A Methodological Framework for the Reconstruction of Contiguous Regions of Ancestral Genomes and its Application to Mammalian Genomes

Eric Tannier, Cedric Chauve

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A Methodological Framework for the Reconstruction of Contiguous Regions of Ancestral Genomes and its Application to Mammalian Genomes

Cedric Chauve — Eric Tannier

N° 6494

Avril 2008

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--6494--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

A Methodological Framework for the Reconstruction of Contiguous Regions of Ancestral Genomes and its Application to Mammalian Genomes

Cedric Chauve*, Eric Tannier†

Thème BIO — Systèmes biologiques
Équipe-Projet Helix

Rapport de recherche n° 6494 — Avril 2008 — 26 pages

Abstract: We describe a general methodological framework for reconstructing ancestral genome segments from conserved syntenies in extant genomes. We show that this problem, from a computational point of view, is naturally related to physical mapping of chromosomes, and benefits from using combinatorial tools developed for this problem. We develop this framework into a new reconstruction method considering conserved gene clusters with similar gene content. We implement and apply it to datasets of mammalian genomes. Compared to other bioinformatics methods for ancestral genome segments reconstructions, this one is stable: it gives convergent results using several kinds of data and different levels of resolution; it is reliable: all predicted ancestral regions are well supported; and it gives results that come very close to cytogenetics studies. It suggests the principle we propose, although based on a bioinformatics approach, relies on fundaments that are close to the ones used to analyze cytogenetic data.

Key-words: genome structure and organisation, ancestral genomes, ancestral chromosomes, PQ-tree, gene teams, consecutive ones property, methodology, algorithmics, genome rearrangements

* Department of Mathematics, Simon Fraser University, 8888 University Drive; V5A 1S6, Burnaby (BC), Canada.Email: cedric.chauve@sfu.ca

† INRIA Rhône-Alpes, Projet Helix ; Université de Lyon, F-69000, Lyon ; Université Lyon 1 ; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622, Villeurbanne, France. Email: eric.tannier@inria.fr

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Cadre méthodologique pour la reconstruction de régions génomiques ancestrales et application à des génomes de mammifères

Résumé : Nous décrivons un cadre méthodologique général pour la reconstruction de segments génomiques ancestraux à partir de synténies conservées dans les espèces actuelles. Nous montrons que ce problème, d'un point de vue algorithmique, est similaire aux questions de cartographie de chromosomes, et profite donc d'outils combinatoires développés dans ce cadre. Nous développons une méthode de reconstruction fondée sur ces outils en regroupant des syntons de gènes conservés dans des génomes de mammifères. Comparée à d'autres études bioinformatiques publiées, cette méthode est stable : elle donne des résultats convergents pour des jeux de données différents et différentes échelles de résolution ; elle est fiable : toutes les synténies ancestrales prédites sont bien soutenues par les données ; enfin, les résultats concordent avec des études cytogénétiques, ce qui suggère que les fondements du principe proposé se rapproche de ceux utilisés pour traiter des données cytogénétiques.

Mots-clés : structure et organisation des génomes, génomes ancestraux, chromosomes ancestraux, PQ-arbre, équipes de gènes, propriété des uns consécutifs, méthodologie, algorithmique, réarrangements de génomes

Introduction

The reconstruction of ancestral karyotypes and gene orders from homologies between extant species is a long-standing problem. In the case of mammalian genomes, it has first been approached using cytogenetics methods [28, 56, 57, 60, 62, 52]. The recent availability of sequenced and assembled genomes has led to the development of bioinformatics methods that address this problem at a much lower resolution, although with much less available genomes; such methods propose in general more detailed ancestral genome architectures than cytogenetics methods (see [13, 14, 16, 45, 39] and reviews in [26, 44, 49]). The comparison of the two approaches was recently investigated and discussed in a series of papers, sometimes with diverging point of views [27, 15, 53].

Among the bioinformatics methods that were applied at the level of mammalian genomes (previous works were limited to small genomes such as organellar genomes [11] or to bacterial genomes [24]), the one based on a global parsimony approach in terms of evolutionary events such as reversals, translocations, fusions and fissions [13, 45], leads to results that are sometimes in disagreement with cytogenetics results [27]. Recent results on this model-based approach point out that the modelization of genome rearrangements probably needs further studies before it can be used for the reconstruction of ancestral genomes (see also [15] where it was suggested that inferring parsimonious rearrangement scenarios is more intended to infer evolutionary dynamics characteristics such as rearrangement rates than ancestral genomes). Another type of approach infers ancestral genome segments, called Contiguous Ancestral Regions (CARs), from syntenic features that are conserved in extant species. This principle, that can be seen as model-free as it does not propose evolutionary scenarios and does not consider any kind of evolutionary event [8, 1]. It can be seen as less ambitious than the model-based approach as it does not propose evolutionary events, neither does it ensure that proposed CARs are ancestral whole chromosomes. However, recently applied on mammalian genomes [39] it gave results more in agreement with cytogenetic methods, while exhibiting few other points of divergence [53].

We describe here a very general model-free framework for the reconstruction of CARs, that formalizes and generalizes the principles used in several computational [39] and cytogenetics [60, 62, 52] studies. This framework takes as input a representation of extant genomes as sequences of homologous genomic markers (synteny blocks or orthologous genes for example), and decomposes into two main steps: we first compute a collection of possible ancestral syntenic groups (in general small groups of genomic markers that were possibly contiguous in the ancestral genome), each of them being weighted according to its conservation in the extant species; from this set of possible ancestral syntenies, we regroup and order the considered genomic markers into one (or several alternative) set(s) of CARs, each of these sets of CARs representing a possible ancestral genome architecture. An important feature of our framework is that we propose the set of all possible genome architectures that agree with the conserved ancestral syntenies. This framework is general in the sense that both steps can be made effective in several ways. For example, during the first phase, the signal for ancestral syntenies can be defined in terms of conserved (in extant species) adjacencies between homologous markers as in [39] or between chromosome segments as in [60, 62, 52]. We propose one possible implementation of this

framework, choosing as ancestral features both conserved adjacencies and gene teams [4, 38], generalizing the approach of Ma *et al* [39] (where only adjacencies were considered), and mimicking the methods employed with cytogenetic data [60, 62, 52] (conserved segments may be formalized as gene teams). The second step, that computes CARs and ancestral genome architectures, benefits from a combinatorial framework, centered around the Consecutive Ones Problem and an ubiquitous combinatorial structure called PQ-tree [12], well known and used in physical mapping [3, 18], and that was recently applied in other comparative genomics problems [36, 7]; in particular, in [8, 1, 61], PQ-trees were already considered to represent ancestral genomes. In our implementation of this second step, we follow the same principle than in [39], and extracts a maximum unambiguous subset of ancestral syntenies.

We apply our method on two datasets. We first consider the case of the ancestral boreoeutherian genome using a dataset obtained from the whole genome alignments available on the UCSC Genome Bioinformatics site [34]; from these alignments, we build sets of synteny blocks at different levels of resolution (we use from 625 to 2741 homolog markers). Our experiments show that the results of our method are constant, in the sense that they are very similar, independently of the chosen resolution. This reinforces the impression that differences in the results of [45, 39] discussed in [27, 15, 53] are more due to the method of reconstruction than to the differences of data acquisition and interdisciplinary problems [53]. Moreover, the results we obtain are very close to the ones towards which cytogenetics methods tend to converge. As these are obtained with much more species and expertise, we take it as a validation of the framework and method we propose. Compared to the recently published method of Ma *et al.* [39], we obtain sets of CARs that are a little less well defined, as they contain less adjacencies, but more supported, as any proposed adjacency is supported by at least one syntenic group that is conserved in at least two extant species. We also reconstruct an ancestral ferungulate genome architecture for the the same data than [45]. This dataset is at a different level of resolution, of quite high resolution as it represents mammalian genomes with only 307 synteny block. On this dataset, our method and the method of Ma *et al.* obtain again similar results that are comparable to the boreoeutherian chromosomes found by cytogenetics studies of the ferungulate chromosomes from E-painting studies [35], while the method of [45] on the same dataset gave divergent results.

The sequel of this paper is organized as follows. In the next section, we describe the general framework and how we implemented it to design a new method for ancestral genome reconstruction. We then describe the results of our method on the two mammalian datasets. We conclude by a discussion and describe several possible extensions.

Results

A general methodological framework and an implementation

We now describe more precisely the two steps of the framework, together with their implementation into an effective method for reconstructing a set of CARs. We separate the general principles from the implementation details to emphasize

that there are many possible implementations (the method of Ma *et al.* [39] can be seen as one possibility).

Input. Species tree. The input of our method is a set of extant genomes, together with a phylogenetic tree T describing the evolutionary relationships between the species to which the genomes belong. The ancestral genome we want to construct is characterized by its position, as an internal node on the phylogenetic tree. Finally, following [39], we assume that there is at least one outgroup species. This implies that the ancestral node has least two branches towards its descendants (exactly two if the tree is resolved) and one branch towards the outgroup species. Additionally we may add branch lengths to indicate the relative *a priori* expected quantity of evolution.

