Ancestral Genome Organization as a Diagnosis Tool for Phylogenomics

Eric Tannier, Adelme Bazin, Adrián Davín, Laurent Guéguen, Sèverine Bérard, Cedric Chauve

▶ **To cite this version:**

Eric Tannier, Adelme Bazin, Adrián Davín, Laurent Guéguen, Sèverine Bérard, et al.. Ancestral Genome Organization as a Diagnosis Tool for Phylogenomics. Scornavacca, Celine; Delsuc, Frédéric; Galtier, Nicolas. Phylogenetics in the Genomic Era, No commercial publisher | Authors open access book, pp.2.5:1–2.5:19, 2020. hal-02535466

HAL Id: hal-02535466
https://hal.science/hal-02535466

Submitted on 10 Apr 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Chapter 2.5 Ancestral Genome Organization as a Diagnosis Tool for Phylogenomics

Eric Tannier
INRIA and Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558,
[Villeurbanne, France]
eric.tannier@inria.fr
https://orcid.org/0000-0002-3681-7536

Adelme Bazin
Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay
[91057 Evry, France]
adelme.bazin@genoscope.cns.fr

Adrián A. Davín
RIKEN Center for Advanced Intelligence Project (AIP)
[36-1 Yoshida Honmachi, Sakyo-ku, Kyoto, Japan]
adrian.arellanodavin@riken.jp
https://orcid.org/0000-0003-4945-4938

Laurent Guéguen
Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558
[Villeurbanne, France]
laurent.gueguen@univ-lyon1.fr

Sèverine Bérard
ISEM, Université de Montpellier, CNRS, IRD, EPHE,
[Montpellier, France]
severine.berard@univ-montellier.fr

Cédric Chauve[1]
Department of Mathematics, Simon Fraser University,
[8888 University Drive, Burnaby (BC), V5A 1S6, Canada]
LaBRI, Université de Bordeaux
[351 Cours de la Libération, 33405 Talence Cedex, France]
cedric.chauve@sfu.ca
https://orcid.org/0000-0001-9837-1878

Abstract

The reconstruction of the chromosomal organization of ancient genomes has many applications in comparative and evolutionary genomics. Here we propose a novel, methodological, use for these predicted ancestral syntenies, directly focused on phylogenomics. It is a way to assess the accuracy of gene trees and species trees. We use a method that reconstructs, from gene trees and extant gene orders, ancestral adjacencies, i.e. the immediate neighborhood between pairs of genes, independently for each pair. This independence allows to split the computations into many independent problems that can each be solved exactly using efficient algorithms, but might result in sets of ancestral adjacencies that are incompatible with the expected linear or circular structure of chromosomes. We show here that this drawback can actually be turned into a useful

[1] This work benefited from the support of ComputeCanada.

© Eric Tannier, Adelme Bazin, Adrián Davín, Laurent Guéguen, Sèverine Bérard and Cédric Chauve.
Licensed under Creative Commons License CC-BY-NC-ND 4.0.
Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 2.5; pp. 2.5:1–2.5:19
A book completely handled by researchers.
No publisher has been paid.

feature. We show on simulated data that the degree of linearity of the reconstructed ancestral gene orders is well correlated to the accuracy of the input gene trees. Moreover, a localized error in the species trees results in a burst of non linearity of ancestral genomes at the wrong node. We eventually show that integrated phylogenomic methods expectedly lead to better linearity scores than methods based on gene alignments only. Allowing a method to output an unrealistic result, but proving that the expected output is closer to realistic when the input is closer to correct, we thus provide an original validation protocol for standard evolutionary studies.

How to cite: Eric Tannier, Adelme Bazin, Adrián A. Davín, Laurent Guéguen, Sèverine Bérard, and Cédric Chauve (2020). Ancestral Genome Organization as a Diagnosis Tool for Phylogenomics. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 2.5, pp. 2.5:2–2.5:19. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Supplement Material `https://github.com/sberard/SAGe`

1 Introduction

Rearrangements of gene organization along chromosomes were discovered long before the molecular structure of DNA (Sturtevant, 1921). The comparison of genetic maps or polytene chromosomes were seen in the first half of the XXth century as a promising approach to reconstruct evolutionary relationships or ancestral configurations (Babcock and Navashin, 1930; Dobzhansky and Sturtevant, 1938), and advance the knowledge on extant and extinct biodiversity. This later motivated the development of genetics, cytogenetics, or bioinformatics techniques aimed at detecting "structural" mutations of chromosomes (Timoshevskiy et al., 2013). The definition of what is a structural mutation largely depends on the observation technique used to detect them. It can consider every mutation involving several contiguous nucleotides (starting with small indels, micro-satellites or micro-inversions), or be limited to very large-scale mutations that affect gene content and orders, such as large inversions or chromosomal translocations (that can be detected by genetics or cytogenetics techniques). As genes are often taken as evolutionary units by bioinformatics and phylogenomics methods, genome rearrangements are usually limited to structural mutations whose breakpoints are located in non-coding regions and that change the organisation of genes along the genome.

Rearrangements have an important role in several evolutionary processes such as adaptation, speciation, sex differentiation, polyploidization (Fuller et al., 2018; Lemaitre et al., 2009). Knowing ancestral configurations can thus inform on conserved structures, functional gene clusters (Abrouk et al., 2010), as well as on patterns and processes of the history of wild or domestic biodiversity (Murat et al., 2012).

1.1 Ancestral gene order reconstruction methods

The reconstruction of ancient genome organization has been called *paleogenomics*, a term shared with ancient genome sequencing (Pont et al., 2019).

Over the last 25 years, there has been an intense research activity in developing computational methods for the reconstruction of ancestral gene orders, that we reviewed extensively by Anselmetti et al. (2018b). We distinguish two main families of methods: chromosome based methods and adjacency based methods.

Chromosome based methods take as input a species phylogeny, the gene orders of extant species in this phylogeny, and a genome rearrangement evolutionary model. Their aim is to infer ancestral gene orders and evolutionary scenarios along the branches of the species phylogeny. Ancestral orders and scenarios can be given a score (parsimony, likelihood) according to the model, and methods can optimize or sample according to this score.

Such methods are natural extensions to gene orders of ancestral sequence reconstruction methods, reviewed for example by Groussin et al. (2016); Joy et al. (2016). However, unlike ancestral sequence reconstruction, ancestral gene order reconstruction is computationally intractable for almost all genome rearrangement models. Indeed, even if gene scale evolution events as duplication and loss are ignored, and if there are only 3 species in the species tree, the parsimony problem is NP-complete (see Tannier et al., 2009; Kovác, 2014, and references there). If duplications are allowed, even the problem of computing the pairwise distance between two gene orders is NP-complete (e.g. see Blin et al., 2007; Angibaud et al., 2009).

To skirt this computational challenge, adjacency based methods model the evolution of the physical link between two consecutive genes only, called *adjacencies*, instead of full chromosomes. In this framework, ancestral adjacencies are reconstructed along the species phylogeny from the pattern of presence/absence of extant adjacencies, using a model allowing gains and losses of adjacencies. The set of inferred ancestral adjacencies for a specific ancestral species then forms an *adjacency graph* whose vertices are the ancestral genes and edges the ancestral gene adjacencies. A side effect of inferring ancestral adjacencies using such an approach that considers the evolution of each adjacency independently from the others is that the adjacency graph may not have the expected structure of a chromosome, that is, a collection of paths and cycles. In order to present a structure compatible with a set of chromosomes, or at least scaffolds, some methods select a subset of the inferred ancestral adjacencies which form a collection of paths and/or cycles. Computationally more tractable, adjacency based approaches can handle unequal gene content and gene duplication, gain and loss, and several methods have been developed that allow ancestral gene orders to contain duplicated genes (Ma et al., 2008; Rajaraman and Ma, 2016; Zhou et al., 2017).

