Duplication, Rearrangement and Reconciliation: a follow-up 13 years later

Cedric Chauve[1,2], Nadia El-Mabrouk[3], Laurent Guéguen[4], Magali Semeria[4], Eric Tannier[4,5]

[1] LaBRI, Université Bordeaux I, Talence, France
[2] Department of Mathematics, Simon Fraser University, Burnaby BC, Canada
[3] DIRO, Université de Montréal, Montréal QC, Canada
[4] LBBE, Université Lyon I Claude Bernard, Lyon, France
[5] INRIA Rhône-Alpes, France

Abstract. The evolution of genomes can be studied at at least three different scales: the nucleotide level, accounting for substitutions and indels, the gene level, accounting for gains and losses, and the genome level, accounting for rearrangements of chromosome organization. While the nucleotide and gene levels are now often integrated in a single model using reconciled gene trees, very little work integrates the genome level as well, and considers gene trees and gene orders simultaneously. In a seminal book chapter published in 2000 and entitled "Duplication, Rearrangement and Reconciliation", Sankoff and El-Mabrouk outlined a general approach, making a step in that direction. This avenue has been poorly exploited by the community for over ten years, but recent developments allow the design of integrated methods where phylogeny informs the study of synteny and vice-versa. We review these developments and show how this influence of synteny on gene tree construction can be implemented.

1 Introduction

Genomes evolve through a wide variety of mechanisms, not all of them well understood, or even known to us. These mechanisms range from small-scale events, such as point mutations or small insertions or deletions at the nucleotide level, to large-scale cataclysmic events such as whole-genome duplications, through segmental duplications or deletions, inversions, transpositions, insertion of mobile elements, translocations, and chromosomes fusions and fissions [1]. While genome evolution is a joint process that combines all such mechanisms, evolutionary studies through computational and statistical methods are generally compartmentalized, as most of them focus on one or few kinds of evolutionary events. Nucleotide level mutations, inferred from alignments, are those considered by phylogenetic methods for gene and species tree constructions. Duplications and other *content-modifying operations* (gains, losses, transfers,...) are considered for the inference of evolutionary histories of gene families. Inversions, transpositions, translocations and other gene order modifying *rearrangements* are the

events considered in synteny evolution studies, which aim at reconstructing ancestral genome organizations or inferring rearrangement based phylogenies. Figure 1 attempts to represent some models for genome evolution according to the type of mutations they handle. For example, the blue and gray dots represent phylogenetic methods from nucleotide or genome level mutations. The gray star and blue square respectively stand for evolutionary studies of gene order and gene content accounting for rearrangement or gene gains and losses.

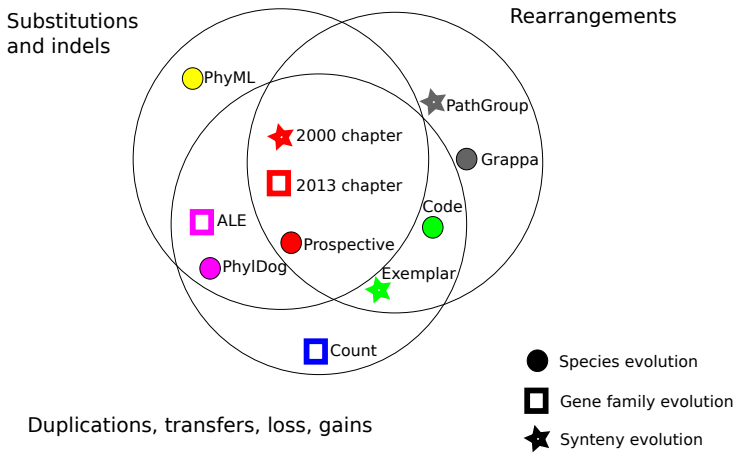

Fig. 1. Each of the three big sets represents one of the three kinds of mutations we are dealing with. Dots, squares and stars are models of genome evolution handling these kinds of mutations, respectively aimed at reconstructing phylogenies, gene content evolution and synteny evolution. If they lie in a set intersection, they integrate several kinds of mutations. Apart from the red area, the names aside the dots, squares and stars are examples of softwares or methods achieving the described goal (PhyML [2], Count [3], ODT [4], PhylDog [5], Exemplar [6], Pathgroup [7], Grappa [8],Code [9]). They are often chosen among a lot of other examples which would have been as relevant. The red area is the core of our chapter: the star refers to the 2000 Sankoff and El-Mabrouk book chapter we are celebrating, the square is achieved in the present chapter, and the dot is the still open problem toward which all integrative methods tend.

In this paper, we review the attempts to integrate this variety of multiple-scale events into a single framework (red region in Fig 1). In addition we contribute to this integration by showing how reconcilied gene trees can be improved using synteny information.

Sequence evolution (substitutions and indels) and chromosome evolution (rearrangement, gene order) are traditionally two separate domains of study. This can be traced back to the early stages of evolutionary studies based on molecular data. Usually, the first molecular phylogenies are dated back to the Pauling and Zuckerkandl series of papers in the early 1960's [10]. Nevertheless thirty years before, Surtevant and Dobzhansky were drawing *Drosophila* phylogenies comparing the structure of polytene chromosomes [11]. Even the mathematics of chromosome rearrangements were already investigated at that time, and computational problems were formally stated [12]. However, these pioneering works have not been followed, and mathematical and computational studies of genome rearrangements have been nearly absent for several decades. Instead methodologists did put a lot of effort into the modelization of the evolution of DNA or protein sequences. Advanced models and algorithms have been developed [13, 14] (see also Chapter 6 in this volume), integrating character substitutions and indels, reconstructing phylogenies and ancestral states by various statistical methods. It is only in the early eighties that formal models of gene order evolution were investigated again, after a nearly 50 year hiatus [15], mainly following Sankoff's efforts [16, 17]. As for today, despite significant progress, the considered models for gene order evolution are still not reaching the sophistication of those for sequence evolution [18] (see also Chapter 7 in this volume).