Implementation. We consider two datasets, focusing on two ancestral nodes of the mammalian clade: the boreoeutherian and ferungulate ancestors. The choice were made according to the possibilities of comparisons of the obtained ancestors with former studies [39, 45, 60, 62, 52, 35]. The phylogenetic tree of all considered species is described in Figure 1, and the branch lengths were taken according to lower bounds from recent studies in paleotonlogical dating [6].

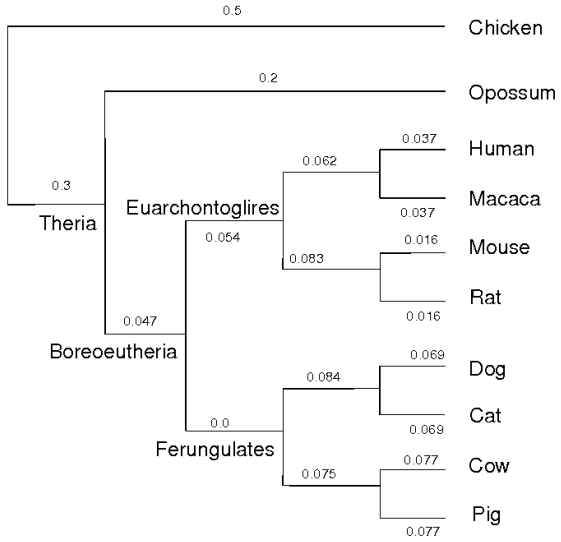

Figure 1: The phylogenetic relationships between studied species.

Input. Representation of extant genomes. Following other approaches for ancestral genome reconstruction, we represent the genome of an extant species by a set of sequences of homologous genomic markers, each of them belonging to a family identified by a unique label. Such families of genomic markers can be defined in several ways: from annotated orthologous genes [16, 1, 46, 10], to whole genome alignments methods [20, 58] as in [16, 39], comparative maps [45] or virtual hybridation [5]. Each extant chromosome is an ordered

sequence of markers, each marker being represented by the label of its family. If there are n family labels, we denote by $\mathcal{L} = \{1, \ldots, n\}$ the set of all family labels (the markers alphabet).

Implementation. We construct several datasets from the pairwise whole genome alignments between the human genome taken as reference, and the rhesus, mouse, rat, cow, dog, chicken, and opossum genomes, available on the UCSC Genome Bioinformatics website [34]. Pairwise synteny blocks between the human genome and the seven other extant genomes were computed from pairwise genome alignments, following the method described in [19, 55], for value of the parameters *max_gap* (the size of ignored micro-rearrangements) of 100kb and of *min_len* (the minimum length of pairwise alignments with the reference genome) ranging from 100kb to 500kb (see details in Material and Methods). Then multispecies synteny blocks were computed using the human genome as a reference. For each value of the parameters *max_gap* and *min_len*, we kept the set of markers that are present in all eight genomes. In order to make several comparisons with published methods, we also use a dataset taken from the supplementary material of Murphy *et al.* [45], based on human-mouse synteny blocks and comparative maps of seven mammalian genomes (human, mouse, rat, pig, cattle, cat, dog).

Additional remarks. Unlike some other approaches, especially model-based approaches, that require unique and sometimes universal markers in extant genomes [8, 16, 45, 47], the framework we propose does not impose in general such constraint: in a given extant genome, the number of markers that belong to a given family, which have the same label and are then considered indistinguishable, can be any number (see the Discussion section for more comments on this matter).

Step 1. Detection of putative ancestral genome segments. The first step consists in detecting groups of genomic markers (i.e. subsets of the set of markers labels) that are candidate to be *contiguous in the ancestral genome*; this point is central in our framework (see Discussion) and is close to cytogenetic methods, though working with different data. The general problem of this first step reduces then to define synteny conservation patterns along the species tree T that indicates a possible ancestral synteny, and to detect such patterns.

Implementation. We chose to follow a simple general principle: a group of genomic markers is possibly contiguous in the ancestor genome if it is contiguous in at least two extant species whose evolutionary path on the phylogenetic tree goes through the considered ancestral node. From then, several synteny conservation models between pairs of genomes can be considered: adjacent pairs of genes with the same orientation, as in [39, 10], or common intervals in [1]. Here we use two kinds of conserved features: gene teams with no gaps (also called maximal common intervals) [38], defined as maximal genome segments that have the same content in terms of genomic markers, and non-ambiguous unsigned adjacencies (see Material and Methods for precise definitions). As such ancestral syntenic groups can have very different conservation patterns in T, we associate to each of them a weight, based on the pattern of occurrence of this set of markers in T and on the branching pattern of T, following the weighting scheme used in [39] (see Material and Methods). This weight can be seen as a way to measure the extent of conservation of a given feature.

Step 2. Structuring ancestral features and PQ-trees. The output of the first phase is a set $\mathcal{S} = \{S_1, \ldots, S_m\}$ of m weighted, and pairwise different, subsets of $\mathcal{L}$, where each subset is a set of genomic markers that are believed to be contiguous in the ancestral genome. The problem is then to regroup the markers of $\mathcal{L}$ into CARs, and to order them inside these CARs, which, from a computational point of view, is very related to physical mapping problems [3, 18][1]. We then use an approach developed, first in the graph theory community and then applied to physical mapping problems, based on the *consecutive ones property* (C1P) and *PQ-trees*.

We encode $\mathcal{S}$ by an $m \times n$ 0/1 matrix $\mathcal{M}$ where row i represents S_i as follows: $\mathcal{M}[i, j] = 1$ if marker j belongs to S_i and 0 otherwise. Ordering markers into CARs consists in finding a permutation of the columns of the matrix $\mathcal{M}$, such that all 1's entries in each row are consecutive (also called a C1P ordering for $\mathcal{M}$). Finding such an order of the columns of $\mathcal{M}$ is not always possible, in particular if there are false positives in $\mathcal{S}$, that is groups of markers that were not contiguous in the ancestral genome. Moreover, if there exists a C1P ordering of the columns of $\mathcal{M}$, there are often several possible (sometimes an exponential number of) such orderings that make all 1's consecutive on each row, that represent several alternative possible ancestral genome architectures.

In the case where there exists a C1P ordering for $\mathcal{M}$, all C1P orderings can be represented in a compact way, using the *PQ-tree* of $\mathcal{M}$, denoted $T(\mathcal{M})$. We now provide a short description of the important properties of this structure with respect to C1P orderings (a complete formal description is given in Material and Methods). $T(\mathcal{M})$ is a tree with three kinds of nodes: leaves, P-nodes and Q-nodes. The leaves are labeled by $\mathcal{L}$, in such a way that each $i \in \mathcal{L}$ labels exactly one leaf of $T(\mathcal{M})$. P-nodes and Q-nodes are internal nodes, both with a total order on their children. The main property of $T(\mathcal{M})$ is that any C1P ordering of $\mathcal{M}$ can be obtained from $T(\mathcal{M})$ by reading, from left to right, the leaves labels of $T(\mathcal{M})$ after choosing for each node N, independently of the other nodes, (1) an arbitrary order for the children of N if N is a P-node, or (2) to reverse or not the order of the children of N if N is a Q-node. In such a PQ-tree, two markers define an *adjacency* if they are consecutive siblings of a Q-node.

Finally, if $\mathcal{M}$ is not C1P, we can still represent some partial information from it using a structure called the *PQR-tree* in [41] or *generalized PQ-tree* in [40], that we also denote by $T(\mathcal{M})$. It contains a fourth kind of nodes, called *degenerate nodes* or *R-nodes* that represent disjoint subsets of $\mathcal{S}$ that are not C1P. Computing $T(\mathcal{M})$ can be done efficiently (see Material and Methods).

[1] In physical mapping problems, markers representing the hybridation of probes are known but their relative order in the mapped genome is not known. What is known, from hybridation with genome fragments, is that some sets of markers need to be contiguous. The problem is then to find an organization of the markers into chromosomes, such that all, or a maximum of subset of $\mathcal{S}$ if it is not possible to handle all markers, are indeed contiguous in the resulting genome. Intuitively, the conserved syntenic groups of $\mathcal{S}$, that represent sets of possibly ancestral contiguous markers, can be seen as ancestral genome fragments that have evolved along T and are observed today.