Here, we consider again such methods, but from another point of view: we make the hypothesis that syntenic conflicts might be caused by errors in the earlier steps of the whole pipeline, especially the construction of the reconciled gene trees (see Chapter 3.2 [Boussau and Scornavacca 2020]). This hypothesis has been considered in several studies (Boussau et al., 2013; Peres and Crollius, 2015; Duchemin et al., 2017; Anselmetti et al., 2018a; Zerbino et al., 2018), but never assessed through experiments. In this paper, we provide a first proof of principle that indeed syntenic conflict in reconstructed ancestral gene orders can be correlated to errors in earlier steps of a phylogenomics pipeline.

1.2 Impact of errors on the linearity of reconstructed genomes

As described above, ancestral gene orders are typically obtained at the end of a multi-step sequential phylogenomics pipeline starting with genome assemblies and leading to the inference of ancestral gene adjacencies, which link consecutive genes in ancestral chromosomes, and ultimately ancestral gene orders. Intermediate steps include gene annotation (Chapter 4.1 [Necsulea 2020]), gene clustering into gene families (Chapter 2.4 [Fernández et al. 2020]), multiple sequence alignment of genes within families (Chapter 2.2 [Ranwez and Chantret 2020]), gene tree and species tree reconstruction (Chapters 1.2 and 1.4 [Stamatakis and Kozlov 2020; Lartillot 2020]) and gene tree reconciliation (Chapter 3.2 [Boussau and Scornavacca 2020]). Each step in such pipelines is susceptible to introduce errors that can propagate further in the pipeline. For example, the effect of errors in multiple alignments for species

tree reconstructions has been explored by Philippe et al. (2017), and the effect or errors in gene tree for reconciliations has been investigated by Hahn (2007). The effect of model choice has recently been investigated by Hoff et al. (2016); Yang and Zhu (2018), as well as the effect of the phylogenetic software choice (Zhou et al., 2018). And even bugs in many standard software can blur the results (Czech et al., 2017). Along these lines of research, we propose to investigate the effect of errors in gene trees and species trees on the *linearity* (or more precisely non-linearity) observed in ancestral gene adjacencies, and propose to use the latter to correct phylogenetic trees.

The notion of linearity is related to the arrangement of genes along chromosomes as defined by ancestral gene adjacencies. In this work we assume that a genome is composed of a set of linear and/or circular molecules carrying genes – chromosomes, organelles, plasmids, . . . – with genes *totally ordered* along each molecule. This implies that a correct adjacency graph, representing an actual gene order, whether it is extant or ancestral, is a collection of paths and/or cycles. This is an approximation as it is common in extant genomes that genes overlap, or are included one in an other, or that the definition of their limits lack precision due to alternative splicing for example. Nevertheless, a vast majority of genes in cellular organisms can be totally ordered along chromosomes. Given this assumption, the hypothesis we investigate is the following: *the amount of non-linearity observed in ancestral adjacency graphs is correlated to the level of errors made by earlier steps of the pipeline leading to these graphs, in particular to the amount of errors in gene trees.* We are not claiming that these errors are the only possible source of non linearity. Indeed our ancestral adjacency reconstruction method can make mistakes itself; in particular it is a parsimony method, as such unable to cope with convergent or reverse evolution. However we expect that, if we are provided with real species tree, gene families and gene trees and if gene order has evolved without much convergent evolution, then reconstructing ancestral adjacencies should result in few false positive adjacencies and the resulting ancestral adjacency graphs should be close to linear, *i.e.* most ancestral genes should have at most two neighbors.

The idea of a correlation between the extent of non linearity in ancestral adjacency graphs and the distance to an ideal situation was first introduced by Bérard et al. (2012) to compare two sets of gene trees, and has been used in several works to compare gene trees (Boussau et al., 2013; Patterson et al., 2013; Peres and Crollius, 2015; Duchemin et al., 2017; Anselmetti et al., 2018a) or species trees (Anselmetti et al., 2018a). However, there is so far no systematic study providing a proof of principle. In particular the hypothesis that a better linearity implies that the input gene trees are more accurate has never been assessed. This is what we propose to do. We use simulations of species tree, gene trees, gene sequences and gene orders in two situations, one where gene families evolve by speciation, duplication and loss and one where gene families evolve by speciation, duplication, loss and horizontal gene transfer (HGT). We then perturb the gene trees and species tree in order to measure the effect of the introduced errors on the linearity of ancestral adjacency graphs.

In our first set of experiments, we observe a very strong correlation between the amount of noise introduced in the gene trees and, on one hand, the number of inferred structural mutations of genomes, and on the other hand, the non-linearity of the ancestral adjacency graphs. This tends to confirm our hypothesis and suggests that predicted ancestral genomes could be used to assess the quality of phylogenomics data and could thus provide an important signal to correct gene trees. Indeed, while predicted ancestral genomic features, such as gene content, can always be explained by – possibly highly non-parsimonious and unrealistic – evolutionary scenarios, the non linearity of gene order can not be justified in any way. So we provide an additional, original, quality measure. In a second set of experiments, we observe

that with moderately perturbed gene trees, a local error in the species tree correlates with a burst of non linearity precisely in the ancestral genomes close to the erroneous branch. This burst is very localized and could be used to give a hint on erroneous parts of species phylogenies. Finally, in a third set of experiments, we reproduce gene tree construction pipelines starting from sequence data; our results suggest, based on the linearity score, that integrated phylogenomics methods, including gene tree species tree reconciliation, lead to more accurate results than gene tree reconstruction methods based only on multiple alignments.

2 Methods

Our experiments are based on the analysis of simulated data, providing a clear ground truth on the evolution of a set of gene orders. We first describe these simulations, then the analyses performed on the simulated data.

2.1 Simulations

We used Zombi (Davín et al., 2019) to perform simulations. This program constructs artificial species tree, gene trees evolving along this species tree, extant and ancestral gene orders evolving through genome rearrangements, and gene DNA sequences. Genomes evolve by duplication of one or several genes, losses, horizontal gene transfer, and inversions of segments of several genes. Duplications are tandem or not, according to a parameter, and transfers either replace a homologous gene or consists of an insertion at a random place in the genome. Zombi is interesting for our purpose because it mixes gene based events and genome based events. Moreover it is the only available software which is able to take into account extinct or unsampled species when performing HGTs.

The set of parameters is fully available in the supplementary material of this paper. We simulated one species tree with 151 leaves, 26 of them being extant species, the others being extinct or unsampled species. The ancestral gene order at the root of the tree is composed of a single circular chromosome of $1,000$ genes, with no in-paralogs. From there we simulated two datasets:

- Dataset 1: gene families evolved through speciation, gene duplication and gene loss;
- Dataset 2: gene families evolved through a more comprehensive model including speciation, gene duplication, gene loss and horizontal gene transfer.

For each dataset, we obtained from Zombi the true gene tree for each gene family, together with the gene orders of all extant and ancestral species and the DNA sequence of all genes.

2.2 Correlating non-linearity of adjacency graphs and errors in gene trees

In this first experiment, we introduced errors in gene trees and species tree and measured how this impacts a non-linearity score recorded in the adjacency graphs of the ancestral species. An overview of the whole process is depicted on Figure 1.

