JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 17, Number 9, 2010
© Mary Ann Liebert, Inc.
Pp. 1097–1112
DOI: 10.1089/cmb.2010.0092

Yeast Ancestral Genome Reconstructions: The Possibilities of Computational Methods II

CEDRIC CHAUVE,[1] HARIS GAVRANOVIC,[2] AIDA OUANGRAOUA,[3]
and ERIC TANNIER[4]

ABSTRACT

Since the availability of assembled eukaryotic genomes, the first one being a budding yeast, many computational methods for the reconstruction of ancestral karyotypes and gene orders have been developed. The difficulty has always been to assess their reliability, since we often miss a good knowledge of the true ancestral genomes to compare their results to, as well as a good knowledge of the evolutionary mechanisms to test them on realistic simulated data. In this study, we propose some measures of reliability of several kinds of methods, and apply them to infer and analyse the architectures of two ancestral yeast genomes, based on the sequence of seven assembled extant ones. The pre-duplication common ancestor of *S. cerevisiae* and *C. glabrata* has been inferred manually by Gordon et al. (*Plos Genet*. 2009). We show why, in this case, a good convergence of the methods is explained by some properties of the data, and why results are reliable. In another study, Jean et al. (*J. Comput Biol.* 2009) proposed an ancestral architecture of the last common ancestor of *S. kluyveri*, *K. thermotolerans*, *K. lactis*, *A. gossypii*, and *Z. rouxii* inferred by a computational method. In this case, we show that the dataset does not seem to contain enough information to infer a reliable architecture, and we construct a higher resolution dataset which gives a good reliability on a new ancestral configuration.

Key words: algorithms, combinatorics, computational molecular biology, genomic rearrangements.

1. INTRODUCTION

The reconstruction of ancestral karyotypes and gene orders from homologies between extant species is a long-standing problem pioneered by Dobzhansky and Sturtevant (1938) on drosophila chromosomes. It helps to understand the large-scale evolutionary mutations that differentiate the present genomes and happened in every lineage of the living world.

Computational methods to handle gene order and propose ancestral genome architectures have a shorter (Sankoff et al., 1996) but prolific (Fertin et al., 2009) history. However, despite the numerous efforts of the

[1] Department of Mathematics, Simon Fraser University, Burnaby, BC, Canada.
[2] Faculty of Natural Sciences, University of Sarajevo, Sarajevo, Bosnia and Herzegovina.
[3] INRIA Lille-Nord-Europe, Université Lille 1, LIFL, UMR CNRS 8022, Villeneuve d'Ascq, France.
[4] INRIA Rhône-Alpes, Université de Lyon, Lyon, and Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France.

computational biology community, two recent rounds of publications have put a doubt on their efficiency. In 2006, comparing ancestral boreoeutherian genome reconstructions made by cytogeneticists on one side and bioinformaticians on the other, Froenicke et al. (2006) found that the manually constructed and expertized cytogenetics one was not acceptably recovered by computational biology. This provoked many comments (Bourque et al., 2006; Rocchi et al., 2006) published the same year. Since then, several bioinformatics teams (Ma et al., 2006; Chauve and Tannier, 2008; Alekseyev and Pevzner, 2009; Kemkemer et al., 2009) have built up computational methods which now all give results on mammalian genomes which agree with the earlier cytogenetic ancestor.

Three years after this first controversy concerning mammalian genome history, Gordon et al. (2009) published the refined configuration of the genome of an ancestor of *Saccharomyces cerevisiae* before it underwent a whole genome duplication (which is approximately as ancient as the protoboreoeutherian), and chose not to use any computational framework, arguing that those are still in development and yet cannot handle the available data. Sankoff (2009) wrote a comment on Gordon et al. (2009)'s article, detailing the deficiencies of computational approaches, yet with an optimistic conclusion for the future. A few weeks later, a publication by the Genolevures consortium (Souciet et al., 2009) of a wide yeast genome comparative study included a reconstruction of an ancestral configuration of some non-duplicated yeast species, with a method by Jean et al. (2009). As no species affected by the whole genome duplication was considered, the ancestor constructed by Gordon et al. (2009) still seems inaccessible by computational approaches. Moreover, the study of Souciet et al. (2009) was partially based on the optimization criterion that caused the divergence with cytogenetic results for mammalian genomes, and shows a large number of potentially optimal ancestral architectures. So this result deserves an analysis similar to the one that followed the debate on mammals.

Hence we raise, for both considered ancestral genomes, the questions of the convergence and reliability of the proposed architectures, and more generally the question of the reliability of computationally inferred ancestral genomes. Right now there still seems to be a gap between computational methods and the application they are designed for. This article intends to explain this gap, and to a certain extent, to fill it. We survey the general principles implemented in computational methods that are applicable to yeast data. In particular, we present in the same framework methods to reconstruct ancestors of all yeasts, whether they underwent a whole genome duplication or not. We define some properties for the data which explain why the different methods can or cannot give reliable and converging results. We give a theoretical basis to the notion of "reliable rearrangements" introduced by Zhao and Bourque (2009), extend it to the whole genome duplication context, and apply it on yeast data. Eventually we combine principles of several methods we found reliable to infer proposals for both the ancestors searched by Gordon et al. (2009) and Jean et al. (2009), and compare our results to theirs. The construction of these two ancestors sheds light on the possibilities of computational methods.

We show that the pre-duplication ancestor of Gordon et al. (2009) can be well approached by all methods, which enforces the confidence in the proposed architecture. For the non-duplicated ancestor, we show that with the data used in Jean et al. (2009), proposals of ancestral architectures obtained with different methods lack good reliability indicators as well as good convergence. For this second case, we found this lack of signal is inherent to the used orthology block set. Indeed, with a de novo construction of another one with higher resolution for the comparison of the same species, it is possible to propose a high confidence ancestral architecture of the same ancestor, which significantly differs from the proposition of Jean et al. (2009). This helps understanding the behavior of automatic methods and making more explicit what can be expected from them at the present time.

This article extends and prolongates a conference version presented at the RECOMB satellite workshop on Comparative Genomics (Tannier, 2009). The scope of the study has largely been extended, dealing with other ancestors and a wider range of methods. The core of the article is the next section, exploring several methodological aspects of ancestral genome reconstructions. Then we present the different yeast ancestors we constructed with those principles, and the comparisons with earlier studies.

All data and results are available at: www.cecm.sfu.ca/~cchauve/SUPP/RCG09-JCB-YEASTS/.

2. METHODS

In this section, we describe a set of methodological principles that were used for ancestral genome reconstructions and how we apply them on yeast data. We give a high level description of the two kinds of methods we use, namely the "physical mapping" and "rearrangement" principles, and we describe properties related to the reliability of the inferrence produced by these two approaches.

2.1. The data and objectives

The considered species and ancestors. We start from a set of genes and homologies in the 7 assembled yeast genomes (*Saccharomyces cerevisiae*, *Candida glabrata*, *Zygosaccharomyces rouxii*, *Kluyveromyces lactis*, *Ashbya gossypii*, *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri*) taken from the Yeast Gene Order Browser (Byrne and Wolfe, 2005). We suppose that the phylogeny is known, and is the one of Gordon et al. (2009). Some of the methods need branch lengths, so we assigned the same number to all small branches and completed with a hypermetric branch length system (Fig. 1).