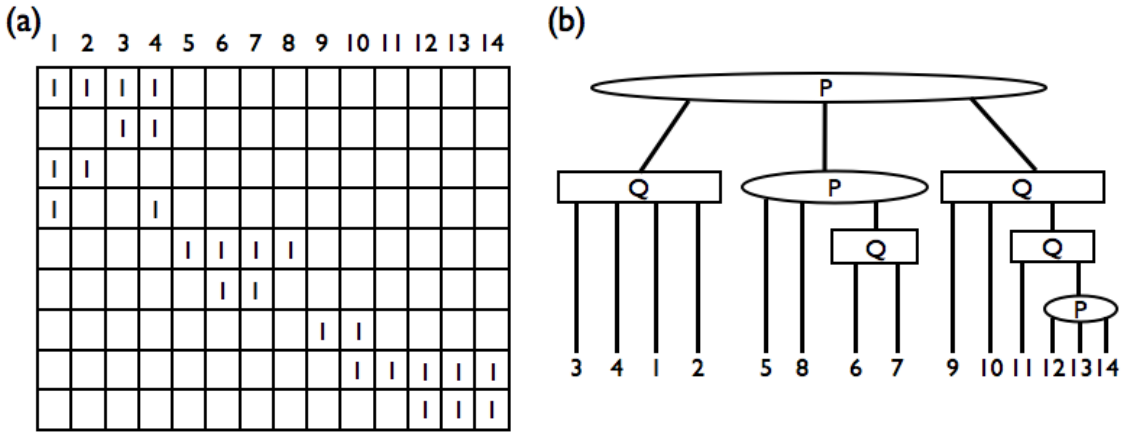

Figure 2: (a) A matrix $\mathcal{M}$ with the consecutive ones property. (b) The corresponding PQ-tree $T(\mathcal{M})$, where P-nodes are rounded and Q-nodes are square. 3 4 1 2 5 6 7 8 9 10 11 12 13 14 and 3 4 1 2 9 10 14 12 13 11 6 7 5 8 are two possible C1P orderings for $\mathcal{M}$, among 13824 possible C1P orderings. 3 4 1 2 5 7 6 8 9 10 11 12 13 14 is not a C1P ordering for $\mathcal{M}$: columns 6 and 7 need to be consecutive as they are consecutive children of a same Q-node.

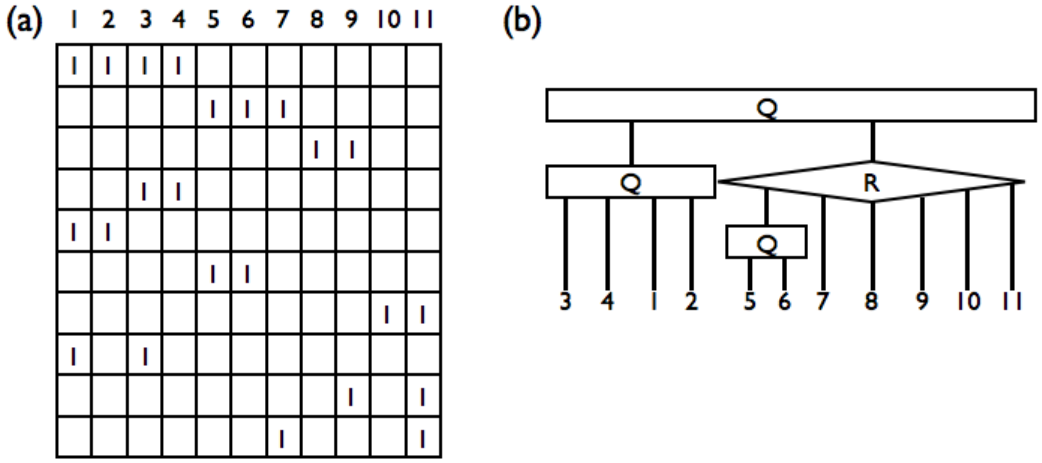

Figure 3: (a) A matrix $\mathcal{M}$ without the consecutive ones property. (b) The corresponding generalized PQ-tree, where there is a single R-node represented by a diamond shape labeled R. The only R-node is due to the rows 1, 2, 6, 7 and 9 of $\mathcal{M}$ that define a sub-matrix that is not C1P, while the submatrix defined by the remaining rows is C1P.

An important property of the framework we describe is that, if all allegated orthologies are true and if all $S'_i s$ are true positive, that is, were indeed contiguous in the ancestral genome, then there exists a C1P ordering of the markers of $\mathcal{L}$. In that case, $T(\mathcal{M})$ encodes in a compact way all possible C1P orderings of the columns of $\mathcal{M}$ and then all alternative genome architectures we can deduce from $\mathcal{S}$: the root of $T(\mathcal{M})$ is a P-node, children of the root represent CARs, where Q-nodes describe fixed orderings, up to a reversal, while P-nodes except the root describe subsets of markers that have to be contiguous but where there is no information to fix a relative order (see Figure 4 for an illustration).

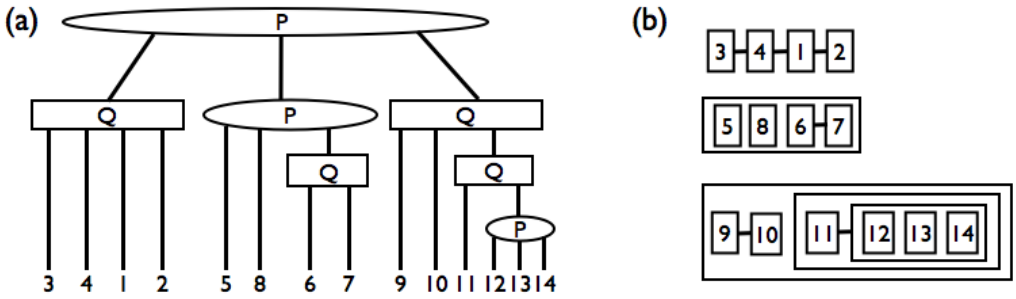

Figure 4: (a) The PQ-tree $T(\mathcal{M})$ of the matrix $\mathcal{M}$ of Figure 2.(a). (b) An equivalent representation of $T(\mathcal{M})$ that highlights all ancestral genome architectures that correspond to C1P orderings for $\mathcal{M}$: each row corresponds to a chromosomal segment represented by a child of the root, two glued blocks have to be adjacent in any ancestral genome architecture and sets blocks that float in the same box have to be consecutive in any genome architecture but their order is not constrained. Here we see three ancestral chromosomal segments: the first one, that contains markers 1 to 4 is totally ordered; the second one contains markers 5 to 8, with only constraint that markers 6 and 7 are adjacent; the third one contains markers 9 to 14, with 9 and 10 being adjacent, 11 being adjacent to a block that contains 12, 13 and 14 with no order between these three markers. Hence, 9 10 11 12 13 14 is a possible order for this last segment, but not 9 10 12 11 13 14 as 11 is inserted inside the block that contains 12, 13 and 14. All 13824 possible C1P orderings (possible ancestral orderings) are visible on this representation.

On the other hand, if $\mathcal{M}$ is not C1P, $T(\mathcal{M})$ extracts parts of $\mathcal{S}$ that are unambiguous and can be used directly to define CARs (the P-nodes and Q-nodes of the generalized PQ-tree), unlike the ambiguous parts of $\mathcal{S}$ that contain non-ancestral features (the R-nodes). It is then a first level of representation of CARs, that contains possible ambiguous information and generalizes the successor and predecessor graphs of [39].

Step 3. Clearing ambiguities and constructing CARs. As pointed above, if $\mathcal{M}$ is C1P, there is no indication that some features of $\mathcal{S}$ are not ancestral, so we directly output the possible ancestral genomes as the PQ-tree

$T(\mathcal{M})$. However, if $\mathcal{M}$ is not C1P, then we know that some sets of markers in $\mathcal{S}$ are false positive and were not contiguous in the ancestral genome. There can be several reasons: errors in constructing homolog markers (paralogies inferred instead of orthologies), incomplete syntenies resulting from convergent loss of markers, convergent fusions of chromosomal segments in several lineages for example. As in physical mapping [29], depending of the kinds of errors that have to be removed, there are several ways to remove ambiguous information present in $\mathcal{S}$ such as discarding some markers or features, or splitting possibly chimeric sets of markers in two or more subsets. After ambiguous information has been removed, there remains a subset $\mathcal{S}'$ of $\mathcal{S}$ that defines a C1P matrix $\mathcal{M}'$ and a PQ-tree $T(\mathcal{M}')$ that represent all possible genome architectures compatible with $\mathcal{M}'$.

Implementation. In our implementation, as we considered DNA alignments at a resolution of at least 100kb, taking care about possible paralogies by eliminating segmental duplications and repeated elements, we did not consider the option of discarding markers. We then clear ambiguities by removing elements from $\mathcal{S}$, i.e. rows from $\mathcal{M}$ that represent possibly non-ancestral syntenies. More precisely, we rely on the following combinatorial optimization problem: find a subset of $\mathcal{S}$ of maximum cumulated weight, such that the matrix of this subset is C1P. This problem, that generalizes the approach used in [39], is NP-hard. We solve it using a branch-and-bound algorithm based on a greedy heuristic inspired from [39] (see Material and Methods).

Reconstructing ancestral mammalian genome architectures

All results discussed in this section are available on a companion website: `http://www.cecm.sfu.ca/~cchauve/SUPP/ANCESTOR08/`.

The boreoeutherian ancestor from UCSC whole genome alignments

We computed five datasets, with parameters $max_gap = 100$kb and $min\_len =$ 100kb, 200kb, 300kb, 400kb and 500kb. Table 1 describes the number of synteny blocks and the covered size of the human genome.

min_len (kb)	Number of synteny blocks	Human coverage (Mb)
100	2741	1963
200	1651	1691
300	1097	1420
400	859	1258
500	625	1147

Table 1: Characteristics of the datasets based on the UCSC alignments. "human coverage" is the portion of the human genome that is covered by the set of considered synteny blocks, expressed in Mb.