2.2.1 Introducing errors in gene trees

For each dataset, we introduced various levels of errors in the true gene trees by applying random local perturbations using Nearest Neighbor Interchanges (NNI) on gene tree branches uniformly at random. The level of noise was controlled by the number of NNI performed,

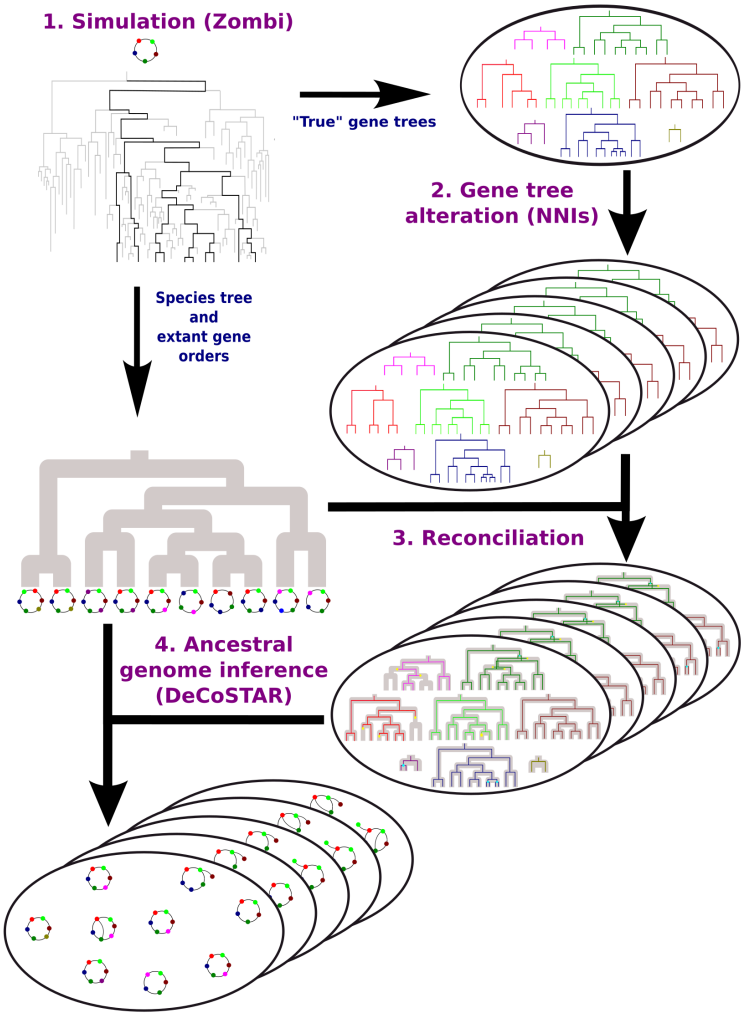

Figure 1 Overview of the simulation/perturbation/reconstruction process and dependencies. We use Zombi to simulate species trees, gene trees, gene orders and gene sequences. We apply some perturbations to gene trees. Then we use DeCoStar to reconcile gene trees (it uses the ecceTERA package) and to construct ancestral adjacencies.

chosen from a Poisson distribution with parameter $\lambda \in \{0.25, 0.5, 1, 2, 3, 5, 7, 10, 20, 30, 50\}$. It follows that we obtained 12 sets of gene trees (the true trees and 11 sets of perturbed trees) for each starting dataset.

For each set of perturbed gene trees we recorded the mean Robinson-Foulds (RF) distance to the true trees. Figure 2 plots the RF distance growing with λ, showing that in this parameter range, there is no saturation of gene tree perturbation, but that the RF distance, more than λ itself, can capture the amount of distortion.

2.2.2 Introducing noise in the species tree

We also looked at the impact of errors in the species tree. To do so, we perturbed the species tree by manually performing a single NNI at an arbitrary branch; we denote by S the true species tree and S_1 the perturbed species tree. Both trees were tested with true and perturbed gene trees.

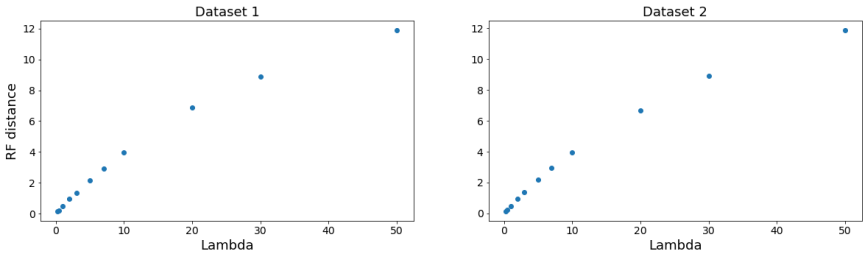

Figure 2 Mean Robinson-Foulds distance as a function of the value of λ.

2.2.3 Reconstructing ancestral gene adjacencies with DeCoStar

DeCoStar (Duchemin et al., 2017) takes as input extant gene orders, gene trees and a species tree. Gene trees are reconciled with the provided species tree using ecceTERA (Jacox et al., 2016), an exact dynamic programming algorithm computing a parsimonious reconciliation. Then ancestral adjacencies are reconstructed for each ancestral species. The principle for this reconstruction is that first extant adjacencies are clustered into families, according to the homology of the corresponding gene extremities. Then for each family of adjacencies, ancestral adjacencies are constructed also with an exact dynamic programming procedure minimizing the cost of gains and breakages of adjacencies; we used default costs for adjacency gain and break (3 and 1).

We added to the program DeCoStar a novel feature, described here for the first time, aimed at handling gene losses without artificially increasing the parsimony cost of an adjacency evolutionary scenario, which is important regarding the linearity score we describe later. This feature consists in iterating the DeCoStar program several times, modifying the costs of creating adjacencies in function of the previous iteration. More precisely, if in the solution computed by the algorithm at iteration i the loss of a gene A, located between two genes B and C, is inferred, then at iteration $i+1$ the gain of an adjacency between genes B and C is free, *i.e.* it does not increase the cost of the evolutionary scenario for the adjacency family containing B and C. This is generalized to any set of consecutive genes located between B and C, being lost concomitantly at iteration i. It can have a significant impact on the linearity score in the case of convergent losses of genes. As we focus on linearity we use this "loss aware" option with two iterations ($i = 2$) in all our experiments.

For each run of DeCoStar, we recorded the number of gene duplications, gene losses and HGTs, as well as the number of gains and breaks of adjacencies.

2.2.4 Non-linearity score

Each run of DeCoStar results in a set of ancestral gene adjacencies. Under the hypothesis that perfect data and a moderate amount of non parsimonious structural evolution will result in linear ancestral genomes, we expect that each gene is the extremity of exactly two adjacencies[2]. In other words it has *degree* 2, where the degree, noted $\deg(g)$ for a gene g, is the number of adjacencies using g as an extremity. Thus we define the *non-linearity score* as the distance from this expectation. For a given ancestral species, the non-linearity score is

[2] This is true for circular chromosomes, and in particular for our current experiments with Zombi. In general chromosomes may be linear and the genes at their extremities are expected to be involved in only one adjacency. However their number is so low compared to a standard gene set that this expectation can be used in practice as well for linear chromosomes

the sum of $|\deg(g) - 2|$ over all vertices g of its ancestral adjacency graph. The non-linearity score for a given experiment is then the sum of the non-linearity scores over all ancestral species.

2.3 Reconstructing gene trees

In our third experiment, for both datasets we reconstructed gene trees for all gene families from the simulated gene sequences, using IQ-TREE (Nguyen et al., 2015) with bootstrap supports on all branches. For Dataset 1 we corrected the IQ-TREE trees with Treerecs (Comte et al., 2020). For Dataset 2 we additionally constructed a sample of gene trees from the sequences using MrBayes (Ronquist et al., 2012) and used the amalgamation option of ecceTERA (Jacox et al., 2016) to construct a single reconciled gene tree from the MrBayes sample, that is, one reconciled gene tree per gene family.

The rationale behind these choices of methodologies is that for gene families evolving under the duplication/loss model (Dataset 1) there are fast methods to obtain reconciled gene trees from IQ-TREE trees, able to correct branch supports, such as Treerecs. For gene families evolving also with HGTs (Dataset 2), where the same problem is NP-complete, we used the amalgamation principle, in a reconciliation framework considering HGTs, which requires to compute a sample of gene trees from a posterior probability.

