. This is a first step toward a proof-of-principle of the following:

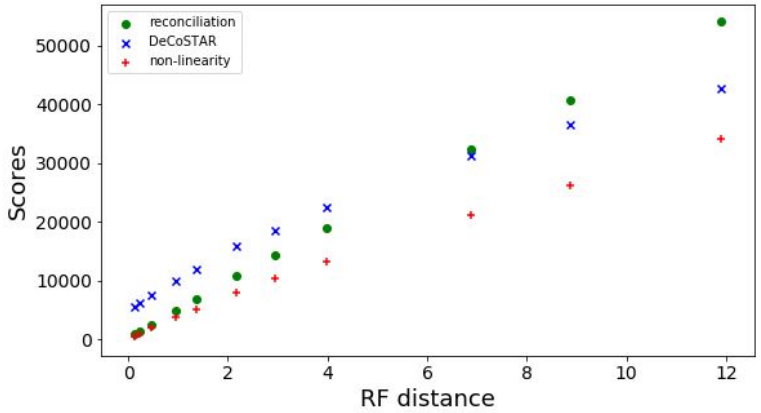
The gene order inconsistencies observed after reconstructing (parsimonious) ancestral gene adjacencies should not be considered only as noise but also as a potentially useful signal to guide the correction of gene trees (and species trees?), complementing extant synteny.

More experiments are needed, covering a wider range of biological conditions and noise sources (assembly, gene families, alignment, ILS, …).

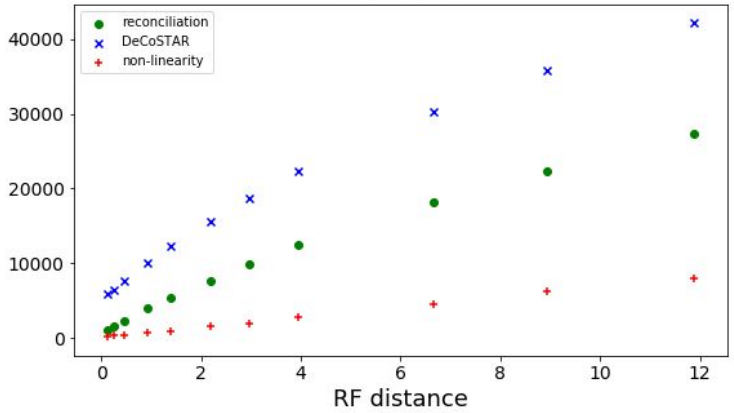
The effect of noise in the gene trees is very different in the two datasets and this suggests there is a need to better understand the difference due to considering HGT in the reconciliation model.

Results: Looking at other intermediate scores

Dataset DL



Dataset DLT



All three scores (reconciliation, DeCoStar, non-linearity) correlate well with the **RF** distance to the true trees.
This is consistent with (Hahn 2007) for the reconciliation score.
The DeCoStar score (parsimony score in terms of gains and breaks for gene adjacency evolutionary histories) follows the same trend.

References

Zombi: A simulator of species, genes and genomes that accounts for extinct lineages. (2018) Adrian A. Davin, Theo Tricou, Eric Tannier, Damien M. de Vienne, Gergely J. Szollosi. *Biorxiv* 339473.

ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. (2016) Edwin Jacox, Cedric Chauve, Gergely J Szollosi, Yann Ponty, Celine Scornavacca. *Bioinformatics* 32(13):2056-2058.

DeCoSTAR: Reconstructing the Ancestral Organization of Genes or Genomes Using Reconciled Phylogenies. (2017) Wandrille Duchemin, Yoann Anselmetti, Murray Patterson, Yann Ponty, Sèverine Bérard, Cedric Chauve, Celine Scornavacca, Vincent Daubin, Eric Tannier. *Genome Biology and Evolution* 9(5):1312-1319.

Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. (2007) Matthew Hahn. *Genome Biology* 8(7):R141.