The two ancestors we attempt to reconstruct are those indicated in Figure 1. One is the genome of the organism immediately pre-dating the whole genome duplication on one branch of the tree, which we call the *pre-duplication* ancestor, and the other is the ancestor of all non-duplicated species except *Zygosaccharomyces rouxii*, which we call the *non-duplicated* ancestor. Jean et al. (2009) and Souciet et al. (2009) aim at reconstructing ancestral genome segments and proto-chromosomes of all non-duplicated species without the phylogenetic assumption; we discuss the impact of this difference in the sequel.

Given two species, their *evolutionary path* is the path between them in the phylogenetic tree. Given three distinct species, their *median point* is the intersection of the three evolutionary paths of the three possible pairs of species.

Ancestral genomic markers. In order to construct the two ancestral genomes, we need a set of homologous markers covering a large part of the extant genomes and which are believed to be present in a unique exemplar in the ancestral genome. All methods use such a set of makers as a starting point. The datasets of markers we consider are illustrated in Figure 2.

Markers for the non-duplicated ancestor. For the non-duplicated ancestor, we use two sets of markers. One is constructed by Jean et al. (2009) and contains 135 markers.[5] It was generated by filtering the largest markers from an output provided by the i-Adhore software (Simillion et al., 2008) running on gene orthologies. We call it the *low resolution marker set*, in comparison to the second one.

Due to the apparent high rate of rearrangements (Souciet et al., 2009), some of them being invisible at too low resolution, a higher resolution marker set was also constructed by retrieving from the Yeast Gene Order Browser (Byrne and Wolfe, 2005) all gene families with exactly one exemplar non intersecting with another gene in each of the 5 non-duplicated species. The method to define markers consisted in joining iteratively every pair of genes that were immediately consecutive in all 5 species, which defined non-overlapping groups of colinear genes in all 5 genomes. Next, groups spanning less than 2kb on at least one genome were removed, and the remaining groups of genes were again joined into larger colinear groups. This resulted in 710 markers that have unique and non-overlapping coordinates on the 5 non duplicated genomes. We call it the *high resolution marker set*.

The low resolution dataset covers from 31% to 35% of the extant genomes, while the high resolution datasets covers from 57% to 66% of the extant genomes. Out of the 710 markers of the high resolution set, 202 map to exactly one of the low resolution set. Conversely, 118 markers of the low resolution set intersect at least one marker of the high resolution set. These 118 markers allow to compare two results coming from the two different marker sets, as we do in Section 3.2.

Markers for the pre-duplication ancestor. For the pre-duplication ancestor, we used the set of 212 ancestral markers computed in Tannier (2009) using an implementation of the "double conserved synteny" principle, that was introduced by Kellis et al. (2004) and Dietrich et al. (2004), discussed in de Peer (2004) and used several times in a whole genome duplication context.

[5] The coordinates of markers were provided by G. Jean. Some markers were overlapping and the coordinates have been modified to remove overlaps in such a way that the order of the markers along chromosomes is the same than in Jean et al. (2009).

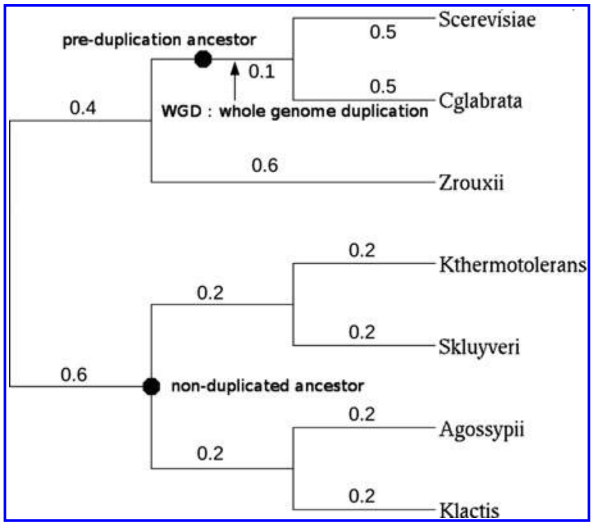

FIG. 1. The phylogeny of the seven considered species, and the positions of the studied ancestors. The whole genome duplication (WGD) is indicated by an arrow, and the black spots point the positions of the *pre-duplication* ancestor, immediately pre-dating the WGD, and the *non-duplicated* ancestor, which is the ancestor of all non-duplicated species, except *Zygosaccharomyces rouxii*, which is an outgroup.

Definition 1. Let G be a non-duplicated genome and D a duplicated genome, together with orthology relationships between the genes of these two genomes. A double conserved synteny (DCS) between G and D is a set S of contiguous genes of G such that

1. the set of genes of D orthologous to a gene in S is composed of two contiguous segments A_D and B_D in D, each of size at least two.
2. the sets A_G and B_G of genes of S respectively orthologous to genes of A_D and B_D ($S = A_G \cup B_G$) span two intersecting segments in G.
3. S is maximal for these properties.

The first condition imposes the presence of one segment in G spanned by the genes in S, and two orthologous segments in D spanned by the genes in A_D and B_D, containing a minimum number of genes. It is the basis of the double synteny signal. The presence of at least two genes avoids the possible presence of one transposed or misannotated gene. The second condition avoids the ambiguous signal of two successive single syntenies.

We need to assign a direction to each marker, since most methods use this information. To do so, we require in addition a condition on the order of the genes: the first or last gene of A_D (and B_D) has to be ortholog to the first or last gene in A_G (B_G). Then it is possible to decide a relative orientation for every marker.

This definition aims at detecting sets of genes that are believed to have been contiguous in the last common ancestor of G and D. They have been modified only by internal rearrangements on the path from this ancestor to G, and duplicated (due to the whole genome duplication) then rearranged internally on the path to D. As a consequence, they were contiguous also in the pre-duplication ancestor. Each DCS then results in one marker which has one occurrence in G and two occurrences in D. We apply this definition to each possible pair of duplicated and non-duplicated genome, and the ancestral markers for the pre-duplication ancestor were given by the comparison of *Saccharomyces cerevisiae* and *Saccharomyces kluyveri*. This gives a set of 212 makers, with sizes ranging from 15kb to 95kb on the *Saccharomyces kluyveri* genome, and covering 96% of its genes.

Representation of genomes with markers. All markers have a reading direction. To take it into account in our methods, we define for each marker its two *extremities*, among which the *tail* is the starting point of the marker, and the *head* is its end point. A *double marker* has two tails and two heads. An *adjacency* is an undirected pair of marker extremities. A (possibly duplicated) ancestral or extant genome is