3 Results

First, we discuss in details our main observation that the non-linearity score is highly correlated with the level of noise introduced in the gene trees. In a second set of experiments, we consider an erroneous species tree and we show that again the non-linearity score increases around the erroneous branch of the species tree, suggesting it could be used to point at species phylogeny errors. Last we consider gene trees reconstructed from sequence data, instead of true gene trees perturbed by random NNIs, and show that reconstruction methods accounting for gene evolution events perform better in terms of non-linearity scores than traditional phylogenetics methods.

3.1 The distance to true trees is highly correlated with the non-linearity score

3.1.1 Overview

Our first result is illustrated in Figure 3: in the two datasets, the non-linearity score grows almost linearly with the mean RF distance between the perturbed gene trees and true trees.

Figure 3 actually represents three scores: the reconciliation score (cost of gene family evolutionary events: gene duplications, gene losses, HGTs), the DeCoStar score (cost of adjacency gains and breakages) and the non-linearity score (see Section 2). We present these three scores, despite the fact that our main interest is in the linearity score, in order to give a broader picture of the impact of noise in the input data on the result of phylogenomic algorithms. In particular, it is interesting to observe that the three scores grow almost linearly with the mean RF between the perturbed gene trees and the true gene trees.

3.1.2 Gene content

An interesting observation is that the reconciliation score grows much faster in Dataset 1 than in Dataset 2. As reconciliation defines the gene content of ancestral species, and it was shown

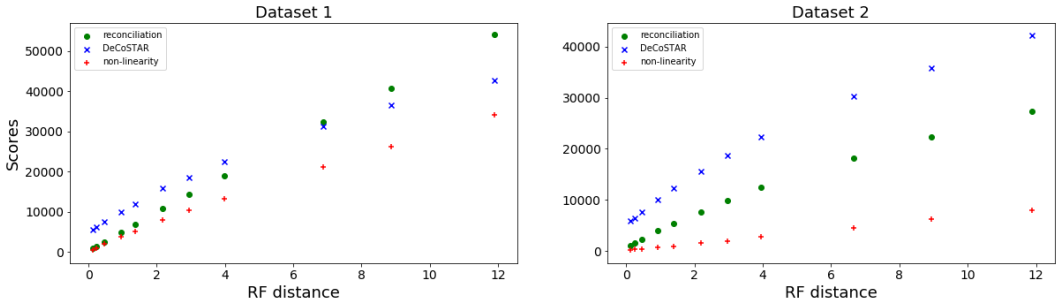

Figure 3 Non-linearity score (red +), DeCoStar score (blue x) and reconciliation score (green circle) as a function of the mean Robinson-Foulds distance between the perturbed gene trees and the true gene trees.

by Hahn (2007) that in a duplication/loss model, errors in gene trees result in an unrealistic gene content of ancestral species, especially for higher nodes of the species phylogeny, we were interested in recording the gene content of ancestral species in both datasets (Figure 4). A somewhat surprising observation is that the patterns of observed gene content deduced from the reconciliations are very different: in Dataset 1, as expected, more ancient ancestral species accumulate genes due to how the parsimonious reconciliation algorithm copes with errors in gene trees, while in Dataset 2, the converse happens, as ancestral species closer to the root of the species phylogeny have less genes, although the variation is less strong than in Dataset 1.

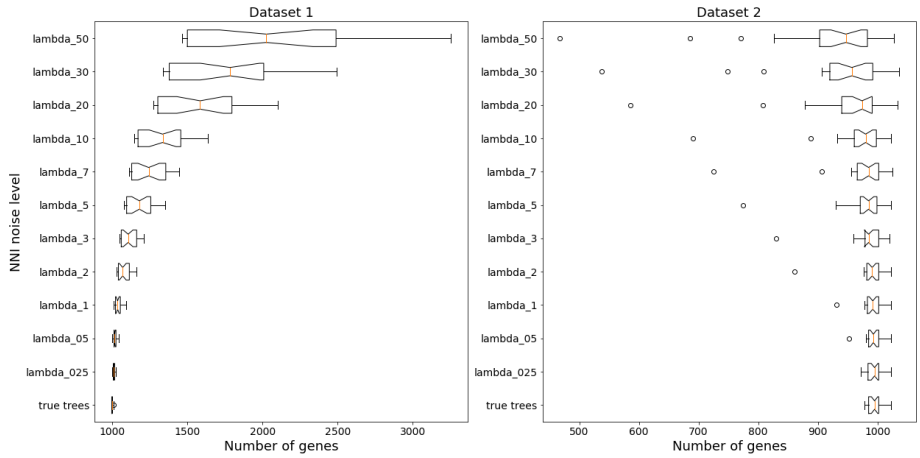

Figure 4 Gene content of ancestral species, in function of the degree of perturbation in gene trees. For Dataset 1 (Left) and Dataset 2 (Right). We see an opposite behavior of DL models (Left) and DTL models (right) with respect to gene content. On one side gene content increases with perturbation, on the other it decreases. In both cases gene content is altered proportionally to the amount of perturbation.

3.1.3 Non-linearity score

We now refine the analysis of the non-linearity score by considering the non-linearity score specific to each internal node of the species tree. We first focus on Dataset 1 and, for the sake

of clarity, we consider only the rates of errors in gene trees of $\lambda \in \{0.25, 0.5, 1\}$ (Figure 5), as they illustrate well the general trend observed for all levels of noise.