Overview of the results with all datasets. In Table 2 below, we see that the number of CARs obtained decreases as *min_len* increases, which is expected

as synteny blocks hide more rearrangements that could prevent ancestral syntenies to be detected. This number of CARs in fact tends to converge towards the accepted number of 23 chromosomes in the ancestral boreoeutherian ancestral genome. We also report the number of adjacencies; this number indicates how well defined are the ancestral genome architectures, as the synteny blocks that are not in an adjacency belong to sets of markers that are children of a P-node and whose relative order is not known. It can be seen in Table 2 that there is a few such synteny blocks (less than 10%), which means that the ancestral genome architectures are very well defined. Finally, we also report chromosomal syntenic associations in the inferred ancestral genome architecture between some human chromosomes. We can also see that the results we obtain are very consistent, and in general do not propose human chromosomal syntenies that disagree with previous cytogenetics studies [44, 27]. The only such difference is the synteny between human chromosomes 1 and 4, seen with $min_len = 100$kb only, and a synteny between human chromosomes 5 and 8, observed only with $min_len = 500$kb. The synteny between human chromosomes 1 and 4 joins a single synteny block of human chromosome 4 (of size 253kb) with 273 synteny blocks of human chromosome 1; this synteny is supported by several gene teams, and it should be assessed if the occurrences of this synteny block in the considered genomes are really orthologous copies. This synteny between chromosomes 5 and 8 is supported by a single gene team, with relatively low weight as it is found only in the rat and opossum genomes and that involves a single synteny block of human chromosome 5, of length 1600kb. In the same time, it is not surprising that the synteny between human chromosomes 4 and 8 disappears with $min_len = 500$kb, as with $min_len = 400$kb, it was supported by a single gene team that contained a single marker of chromosome 8 (and three markers of chromosome 4), that is not present in the chicken genome with $min\_len = 500kb$. Other differences between the results obtained with the different values of min_len mostly involve the number of CARs corresponding to human chromosomes 1, 2 and 5.

min_len (kb)	CARs	Adjacencies	Human chromosomal syntenies
100	33	2638	1-4, 3-21, 4-8, 7-16, 12-22, 12-22
200	29	1552	3-21, 4-8, 7-16, 12-22, 12-22, 14-15, 16-19
300	28	1012	3-21, 4-8, 7-16, 12-22, 12-22, 14-15, 16-19
400	26	774	3-21, 4-8, 7-16, 12-22, 12-22, 14-15, 16-19
500	24	564	3-21, 5-8, 7-16, 12-22, 12-22, 14-15, 16-19

Table 2: Characteristics of the reconstructed genome architectures with our method and the universal datasets based on the UCSC alignments. In the third column, a set of numbers linked by - indicate a CAR that contains markers that belong to the corresponding human genomes.

Results with $max\_gap = 100kb$ and $min\_len = 400kb$. With values of $max\_gap = 100kb$ and $min\_len = 400kb$, a similar resolution level than the one used in in [16, 45], we obtain the 26 CARs presented in Figure 5.

We can compare the obtained CARs with the previously published boreoeutherian ancestors, in the light of some recent discussions on these results [27].

We recover ancestral segments that are very close to cytogenetic studies: all the 26 segments of the $max\_gap = 100$, $min\_len = 400$ dataset are indeed segments with which all cytogenetic publications agree [28, 56, 57, 60, 62, 52], and this is the first reported bioinformatics study which verifies this. We just miss two or three adjacencies according to the studies: some are probably due to an insufficiency of our data (human chromosome 1 is cut into three pieces in our reconstruction, whereas it was probably a unique piece in the ancestor; note however that in the earliest studies [52], human chromosome 1 was actually cut into two in the mammalian ancestor), and others are debated in the community (adjacency between human arm 10p and an ancestral chromosome 12-22 [60, 62], or between chromosome 1 and a segment from chromosome 19 [62]).

Computational characteristics of the CARs inference method. From a computational point of view, we can notice that these five datasets seem to contain very little ambiguity. For example, with max_gap = 100kb and, min_len = 400kb, only 14 ancestral syntenies detected during the first step needed to be discarded to clear all ambiguities in the 0/1 matrix, over a total of 1498 detected ancestral syntenies. The heuristic discarded a set of 17 ancestral syntenies, that was reduced to 14 by the branch-and-bound algorithm, which finds a provably optimal solution in a very small amount of time. The generalized PQ-tree contained 25 CARs, 9 of them were represented by R-nodes (children of the root of the generalized PQ-tree), and were then ambiguous, and discarding these 14 ancestral syntenies broke one of these ambiguous CARs into 2 CARs. With other values of min_len, the computational characteristics were similar (very few ancestral syntenies need to be discarded to clear, ambiguities), with the only difference that with min_len = 100kb, the branch-and-bound algorithm did not terminate in a reasonable time and was stopped before it finds the optimal result.

Comparison with the method of Ma *et al.* Up to date, the method developed by Ma *et al.* in [39] seems to be the bioinformatics method that proposes CARs that agree the most with cytogenetice methods. Moreover, it is a possible implementation of the general model-free framework we propose, based on syntenic characters that are oriented adjacencies, detected using a Fitch-like approach, and where ambiguities are discarded using a local parismony heuristic that is a particular case of the one we used in our method. Note however that, unlike our method, the method of Ma *et al.* proposes an orientation for the synteny blocks in the reconstructed CARs; our method could easily be completed by a post-processing phase to compute these orientations, using a parsimony approach for example.

We report in Table 3 the results of the method described in [39] on the constructed datasets, where, for every proposed adjacency between two markers i and j, we say that it is *weakly supported* if there is no ancestral synteny in $\mathcal{S}$ that contains both i and j (by construction, all the adjacencies reported by our method is supported by at least one ancestral synteny). We also say that an adjacency is *common*, if it is also present in the CARs obtained with our method.

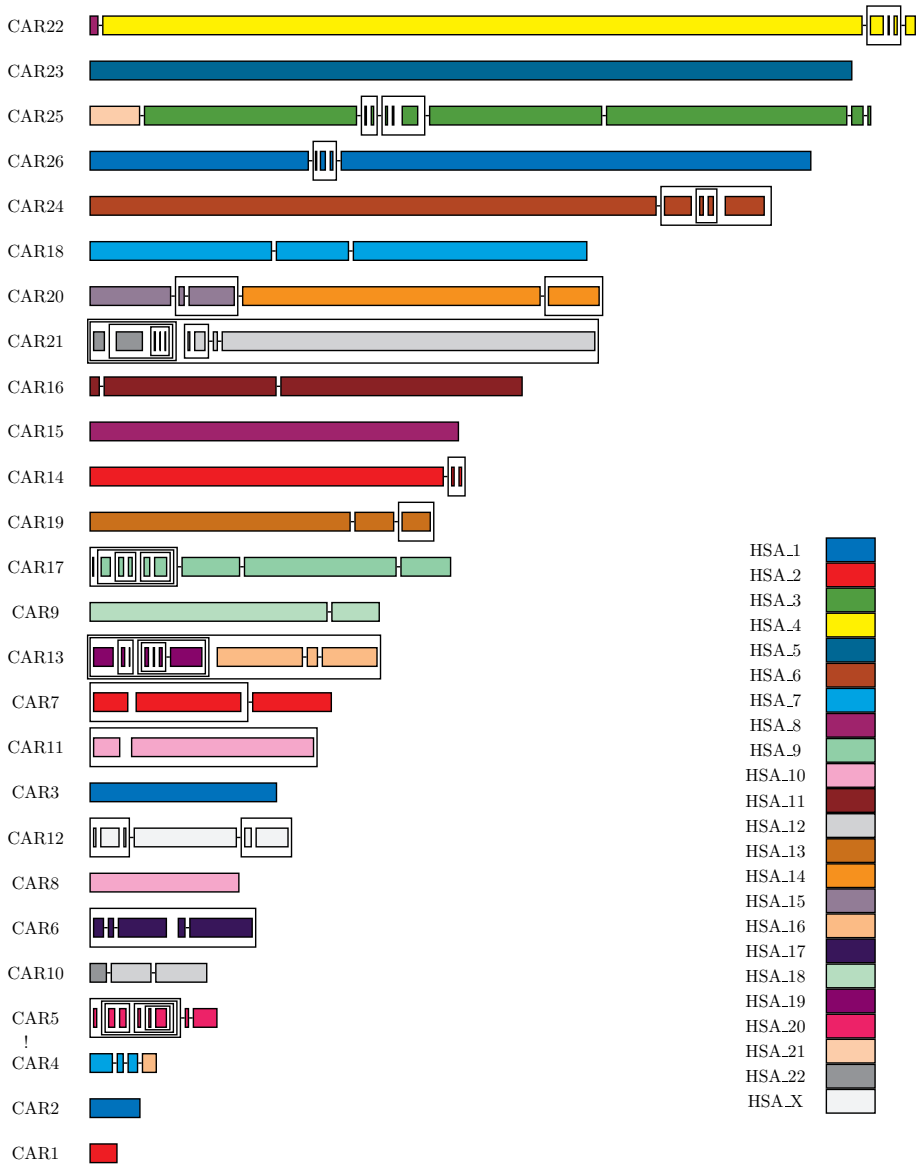

Figure 5: The ancestral genome architecture obtained with the dataset constructed from the UCSC whole genome alignements, with parameters max_gap = 100kb and min_len = 400kb. Segments of a given color represent sequences of genomic markers that are colinear in the inferred CARs and in a human chromosome (colors correspondences are given in the bottom right part of the figure), called *conserved segments*. There are 111 such conserved segments. The size of conserved segments in the figure is proportional to the sum of the sizes, in the human genome, of the synteny blocks they contain. The nodes of the PQ-tree are represented: children of a linear (Q) node are linked by a small segment, while children of a prime (P) node are grouped together with a rectangular frame.