For a long time, in most phylogenetic studies based on sequence, only genes with apparently simple histories, typically those present in a single copy in every genome, were considered [19]. This aspect has changed during the last twenty years, driven by the gene tree/species tree reconciliation studies pioneered by Goodman *et al.* [20]. Reconciliation gives a way of integrating gene family evolution into models of sequence evolution. Recently, many sophisticated methods for gene tree and/or species tree inference, integrating gene sequence evolution and gene insertion, duplication, loss or transfer, into a unified model have been developed [4, 5, 21–25]. This integration is represented by the purple square and dot in Figure 1 (see also Chapter 12 in this volume).

In parallel, genome rearrangement studies were at first developed in a context where genomes were also assumed to have exactly the same set of genes, with exactly one copy per genome. Such an assumption is reasonable for specific data sets such as animal mitochondrial genomes [16], or in a more general context provided an appropriate pre-processing of genomes [26–30]. In this context of single copies, two kinds of models have been investigated: a "global model" where a genome is encoded by a unique object (permutation, string, or a variant) with a value space of size $\geq n!$ where n is the number of genes, and a "local model" in which a genome is decomposed into $O(n)$ characters as adjacencies evolving independently, each taking two possible values (present/absent). For global models it was shown that comparing two genomes can be done pretty efficiently [31,32], while almost all attempts to compare more than two genomes lead to intractable problems (survey in [18]). The local model gave rise to easier problems [27–30, 33] (see also Chapter 7 in this volume), the drawback being that the independence hypothesis between adjacencies lead to ancestral states

that are not necessarily compatible with linear or circular chromosomal structures, leading again to difficult linearization problems (although few exceptions exist [34, 35]).

Integrating duplications and more generally gene families with complex histories into the study of synteny evolution (the green star in Figure 1) has been initiated by David Sankoff with the so-called "exemplar approach" [6], which consists in encoding genomes as strings instead of permutations, allowing for the representation of a single gene many times in a genome. In this spirit an insight on gene content evolution can be infered from synteny by the detection of orthology relationships [36–38]. But this direction was hampered by ubiquitous intractability results even for the comparison of two genomes [18, 39, 40] (see also Chapter 9 of this volume). The local model has also allowed to overcome the computational complexity in that direction [9] (red dot in Figure 1, see also Chapter 7 of this volume).

In 2000, a conference was held named *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families* [41], organized by Sankoff and Nadeau. The title of the conference, and of the companion book was already a manifest towards an integration of gene evolution and genome evolution. The present chapter, the volume it is included in, and the 2013 MAGE conference are also, in some ways, attempts to follow up on this event. In particular, the book published in 2000 contained the "Duplication, Rearrangement and Reconciliation" chapter by Sankoff and El-Mabrouk [42], that we revisit, 13 years later.

That chapter was one among several examples (see also Chapter 2 of the present volume for example) of a work in which David Sankoff has laid the basis of a research avenue several years before it was explored by the scientific community. We feel the exploration of this avenue really starts now, 13 years later. To advocate this, we first summarize the key concepts used by Sankoff and El-Mabrouk [42] (Section 2). Then we describe the two lines of research that have been built on these initial ideas: using phylogenetic information, by means of reconciliation, to study gene order evolution (Section 3, red star in Figure 1), and using gene order information to study gene family evolution (Section 4, red square in Figure 1). We give a contribution to this latter part by constructing an accurate method of synteny-aware gene tree correction. We conclude by some discussion points and perspectives on possible integration of phylogenies, syntenies and histories in a unified framework for studying genome evolution (Section 5, red dot in Figure 1).

2 Duplication, Rearrangement and Reconciliation

In this section we revisit the 2000 Sankoff and El-Mabrouk chapter [42]. It is also the occasion to introduce concepts and objects related to reconciliations and rearrangements.

Evolution of species and genes. Species evolve through *speciation*, which is the separation of one species into two distinct descendant species. The result of this

evolution is a set Σ of n extant species. A *species tree* on Σ is a binary rooted tree whose leaves are in bijection with Σ, representing the evolutionary relationship between the species of Σ: an internal node is an ancestral species at the moment of a speciation, and its two children are the descendent species.

Species are identified with their genomes, and a genome is a set of genes (plus a structure for gene order detailed later). Genes undergo speciation when the species to which they belong do. Within a species, genes can be *duplicated*, *lost* or *gained*. Various mechanisms lead to duplications of various sizes, ranging from one single gene (or segment of genes) to the whole genome [43]. Gene losses arise through the pseudogenization of previously functional genes or the outright deletion of chromosomal fragments. There are other gene level events like transfers, but here we stay with only duplication and losses as it was the context of the 2000 chapter [42].

A *gene tree*, representing the evolution of a *gene family*, is a binary rooted tree where each leaf is labeled by a gene, belonging to a species in Σ. Each internal node of a gene tree refers to an ancestral gene at the moment of an event (either speciation or duplication) resulting in two copies of this gene. The *lowest common ancestor* (LCA) of nodes x and y in a tree T, written $lca_T(x, y)$, is the internal node of T that is both an ancestor of x and y and is the farthest from the root of T.

In a gene tree, losses are invisible, and speciation and duplication events are not distinguishable, unless we *reconcile* it with a species tree.