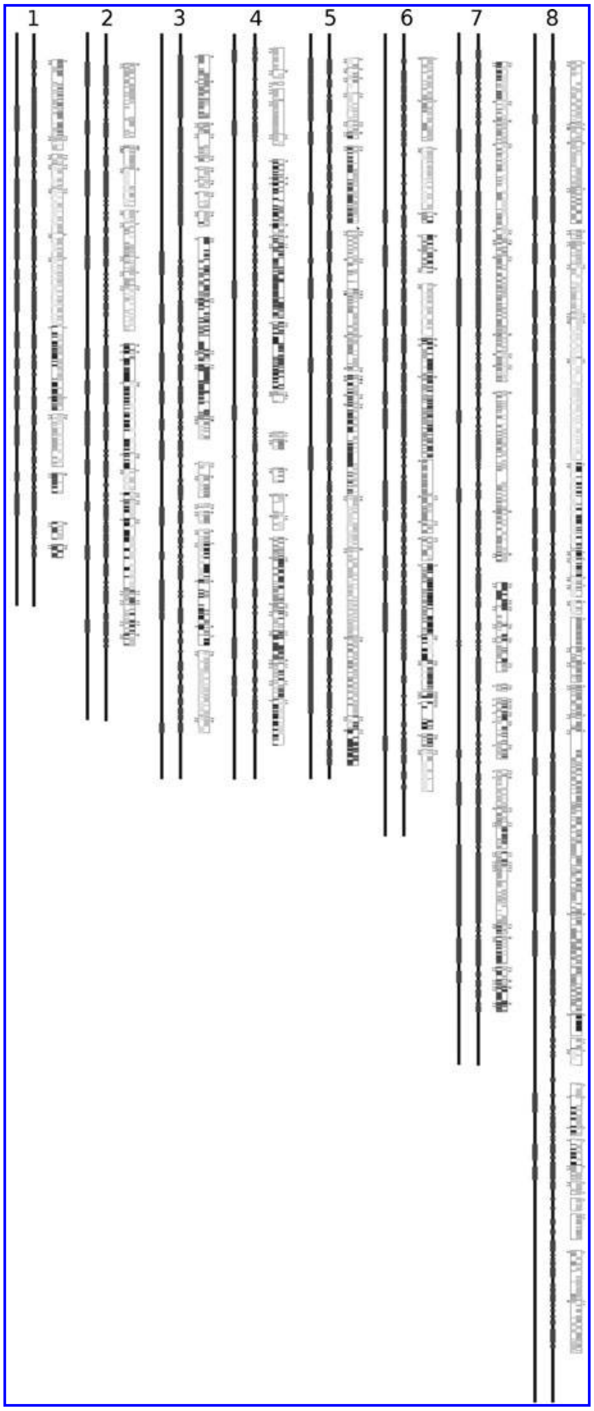

FIG. 2. Ancestral markers mapped onto the *Saccharomyces kluyveri* genome. There are three lines for each *Saccharomyces kluyveri* chromosome. On the leftmost line are indicated the markers from the low resolution set. On the middle line are indicated the markers from the high resolution set. On the rightmost one are mapped the double syntenies with *Saccharomyces cerevisiae*: in each box are represented the genes, mapping either on one segment of *Saccharomyces cerevisiae* (left side of the box), or the other (right side of the box).

then defined as a set of adjacencies on a set of (possibly double) markers extremities, such that an extremity is contained in at most one adjacency. The adjacencies of a genome define the *chromosomes*, which can be linear or circular. A genome with only linear chromosomes is said to be *linear*, and if both are allowed, it is said to be *mixed*. (We recall a few definitions here, with the formalism of Tannier et al. [2009]. For an extensive introduction, the reader may refer to this article.)

2.2. Two computational principles: physical mapping versus rearrangements

All the methododological principles we present here take the directed ancestral markers and their arrangements in the extant genomes as input, and outputs one or several possible ancestral arrangements of the same markers. Physical mapping techniques consist in gathering information on the adjacency, contiguity or distance between markers on the ancestral genome, and assembling this information into a genome afterwards. Rearrangement methods rely on the definition of a distance that compares two genomes. Ancestral genomes are inferred by searching for arrangements that minimize the sum of the distances along a phylogenetic tree (i.e., they are the "Steiner points" in the metric space defined by the chosen distance). Both methods rely on a parsimony principle, but sometimes infer divergent results (Tannier, 2009).

Rearrangement methods. The principle has first been stated by Sankoff et al. (1996), and since then many developments have been published, according to the chosen distance and the kind of genome allowed (Fertin et al., 2009). Here, we consider mixed genomes,[6] and the *distance* between two genomes is the number of rearrangements that are necessary to transform one genome into the other. The classical possible rearrangements are inversions, fusions, fissions, translocation, and block interchanges, all modeled using the double cut-and-join (DCJ) operation.[7]

Given a distance d on the set of genomes, a *median* of three genomes G_1, G_2 and G_3 is a fourth one which minimizes the sum of its distances to G_1, G_2 and G_3. When searching for a median, one can either require all its chromosomes to be linear, in which case we call it a linear median, or allow a mix of linear and circular chromosomes, in which case we call it a mixed median. For k genomes, the *multiple genome rearrangement* problems consist in finding a tree, whose leaves are the genomes, and one additional genome per internal node of the tree, which minimizes the sum of the distances between all pairs of genomes that are the extremities of an edge of the tree.

Such methods have been applied to mammalian and non-duplicated yeast genomes (Alekseyev and Pevzner, 2009; Jean et al., 2009). The possibility of whole genome duplications has been added and applied to plant and yeast genomes (Zheng et al., 2008), though all programs handling this possibility are currently limited to the comparison of two genomes. Zheng et al. (2008) solve a generalization of the genome median problem, which is called the *guided genome halving* (GGH) problem. It is defined as follows. A *doubled* genome $G \oplus G$ is a duplicated genome which is constructed from an ordinary one G by duplicating all markers and all adjacencies. For one genome G there are several possible doubled genomes, and the *double distance* $dd(G, D)$ between an ordinary genome and a duplicated genome is the smallest number among all $d(G \oplus G, D)$. The *Guided Genome Halving* problem consists in, given one ordinary genome G and one duplicated genome D, find a genome M which minimizes $d(M, G) + dd(M, D)$ (Tannier et al., 2009).

Reliability issues of rearrangement based methods. Problems of rearrangement methods were compiled by Gordon et al. (2009) and Sankoff (2009). The main one is probably the multiplicity of solutions, as was first pointed by Eriksen (2007). Almost all rearrangement methods which have been used on eukaryotic nuclear genomes are based on parsimony (some statistical methods have been used on mitochondrial or bacterial genomes, see Darling et al. (2008) for example). A possible saturation of the gene order signal, due to a high rate of genome rearrangements, sometimes gives a huge number of equally optimal solutions, lowering the reliability of features that appear in a solution but are not shared by others. Several studies have tried to restrain this multiplicity in order to find more reliable solutions, or at least partial reliable solutions. The fundamental principle we follow (it is the basis of the median algorithm of Xu (2009), it is also similar to Swenson and Moret (2009), and present in Murphy et al. (2005), though without any theoretical assessment) is that *properties of occurrences of adjacencies in the input genomes can translate into properties of occurrences in all optimal solutions of the median problem*. Features that are shared by all or many optimal medians can then be considered as reliable, at least from the point of view

[6] It is an approximation of linear ones that simplifies some theoretical issues. The same approximation is made by Xu (2009). For yeast genomes, we verify that it is a good one, as all the mixed median or halving solutions we get in the sequel are linear.

[7] Note that we do not *a priori* consider DCJ as a good model for yeast genome evolution, but we test its efficiency on the data we analyse.

of multiplicity of the solutions.[8] We give several results and conjectures, present in the litterature or formally stated here for the first time, which assess the reliability of ancestral features, whether there is a duplicated genome in the data or not.

The case of unduplicated genomes

Definition 2. Let G_1, G_2, G_3 be three genomes. An adjacency which is present in at least two genomes out of these three is called *supported* in the median point of G_1, G_2, G_3. More generally, an adjacency is *supported* in an ancestor A if it is present in two genomes which evolutionary path contains A.

Using the framework of adequate subgraphs of breakpoint graphs, Xu (2009) proved the following result, for which we give a simple proof for the sake of completeness.