It can be seen that most of the differences between the two methods is due to adjacencies that are obtained with the method described in [39] but are not supported by an ancestral synteny as we define them. We also notice that a very small number of differences may have some important implications in terms of inferred chromosomal syntenic associations between human chromosomes in the ancestral genome.

min_len (kb)	CARs	Weak adj.	Common adj.	Human chromosomal syntenies
100	33	9	2629	1-16-19, 3-21, 4-8, 12-22, 12-22, 12-22
200	33	8	1546	1-10, 1-17, 3-21, 4-8, 7-16, 12-22, 12-22
300	36	9	1004	1-10, 1-17, 3-12-21, 4-8, 12-22, 12-22
400	36	7	763	2-4, 2-22, 3-12-21, 12-22
500	36	6	552	2-4, 2-22, 3-12-21, 12-22

Table 3: Characteristics of the reconstructed genome architectures with the method of Ma *et al.* [39] and our synteny blocks.

It is interesting to see that both methods agree on a large majority of syntenic features. This suggests that using a more strict way to define ancestral syntenic features (as we exclude adjacencies that are present only in one group of extant genomes and we exclude conflicting adjacencies), is compensated by considering larger syntenies (common intervals instead of only oriented adjacencies), at the price of a slightly less well defined ancestral architecture (due to the P-nodes of the PQ-tree). Moereover, this suggests that the different architectures we propose should be compared mostly around the few adajcencies that are not in agreement, and are involved in several important human chromosomal syntenies.

The ferungulate ancestor from Murphy *et al.* synteny blocks

We also tested our framework on the ferungulate ancestor on the dataset of Murphy *et al.* [45]. This dataset contains seven genomes, that are represented by 307 synteny blocks that cover 1343Mb of human genome [45, Table S2]. It is hazardous to reconstruct boreoeutherian ancestors with this dataset, because there is no outgroup for the boreoeutherian clade here, but it is interesting to use this dataset to make a comparison between several methods on a dataset we did not construct. We ran both our method and Ma et al [39] method on this dataset and compared the three inferred genome architectures (including the results obtained by Murphy *et al.* [45] on the same dataset, and those of Kemkemer *et al.* [35] obtained independently by a method called E-painting, see Table 4). The ancestral genome architecture we propose is based on 457 ancestral syntenies from an initial number of 461, and here again the dataset seems to contain very little ambiguity.

We can comment on some differences between the results obtained with our method and by the other methods, especially in terms of syntenies that seem to be ferungulate-specific. The synteny between human chromosomes 5 and 19 is inferred only by Murphy *et al.* (where it is not marked as weak, which means that it was found in all alternative genome architectures) but not by our method. However, it is due to an adjacency between two synteny blocks that is not found in any of the ancestral synteny we detected in the first step of our

Method	CARs	Adjacencies	Human chromosomal syntenies
New method	24	250	1-10, 3-21, 4-8, 7-16, 12-22, 12-22, 14-15, 16-19
Ma *et al.*	38	269	2-7-16, 3-21, 4-8, 12-22, 14-15, 16-19
Murphy *et al.*	24	283	1-10, 1-22, 2-20, 3-21, 4-8, 5-19, 7-16, 12-22, 14-15, 16-19
Kemkemer *et al.*	23	-	1-3-19-21, 4-8, 7-16, 12-22, 14-15, 16-19

Table 4: Characteristics of three inferred ferungulate genomic architectures.

method, and is in fact found only in the pig genome. The synteny between human chromosomes 1 and 22 is inferred only by Murphy *et al.*, where it is marked as weak. It is due to an adjacency that is found in no genome, neither supported by none of our ancestral syntenies. The same holds for the synteny between human chromosomes 2 and 20 (that is not weak according to Murphy *et al.*), and seems to be more rodent-specific in fact. The synteny between human chromosomes 1 and 10 was inferred by the two methods, and considered weak by Murphy *et al.*, and is supported by three of our ancestral syntenies that have significant weights. The synteny between human chromosomes 2 and 7, that is found only by the method of Ma *et* al. is due to an adjacency that is found only in the pig and is not supported by any of our ancestral syntenies. We can also note that among the 250 adjacencies inferred by our method, only 196 are common with the results obtained with the methods of Ma *et al.* and Murphy *et al.*, while 240 are common with the ancestor obtained with the method of Ma *et al.* and 204 are common with the ancestor proposed by Murphy *et al.*. We have only the boreoeutherian syntenies in common with Kemkemer *et al.* [35], and those that are supposed to be ferungulate specific all disagree (we don't recover the giant chromosome 1-19-3-21, and recover 1-10 instead). But in spite of these divergences, it is still the closest proposed ancestor from ours.

Discussion

We proposed a general model-free framework for reconstructing ancestral genome architectures from current genomic markers orders. We implemented this framework in a method that considers adjacencies and gene teams in extant genomes with no duplicated markers and applied our method on two ancestral genome reconstruction problems: the boreoeutherian ancestor, from a set of synteny blocks we computed from UCSC whole genome alignments [34], and the ferungulate ancestor from the synteny blocks defined in [45]. We believe that these experimental results we obtain mark a progress compared to previous bioinformatics studies.

Convergences and divergences of the ancestral genome reconstruction methods

In [27, 15, 53], a controversy was engaged, about the divergences between so-called *bioinformatics* and *cytogenetics* methods. Based on the described bioinformatics framework, we would like to emphasize here that those divergences are probably not due to disciplinary problems, or to the differences in data acquisitions, but in the methodologies employed to treat genomic data. The

comparisons we made argue for this: all bioinformatics reconstructions are different, and we are often much closer to cytogenetics results than to other bioinformatics studies. Indeed, for the boreoeutherian ancestor, Ma *et al.* [39], with their own set of synteny blocks (called there orthology blocks) recovered 29 CARs, with several "weak adjacencies". Those adjacencies correspond to features that are not supported by at least two species which evolutionary path along the phylogenetic tree goes through the boreoeutherian ancestor. This means several adjacencies are only present in human and mouse for example, which would more account for an euarchontoglire feature, or even only in human (as the junction of both parts of human chromosomes 10 or 16 for example, with $min_len = 400$kb). In contrast, we infer 26 CARs, which is comparable, but with no such weakly supported adjacency, and this is clearly visible as all our chromosomal syntenies are also supported by cytogenetic studies, but the fusion of a synteny block of human chromosome 4 with a segment of human chromosome 1 we see when $min_len = 100$kb, that probably results from a false ortholog due to the low resolution with this value of min_len. Moreover, the method of Ma *et al.* gives 35 CARs on our dataset, with a significant number of weak adjacencies. We think that this points out that the difference between our two methods is more due to methodological reasons, mostly the way ancestral syntenies are defined (through a Fitch-like approach in [39]), than to the dataset itself (the way we compute synteny blocks are very similar, even if we conserve only blocks that are present in all genomes). It is not surprising since both methods are well comparable: we use less adjacencies (we ask for more support and less conflicts) but more features (we add gene teams as ancestral syntenies), that have to be supported by at least two species. In our opinion, this comparison between our method and Ma *et al.* method suggests that the framework we propose is a useful tool to compare different methods.

To assess if the differences in the published results, that were discussed in [27, 15, 53] for example, are more due to methods than to datasets, we also tested our method on the ferungulate ancestor and compared our results with the ancestor inferred through a model-based method in Murphy *et al.* [45]. With the method Murphy *et al.* used, based on a genome rearrangement model and MGR [13], the results diverged from the cytogenetics data and provoked the discussion in [27, 15, 53]. Using the same synteny blocks than Murphy *et al.*, we found 24 CARs, all of which are chromosomes of the boreoeutherian ancestor, except a fusion of the homologs of human chromosomes 1 and 10, which seem to be ferungulate-specific, and was also inferred by MGR. None of the other chromosomal syntenies proposed by [45] were recovered by our method, or Ma *et al.* method. However, the number of common inferred ancestral adjacencies points out that our method and the method of Ma *et al.* compute quite similar ancestral genome architectures, that are quite different from the one proposed by MGR, despite the fact that this last one has 24 CARs, as with our method. We believe that this three-way comparison clearly indicates that the differences discussed in [27, 15] are more due to the methods themselves, and more precisely to the fact that MGR is a rearrangement-based method, and should then be considered carefully, at least when it comes to propose ancestral genome architectures.

Methodological comments

We now summarize the main methodological features of the framework we propose, and discuss them and possible extensions. We propose to decompose the process of ancestral genome architecture inference into three steps: detection and weighting of ancestral syntenies, representation as a 0/1 matrix and a generalized PQ-tree, clearing ambiguities and representation of a set of alternative genome architectures as a PQ-tree. Although these three steps are performed independently, the implementation choices for each of them can have important consequences on the other ones, as we discuss below. We implemented this method using (1) unique and universal synteny blocks, that appear once in each genome, (2) ancestral syntenies defined as unambiguous adjacencies and maximum common intervals (or gene teams) that are present in at least two genomes whose last common ancestor is the considered ancestral species and (3) a natural combinatorial optimization approach, based on the Consecutive Ones Submatrix Problem, to clear ambiguities.