Reconciliation. A *reconciliation* of a gene tree T with a species tree S consists in assigning to each internal gene g of T a species $M(g) = s$, which is a node of S (either extant or ancestral), indicating that gene g belongs to species s, and an evolutionary event $E(g) \in \{\text{speciation,duplication}\}$. This is done in a way ensuring that the evolutionary history of the gene family described by the reconciliation is in agreement with the species evolution described by S.

The reconciliation of T with S gives information about the gene family history. In particular, it defines the gene content of an ancestral species s at the time of speciation. A reconciliation also implies the orthology and paralogy relationships between genes: Two genes g and g' of T are said to be orthologous if $E(lca_T(g, g')) = Spec$; g and g' are paralogous if $E(lca_T(g, g')) = Dup$. They are said to be *ohnologous* if they are paralogous and the duplication event at $lca_T(g, g')$ is due to a whole genome duplication.

Since the work of Page [44, 45] in the beginning of the 90s, and with an increasing interest in the last decade, several approaches have been developed to reconcile a gene tree with a species tree. The main guiding principle has been to optimize a given criterion such as parsimony in terms of duplications and/or losses or maximum likelihood (see [46] for a recent review). Recent methods aim at reconstructing gene trees and reconciliations simultaneously [22, 23, 25].

Gene order and genome rearrangements. We defined a genome as a set of genes and a structure on these genes. This structure can be for example a *signed permutation* which gives a total order to the genes, and a direction $(+/-1)$ to each

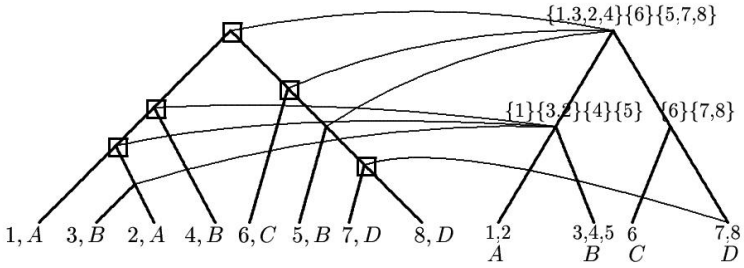

Fig. 2. (Copied from [42]). A reconciliation of a gene tree (on the left) with a species tree (on the right). The genome (letter) to which each gene (number) belongs is indicated in the label of the corresponding leaf in the gene tree. The mapping M is indicated by links between the two trees. The mapping E is indicated on the gene tree: squares are duplication nodes, while other internal nodes are speciations. In the species tree, ancestral nodes are labeled by their gene contents. Each set corresponds to a single ancestral gene.

gene. When two genomes have equal gene content (gene evolution is ignored), a rearrangement scenario is a sequence of operations on the permutation which transforms one genome into the other. In that case efficient algorithms exist for many genome rearrangement distances, often based on analyzing the structure of permutations and of their breakpoint graph. They hardly generalize to more than two genomes: *median* problems, considering three genomes, are often intractable, and hardly allow the exploration of solution spaces [18, 47].

When gene families may have several copies, a natural generalization of signed permutations is given by *signed strings* over the gene families alphabet (see Figure 3). Each family is assigned an integer and each occurrence of this integer indicates an occurrence of a gene from the corresponding family. The comparison of two such strings can be achieved by finding an orthology assignment between gene copies from the same family. Keeping only ortholog pairs of genes transforms a signed string into a signed permutation, on which known algorithms apply. The orthology assignment is a matching in the bipartite graph over the string elements, in which there is an edge between two genes of the same family in different genomes (see Figure 3). In his 1999 paper [6], Sankoff introduced the notion of *exemplar matching* that assumes all duplications are posterior to the speciation between the two genomes. This corresponds to taking only one edge per family in the bipartite graph. But this also leads to hard problems: computing a parsimonious exemplar matching is hard, even to approximate for the simplest distances (see [39,48] and references therein, as well as Chapter 9 in this volume). Other notions of matchings were introduced later, not assuming the precedence of speciation over all duplications [36], also leading to hard optimization problems [18, 49]. Although reasonably efficient exponential time algorithms have been

developed [50], it is still an open question as to whether these approaches will scale efficiently to more than three genomes (see Chapter 13 in this volume).

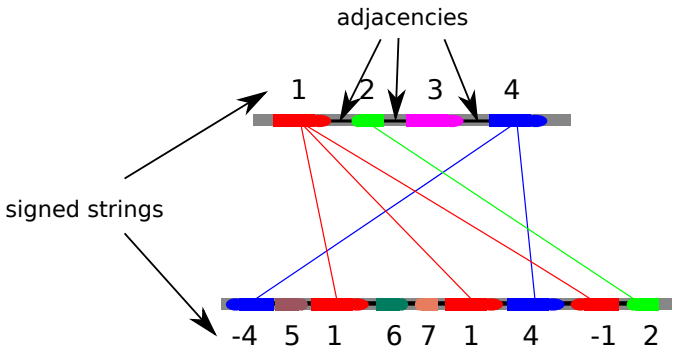

Fig. 3. Gene order comparison of two genomes with duplications. Each genome is a signed string on the gene family alphabet. The direction of each gene is written according to the relative orientation on the two genomes. Homology relationships are edges between genes of the two different genomes, and the comparison is achieved with matching problems on this bipartite graph. Black links indicate adjacencies which are another way to encode the strings. Focusing on one adjacency independently of the neighboring ones makes the comparison computations tractable, but the linear structure of the ancestral genomes might be lost.