Property 1. [(Xu, 2009)] A supported adjacency is present in every solution of the mixed DCJ median of G_1, G_2, G_3.

Proof. The *breakpoint graph* of two genomes G_1 and G_2 is the graph which has the marker extremities as vertices, and the adjacencies in both genomes as edges. By definition, it is a set of disjoint paths and cycles. The DCJ distance between two genomes is $d(G_1, G_2) = n - (c + \frac{pe}{2})$, where n is the number of markers, c is the number of cycles in the breakpoint graph, and pe is the number of paths with an even number of edges (including trivial paths with no edges) (Bergeron et al., 2006). For three genomes G_1, G_2, G_3, the *median score* of a fourth genome M is $d(G_1, M) + d(G_2, M) + d(G_3, M)$.

Assume xy is a supported adjacency, let us say it is present in G_1 and G_2. If a genome M does not contain xy, it is easy to construct a genome M' containing xy which has a smaller median score. Indeed, if M contains two adjacencies xs and yt, let $M' = M \setminus \{xs, yt\} \cup \{xy, st\}$. If M contains an adjacency xs and no adjacency containing y, let $M' = M \setminus \{xs\} \cup \{xy\}$. If M has no adjacency containing either x or y, let $M' = M \cup \{xy\}$. It is easy to see that in each case, the number of cycles of the breakpoint graph of M and G_1 or G_2 increases by one while the number of even paths is unchanged, whereas the distance to G_3 may increase by at most one. So the median score decreases by at least one. In conclusion, M can then never be an optimal median, and any optimal median contains xy. ■

Zhao and Bourque (2009), Alekseyev and Pevzner (2009), and Swenson and Moret (2009) have tried to detect adjacencies that are not necessarily supported but can nevertheless be considered as reliable, and based on their results, we introduce the notion of *reliable adjacency*.

Definition 3. Let G_1, G_2, G_3 be three genomes. An adjacency uv is *reliable* in the median point of G_1, G_2, G_3 if G_1 and G_2 contain an adjacency xy, G_1 contains the adjacency uv, G_3 contains the two adjacencies xu and yv. More generally, an adjacency uv is said to be *reliable* in an ancestor A if there are three genomes G_1, G_2 and G_3 such that the path from their median point M to G_1 contains A, and uv is reliable in M.

This definition is based on the notion of reliable rearrangements defined by Zhao and Bourque (2009), and is illustrated in Figure 3a. Indeed, the presence of this pattern of adjacencies is the sign of a rearrangement, a DCJ which can be a reversal or a translocation for example, located on the branch leading to G_3, that cut adjacencies xy and uv and joined the adjacencies xu and yv. The underlying justification of reliability of this rearrangement is the presence of the adjacency uv in the median point of the three genomes, as it is underlying in the work of Xu (2009) and Alekseyev and Pevzner (2009).

[8] Whether we may really trust these features in all cases is of course not certain. It depends on the number of rearrangements which have occured, and the probability that a realiable feature happends by chance. If this probability is high, then the reconstruction of ancestral features becomes hopeless. We believe the results presented in the sequel (convergence of the methods, proportion of reliable features) are indicators that we are still in a context where the reconstruction is possible.

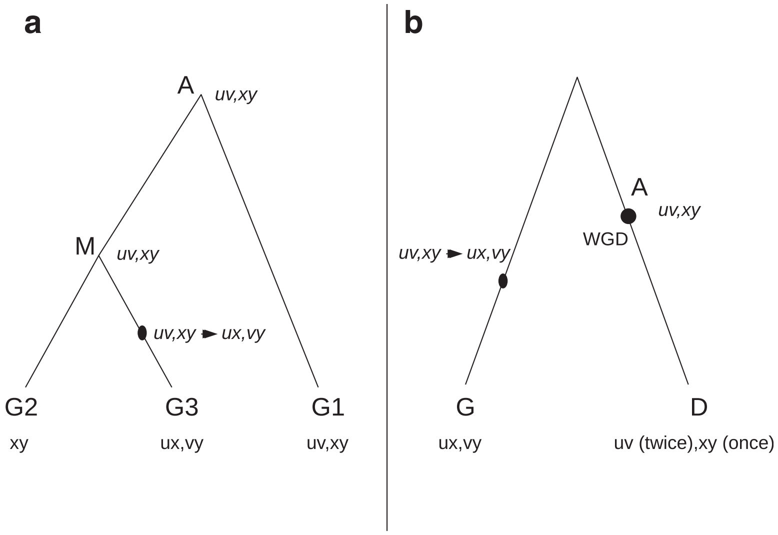

FIG. 3. An example of inferrence of a reliable adjacency, in the absence of duplicated genome (**a**), or in the presence of a duplicated genome (**b**). Extant adjacencies are written at the leaves, while inferred adjacencies and rearrangements are written in italic at the internal nodes. A is the ancestral genome which is to be reconstructed, and M is the median point of the three genomes in (a). In this case, A is an ancestor of M, but any case where A is a descendant or an outgroup of M can be considered, provided A is on the path between G_1 and M.

Property 2. A reliable adjacency uv is present in every solution of the mixed DCJ median of G_1, G_2 and G_3.

Proof. Let uv be a reliable adjacency and M be an optimal median genome that does not contain uv. First, by Property 1, M contains the adjacency xy, as it is supported. If M contains two adjacencies us and vt, let $M' = M \setminus \{us, vt\} \cup \{uv, st\}$. The number of cycles of the breakpoint graph of M' and G_1 increases by one, as in the breakpoint graph of M' and G_3. And comparing M' to G_2, the number of cycles decreases by at most one. The number of even paths is unchanged in all cases, so M' has a strictly lower median score. In the other cases, whether M contains an adjacency xs and no adjacency containing y, or M has no adjacency containing either x or y, it is easy to check that choosing M' in the same way than in the proof of Property 1 leads to a better median, proving that uv is in all optimal median solutions. ■

In consequence, as uv is present in M and G_1, it is probably present in A, which is on the path from M to G_1, and that is why we call it reliable. Note in addition that we make no hypothesis on which species is a descendant of which. This property is valid, whether M is a descendant, an ancestor (Fig. 3a), or an outgroup for A.

Handling duplications. It is possible to extend these notions to the duplication context, and this, while preliminary, is a contribution of the present paper.

Definition 4. *Let G be a non-duplicated genome and D be a duplicated genome. An adjacency is called* supported *if it is present in both genomes, or twice in D.*

Definition 5. *Let G be a non-duplicated genome and D be a duplicated genome. An adjacency uv is called* reliable *if either*

- *G and D contain an adjacency xy, G or D contain the adjacency uv and D contains the two adjacencies xu and yv, or*
- *D contains twice an adjacency xy and D contains at least once the adjacency uv and G contains the two adjacencies xu and yv.*

The notion of reliable adjacency in this case is illustrated in Figure 3b.

We propose the analogue of Property 1 and Property 2 for the guided halving, which justifies the use of supported and reliable adjacencies, and shows that a single framework can be used for duplicated genomes as well as for non-duplicated ones.

Property 3. Let G be a non-duplicated genome and D be a duplicated genome. Any supported adjacency is present in every mixed guided halving optimal solution.