Handling duplicated and non universal markers. Though we do not use this possibility here, the framework we propose does not forbid using duplicated or genomic markers that are not present in every extant genome. Indeed, the only question that is raised by having duplicated markers is the question of detecting possible ancestral syntenies. There are several algorithms that allow to compute efficiently conserved syntenic groups between pairs of genomes with duplicated markers (see a survey in [9] for example), or duplicated segments followed by intensive losses in both copies (see [21]), that could be used instead of the algorithm to detect gene teams we used. However, what is compulsory is that each marker appears at most once in the wished ancestor; indeed, otherwise we cannot use anymore tools such as the notion of consecutive ones property of 0/1 matrices and PQ-trees, which are central in our framework. From that point of view, it would be interesting to extend our approach to problems of inferring a pre-duplication ancestral genome architecture, that has been considered in some model-based recent works [25, 2, 54]. Our final dataset contains only the markers that are present in every species of the study (the universal markers). This gives better results than taking all markers, or markers that are present in every species except outgroups. Indeed, the optimization step is more consuming, and the syntenic associations between human chromosomes 7-16, and 9-16 are not recovered in the boreoeutherian ancestor (experimental results are available on the companion website). If markers are missing in some species, probably a framework that would allow more flexibility as gaps in the reconstruction would be more suitable. Here we lose some coverage of the genomes, but gain in accuracy.

Detecting ancestral syntenies. We emphasize that, in our opinion, the first step, that aims at computing a set of syntenic groups that are possibly ancestral, is essentially a detection phase and does not require to rely on combinatorial optimization. This is not the case of existing methods that rely on methods inspired from the Fitch-Hartigan algorithm, as in [8, 1, 39]. These methods implicitly try to minimize the number of gains and losses of features along the species tree T, following then a parsimony model of evolution that can be very

sensitive to the branching pattern of T. Weighting characters is a possibly more flexible approach to assess the conservation of syntenic characters.

Definition of ancestral syntenies, 0/1 matrices and PQ-trees. The link between the combinatorics of PQ-trees and 0/1 matrices is the main limitation of our approach, as it prevent some flexibility for the detection of ancestral syntenic features. For example, some common features of extant species are not captured by common intervals (gene team [4] with no gaps). We would probably detect a significant amount of approximate ancestral syntenies by considering some amount of gaps in the detection phase [48, 9]. But the combinatorial nature of the reconstruction phase radically changes in this case, as naturally we would like then to consider possible gaps in the rows of the 0/1 matrix that represents ancestral syntenies after reordering the columns of this matrix. When considering only 0/1 matrices, related problems have been considered as in [23], but they are not related any more to PQ-trees, that are important as they represent a set of alternative ancestral genome architectures, an important property of the framework we propose. The decision problem of "consecutive ones with allowed gaps", where each line of the matrix has to have consecutive ones except that between each pair of ones, a fixed number of zeros are allowed, is the one that is closer to the gene teams formalism, and is still open. It relates to bandwidth in graphs [17], where it has a polynomial solution for maximum gaps of 2, but no generalization is known. There is then still an important theoretical work to do on the combinatorics of PQ-trees and of their extension to non-contiguous ancestral syntenies, that would be important to implement the framework we propose in order to handle more ancient and more rearranged genomes.

Clearing ambiguities in ancestral syntenies. In the method we propose, we decided to clear ambiguities in the set of detected ancestral syntenies by discarding the minimum amount (in terms of weight) of such syntenies in order to have a C1P matrix and then a PQ-tree. In fact we then made two choices: removing the minimum amount of information, and considering only rows of the matrix for being discarded.

The bias induced by choosing to apply a combinatorial optimization approach (that can also be seen as following a local parsimony principle), is that we are likely to conserve, in the resulting matrix, false positive ancestral syntenies (for example if there are two false positive ancestral syntenies that have the same weight, and the presence of both contradicts the consecutive ones property, but not the presence of either of the two). Another approach was described in [8], where the notion of *conflicting set* of syntenies was defined as a set of syntenies that is ambiguous but such that discarding any of them leaves a non-ambiguous set of syntenies. It was then proposed to discard all syntenies of such a group. This is what we do with adjacencies in the first step of our method, mostly because such conflicting sets are easy to detect with adjacencies, unlike with common intervals, and because we expect that true ancestral adjacencies should also be supported by larger syntenies that will be detected as maximum common intervals. With our data, such an approach would have been very extreme, as a preliminary studies of ancestral syntenies that belong to the R-nodes of the generalized PQ-tree showed that almost half of such ancestral

syntenies belonged to at least one conflicting set (data not shown). However, using a sampling method, it seems that only very few of these syntenies belong to many conflicting sets. It would then be interesting to apply a cut-off approach where all ancestral syntenies that belong to a large proportion of the conflicting sets present in a given R-node are discarded. However, to implement such an approach, the combinatorics of conflicting sets with general 0/1 matrices needs to be better understood (work in progress).

The second choice we made is the optimization criterion. There are several ways to handle conflicts in a 0/1 matrix that is not C1P (see [22] for example): removing rows (i.e. ancestral syntenies), columns (genomic markers), splitting rows (to account for possible chimeric ancestral syntenies) or even reverting some cells from 0 to 1 or 1 to 0 (to account for approximate syntenies). It is important to notice that choosing one of these approaches should be related to the nature of the errors expected to be found in the set of ancestral syntenies (see [29] for an example of this principle in the case of physical mapping). Based on our definition of genomics markers as synteny blocks computed from whole genome alignments using quite stringent criteria, we considered that orthology relations were correct (even if we found one possible false positive with min_len = 100kb), which did not justify to remove columns. Similarly using maximum common intervals, that is genome segments with the same content, prevents from expecting to have to deal with reverting cells of the matrix. Finally, in the case of chimeric ancestral syntenies (i.e. groups of two or more syntenies joined by convergent evolution), we expect that the individual syntenies that compose them will be detected as well, and then we just need to remove the row corresponding to a chimeric synteny. However, depending on the nature of the data, one could very well consider other optimization criteria: for example, with genomic markers defined using virtual hybridation [5], or when considering duplicated genomic markers that represent ambiguous orthology relations, it would be natural to consider discarding columns of the matrix.

Sensitivity to parameters. The first step of the method (detecting ancestral syntenies) captures more information as the resolution goes down (from 100kb to 500kb). So we are able to handle a resolution of 100kb, but our best results are obtained for $max\_gap = 100kb$ and $min\_len = 400kb$. This is probably because at low resolution, the orthology and synteny signals are still perturbed by all kinds of duplications and repetitions. Increasing the parameter min_len makes the number of CARs decrease, but apart from this, the method is stable, in the sense that it recovers the same basic set of adjacencies for all choices of markers. We also tested the sensitivity to branch lengths, and no results were altered by taking for example the branch lengths proposed by Ma *et al* [39], based on an *a priori* amount of rearrangements that is expected in each branch. The method of Ma *et al* [39], which we tested with the same parameter variability, was not as stable, due to the importance of its optimization step, which may give very different results with similar values.

Material and methods

Computing orthologous markers from whole genome alignments. We construct several datasets, by a unique method depending on two parameters, *max_gap* and *min_len*. This method, or very similar ones, are often used to construct synteny blocks from genomic alignments [19, 55, 13].

- We first downloaded the chained netted pairwise alignments from the UCSC Genome Bioinformatics site [34] and the coordinates of all the alignments of the human genome (build hg18, March 2006 [32]) against respectively macaca (build rheMac2, January 2006 [51]), mouse (build mm9, July 2007 [43]), rat (build rn4, November 2004 [50]), cow (build bosTau3, August 2006), dog (build canFam2, May 2005 [37]), chicken (build galGal3, May 2006 [33]) and opossum (build monDom4, January 2006 [42]);
- For each set of alignments between the human genome and another genome, a graph is built, with vertices being the above alignments and edges joining two alignments if they have the same direction, and if they are not more distant than *max_gap*, a user-defined parameter (here 100kb), in both genomes;
- Pairwise synteny blocks were defined as connected components of the above graphs that span a size of at least *min_len* of both genomes;
- The previous steps give a collection of pairwise breakpoints, with coordinates in the human genome. By considering all these breakpoints together, taking the union of those that intersect, we ended up with markers common to subsets of species, with their coordinates on the human genome and arrangements in all species, as sequences of markers (the chromosomes). We discarded the alignments that spanned less than 50kb of the human genome, and those which were at least 80% covered by segmental duplications. The coordinates of segmental duplications were also downloaded from the UCSC Genome Browser [34].