Reconciliations and rearrangements. Interestingly, gene order provides formal methods for inferring "positional homology" [37], which can be applied to the detection of orthology [51] or ohnology [52, 53]. This places gene order information as a concurrent of reconciliation for orthology or ohnology detection. This, among other reasons, calls for the use of reconciliation and gene order in the same framework, because both carry information on gene family evolution. Synteny and reconciled phylogenies have sometimes been used together to detect orthology or ohnology [52–54] but rarely in a whole genome evolution model integrating duplications and rearrangements.

This is what was proposed by Sankoff and El-Mabrouk [42]. In their framework, an arbitrary number of genomes are given, along with a species tree describing their evolution. In addition, reconciled gene trees are given for all gene families and the genes at the leaves of these gene trees are ordered in the genomes.

Then, as orthology is known from reconciliation, considerations on rearrangement distances between genomes can include duplications in the permutation model, tending to lower the additional algorithmic complexity. Nevertheless the general problem still contains two difficult problems, the median and the exemplar problems, as particular cases.

A solution was proposed in that paper to infer gene content and order at each internal node of a species tree, in a way that minimizes a total breakpoint distance on the species tree. Handling multicopy gene families was undertaken by integrating exemplar matching for the duplications that were identified by reconciliation to be posterior to speciations. A heuristic was proposed for solving the general problem. It was never applied on data, partly because it was developed before data was available in a usable form, and partly because the aim of this article was to open a research avenue more than to close a problem.

For several years this avenue has remained almost unexplored. But several recent publications have followed this research program, at least in some way, with updated genome rearrangement and reconciliation methods and some biological results.

3 Gene tree reconciliations inform synteny evolution

Genome rearrangement studies with permutation or string models, *i.e.* global models, usually do not handle a large number of genomes or events. After the seminal article by Sankoff and El-Mabrouk [42], we may mention a remarkable attempt of Ma *et al.* [55] to devise an integrated global model of genome evolution under certain restrictive conditions. But the computational complexity of global models usually restrains them to the study of small gene clusters, while paradoxically the histories of whole genomes are often inferred with local models of evolution. We describe the two possibilities, the first with a survey of existing literature and the second with a contribution to synteny-aware gene phylogeny construction.

3.1 Evolution of Gene Clusters with Global Models

A large fraction of duplications affecting genome organization consists of local duplications, mainly caused by unequal crossing-over during meiosis. As this phenomenon is favored by the presence of repetitive sequences, a single duplication can induce a chain reaction leading to further duplications, eventually creating large repetitive regions. When those regions contain genes, the result is a *Tandemly Arrayed Gene (TAG) cluster*: a group of paralogous genes that are adjacent on a chromosome. In the 70s, Fitch [56] introduced the first evolutionary model for TAGs accounting for *tandem duplication*, in which the two descendent copies of a duplicated gene are adjacent. Since then, several studies have considered the problem of inferring an evolutionary history for TAG clusters [57–60]. These are essentially phylogenetic inference methods using the additional constraint that the resulting tree should be a *duplication tree*, i.e.

induces a duplication history according to the given gene order. However, due to the occurrence of mechanisms other than tandem duplications (losses, rearrangements), it is often impossible to reconstruct a duplication tree [61].

In a series of papers [62–65], a solution is proposed to retrace the history of gene clusters subject to tandem duplications, losses and rearrangements. The latest developments allow us to study the evolution of orthologous TAG clusters in different species, subject to tandem duplications, inverted tandem duplications, inversions and deletions, each event involving one or a set of adjacent genes. Given the gene trees, the species tree, and the order of genes in TAG clusters, the method for inferring ancestral clusters combines reconciliation with gene order information. First, ignoring gene order, reconciliation is used to infer ancestral gene contents. Second, ancestral gene orders are inferred that are consistent with minimizing the number of rearrangement events required to obtain a duplication tree. Due to the NP-hardness of the problem, exact approaches can hardly be envisaged, except for the special case of clusters in a single species subject to simple duplications (duplications of single genes) [62]. The DILTAG software [64,65] developed for the most general case is a heuristic, showing good performance in practice for the inferrence of recent evolutionary events. Despite the uncertainty associated with the deeper parts of the reconstructed histories, it can be used to infer the duplication size distribution with some precision.

3.2 Evolution of whole genome with adjacency models

The structure of the genomes, as seen in Figure 3, can be described by a set of adjacencies instead of signed permutations or strings. Adjacencies are edges linking gene extremities. If we wish to formalize the linear or circular structure of genomes, the set of adjacencies should be a matching over gene extremities. This models all types of genomes, from linear multichromosomal eukaryote genomes to circular bacterial genomes, possibly including circular or linear plasmids.

The switch to local models is achieved by comparing adjacencies instead of genomes. When extant genomes all have the same gene content, it is possible to rapidly and accurately reconstruct ancestral genomes [27, 30, 66] or even species phylogenies [67]. It is even possible to include no equal gene contents [9] for the same purposes (see Chapter 7 in this volume).

If gene families are described with reconciled trees with duplications, the software DupCAR [68] proposes the reconstruction of ancestral adjacencies. Nevertheless, its possible applications are rather limited as it does not handle losses and requires fully dated gene trees and species tree, in order to compute reconciliations that are compatible with the provided date information. Such precise information about gene trees is rare.

The joint use of adjacencies and reconciled gene trees has really been exploited in the last three years. Agora [66] or the method of El-Mabrouk and colleagues [69, 70] reconstruct ancestral adjacencies with a sort of Dollo parsimony principle, by pairwise comparisons of extant gene orders. Methods designed initially to handle equal gene content [27, 30] can also naturally be extended to follow this principle, by using orthology/paralogy information obtained from

reconciliations. So adjacencies are reconstructed but no evolutionary scenario is proposed to explain them. In all such methods, linearization routines, based on the Travelling Salesman Problem [69, 70] or on path/cycles graph covering techniques [27, 35] are used to linearize ancestral genomes, that is, to remove a posteriori some of the proposed adjacencies so that every gene (or gene extremity) has at most two (or one) neighbor.