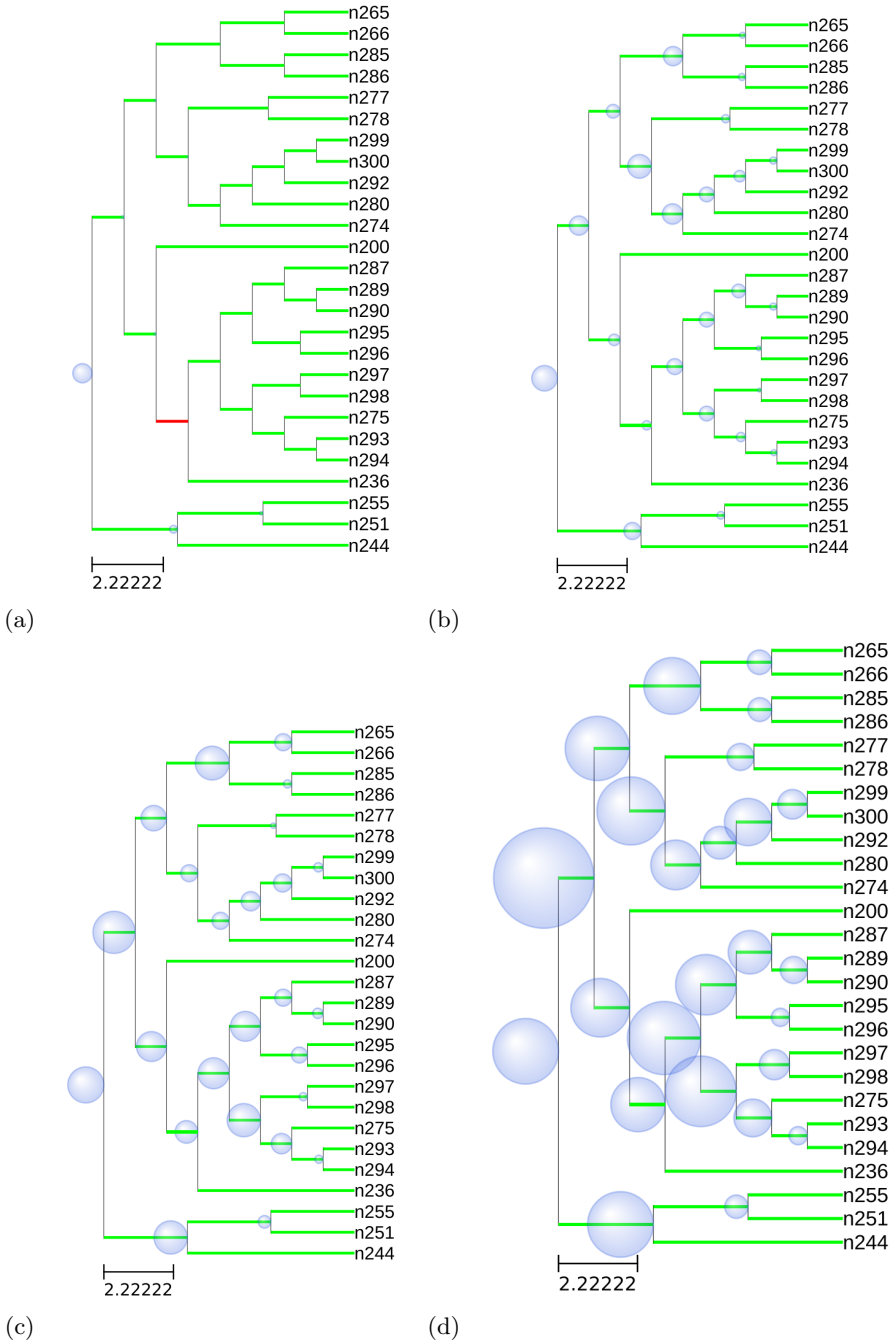

Figure 5 Non-linearity score for Dataset 1 with the true species tree S, with the true gene trees (a), and with λ having value 0.25 (b), 0.5 (c) and 1 (d). The radius of the disks at the internal nodes are proportional to the non-linearity scores. Red branches are the ones that are perturbed (see next section).

The main observation from Figure 5 is that there is a general trend that the non-linearity score increases toward higher nodes of the species tree. It is also interesting to notice that

even with a low level of noise, some lower internal node, such as the roots of cherries (subtrees composed of two leaves) show a non-zero non-linearity score. This suggests that few errors in gene trees are sufficient to create conflicting ancestral adjacencies.

When considering Dataset 2, the effect is rather different, with the root node capturing most of the non-linearity score (Figure 6).

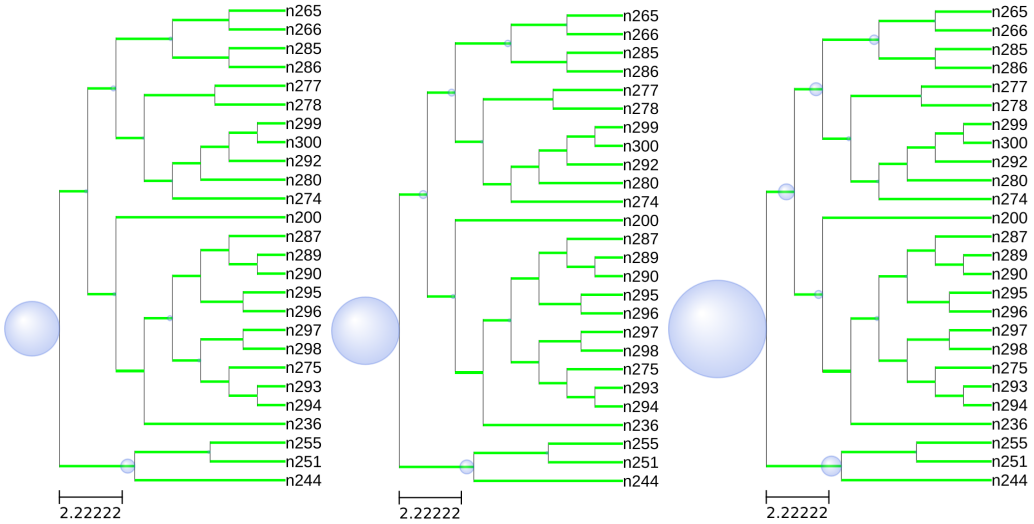

Figure 6 Non-linearity score for Dataset 2, with λ having value 0.25 (Left), 0.5 (Middle) and 1 (Right). The radius of the disks at the internal nodes are proportional to the non-linearity scores. The red branches should be ignored.

Figure 7 below provides another illustration of the difference in terms of non-linearity score variation, as we can observe a much lower magnitude of the score in Dataset 2, as well as a lesser variation compared to Dataset 1.

3.1.4 Discussion

Regarding the interpretation of our observations, an important element is the applicability toward correcting erroneous gene trees. The work described by Hahn (2007) was already a first result toward our hypothesis that scores of phylogenomic algorithms can be correlated to errors in data. Our experiments allow us to go one step further. Indeed, while largely inflated ancestral genomes can be highly unrealistic, one can always consider that there is a non-zero probability that they are correct. A similar remark could apply to the DeCoStar algorithm, that considers individual adjacencies outside of their wider genomic context: adjacency evolutionary scenarios involving a high number of gains and/or breaks could be seen as unrealistic, but not impossible. On the contrary, under the assumptions we outlined in Introduction, a non-zero non-linearity score without false positives ancestral adjacencies is impossible, as genes are linearly or circularly arranged along chromosomes. So if methods are developed with the aim to correct gene trees guided by the reduction of some score, the non-linearity score is a good candidate since its ideal value is known – the closer to zero, the better the gene trees.

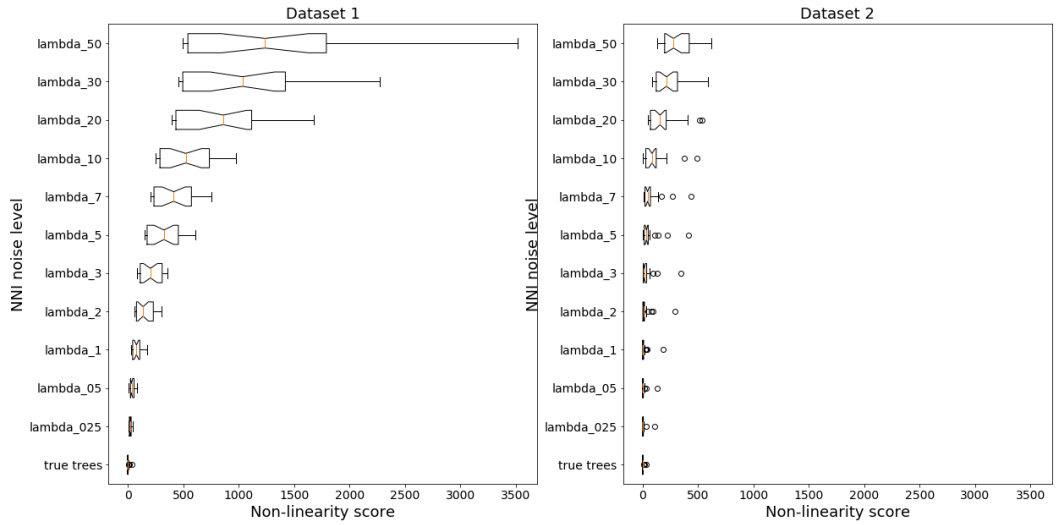

Figure 7 Distribution of the non-linearity score of ancestral species, as a function of the perturbation on gene trees. On the left pannel, for the experiment with only duplications and losses, and on the right pannel, for duplications, transfers and losses. The two pannels have the same scale, in order to illustrate the effect of transfers, in the presence of which the perturbations have a lower effect.

3.2 Non-linearity point at erroneous branches of the species trees

When considering the species tree S_1 differing from the true species tree by a single NNI, the results we obtained in terms of the correlation of the scores of the different steps of our pipeline (reconciliation, DeCoStar, non-linearity) with the level of noise in the gene trees were similar to the ones described above (Figure 8).