Proof. Suppose first the adjacency xy is contained twice in D. Let M be a genome which does not contain xy. We construct M' containing xy such that $d(M', G) + dd(M', D) < d(M, G) + dd(M, D)$. Suppose M contains two adjacencies xs and yt. The breakpoint graph between $M \oplus M$ and D contains two cycles or paths containing the (doubled) extremities x, y, s, t for any choice of $M \oplus M$. The genome $M' = M \setminus \{xs, yt\} \cup \{xy, st\}$ increases the number of cycles by two for a good choice of $M \oplus M$, so decreases the GGH score. The other cases are similar. ■

Conjecture 1. Let G be a non-duplicated genome and D be a duplicated genome. Any reliable adjacency is present in every mixed guided halving optimal solution.

The proof of this conjecture would imply a deeper involvement in the theory of double distance and genome halving, and is postponed to a future work.

The reliability of an adjacency is thus, like in the non-duplicated case, based on the presence of a signal for a DCJ rearrangement on some branch of the phylogenetic tree: either on the branch leading to G, or on one of the two copies of the duplicated genome D. It generalizes the principle invented by Zhao and Bourque (2009) to a wider range of possible genomes.

Conversely, the question can be asked if any adjacency which is present in any optimal solution of the rearrangement method can be considered reliable, in the absence of any signal for a precise rearrangement. We chose not to make this extension in this paper, and postpone a deeper study of these features to a future work.

Physical mapping techniques. Mapping the genes on a genome whose sequence is not known or not assembled often consists in trying to guess which pairs or groups of genes should be close to each other, and propose a mapping in which as much as possible of these features are satisfied (Alizadeh et al., 1995). The same approach can be used for ancestral genomes. It is recognizable in cytogenetics studies on mammalian genomes (Froenicke et al., 2006) as well as on genomic studies on yeast genomes (Gordon et al., 2009). The principles have been computationally implemented first by Ma et al. (2006), and then formalized and generalized in Chauve and Tannier (2008) and Ouangraoua et al. (2009), and generalized to a whole genome duplication context by Ma et al. (2008) and Tannier (2009). It is also used as a first phase by Jean et al. (2009), though no phylogenetic signal is used.

Given a set of ancestral markers and a phylogenetic tree, we define an *ancestral synteny* as a subset of ancestral marker extremities that are believed to be contiguous in the ancestral genome. An ancestral synteny of size two is called an adjacency. It is clear that for any ancestral directed marker, the adjacency containing its two extremities is an ancestral synteny.

An *interval* of a genome is a set of markers that are contiguous on this genome. A total ordering of the ancestral marker extremities is said to *satisfy* an ancestral synteny $\mathcal{AS}$ if the elements of $\mathcal{AS}$ form an interval of the ordering.

A set of ancestral syntenies $\mathcal{S}$ are said to be in *conflict* if there is no total ordering of the ancestral markers which satisfies all ancestral synteny of $\mathcal{S}$. For example, if X, Y and Z are ancestral marker extremities and $\{X, Y\}$, $\{X, Z\}$, $\{Z, Y\}$ are three ancestral syntenies, then there is no linear ordering of X, Y and Z so that the three pairs are contiguous.

Our implementation of the *physical mapping* principle consists in two steps:

(1) searching for *ancestral syntenies*, and weighting these ancestral syntenies according to the confidence put in their presence in the ancestral genome, possibly guided by its phylogenetic signal;
(2) assembling the ancestral markers into *Contiguous Ancestral Regions* (CAR), satisfying the ancestral syntenies as much as possible.

The two steps can be implemented in several ways. We describe our implementation of the first step in the Results section. The second step, the assembly, can benefit from a combinatorial framework based on the consecutive ones property of binary matrices and PQ-trees of weakly partitive set families (Chauve and Tannier, 2008) and we refer the reader to this reference for a detailed description.

Reliability issues in physical mapping techniques. In this approach, the contiguous ancestral regions are built from the set of ancestral syntenies. That is, the assembly part only deals with conflict among the set of ancestral syntenies and does not propose additional features. So depending on the method to infer ancestral syntenies, this method cannot put forward contiguous ancestral regions as the chromosomes of the ancestor, but they are chromosome segments. And actually this kind of method often finds slightly more contiguous ancestral regions than the believed number of chromosomes (Ma et al., 2006; Chauve and Tannier, 2008). It is also the case for the present study on yeasts (see Section 3). It is a weakness because such methods are often not able to reconstruct full chromosomes, but it is also a strength because every adjacency is well supported, and then is easily examined by manual expertise. So a stringent definition of ancestral synteny will often produce a high number of CARs, which may be far from definitive assembled chromosomes, whereas a less stringent definition will lead to less CARs which contain less reliable features. The reliability was also a point of the study on mammalian genomes of Ma et al. (2006), who inferred a reliability of ancestral adjacencies based on a statistical model. The number of CARs was small, at the cost of several unsupported adjacencies.

Another source of unreliability is that an optimization step is required in order to choose a subset of ancestral syntenies that can be satisfied. The more conflict there is, the lower proportion of ancestral syntenies kept, and the more arbitrary choices among optimal solutions make big differences on the final result. Preliminary results on conflicts and maximal solutions for the Consecutive Ones Property might be useful in order to better understand this issue (Chauve et al., 2009).

3. RESULTS AND DISCUSSION

We implemented the rearrangement and physical mapping principles on the two yeast ancestors. But we also tried to initiate a combination of the two principles by using supported and reliable adjacencies, which take their reliability from rearrangement methods, as ancestral syntenies for physical mapping. We also see this as a methodological contribution of this article: we point out the reliable principles from both sides and use them in a single method, which can contribute to the convergence between the two approaches.

3.1. The pre-duplication ancestor

Recall we dispose in this case of 212 ancestral markers, with coordinates on *Saccharomyces cerevisiae* and *Saccharomyces kluyveri*. 35 adjacencies were present both in *S. kluyveri* and twice in *S. cerevisiae*, so they may be immediately joined, which makes a total of 177 markers.

We counted 134 supported adjacencies, and 19 reliable adjacencies. They will be present in all physical mapping and rearrangement solutions (provided Conjecture 1 is true), and account for 86% of the number of adjacencies in a solution. This high number explains the good convergence of all methods that we describe below.

To implement the first step of the physical mapping, we took as ancestral syntenies the supported adjacencies, reliable adjacencies and maximal *supported intervals*, that are intervals present both in *S. kluyveri* and in *S. cerevisiae*, or twice in *S. cerevisiae*.

We also computed the sets of double conserved syntenies (DCS, as in Definition 1) for *Saccharomyces cerevisiae* and all non-duplicated species, or between *Saccharomyces kluyveri* and all duplicated species. And, for each such DCS, we defined an ancestral syntenies by the sets of markers intersecting this DCS on either *S. kluyveri* or *S. cerevisiae*.

The result of the physical mapping method is a set of 11 CARs,[9] depicted on Figure 4b.

The ancestral arrangement is given by the manual study of Gordon et al. (2009), and drawn on Figure 4a. Almost every CAR is included in an ancestral chromosome. The main differences between the two are that

[9] The difference with the conference version (Tannier, 2009) is the addition of the reliable adjacencies in the pool of ancestral syntenies, and in consequence the decrease from 14 to 11 CARs.