DeCo [71] is an algorithm and a software which models the evolution of adjacencies, and reconstructs ancestral adjacencies by minimizing the number of gains and losses of adjacencies (due to rearrangements) along the species tree. It is based on a generalization of the Sankoff-Rousseau algorithm (see Chapter 3 in this volume) adapted to the presence/absence of adjacencies and to reconciled gene trees. It has recently been extended to include tranfers (Patterson et al, in preparation). But here again, the resulting ancestral adjacencies might not be compatible with a genome structure and require to be processed a posteriori.

4 Synteny informs gene family evolution

Synteny as a control. Synteny is usually a good way to infer orthologs [72]. As reconciled gene trees also yield orthology relationships, it is possible to compare the results from both independent methods. This is what is often done to assess the quality of gene trees [73, 74]. Orthologs obtained from synteny are assumed to represent the truth, and can be compared with orthologs obtained from reconciliations.

An alternative idea is to use the structure of reconstructed ancestral genomes as a quality index. We have seen that in the extant genomes, each gene shares two adjacencies with other genes (except for chromosome extremities). Theoretically it should also be the same in the ancestral genomes. But due to errors in gene trees, in the species tree, or in the inference algorithms, in practice there are a lot of exceptions. And the number of exceptions should be correlated with the quality of gene trees. So the quality of gene trees can be measured by the number of ancestral genes with more or less than two adjacencies with other genees [5, 71].

These quality tests are first steps towards integrating synteny information into the construction of gene trees. Indeed, if the quality of gene trees can be measured with synteny, the next step is to integrate synteny into the objective function when computing gene trees.

Synteny as an input data for gene trees. Very few studies use synteny information to construct gene trees. The only ones we are aware of are the ones of Wapinski *et al* [75,76]. In these papers, family clustering as well as gene trees are constructed with a "synteny score" as well as a "sequence score".

In the present book three different contributions to this direction are given. Chapter 13 deals with the construction of gene families with synteny information. Chapter 12 proposes a method to detect inconsitencies in gene trees based on synteny information.

In the present chapter we show how, by a simple procedure using available software, it is possible to guide the construction of gene trees with gene order information. We retrieved all gene trees from the Ensembl database [73], version 70. Then we applied the DeCo software [71] to infer ancestral adjacencies in a mammalian species tree. As said before, DeCo does not guarantee that ancestral genes have at most two neighbors, as extant genes. This apparent weakness was turned into a strength as it was used for a quality control for gene trees. We show that it can be used to construct better gene trees.

Take an ancestral gene g with three neighbors, such that two of them, $g1$ and $g2$, are speciation nodes of the same gene tree, and are the two children of a duplication node d in this gene tree (see Figure 4). Let then $g11$, $g12$ (resp $g21$ and

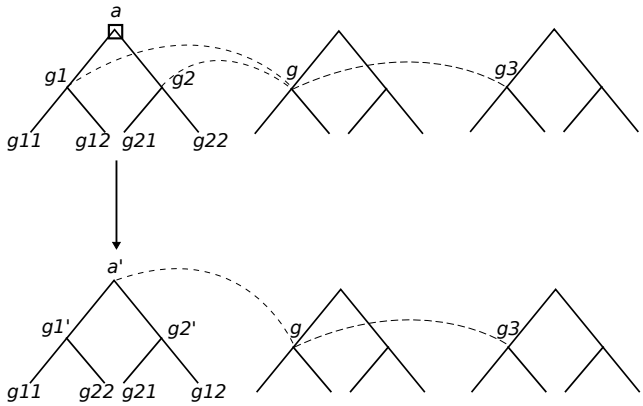

Fig. 4. Modifying a tree according to synteny information. The fact that gene g has three neighbors according to DeCo points to a possible error in gene trees (every gene should have at most two neighbors). In this situation we suspect the duplication d to be erroneous, as having only one gene instead of the two ($g1$ and $g2$) would decrease the number of neighbors of g. So we propose the correction in which there is only one gene, by rearranging the subtree rooted at d.

$g22$) be the children of $g1$ (resp $g2$). Transform the subtree $((g11, g12), (g21, g22))$ to $((g11, g22), (g21, g12))$ whenever this switches d to a speciation node in the reconciliation (1519 trees out of 13132 Ensembl trees that contain mammalian genes can be modified this way). This transformation is illustrated in Figure 4.

For all 1519 trees, we retrieved the gene family alignment from Ensembl. With PhyML [2] we computed the likelihood of two trees, before and after the transformation, given this alignment. These two likelihoods were compared with Consel [77]. For a majority of trees (773), the likelihood of the corrected tree

is higher than the likelihood of the initial tree. And the correction is rejected (probability of the corrected tree < 0.05) for only 281 of them (Figure 5).

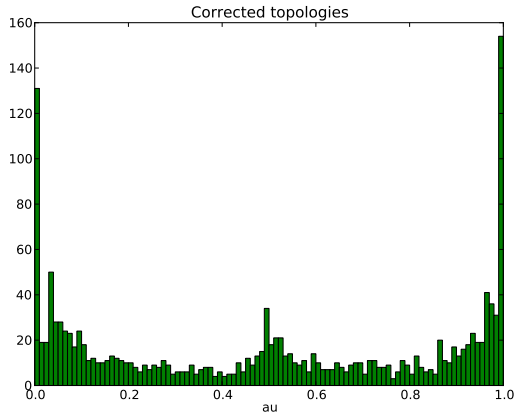

Fig. 5. Each green bar gives, for a given interval of p-values (x-axis), the number of gene families (y axis), among the 1519 whom we corrected the tree, for which the corrected tree should be prefered with a significance in the p-value interval. The shape of this graph shows that Ensembl trees are in general not significantly prefered, showing the accuracy of most corrections.