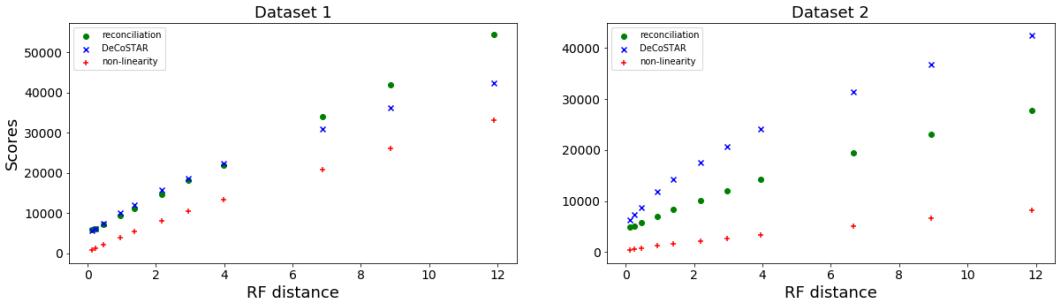

Figure 8 Non-linearity score (red +), DeCoStar score (blue x) and reconciliation score (green circle) as a function of mean Robinson-Foulds distance of perturbed gene trees to true gene trees, with a perturbed species tree.

Moreover, similar to the phenomenon observed when using the true gene trees for Dataset 1, we can observe in Figure 9 that the non-linearity score is greatly inflated around the branch where the NNI was performed, compared to the neighbouring nodes, especially with lower levels of errors in gene trees ($\lambda \in \{0.25, 0.5\}$). This suggests that the non-linearity score can also capture an important signal regarding the accuracy of the species tree.

Next, for Dataset 1 and the true gene trees, we can make two interesting observations by

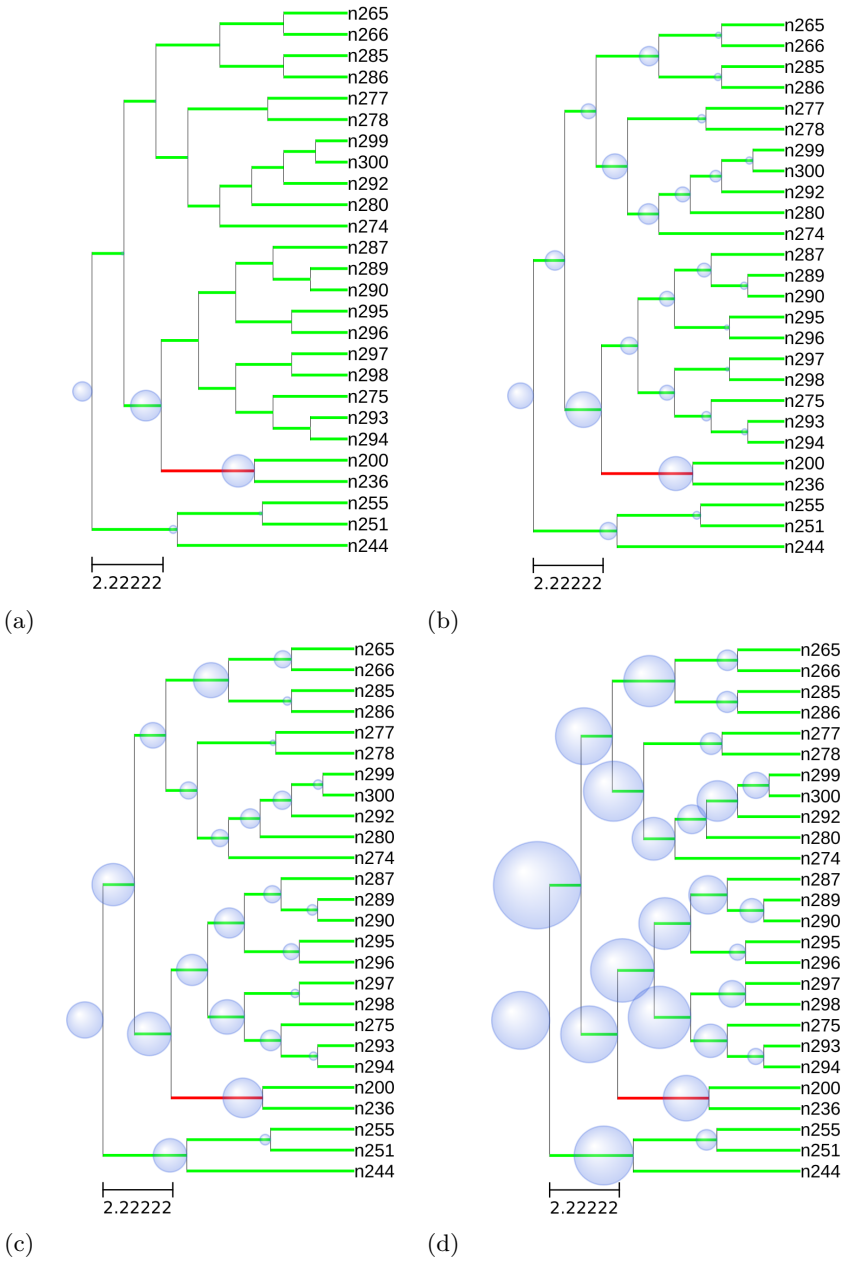

Figure 9 Non-linearity score for Dataset 1 and the species tree S_1, with the true gene trees (a) and with λ having value 0.25 (b), 0.5 (c) and 1 (d). The radius of the disks at the internal nodes are proportional to the non-linearity scores. The branch that has undergone the NNI is shown in red.

comparing the level of non-linearity at each node of the true species tree S, Figure 5(a), and the modified one S_1 (one NNI away), Figure 9(a). First, our assumption that with perfect data from the ground truth (gene families, gene trees, species tree), ancestral adjacency graphs are almost linear, is confirmed. Second, when considering the experiment with the species tree S_1, we can observe a much higher level of conflict, at the branch where the NNI was done, and at its parent.

Last, on Figure 10 we present the same information for the dataset where gene trees

evolved with HGT (Dataset 2). We can observe a similar trend of a level of conflict increasing in higher nodes in the species tree and a strong impact of the NNI performed onto S to obtain S_1, in terms of conflict around the NNI branch.

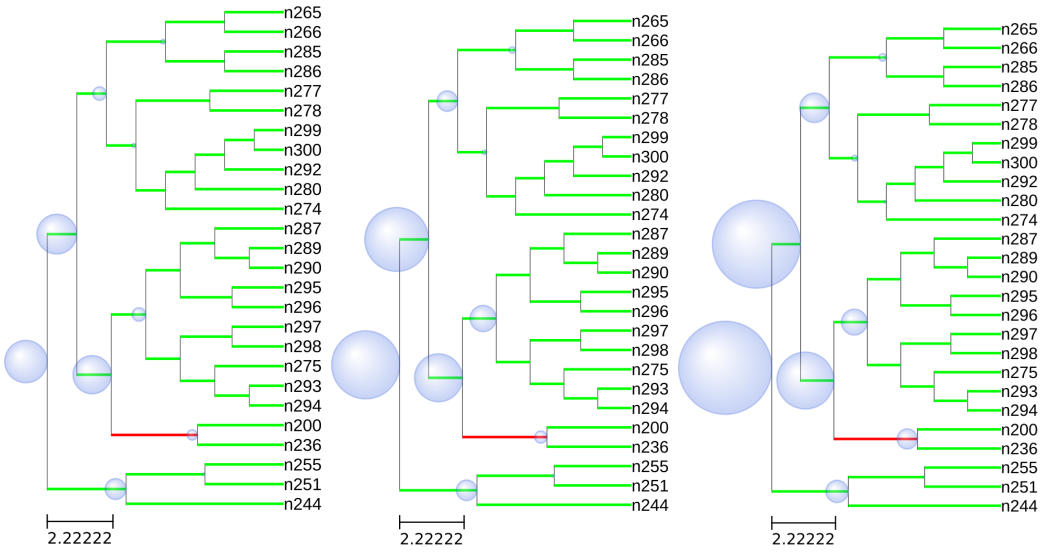

Figure 10 Non-linearity score for Dataset 2 with species tree S_1, with λ having value 0.25 (Left), 0.5 (Middle) and 1 (Right)

Note that these linearity scores are obtained with an error in the species tree uncorrelated with potential errors in gene trees. This is an unrealistic assumption because we can expect that the cause of errors in a species tree are seen as well in the gene trees, or even come from correlated errors in the gene trees. Further tests are needed to control for this effect. Nonetheless some previous observations on *Drosophila* showed that the linearity score was the highest precisely at the most debated node (Semeria et al., 2015), or in *Anopheles* that the linearity score was lower with a species phylogeny agreeing with the gene trees (Anselmetti et al., 2018b). There is an agreeing body of arguments showing that the linearity score is a promising proxy for species phylogeny.