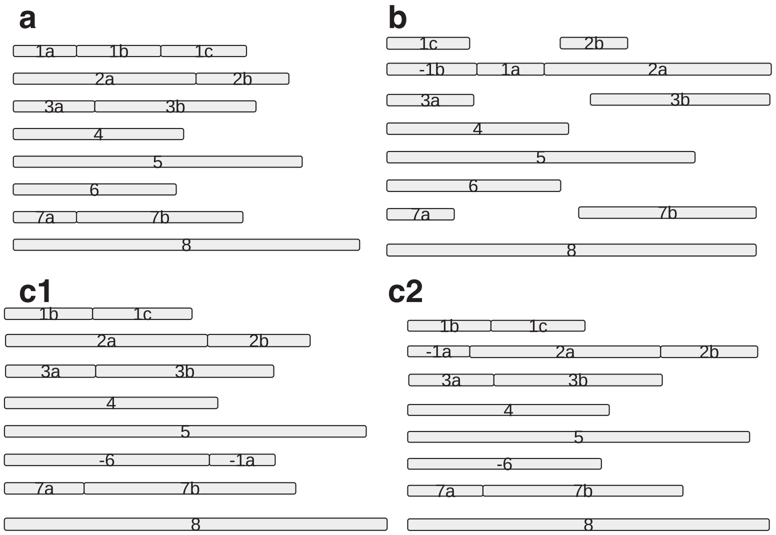

FIG. 4. The arrangements of the pre-duplication yeast ancestral genome obtained as follows: (**a**) by Gordon et al. (2009), (**b**) by the physical mapping method, and (**c1**), (**c2**) by the rearrangement method in Gavranovic and Tannier (2010) (two different solutions were reported). Signs on the figures mean the relative orientation of the markers, and numbers refer to the chromosome to which the blocks belong in Gordon et al. (2009). Blocks correspond to segments of DCS which are found together by all three methods. The presence of only 13 blocks grouping 212 markers shows the good convergence of the methods. Both solutions of the rearrangement method differ from the manual ancestor by one DCJ, while 6 rearrangements are needed to transform the physical mapping method solution into the manual ancestor. They are mainly fusions of CARs.

several contiguous ancestral regions are not assembled into chromosomes, and that one contiguous ancestral region fuses segments that are believed to belong to two different chromosomes in Gordon et al. (2009).

For the rearrangement method, the algorithm and solutions of Gavranovic and Tannier (2010) were used (Fig. 4c1, c2). As pointed out in Gavranovic and Tannier (2010), only two different optimal solutions were found and were both one DCJ away from the standard of Gordon et al. (2009). It is proved that both solutions reach optimality of the GGH objective function, thanks to a lower bound based on the double distance computation.

All indicators are telling the same story for this instance: there is a high proportion of supported and reliable adjacencies, a low diversity of GGH optimal solutions, the optimality is reached by the heuristic, and all methods converge towards a few number of alternative configurations on only 13 common blocks.

So this ancestor is well accessible by computational methods, and it is possible to trust the gross part of the results, focusing on the few difficulties it seems to present, mainly the position of the block named 1*a* on Figure 4. This one deserves manual expertise, and the configuration of Gordon et al. (2009), while not optimal in terms of rearrangements, can still be considered as a gold standard.

3.2. The non-duplicated ancestor

The physical mapping approach. To implement the physical mapping approach, we defined as ancestral syntenies the supported adjacencies, reliable adjacencies, and maximal *supported intervals* that are sets of markers which are an interval of two genomes which evolutionary path contains the ancestor. Moreover, to take advantage of the presence of duplicated outgroups, we computed all DCS (according to Definition 1) between the four ingroups and the two duplicated outgroups, and for each such DCS we added an ancestral synteny containing all markers of the ingroup included in the DCS; we call such intervals *DCS-supported intervals*. The idea of using DCS-supported intervals was introduced in Ouangraoua et al. (2009) for reconstructing the amniote ancestral genome and gives the possibility of using the homology with duplicated species, and gathering all the information from the seven assembled species for the construction of this ancestor.

Ancestors obtained with the low resolution dataset. With the low resolution marker set, we obtained 35 CARs, from a set of 267 ancestral syntenies: 69 adjacencies are supported and two are reliable, which accounts for 41% of the possible adjacencies are in all solutions of the median problem, and contrasts with the 86% of the pre-duplication instance.

Jean et al. (2009) present an ancestor with 8 chromosomes, which contrasts with our 35 CARs. To be more precise, the ancestor proposal of Jean et al. (2009) consists of first a set of around 30 "super-blocks" (a concept similar to CARs) obtained from adjacencies conserved in at least two extant genomes that are then fusioned into 8 proto-chromosomes by solving a median problem on these CARs. Jean et al. (2009) note the stability of the different sets of super-blocks they obtain and the large number (at least 90) of optimal solutions to the median problem. This means that a significant number of adjacencies in the protogenome of Jean et al. (2009) are not supported, reliable, included in supported intervals nor supported by a double homology with the outgroups.

Note that Jean et al. (2009) infer ancestral genomic features of all five non-duplicated yeasts without using phylogenetic signal. Our approach is different since *Z. rouxii* is used as an outgroup and we use the known phylogenetic information for the five non-duplicated genomes. This methodological difference explains some characteristics of their result. Twenty-six ancestral adjacencies of Jean et al. (2009) are present in one or no extant genome: they are retrieved by a rearrangement method, more precisely a median of all five extant species. None of them is reliable, and they appear more as artifacts due to the absence of a phylogeny. Thirteen ancestral adjacencies are only present in *S. kluyveri* and *K. thermotolerans*, which are the closest species (in terms of rearrangement distance) according to the final phylogeny given by Souciet et al. (2009). Hence there is, with the currently available data, no conclusive evidence that they are ancestral and not derived.

These additional adjacencies that we do not retrieve are thus very questionable, as well as the possibility to infer a better ancestor than about 35 chromosome segments, while the expected number of proto chromosomes is aroud 10. That is why we experimented the reconstruction with the high resolution marker set.

Ancestors obtained with the high resolution dataset. Using the 710 markers of the high resolution dataset and the physical mapping approach, we obtained 14 CARs, and counted 720 supported and 35 reliable adjacencies, which makes 86% of the ancestor defined by adjacencies present in all rearrangement solutions, which is comparable to the proportion of reliable adjacencies observed in the pre-duplication ancestor. The first interesting fact is that, with the same method than we used with the low resolution dataset, we obtain a number of CARs that is closer to the expected number of ancestral chromosomes. This suggests that some supported or reliable features vanish when the resolution is lowered, showing that the traces of some rearrangements are only visible with our high resolution marker set. This fact is consistent with the expected high rate of genome rearrangement observed in the considered species in Souciet et al. (2009). This hypothesis is reinforced by an analysis of the same dataset with the EMRAE program of Zhao and Bourque (2009), that detects only 5 reliable rearrangements along all branches of the considered phylogenetic tree with the low resolution marker set, and 148 reliable rearrangements with the high resolution marker set. The CARs were obtained from a set of 3106 ancestral adajcencies, 11 of them being discarded to remove the conflicts. These CARs are represented on Figure 5, with the genome of *S. Kluyveri* as a reference.

It is still possible to compare them to the ancestor of Jean et al. (2009) by mapping the 14 CARs to the low resolution dataset by considering only the 118 markers of the low resolution set that are in correspondence with at least one marker from the high resolution set. This allows to construct an ancestor that contains 118 markers and 11 CARs (3 CARs intersect no marker from the low resolution set), which is close to the 8 chromosomes found by Jean et al. (2009). The comparison is illustrated on Figure 6a, b, together with the projection on these 118 markers of the 35 CARs of the physical mapping method on the low resolution set (Fig. 6c).