In conclusion, the synteny signal can be used to choose among the numerous trees that are statistically equivalent according to the sequence signal. This choice is sometimes in agreement with other reconciliation-based tree correction methods, but sometimes adds additional information. This provides a step towards synteny aware gene tree construction methods (red square in Figure 1).

5 Towards and integrated model

Boussau and Daubin [19] call for models of molecular evolution that would integrate all kinds of mutations and find likely ones according to a mixture of objectives. Because the species tree, gene trees and rearrangements all depend on each other, an iterative method, computing these objects one after another would not find an optimal solution. Integrated models of species and gene trees are already working [5, 25] (purple region in Figure 1).

But despite the efforts we described here to mix gene trees and rearrangement in the same framework, the three-way influence is far from reached. Some attempts can be mentioned, as Ma *et al*'s paper [55] gives a global algorithm under some very strict conditions (exact molecular clock, no convergent evolution, no breakpoint re-use, no gene loss).

Some other are less ambitious, like Kahn and colleagues who argued in a series of papers (see [78] and references therein) that reconciled trees cannot describe properly the evolutionary relationships and propose an extension to DAGs. They give insights into handling both evolutionary relationships and synteny, to trace the history of segmental duplications.

But 13 years after the paper of Sankoff and El-Mabrouk, which marked a first star in the three-way intersection of Figure 1, and accounts for an increasing interest in this area over recent years, the existence of a phylogenetic method over all events is still an open question (red dot in Figure 1).

Acknowledgements

This work is funded by the Agence Nationale pour la Recherche, Ancestrome project ANR-10-BINF-01-01.

References

1. Graur, D., Li, W.H.: Fundamentals of Molecular Evolution, second edition. Sinauer Associates, Inc. (2000)
2. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O.: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. Syst Biol **59**(3) (May 2010) 307–321
3. Csurs, M.: Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics **26**(15) (Aug 2010) 1910–1912
4. Szllosi, G.J., Boussau, B., Abby, S.S., Tannier, E., Daubin, V.: Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc Natl Acad Sci U S A **109**(43) (Oct 2012) 17513–17518
5. Boussau, B., Szllosi, G.J., Duret, L., Gouy, M., Tannier, E., Daubin, V.: Genome-scale coestimation of species and gene trees. Genome Res **23**(2) (Feb 2013) 323–330
6. Sankoff, D.: Genome rearrangement with gene families. Bioinformatics **15**(11) (1999) 909–917
7. Zheng, C.: Pathgroups, a dynamic data structure for genome reconstruction problems. Bioinformatics **26**(13) (Jul 2010) 1587–1594
8. Tang, J., Moret, B.M.E.: Scaling up accurate phylogenetic reconstruction from gene-order data. Bioinformatics **19 Suppl 1** (2003) i305–i312
9. Lin, Y., Hu, F., Tang, J., Moret, B.: Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. In: Pacific Symposium on Biocomputing. (2013)
10. Zuckerkandl, E., Pauling, L.: Molecules as documents of evolutionary history. J Theor Biol **8**(2) (Mar 1965) 357–366
11. Sturtevant, A., Dobzhansky, T.: Inversions in the third chromosome of wild races of drosophila pseudoobscura, and their use in the study of the history of the species. Proc Natl Acad Sci U S A **22** (1936) 448–450
12. Sturtevant, A., Novitski, E.: The homologies of chromosome elements in the genus drosophila. Genetics **26** (1941) 517–541
13. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.J.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press (1998)