3.3 Phylogenomic methods to reconstruct gene trees

Finally we compared two ways of constructing gene tree sets, in terms of statistics discussed in the previous sections. First we reconstructed gene trees, using IQ-TREE, from multiple sequence alignments of the gene families simulated with Zombi. Next we used integrated phylogenomic methods including the principle of reconciliation to reconstruct gene trees (see Chapter 3.2 [Boussau and Scornavacca 2020]). We refer to Section 2.3 for a description of the methods used with each dataset.

Figure 11 shows the ancestral gene content (number of ancestral genes by species) distribution and the linearity of ancestral genomes, according to different sets of gene trees (true trees, IQ-TREE trees and trees reconstructed with a reconciliation method).

As in our previous experiments, we can observe that the behavior is different in the duplication/loss (Dataset 1) and duplication/loss/HGT (Dataset 2) cases. For Dataset 1, the gene content is unrealistically higher with IQ-TREE trees, confirming the remark by Hahn (2007) discussed above that errors in gene trees could affect the gene content of ancestral

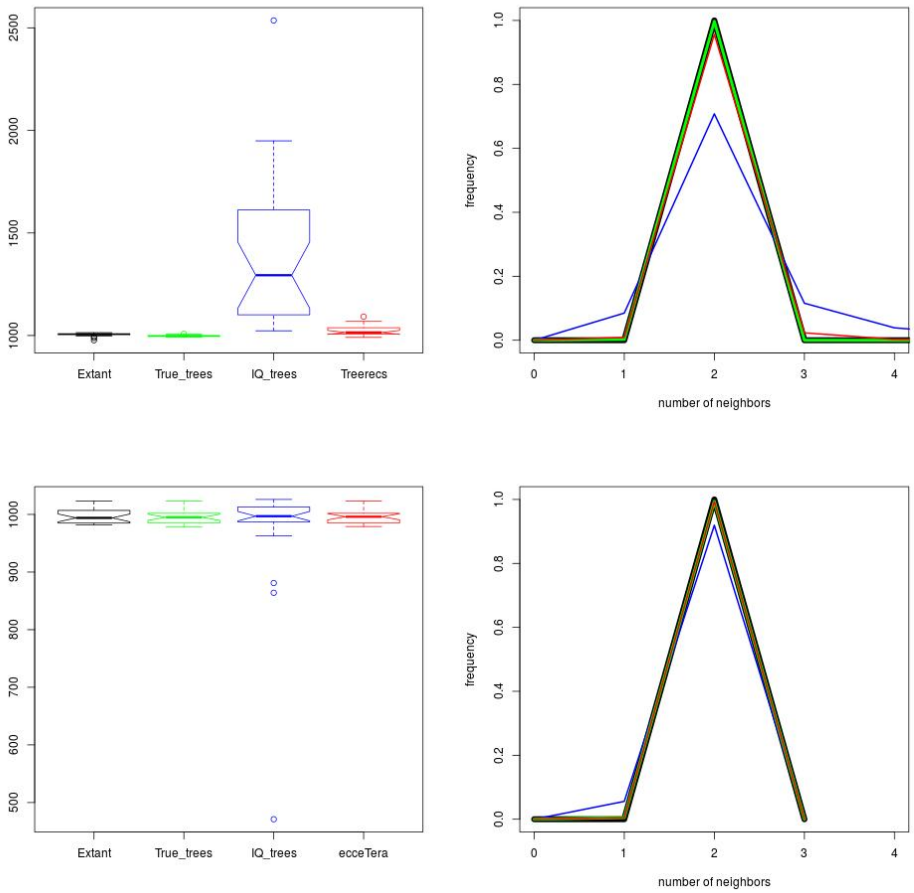

Figure 11 (Left) Distribution of extant (black) and ancestral (other colors) gene contents, computed with true trees (green), IQ-TREE (blue) and Treerecs (red). (Right) Frequency of degrees (number of neighbors) of ancestral genes. (Top) Dataset 1 (with duplications and no transfer), the number of ancestral genes is vastly overestimated and the linearity is lowered if trees are computed from sequences only (IQ-TREE). Both are improved by the reconciliation method. (Bottom) Dataset 2 (with transfers and duplications), the number of ancestral genes is underestimated and the linearity is slightly lowered if trees are computed from sequences only (IQ-TREE). Both are improved by the reconciliation methods.

species. The linearity is almost the same for true trees and reconciled trees, and much worse for IQ-TREE trees. This confirms the intuition present in several former papers (Boussau et al., 2013; Anselmetti et al., 2018b) that the linearity and gene content could serve as an indicator of the quality of gene trees. For Dataset 2 the results are similar but present significant and interesting differences. First, contrary to Dataset 1, gene content is lower for low quality trees, instead of higher. The linearity differences, if present, are much less marked. It seems that the possibility of HGTs in the reconciliation methods can "correct" the errors in gene tree topologies and gives nonetheless almost correct gene numbers and ancestral genome linearity, which, if true, would be an interesting case of robustness of a pipeline to errors in preliminary steps.

4 Conclusion

The present work was motivated by the observation, in previous works from our group, that our efforts to improve species trees or gene tree sets had effects on the linearity of ancestral genomes. We thus formulated the hypothesis that the non-linearity would be a good indicator of the quality of the input data, especially gene trees. In order to explore this hypothesis, we designed a set of experiments on simulated data where the level of noise in the considered trees (gene trees and species tree) is controlled. This allowed us to test our starting hypothesis, and we indeed observe that there is a strong correlation between the non-linearity score and the level of noise. As discussed above, this observation could have practical applications, where non-linear structures in the adjacency graphs of ancestral gene orders could provide starting points to correct gene trees or the species tree.

From a methodological point of view, our general idea can be described as follows. Facing a computationally intractable problem (reconstructing ancestral gene orders in a parsimony framework), one can relax some biological constraints (here the fact that chromosomes are linear or circular gene orders) in order to gain computational tractability; then the inconsistencies observed in the obtained solution with regard to the relaxed biological constraints open a window toward improving the input data. A few examples exist of this kind of serendipitous approaches. For example, one can think to horizontal gene transfers: biology would impose time-consistency on reconstructed transfers, however the problem of inferring time-consistent transfers is NP-hard (Hallett et al., 2004). Finding transfers while allowing them to be time-inconsistent can be solved polynomial in polynomial time (Jacox et al., 2016; Bansal et al., 2018). And, as shown by Chauve et al. (2017), it seems that, similarly to the way we interpret the non-linearity of ancestral gene orders, the level of time inconsistency is correlated with the quality of the input data.

Our work is limited to this proof of principle. We devised experiments only within a small range of parameters, that were chosen to show the possibility of using linearity as diagnosis and its limits. We do not cover all biological conditions. In particular the effect of errors in alignments, gene clustering or annotations have not been investigated, and can be the object of future work.

References

Abrouk, M., Murat, F., Pont, C., Messing, J., Jackson, S., Faraut, T., Tannier, E., Plomion, C., Cooke, R., and Feuillet, C. (2010). Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends in Plant Science*, 15(9):479–487.

Angibaud, S., Fertin, G., Rusu, I., Thévenin, A., and Vialette, S. (2009). On the approximability of comparing genomes with duplicates. *Journal of Graph Algorithms and Applications*, 13(1):19–53.