Ancestral features: gene teams and adjacencies. We first use the notion of "teams of markers" [38]. This notion relies on a parameter δ, a positive integer. In a genome, the *position* of a marker is its rank from on the sequence of its chromosome. That is, the first marker on a chromosome has rank 1, the second has rank 2, and so on. The position of a marker m on a chromosome is denoted by $p(m)$. Two markers m_1 and m_2 are said to be *close* to each other in a genome, for the parameter δ, if they lie on the same chromosome, and $|p(m_1) - p(m_2)| \leq \delta$. A subset of markers M is said to be a *team* for a genome if for any two markers a, b from M, there exists a sequence $S = a, a_1, \ldots, a_k, b$ of markers from M, such that any two consecutive markers in S are close to each other. Given two genomes X and Y, a *team* S common to X and Y is a set of markers labels (a subset of Σ the alphabet of markers) that is a team in both genomes X and Y. Such a team S is *maximal* if no other team is common to X and Y and contains S. Maximal common intervals are maximal common teams for $\delta = 1$. Maximal common teams can be computed efficiently thanks to an algorithm by Beal and al. [4] and a software described in [38]. We collect a set of teams, representing possible ancestral syntenies, by computing all maximal common teams of pairs of species which evolutionary path contains the wished ancestor.

As teams rely only on similarity in markers content, and do not involve any markers order constraints, we added to this set of ancestral syntenies a set the set of putative ancestral adjacencies, defined as pairs of markers that are consecutive in at least two genomes which evolutionary path contains this ancestor and do not belong to a conflict. A conflict is defined as follows [39, Figure 7]: an adjacency $\{i, j\}$ belongs to a conflict if, in the graph G whose vertices are the markers ($V(G) = \Sigma$) and the edges are the conserved adjacencies, either i or j has degree more than 2, or the edge $\{i, j\}$ belongs to a cycle.

Each of these ancestral syntenies was weighted following the same principle than in [39]. Let S be a subset of Σ that represents a possible ancestral synteny. In any leaf X of the species tree, if S is a team in X, the weight of S in X is $w_X(S) = 1$, otherwise, $w_X(S) = 0$. Then, in any internal node N of T (other than the ancestral node A) having two children R and L, $w_N(S)$ is defined recursively by the formula

$$w_N(S) = \frac{d_L w_R(S) + d_R w_L(S)}{d_L + d_R}$$

where d_L and d_R are respectively the length of the branch between N and L and N and R. The weight of S in A is then defined by

$$w_A(S) = \frac{1}{3}\left(\frac{d_{A_1} w_{A_2}(S) + d_{A_2} w_{A_1}(S)}{d_{A_1} + d_{A_2}} + \frac{d_{A_1} w_{A_3}(S) + d_{A_3} w_{A_1}(S)}{d_{A_1} + d_{A_3}} + \frac{d_{A_2} w_{A_3}(S) + d_{A_3} w_{A_2}(S)}{d_{A_2} + d_{A_2}}\right)$$

where A_1, A_2 and A_3 are the three neighbors of the ancestral node A in T, and d_{A_1}, d_{A_2} and d_{A_3} are the respective length of the branch between A and A_1, A and A_2 and A and A_3.

Construction of the generalized PQ-tree. Recall $\mathcal{L}$ is the set of homologous markers, $\mathcal{S}$ is the set of subsets of $\mathcal{L}$ that represent possible ancestral syntenies and $\mathcal{M}$ the corresponding 0/1 matrix.

We say that two elements S_i and S_j of $\mathcal{S}$ *overlap* if their intersection is not empty, but none is included in the other. Let $\mathcal{N}(\mathcal{S})$ be the family of all subsets of $\mathcal{L}$ that do not overlap with any member of $\mathcal{S}$; in other words, given X an element of $\mathcal{N}(\mathcal{S})$, any S_i of $\mathcal{S}$ either contains all elements of X or contains no element of X. Among the subsets of $\mathcal{N}(\mathcal{S})$, call *strong* the elements that do not overlap any other elements of $\mathcal{N}(\mathcal{S})$. The *inclusion tree* of the strong elements of $\mathcal{N}(\mathcal{S})$, denoted $I(\mathcal{N}(\mathcal{S}))$, is a tree where each strong element of $\mathcal{N}(\mathcal{S})$ corresponds to a single node and the node corresponding to a strong subset X is an ancestor of the node corresponding to a strong subset Y if and only if X contains Y as a subset.

Given a node N of $I(\mathcal{N}(\mathcal{S}))$, we associate to it the subset $s(N)$ of the elements of $\mathcal{S}$ defined as all S_i's that are included in N but in none of its children. The PQ-tree $T(\mathcal{M})$ is defined from $I(\mathcal{N}(\mathcal{S}))$ as follows: an internal node N such that $s(N) = \emptyset$ is a P-node, while an internal node N such that $s(N) \neq \emptyset$ is a Q-node if $s(N)$ can be partitioned by a partition refinement process [30] and a R-node otherwise. The construction of $T(\mathcal{M})$ can be achieved in optimal $O(n + m)$ time where $|\mathcal{L}| = n$ and $|\mathcal{S}| = m$, as described in [40].

Algorithms for clearing ambiguities in ancestral syntenies. In the last step, we want to remove the minimal amount (in terms of weight) of ancestral syntenies from $\mathcal{S}$ in order that the resulting matrix $\mathcal{M}'$ is C1P. This problem, that is known as the Consecutive Ones Submatrix Problem generalizes the Maximum Path Cover Problem used in [39] and is known to be NP-hard [31] even for sparse matrices [59], which is the case of the matrices we obtain. However, using the structural information given by the PQ-tree $T(\mathcal{M})$, it is possible to design an efficient branch-and-bound algorithm.

More precisely, it follows immediately from the definition of $T(\mathcal{M})$ that ambiguous information that prevent a matrix $\mathcal{M}$ to be C1P can only be located in the submatrices defined by the subsets $s(N)$ of $\mathcal{S}$ for the degenerate nodes of $T(\mathcal{M})$. Hence each of these subsets of $\mathcal{S}$ can be processed independently of the remaining of $\mathcal{S}$. For such a subset, say $s(N) = \{S_{i_1}, \ldots, S_{i_k}\}$, we first compute an upper bound on the maximum subset S of $s(N)$ that defines a matrix that is C1P, using the same approach than in [39]: start with $S = \emptyset$ and, for each element S_{i_j} of $s(N)$, taken in decreasing order of weight, if adding S_{i_j} to S defines a matrix that is not C1P (which can be tested

using the efficient algorithms described in [40, 30]), then discard it, else leave it in S. From that upper bound, using the same principle, we use a classical branch-and-bound algorithm that looks for a better subset of $s(N)$ that defines a C1P matrix.

Acknowledgments

Eric Tannier is funded by the Agence Nationale de la Recherche (GIP ANR JC05_49162 and NT05-3_45205) and by the Centre National de la Recherche Scientifique. Cedric Chauve is funded by a Discovery grant from National Science and Engineering Research Council (NSERC) of Canada and a Simon Fraser University Startup Grant.

References

[1] Z. Adam, M. Turmel, C. Lemieux, and D. Sankoff. Common intervals and symmetric difference in a model-free phylogenomics, with an application to streptophyte evolution. *Journal of Computational Biology*, 14:436–445, 2007.

[2] M. Alekseyev and P. Pevzner. Colored de bruijn graphs and the genome halving problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4:98–107, 2007.

[3] F. Alizadeh, R. Karp, D. Weisser, and G. Zweig. Physical mapping of chromosomes using unique probes. *Journal of Computational Biology*, 2:159–184, 1995.

[4] M.-P. Beal, A. Bergeron, S. Corteel, and M. Raffinot. An algorithmic view of gene teams. *Theoretical Computer Science*, 320:395–418, 2004.

[5] M. Belcaid, A. Bergeron, A. Chateau, C. Chauve, Y. Gingras, G. Poisson, and M. Vendette. Exploring genome rearrangements using virtual hybridization. In D. Sankoff, L. Wang, and F. Chin, editors, *Proceedings of the 5th Asia-Pacific Bioinformatics Conference*, volume 5 of *Series on Advances in Bioinformatics and Computational Biology*, pages 205–214. Imperial College Press, 2007.

[6] M. J. Benton and P. C. J. Donoghue. Paleontological evidence to date the tree of life. *Molecular Biology and Evolution*, 24:26–53, 2007.

[7] S. Bérard, A. Bergeron, C. Chauve, and C. Paul. Perfect sorting by reversals is not always difficult. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4:4–16, 2007.

[8] A. Bergeron, M. Blanchette, A. Chateau, and C. Chauve. Reconstructing ancestral genomes using conserved intervals. In I. Jonassen and J. Kim, editors, *Algorithms in Bioinformatics*, volume 3240 of *Lecture Notes in Computer Science*, pages 14–25. Springer-Verlag, 2004.

[9] A. Bergeron, C. Chauve, and Y. Gingras. *Bioinformatics Algorithms: Techniques and Applications (A. Zelikovsky and I. Mandoiu, editors)*, chapter

Formal models of gene clusters. Wiley Series on Bioinformatics: Computational Techniques and Engineering. Wiley Interscience, to appear.

[10] A. Bhutkar, W. Gelbart, and T. Smith. Inferring genome-scale rearrangement phylogeny and ancestral gene order: a *Drosophilia* case study. *Genome Biology*, 8:R236, 2007.

[11] M. Blanchette, G. Bourque, and D. sankoff. Breakpoint phylogenies. *Genome Informatics Series Workshop Genome Informatics*, 8:25–34, 1997.

[12] K. Booth and G. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using *PQ*-tree algorithms. *Journal of Computer and System Science*, 13:335–379, 1976.