14. Felsenstein, J.: Inferring phylogenies. Sinauer Associates, Inc. (2004)
15. Watterson, G.A., Ewens, W.J., Hall, T.E., Morgan, A.: The chromosome inversion problem. Journal of Theoretical Biology **99** (1982) 1–7
16. Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F., Cedergren, R.: Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc Natl Acad Sci U S A **89**(14) (1992) 6575–6579
17. Sankoff, D.: Edit distances for genome comparisons based on non-local operations. In: A. Apostolico, M. Crochemore, Z. Galil and U. Manber (eds). Combinatorial Pattern Matching, Third Annual Symposium, CPM 92, Tucson, Arizona, USA, April 29 - May 1, 1992, Proceedings. Volume 644 of Lecture Notes in Computer Science., Springer (1992) 121–135
18. Fertin, G., Labarre, A., Rusu, I., Tannier, E., Vialette, S.: Combinatorics of Genome Rearrangements. MIT Press (2009)
19. Boussau, B., Daubin, V.: Genomes as documents of evolutionary history. Trends Ecol Evol **25**(4) (Apr 2010) 224–232
20. Goodman, M., Czelusniak, J., Moore, G., Romero-Herrera, A., Matsuda, G.: Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Systematic Zoology **28** (1979) 132–163
21. Durand, D., Halldórsson, B.V., Vernot, B.: A hybrid micro-macroevolutionary approach to gene tree reconstruction. Journal of Computational Biology **13**(2) (2006) 320–335
22. Akerborg, O., Sennblad, B., Arvestad, L., Lagergren, J.: Simultaneous bayesian gene tree reconstruction and reconciliation analysis. Proceedings of the National Academy of Sciences of the United States of America **106**(14) (2009) 5714–5719
23. Rasmussen, M.D., Kellis, M.: A bayesian approach for fast and accurate gene tree reconstruction. Molecular Biology and Evolution **28**(1) (2011) 273–290
24. Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., Durand, D.: Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics **28**(18) (Sep 2012) i409–i415
25. Szöllosi, G.J., Rosikiewicz, W., Bousseau, B., Tannier, E., Daubin, V.: Efficient exploration of the space of reconciled gene trees. Submitted (2013)
26. Pevzner, P.A., Tesler, G.: Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Research **13**(1) (2003) 37–45
27. Ma, J., Zhang, L., Suh, B.B., Raney, B.J., Burhans, R.C., Kent, W.J., Blanchette, M., Haussler, D., Miller, W.: Reconstructing contiguous regions of an ancestral genome. Genome Res **16**(12) (Dec 2006) 1557–1565
28. Chauve, C., Tannier, E.: A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. PLoS Comput Biol **4**(11) (Nov 2008) e1000234
29. Chauve, C., Gavranovic, H., Ouangraoua, A., Tannier, E.: Yeast ancestral genome reconstructions: the possibilities of computational methods ii. J Comput Biol **17**(9) (Sep 2010) 1097–1112
30. Jones, B.R., Rajaraman, A., Tannier, E., Chauve, C.: Anges: reconstructing ancestral genomes maps. Bioinformatics **28**(18) (Sep 2012) 2388–2390
31. Hannenhalli, S., Pevzner, P.A.: Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. In: F. Thomson Leighton, A. Borodin (eds). Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing, 29 May-1 June 1995, Las Vegas, Nevada, USA, ACM (1995) 178–189

32. Hannenhalli, S., Pevzner, P.A.: Transforming men into mice (polynomial algorithm for genomic distance problem). In: 36th Annual Symposium on Foundations of Computer Science, Milwaukee, Wisconsin, 23-25 October 1995, IEEE Computer Society (1995) 581–592
33. Zhang, Y., Hu, F., Tang, J.: A mixture framework for inferring ancestral gene orders. BMC Genomics **13 Suppl 1** (2012) S7
34. Feijão, P., Meidanis, J.: Scj: A breakpoint-like distance that simplifies several rearrangement problems. IEEE/ACM Transactions on Computational Biology and Bioinformatics **8**(5) (2011) 1318–1329
35. Mauch, J., Patterson, M., Wittler, R., Chauve, C., Tannier, E.: Linearization of ancestral multichromosomal genomes. BMC Bioinformatics **13 Suppl 19** (2012) S11
36. Fu, Z., Chen, X., Vacic, V., Nan, P., Zhong, Y., Jiang, T.: MSOAR: A high-throughput ortholog assignment system based on genome rearrangement. Journal of Computational Biology **14**(9) (2007) 1160–1175
37. Dewey, C.N.: Positional orthology: putting genomic evolutionary relationships into context. Briefings in Bioinformatics **12**(5) (2011) 401–412
38. Doerr, D., Thévenin, A., Stoye, J.: Gene family assignment-free comparative genomics. BMC Bioinformatics **13**(Suppl 19) (2012) S3
39. Zhu, B.: Approximability and fixed-parameter tractability for the exemplar genomic distance problems. In: J. Chen and S. B. Cooper (eds). Theory and Applications of Models of Computation, 6th Annual Conference, TAMC 2009, Changsha, China, May 18-22, 2009. Proceedings. Volume 5532 of Lecture Notes in Computer Science., Springer (2009) 71–80
40. El-Mabrouk, N., Sankoff, D.: Analysis of gene order evolution beyond single-copy genes. Methods Mol Biol **855** (2012) 397–429
41. Sankoff, D., Nadeau, J., eds.: Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families. Kluwer academic publishers (2000)
42. Sankoff, D., El-Mabrouk, N.: Duplication, rearrangement and reconciliation. In: Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map alignment and the Evolution of Gene Families. Volume 1 of Computational Biology Series. Kluwer Academic Publishers (2000) 537–550
43. Gregory, T.R., ed.: The evolution of the genome. Elsevier Academic Press (2004)
44. Page, R.: Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Systematic Biology **43** (1994) 58–77
45. Page., R.: Genetree: comparing gene and species phylogenies using reconciled trees. Bioinformatics **14** (1998) 819–820
46. Doyon, J.P., Ranwez, V., Daubin, V., Berry, V.: Models, algorithms and programs for phylogeny reconciliation. Briefings in Bioinformatics **12**(5) (2011) 392–400
47. Miklos, I., Tannier, E.: Approximating the number of double cut-and-join scenarios. Theoretical Computer Science **439** (2012) 30–40
48. Bulteau, L., Jiang, M.: Inapproximability of (1, 2)-exemplar distance. In: L.G. Bleris and I.I. Mandoiu and R. Schwartz and J. Wang (eds). Bioinformatics Research and Applications - 8th International Symposium, ISBRA 2012, Dallas, TX, USA, May 21-23, 2012. Proceedings. Volume 7292 of Lecture Notes in Computer Science., Springer (2012) 13–23
49. Angibaud, S., Fertin, G., Rusu, I., Thévenin, A., Vialette, S.: On the approximability of comparing genomes with duplicates. Journal of Graph Algorithms and Applications **13**(1) (2009) 19–53