The rearrangement approach. To analyze the non-duplicated instance and compare its reliability to the pre-duplication one, we used the genome median on three genomes which median point is the searched ancestor, *Zygosaccharomyces rouxii*, *Ashbya gossypii*, and *Saccharomyces kluyveri*.

On the low resolution marker set, an optimal median was computed with the program of Xu (2009) in 3754 seconds.[10] The optimal median score is close to the median score of the 35 CARs, even if the two ancestral proposals differ significantly (Fig. 6).

[10] Though it solves the mixed genome median problem, all chromosomes in the solution were linear.

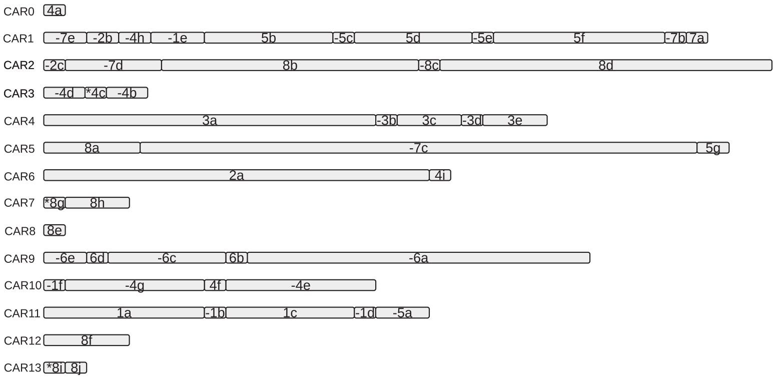

FIG. 5. The CARs of the non-duplicated ancestor with a high resolution marker set. Chromosome segments represent segments colinear with the genome of *S. Kluyveri*, chosen because it is the less rearranged genome since this ancestor. Segments with an ambiguous orientation in the ancestor are identified by the symbol *. Segment sizes are roughly proportional to the colinear segments in *S. Kluyveri*, but for very short segments whose relative size is larger. It is computed with a physical mapping method with reliable ancestral syntenies, and we feel it can be considered as the new standard for this ancestor.

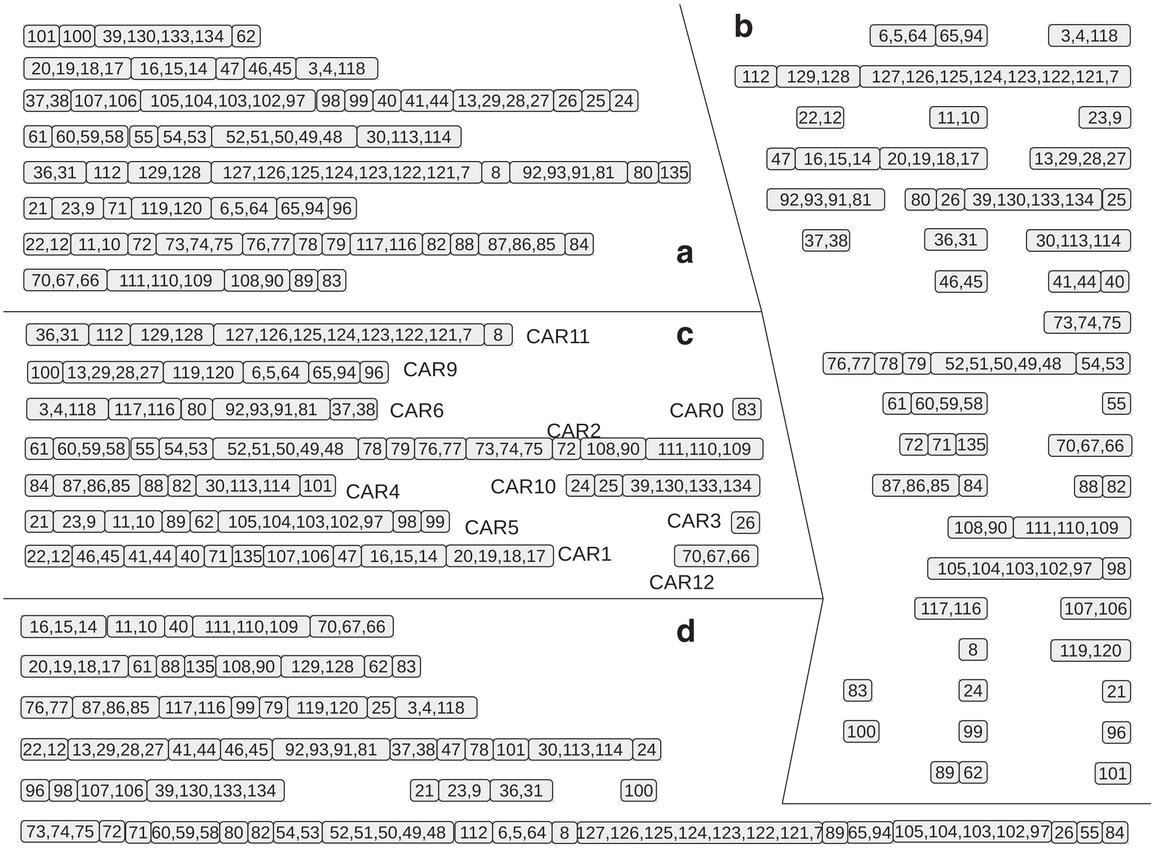

FIG. 6. The arrangements of the non-duplicated yeast ancestral genome obtained as follows: (**a**) by Jean et al. (2009), (**b**) by the physical mapping method on the low resolution marker set, (**c**) by a projection on the low resolution set of the physical mapping method on the high resolution set, and (**d**) by a median solver on the low resolution set. There are 57 blocks of markers common to the four proposed ancestor, contrasting with the 13 ones in the pre-duplication study. Numbers in a block correspond to the markers from the low resolution set which are contained in the block. The four arrangements are very different, showing a high divergence of the methods, explained by the low rate of reliable and supported features in the low resolution marker set. The most reliable arrangement is the (c), as it is achieved on the high resolution set, which has better properties in terms of reliable features. The CAR numbers are reported, and correspond to those in Figure 5 (CARs 7, 8, and 13 don't contain markers of the low resolution set).

All indicators coherently warn against the reliability of a result based on this low resolution instance: large number of CARs, low proportion of reliable adjacencies, divergence of the methods, gap between the median solution and the lower bound.

Next, we computed the same indicators on the high resolution marker set. The physical mapping and rearrangement approaches are not convergent. This suggests that, despite a high proportion of adjacencies common to all optimal solutions, there is still a large number of optimal solutions, probably due to a high number of genome rearrangements during the evolution of these genomes. Nevertheless, even if the rearrangement method faces the saturation of the signal and the multiplicity of solutions, we think that the results we obtain with this marker set illustrate the good possibilities for computational methods, confirmed by the low number of CARs found by the physical mapping method, that we consider as the most reliable ancestor.

4. CONCLUSION

We attempted the reconstruction of two ancestral genomes of a yeast clade in which 7 genomes have been sequenced and assembled. Two of the considered species have undergone a whole genome duplication. We used several methodological principles and combined them to retrieve as much information as possible from the data.

The conference version of this article (Tannier, 2009) showed that computational methods are able to handle data in a whole genome duplication context and to construct ancestors of duplicated species with a high accuracy. On the ‘‘pre-duplication ancestor,’’ physical mapping and rearrangement methods provide results close to the biological standard.