[13] G. Bourque and P. Pevzner. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Research*, 12:26–36, 2002.

[14] G. Bourque, P. Pevzner, and G. Tesler. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse and rat genomes. *Genome Research*, 14:507–516, 2004.

[15] G. Bourque, G. Tesler, and P. Pevzner. The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction. *Genome Research*, 16:311–313, 2006.

[16] G. Bourque, E. Zdobnov, P. Bork, P. Pevzner, and G. Tesler. Comparative architectures of mammalian and chicken genomes reveal highly rates of genomic rearrangements across different lineages. *Genome Research*, 15:98–110, 2005.

[17] A. Caprara, F. Malucelli, and D. Pretolani. On bandwidth-2 graphs. *Discrete Applied Mathematics*, 117:1–13, 2002.

[18] T. Christof, M. Jnger, J. Kececioglu, P. Mutzel, and G. Reinelt. A branch-and-cut approach to physical mapping of chromosome by unique end-probes. *Journal of Computational Biology*, 4:433–447, 1997.

[19] Ensembl comarative genomics database. http://www.ensembl.org/info/about/docs/compara/index.html.

[20] A. Darling, B. Mau, F. Blattner, and N. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14:1394–1403, 2004.

[21] Y. V. de Peer. Computational approaches to unveiling ancient genome duplications. *Nature Reviews*, 5:752–763, 2004.

[22] M. Dom, J. Guo, and R. Niedermeier. Approximability and parameterized complexity of consecutive ones submatrix problems. In J. Cai, S. Cooper, and H. Zhu, editors, *Theory and Applications of Models of Computation, 4th International Conference, TAMC 2007, Shanghai, China, May 22-25, 2007, Proceedings*, volume 4484 of *Lecture Notes in Computer Science*, pages 680–691. Springer, 2007.

[23] M. Dom, J. Guo, R. Niedermeier, and S. Wernicke. Minimum membership set covering and the consecutive ones property. In L. Arge and R. Freivalds, editors, *Algorithm Theory - SWAT 2006*, volume 4059 of *Lecture Notes in Computer Science*, pages 339–350. Springer-Verlag, 2006.

[24] J. Earnest-DeYoung, E. Lerat, and B. Moret. Reversing gene erosion: Reconstructing ancestral bacterial genomes from gene-content and order data. In I. Jonassen and J. Kim, editors, *Algorithms in Bioinformatics*, volume 3240 of *Lecture Notes in Computer Science*, pages 1–13. Springer-Verlag, 2004.

[25] N. El-Mabrouk and D. Sankoff. The reconstruction of doubled genomes. *SIAM Journal on Computing*, 32:754–792, 2003.

[26] T. Faraut. Addressing chromosome evolution in the whole-genome sequence era. *Chromosome Research*, 16:5–16, 2008.

[27] L. Froenicke, M. G. Caldés, A. Graphodatsky, S. Mueller, L. Lyons, T. Robinson, M. Volleth, F. Yang, and J. Wienberg. Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes? *Genome Research*, 16:306–310, 2006.

[28] L. Froenicke, J. Wienberg, G. Stone, L. Adams, and R. Stanyon. Towards the delineation of the ancestral eutherian genome organization: comparative genome maps of human and the African elephant (loxodonta africana) generated by chromsome painting. *Proceeding of the Royal Society of London*, 270:1331–1340, 2003.

[29] P. Goldberg, M. Golumbic, H. Kaplan, and R. Shamir. Four strikes against physical mapping of dna. *Journal of Computational Biology*, 2:139–152, 1995.

[30] M. Habib, R. McConnell, C. Paul, and L. Viennot. Lex-bfs and partition refinement, with applications to transitive orientation. *Theoretical Computer Science*, 234:59–84, 2000.

[31] M. Hajiaghayi and Y. Ganjali. A note on the consecutive ones submatrix problem. *Information Processing Letters*, 83:163–166, 2002.

[32] Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[33] International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432:695–716, 2004.

[34] D. Karolchik, R. Kuhn, R. Baertsch, G. Barber, H. Clawson, M. Diekhans, B. Giardine, R. Harte, A. Hinrichs, and F. H. et al. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res*, 36:D773–D779, 2007.

[35] C. Kemkemer, M. Kohn, H. Kehrer-Sawatzki, P. Minich, J. Högel, L. Froenicke, and H. Hameister. Reconstruction of the ancestral ferungulate karyotype by electronic chromosome painting (E-painting). *Chromosome Research*, 14:899–907, 2006.

[36] G. Landau, L. Parida, and O. Weimann. Gene proximity analysis across whole genomes via pq trees. *Journal of Computational Biology*, 12:1289–1306, 2005.

[37] K. Lindblad-Toh, C. Wade, T. Mikkelsen, E. Karlsson, D. Jaffe, M. Kamal, M. Clamp, J. Chang, E. K. III, and M. Z. et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438:803–819, 2005.

[38] N. Luc, J.-L. Risler, A. Bergeron, and M. Raffinot. Gene teams: a new formalization of gene clusters for comparative genomics. *Computational Biology and Chemistry*, 27:59–67, 2003.

[39] J. Ma, L. Zhang, B. Suh, B. Rany, R. Burhans, W. Kent, M. Blanchette, D. Haussler, and W. Miller. Reconstructing contiguous regions of an ancestral genome. *Genome Research*, 16:1557–1565, 2006.

[40] R. McConnell. A certifying algorithm for the consecutive-ones property. In J. Munro, editor, *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 761–770. Society for Industrial and Applied Mathematics, 2004.

[41] J. Meidanis, O. Porto, and G. Telles. On the consecutive ones property. *Discrete Applied Mathematics*, 88:325–354, 1998.

[42] T. Mikkelsen, M. Wakefield, B. Aken, C. Amemiya, J. Chang, S. Duke, M. Garber, A. Gentles, L. Goodstadt, and A. H. et al. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature*, 447:167–177, 2007.

[43] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

[44] M. Muffato and H. R. Crollius. Paleogenomics, or the recovery of lost genomes from the mist of times. *BioEssays*, 30:122–134, 2008.

[45] W. Murphy, D. Larkin, A. E. van der Wind, G. Bourque, G. Tesler, L. Auvil, J. Beever, B. Chowdhary, F. Galibert, and L. G. et al. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309:613–617, 2005.

[46] Y. Nakatani, H. Takeda, and S. Morishita. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Research*, 17:1254–1265, 2007.

[47] L. Parida. Using PQ structures for genomic rearrangement phylogeny. *Journal of Computational Biology*, 13:1685–1700, 2006.

[48] S. Pasek, A. Bergeron, J.-L. Risler, A. Louis, E. Ollivier, and M. Raffinot. Identification of genomic features using microsyntenies of domains: Domain teams. *Genome Research*, 15:867–874, 2005.

[49] V. L. Rascol, P. Pontarotti, and A. Levasseur. Ancestral animal genomes reconstruction. *Current Opinions in Immunology*, 19:542–546, 2007.

[50] Rat Genome Sequencing Consortium. Genome sequence of the Brown Norway rat yields insight into mammalian evolution. *Nature*, 428:493–521, 2004.

[51] Rhesus Macaque Genome Sequencing and Analysis Consortium. Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, 316:222–234, 2007.

[52] F. Richard, M. Lombard, and B. Dutrillaux. Reconstruction of the ancestral karyotype of eutherian mammals. *Chromosome Research*, 11:605–618, 2003.

[53] M. Rocchi, N. Archidiacono, and R. Stanyon. Ancestral genome reconstruction: An integrated, multi-disciplinary approach is needed. *Genome Research*, 16:1441 – 1444, 2006.

[54] D. Sankoff, C. Zheng, and Q. Zhu. Polyploids, genome halving and phylogeny. *Bioinformatics*, 23:i433–i439, 2007.

[55] A. Sinha and J. Meller. Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, 8(82), 2007.

[56] M. Svartman, G. Stone, J. Page, and R. Stanyon. A chromosome painting test of the basal eutherian karyotype. *Chromosome Research*, 12:45–53, 2004.

[57] M. Svartman, G. Stone, and R. Stanyon. The ancestral eutherian karyotype is present in xenarthra. *PLoS Genetics*, 2:e109, 2006.

[58] F. Swidan, E. Rocha, M. Shmoish, and R. Pinter. An integrative method for accurate comparative genome mapping. *PLoS Computational Biology*, 2:e75, 2006.

[59] J. Tang and L. Zhang. The consecutive ones submatrix problem for sparse matrices. *Algorithmica*, 48:287–299, 2007.

[60] J. Wienberg. The evolution of eutherian chromosomes. *Current Opinion in Genetics and Development*, 14:657–666, 2004.

[61] R. Wittler. ROCI; - reconstruction of conserved intervals. http://bibiserv.techfak.uni-bielefeld.de/roci/.

[62] F. Yang, E. Alkalaeva, P. Perelman, A. Pardini, W. Harrison, P. O'Brien, B. Fu, A. Graphodatsky, M. Ferguson-Smith, and T. Robinson. Reciprocal chromsome painting among human, aardvark, and elephant (superorder afrotheria) reveals the likely eutherian ancestral karyotype. *Proceedings of the National Academy of Science of the United States of America*, 100:1062–1066, 2003.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399