50. Angibaud, S., Fertin, G., Rusu, I., Thévenin, A., Vialette, S.: Efficient tools for computing the number of breakpoints and the number of adjacencies between two genomes with duplicate genes. Journal of Computational Biology **15**(8) (2008) 1093–1115
51. Goodstadt, L., Ponting, C.P.: Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput Biol **2**(9) (Sep 2006) e133
52. Ouangraoua, A., Tannier, E., Chauve, C.: Reconstructing the architecture of the ancestral amniote genome. Bioinformatics **27**(19) (2011) 2664–2671
53. Makino, T., McLysaght, A.: Positionally biased gene loss after whole genome duplication: evidence from human, yeast, and plant. Genome Res **22**(12) (Dec 2012) 2427–2435
54. Cai, B., Yang, X., Tuskan, G.A., Cheng, Z.M.: Microsyn: a user friendly tool for detection of microsynteny in a gene family. BMC Bioinformatics **12** (2011) 79
55. Ma, J., Ratan, A., Raney, B.J., Suh, B.B., Miller, W., Haussler, D.: The infinite sites model of genome evolution. Proceedings of the National Academy of Sciences of the United States of America **105**(38) (2008) 14254–14261
56. Fitch, W.: Phylogenies constrained by cross-over process as illustrated by human hemoglobins and a thirteen-cycle, eleven amino-acid repeat in human apolipoprotein A-I. Genetics **86** (1977) 623–644
57. Bertrand, D., Gascuel, O.: Topological rearrangements and local search method for tandem duplication trees. IEEE/ACM Transactions on Computational Biology and Bioinformatics **2** (2005) 15–28
58. Elemento, O., Gascuel, O., Lefranc, M.P.: Reconstructing the duplication history of tandemly repeated genes. Molecular Biology and Evolution **19**(3) (2002) 278–288
59. Tang, M., Waterman, M., Yooseph, S.: Zinc finger gene clusters and tandem gene duplication. In: Research in Molecular Biology (RECOMB 2001). (2001) 297–304
60. Zhang, L., Ma, B., Wang, L., Xu, Y.: Greedy method for inferring tandem duplication history. Bioinformatics **19** (2003) 1497–1504
61. Gascuel, O., Bertrand, D., Elemento, O.: Reconstructing the duplication history of tandemly repeated sequences. In Gascuel, O., ed.: Mathematics of Evolution and Phylogeny, Oxford (2005) 205–235
62. Lajoie, M., Bertrand, D., El-Mabrouk, N., Gascuel, O.: Duplication and inversion history of a tandemly repeated genes family. Journal of Computational Biology **14**(4) (2007) 462–478
63. Bertrand, D., Lajoie, M., El-Mabrouk, N.: Inferring ancestral gene orders for a family of tandemly arrayed genes. Journal of Computational Biology **15**(8) (2008) 1063–1077
64. Lajoie, M., Bertrand, D., El-Mabrouk, N.: Inferring the evolutionary history of gene clusters from phylogenetic and gene order data. Molecular Biology and Evolution **27**(4) (2010) 761–772
65. Tremblay-Savard, O., Bertrand, D., El-Mabrouk, N.: Evolution of orthologous tandemly arrayed gene clusters. BMC Bioinformatics **12**(Suppl 9) (2011) S2
66. Muffato, M., Louis, A., Poisnel, C.E., Crollius, H.R.: Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. Bioinformatics **26**(8) (2010) 1119–1121
67. Hu, F., Gao, N., Zhang, M., Tang, J.: Maximum likelihood phylogenetic reconstruction using gene order encodings. In: 8th Annual IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'11). (2011)

68. Ma, J., Ratan, A., Raney, B.J., Suh, B.B., Zhang, L., Miller, W., Haussler, D.: Dupcar: reconstructing contiguous ancestral regions with duplications. Journal of Computational Biology **15**(8) (2008) 1007–1027
69. Bertrand, D., Gagnon, Y., Blanchette, M., El-Mabrouk, N.: Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss. In: V. Moulton and M. Singh (eds). Algorithms in Bioinformatics, 10th International Workshop, WABI 2010, Liverpool, UK, September 6-8, 2010. Proceedings. Volume 6293 of Lecture Notes in Computer Science., Springer (2010) 78–89
70. Gagnon, Y., Blanchette, M., El-Mabrouk, N.: A flexible ancestral genome reconstruction method based on gapped adjacencies. BMC Bioinformatics **13 Suppl 19** (2012) S4
71. Bérard, S., Gallien, C., Boussau, B., Szöllősi, G.J., Daubin, V., Tannier, E.: Evolution of gene neighborhoods within reconciled phylogenies. Bioinformatics **28**(18) (2012) i382–i388
72. Jun, J., Mandoiu, I.I., Nelson, C.E.: Identification of mammalian orthologs using local synteny. BMC Genomics **10** (2009) 630
73. Vilella, A., Severin, J., Ureta-Vidal, A., Heng, L., Birney, E.: EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Research **19**(2) (2009) 327–335
74. Wu, Y.C., Rasmussen, M.D., bansal, M.S., Kellis, M.: Treefix: Statistically informed gene tree error correction using species trees. Systematic Biology **62**(1) (2013) 110–120
75. Wapinski, I., Pfeffer, A., Friedman, N., Regev, A.: Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics **23**(13) (2007) i549–i558
76. Wapinski, I., Pfeffer, A., Friedman, N., Regev, A.: Natural history and evolutionary principles of gene duplication in fungi. Nature **449** (2007) 54–61
77. Shimodaira, H., Hasegawa, M.: Consel: for assessing the confidence of phylogenetic tree selection. Bioinformatics **17**(12) (Dec 2001) 1246–1247
78. Kahn, C.L., Hristov, B.H., Raphael, B.J.: Parsimony and likelihood reconstruction of human segmental duplications. Bioinformatics **26**(18) (2010) i446–i452