Here we try in addition to understand why in this case the results were good, while previous attempts have put a doubt on the possibilities of computational methods to reconstruct reliable ancestors. To this aim, we introduce some notions of reliability and support in the ancestral configuration, based on the methodological principles of two different types of methods. The study of the ‘‘non-duplicated ancestor’’ provides some hints on this question. We tested two orthology block sets for the same ancestor, one of low resolution published by Jean et al. (2009) and one of high resolution that we constructed here. For the low resolution set, physical mapping and rearrangement methods find significantly different results, and none are close to the one published by Jean et al. (2009). The number of supported and reliable adjacencies is low. So it is likely that the physical mapping method gives a higher number of contiguous ancestral regions than the probable number of chromosomes, while the rearrangement method gives unreliable results. The conclusion is simply that this dataset does not allow for the reconstruction of a reliable ancestor due to a lack of signal. Switching to the high resolution marker set, we have better hopes that it contains more signal, since the number of CARs become lower, and the number of reliable adjacencies higher. The median score of the physical mapping solution is still far from being optimal, putting doubts on the ability of rearrangement methods to solve this dataset, while we have a good confidence in the results of the physical mapping method.

To conclude, we claim that some combinatorial characteristics of the instances are good indicators of the possible reliability of the results given by a rearrangement method. This study helps to understand why and in which context some methods can be trusted and suggests that deficiencies of automatic computational methods can, at least partially, be addressed by a better understanding of the mathematical properties of such methods. The recent trend of research that investigates properties of the DCJ median and related problems (Zhao and Bourque, 2009; Alekseyev and Pevzner, 2009; Tannier et al., 2009; Xu, 2009; Swenson and Moret, 2009) proved to be useful in the present work and shed new light on the reconstruction of ancestral yeasts genomes.

5. ACKNOWLEDGMENTS

We thank Géraldine Jean and Macha Nikolski for providing the low-resolution marker set, Agence Nationale pour la Recherche (grant ANR-08-GENM-036-01 ‘‘CoGeBi’’), Centre National de la Recherche Scientifique (CNRS), Ministère des Affaires étrangères et européennes (programmes ECO-NET et collaboration France-Canada), and Natural Sciences and Engineering Research Council of Canada (NSERC).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Alekseyev, M.A., and Pevzner, P.A. 2009. Breakpoint graphs and ancestral genome reconstructions. *Genome Res.* 19, 943–957.

Alizadeh, F., Karp, R.M., Weisser, D.K., et al. 1995. Physical mapping of chromosomes using unique probes. *J. Comput. Biol.* 2, 159–184.

Bergeron, A., Mixtacki, J., and Stoye, J. 2006. A unifying view of genome rearrangements. *Lect. Notes Comput. Sci.* 4175, 163–173.

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

Byrne, K.P., and Wolfe, K.H. 2005. The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* 15, 1456–1461.

Chauve, C., Haus, U., Stephen, T., et al. 2009. Minimal conflicting sets for the consecutive ones property in ancestral genome reconstruction. *Lect. Notes Comput. Sci.* 5817, 48–58.

Chauve, C., and Tannier, E. 2008. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genome. *PLoS Comput. Biol.* 4, e1000234.

Darling, A.E., Mikls, I., and Ragan, M.A. 2008. Dynamics of genome rearrangement in bacterial populations. *PLoS Genet.* 4, e1000128.

de Peer, Y.V. 2004. Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.* 5, 752–63.

Dietrich, F., Voegeli, S., Brachat, S., et al. 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* 304, 304–307.

Dobzhansky, T., and Sturtevant, A.H. 1938. Inversions in the chromosomes of *Drosophila pseudoobscura. Genetics* 23, 28–64.

Eriksen, N. 2007. Reversal and transpositiom medians. *Theoret. Comput. Sci.* 374, 111–126.

Fertin, G., Labarre, A., Rusu, I., et al. 2009. *Combinatorics of Genome Rearrangements*. MIT Press, Cambridge, MA.

Froenicke, L., Caldés, M.G., Graphodatsky, A., et al. 2006. Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes? *Genome Res.* 16, 306–310.

Gavranovic, H., and Tannier, E. 2010. Guided genome halving: provably optimal solutions provide good insights into the preduplication ancestral genome of *Saccharomyces cerevisiae*. *Proc. Pac. Symp. Bioinform. 2010* 21–30.

Gordon, J.L., Byrne, K.P., and Wolfe, K.H. 2009. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genet.* 5, e1000485.

Jean, G., Sherman, D., and Nikolski, M. 2009. Mining the semantics of genome superblocks to infer ancestral architechtures. *J. Comput. Biol.* 16, 1267–1284.

Kellis, M., Birren, B., and Lander, E.S. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae. Nature* 428, 617–624.

Kemkemer, C., Kohn, M., Cooper, D.N., et al. 2009. Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. *BMC Evol. Biol.* 24, 84.

Ma, J., Ratan, A., Raney, B.J., et al. 2008. Dupcar: reconstructing contiguous ancestral regions with duplications. *J. Comput. Biol.* 15, 1007–1027.

Ma, J., Zhang, L., Suh, B., et al. 2006. Reconstructing contiguous regions of an ancestral genome. *Genome Res.* 16, 1557–1565.

Murphy, W., Larkin, D., van der Wind, A.E., et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309, 613–617.

Ouangraoua, A., Boyer, F., Tannier, E., et al. 2009. Prediction of contiguous ancestral regions in the amniote ancestral genome. *Lect. Notes Comput. Sci.* 5542, 173–185.

Rocchi, M., Archidiacono, N., and Stanyon, R. 2006. Ancestral genome reconstruction: an integrated, multi-disciplinary approach is needed. *Genome Res.* 16, 1441–1444.

Sankoff, D. 2009. Reconstructing the history of yeast genomes. *PLoS Genet.* 5, e1000483.

Sankoff, D., Sundaram, G., and Kececioglu, J. 1996. Steiner points in the space of genome rearrangements. *Int. J. Found. Comput. Sci.* 7, 1–9.

Simillion, C., Janssens, K., Sterck, L., et al. 2008. i-adhore 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* 24, 127–128.

Souciet, J.-L., Dujon, B., Gaillardin, C., et al. 2009. Comparative genomics of protoploid Saccharomycetaceae. *Genome Res.* 19, 1696–1709.

Swenson, K.M., and Moret, B.M.E. 2009. Inversion-based genomic signatures. *BMC Bioinform.* 10, Suppl 1, S7.

Tannier, E. 2009. Yeast ancestral genome reconstruction: the possibilities of computational methods. *Lect. Notes Comput. Sci.* 5817, 1–12.

Tannier, E., Zheng, C., and Sankoff, D. 2009. Multichromosomal median and halving problems under different genomic distances. *BMC Bioinform.* 10, 120.

Xu, A. 2009. Dcj median problems on linear multichromosomal genomes: graph representation and fast exact solutions. *Lect. Notes Comput. Sci.* 5817, 70–83.

Zhao, H., and Bourque, G. 2009. Recovering genome rearrangements in the mammalian phylogeny. *Genome Res.* 19, 934–942.

Zheng, C., Zhu, Q., Adam, Z., et al. 2008. Guided genome halving: hardness, heuristics and the history of the hemiascomycetes. *Bioinformatics* 24, i96–i104.

Address correspondence to:
Dr. Eric Tannier
INRIA Rhône-Alpes
Université de Lyon
F-69000, Lyon

E-mail: Eric.Tannier@inria.fr