Anselmetti, Y., Duchemin, W., Tannier, E., and Bérard, S. (2018a). Phylogenetic signal from rearrangements in 18 Anopheles species by joint scaffolding extant and ancestral genomes. *BMC Genomics*, 19(2):96.

Anselmetti, Y., Luhmann, N., Bérard, S., Tannier, E., and Chauve, C. (2018b). Comparative methods for reconstructing ancient genome organization. In Setubal, J. C., Stoye, J., and Stadler, P. F., editors, *Comparative Genomics: Methods and Protocols*, volume 1704 of *Methods in Molecular Biology*, pages 343–362. Springer New York.

Babcock, E. B. and Navashin, M. S. (1930). *The Genus Crepis*, volume 6. Bibliographia Genetica.

Bansal, M. S., Kellis, M., Kordi, M., and Kundu, S. (2018). RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*, 34(18):3214–3216.

Bérard, S., Gallien, C., Boussau, B., Szöllősi, G. J., Daubin, V., and Tannier, E. (2012). Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics*, 28(18):i382–i388.

Blin, G., Chauve, C., Fertin, G., Rizzi, R., and Vialette, S. (2007). Comparing genomes with duplications: A computational complexity point of view. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(4):523–534.

Boussau, B. and Scornavacca, C. (2020). Reconciling gene trees with species trees. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.2, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.

Boussau, B., Szöllősi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Research*, 23:323–30.

Chauve, C., Rafiey, A., Davin, A. A., Scornavacca, C., Veber, P., Boussau, B., Szöllősi, G. J., Daubin, V., and Tannier, E. (2017). MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers. *Peer Community In Evolutionary Biology*.

Comte, N., Morel, B., Hasic, D., Guéguen, L., Penel, S., Boussau, B., Daubin, V., Scornavacca, C., Gouy, M., Stamatakis, A., Tannier, E., and Parsons, D. (2020). Treerecs: an integrated phylogenetic tool, from sequences to reconciliations. submitted.

Czech, L., Huerta-Cepas, J., and Stamatakis, A. (2017). A critical review on the use of support values in tree viewers and bioinformatics toolkits. *Molecular Biology and Evolution*, 34(6):1535–1542.

Davín, A. A., Tricou, T., Tannier, E., de Vienne, D. M., and Szöllősi, G. J. (2019). Zombi: A phylogenetic simulator of trees, genomes and sequences that accounts for dead lineages. *Bioinformatics*, 36(4):1286–1288.

Dobzhansky, T. and Sturtevant, A. H. (1938). Inversions in the chromosomes of Drosophila Pseudoobscura. *Genetics*, 23(1):28–64.

Duchemin, W., Anselmetti, Y., Patterson, M., Ponty, Y., Bérard, S., Chauve, C., Scornavacca, C., Daubin, V., and Tannier, E. (2017). Decostar: Reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. *Genome Biology and Evolution*, 9(5):1312–1319.

Fernández, R., Gabaldón, T., and Dessimoz, C. (2020). Orthology: Definitions, prediction, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.4, pages 2.4:1–2.4:14. No commercial publisher | Authors open access book.

Fuller, Z. L., Koury, S. A., Phadnis, N., and Schaeffer, S. W. (2018). How chromosomal rearrangements shape adaptation and speciation: Case studies in Drosophila pseudoobscuraand its sibling species Drosophila persimilis. *Molecular Ecology*, 28(6):1283–1301.

Groussin, M., Daubin, V., Gouy, M., and Tannier, E. (2016). Ancestral reconstruction: Theory and practice. In *Encyclopedia of Evolutionary Biology*, pages 70–77. Elsevier.

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

Hallett, M. T., Lagergren, J., and Tofigh, A. (2004). Simultaneous identification of duplications and lateral transfers. In *Proceedings of the Eighth Annual International Conference on Computational Molecular Biology, 2004, San Diego, California, USA, March 27-31, 2004*, pages 347–356.

Hoff, M., Orf, S., Riehm, B., Darriba, D., and Stamatakis, A. (2016). Does the choice of nucleotide substitution models matter topologically? *BMC Bioinformatics*, 17(1):143.

Jacox, E., Chauve, C., Szöllősi, G. J., Ponty, Y., and Scornavacca, C. (2016). eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058.

Joy, J. B., Liang, R. H., McCloskey, R. M., Nguyen, T., and Poon, A. F. Y. (2016). Ancestral reconstruction. *PLOS Computational Biology*, 12(7):1–20.

Kovác, J. (2014). On the complexity of rearrangement problems under the breakpoint distance. *Journal of Computational Biology*, 21(1):1–15.

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lemaitre, C., Braga, M. D. V., Gautier, C., Sagot, M.-F., Tannier, E., and Marais, G. A. B. (2009). Footprints of Inversions at Present and Past Pseudoautosomal Boundaries in Human Sex Chromosomes. *Genome Biology and Evolution*, 1:56–66.

Ma, J., Ratan, A., Raney, B. J., Suh, B. B., Zhang, L., Miller, W., and Haussler, D. (2008). DUPCAR: reconstructing contiguous ancestral regions with duplications. *Journal of Computational Biology*, 15(8):1007–1027.

Murat, F., Peer, Y. V. d., and Salse, J. (2012). Decoding Plant and Animal Genome Plasticity from Differential Paleo-Evolutionary Patterns and Processes. *Genome Biology and Evolution*, 4(9):917–928.

Necsulea, A. (2020). Phylogenomics and genome annotation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.1, pages 4.1:1–4.1:26. No commercial publisher | Authors open access book.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Patterson, M., Szöllosi, G. J., Daubin, V., and Tannier, E. (2013). Lateral gene transfer, rearrangement, reconciliation. *BMC Bioinformatics*, 14(S-15):S4.

Peres, A. and Crollius, H. R. (2015). Improving duplicated nodes position in vertebrate gene trees. *BMC Bioinformatics*, 16(3):A9.

Philippe, H., de Vienne, D. M., Ranwez, V., Roure, B., Baurain, D., and Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, 283.

Pont, C., Wagner, S., Kremer, A., Orlando, L., Plomion, C., and Salse, J. (2019). Paleogenomics: reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biology*, 20(1):29.

Rajaraman, A. and Ma, J. (2016). Reconstructing ancestral gene orders with duplications guided by synteny level genome reconstruction. *BMC Bioinformatics*, 17(S-14):201–212.

Ranwez, V. and Chantret, N. (2020). Strengths and limits of multiple sequence alignment and filtering methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.2, pages 2.2:1–2.2:36. No commercial publisher | Authors open access book.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542.

Semeria, M., Tannier, E., and Guéguen, L. (2015). Probabilistic modeling of the evolution of gene synteny within reconciled phylogenies. *BMC Bioinformatics*, 16(Suppl 14):S5.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Sturtevant, A. H. (1921). A case of rearrangement of genes in Drosophila. *Proceedings of the National Academy of Sciences U S A*, 7(8):235–237.

Tannier, E., Zheng, C., and Sankoff, D. (2009). Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics*, 10.

Timoshevskiy, V. A., Severson, D. W., deBruyn, B. S., Black, W. C., Sharakhov, I. V., and Sharakhova, M. V. (2013). An integrated linkage, chromosome, and genome map for the yellow fever mosquito aedes aegypti. *PLOS Neglected Tropical Diseases*, 7(2):1–11.

Yang, Z. and Zhu, T. (2018). The good, the bad, and the ugly: Bayesian model selection produces spurious posterior probabilities for phylogenetic trees. arXiv preprint `https://arxiv.org/abs/1810.05398`.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A., and Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761.

Zhou, L., Lin, Y., Feng, B., Zhao, J., and Tang, J. (2017). Phylogeny analysis from gene-order data with massive duplications. *BMC Genomics*, 18(7):760.

Zhou, X., Shen, X.-X., Hittinger, C. T., and Rokas, A. (2018). Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Molecular Biology and Evolution*, 35(2):486–503.