BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 19 2011, pages 2664–2671
doi:10.1093/bioinformatics/btr461

Phylogenetics

Advance Access publication August 16, 2011

Reconstructing the architecture of the ancestral amniote genome

Aïda Ouangraoua[1], Eric Tannier[2] and Cedric Chauve[3,*]

[1]INRIA Lille-Nord-Europe, Université Lille 1, LIFL, UMR CNRS 8022, Villeneuve d'Ascq, [2]INRIA Grenoble Rhône-Alpes, Université Lyon 1, LBBE, UMR CNRS 5558, Villeurbanne, France and [3]Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada

Associate Editor: David Posada

ABSTRACT

Motivation: The ancestor of birds and mammals lived approximately 300 million years ago. Inferring its genome organization is key to understanding the differentiated evolution of these two lineages. However, detecting traces of its chromosomal organization in its extant descendants is difficult due to the accumulation of molecular evolution since birds and mammals lineages diverged.

Results: We address several methodological issues for the detection and assembly of ancestral genomic features of ancient vertebrate genomes, which encompass adjacencies, contiguous segments, syntenies and double syntenies in the context of a whole genome duplication. Using generic, but stringent, methods for all these problems, some of them new, we analyze 15 vertebrate genomes, including 12 amniotes and 3 teleost fishes, and infer a high-resolution genome organization of the amniote ancestral genome, composed of 39 ancestral linkage groups at a resolution of 100 kb. We extensively discuss the validity and robustness of the method to variations of data and parameters. We introduce a support value for each of the groups, and show that 36 out of 39 have maximum support.

Conclusions: Single methodological principle cannot currently be used to infer the organization of the amniote ancestral genome, and we demonstrate that it is possible to gather several principles into a computational paleogenomics pipeline. This strategy offers a solid methodological base for the reconstruction of ancient vertebrate genomes.

Availability: Source code, in C++ and Python, is available at http://www.cecm.sfu.ca/~ cchauve/SUPP/AMNIOTE2010/

Contact: cedric.chauve@sfu.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 17, 2010; revised on June 17, 2011; accepted on August 2, 2011

1 INTRODUCTION

The reconstruction of ancestral karyotypes and gene orders from homologies between extant species contributes to understanding the large-scale evolutionary mutations that led to current genomes from their last common ancestor. For example, given ancestral genomes or gene orders, distances and evolutionary scenarios between genomes can be computed along branches of the considered phylogenetic tree (Miklos and Tannier, 2010) and not only between pairs of extant species. More generally, one can hope to work within an 'evolutionary framework combining ancestral and extant genomes in a robust phylogenetic tree' (Muffato and Roest-Crollius, 2008).

The ancestral reconstruction problem has been approached for eukaryotic genomes using cytogenetics techniques pioneered by Dobzhansky and Sturtevant (1938) and recently applied to mammalian genomes (Froenicke, 2005; Froenicke *et al.*, 2003; Richard *et al.*, 2003; Wienberg, 2004; Yang *et al.*, 2003). However, when dealing with older ancestors and larger evolutionary distances, homologies are less visible by cytogenetics methods. It is only with the recent availability of sequenced and assembled genomes that bioinformatics methods can handle the problem of predicting the deep past of metazoan chromosomes [reviews by Faraut (2008); Muffato and Roest-Crollius (2008); Rascol *et al.* (2007)]. These methods address the problem at a higher resolution than cytogenetics techniques, as they require assembled genomes, either fully or at least in large contigs and scaffolds. The results obtained on mammalian genomes, based on genome rearrangement models (Alekseyev and Pevzner, 2009; Bourque *et al.*, 2004; Murphy *et al.*, 2005) or on physical mapping methods (Chauve and Tannier, 2008; Kemkemer *et al.*, 2009; Ma *et al.*, 2006; Mikkelsen *et al.*, 2007; Muffato *et al.*, 2010), slowly converge toward the cytogenetics ones (Ferguson-Smith and Trifonov, 2007).

Prospective methods have attempted the reconstruction of more ancient animal proto-genomes: amniotes (Kohn *et al.*, 2006; Nakatani *et al.*, 2007), bony fishes (Catchen *et al.*, 2008; Jaillon *et al.*, 2004; Woods *et al.*, 2005), vertebrates (Kohn *et al.*, 2006; Nakatani *et al.*, 2007; Naruse *et al.*, 2004), chordates (Putnam *et al.*, 2008) or even eumetazoa (Putnam *et al.*, 2007). However, the accuracy of these *ad hoc* methodologies has not been studied as thoroughly as for more generic methods. The ability of the latter to produce reliable results on mammalian or yeast genomes was evaluated either by comparisons to cytogenetics or manual reconstructions (Chauve and Tannier, 2008; Chauve *et al.*, 2010). The validation of the results is indeed a major issue, since ancient genomes can not be sequenced due to DNA decay. Simulations, although used in a few studies (Bertrand *et al.*, 2010; Ma *et al.*, 2006; Muffato *et al.*, 2010), face the currently limited understanding of large-scale evolutionary processes such as genome rearrangements and gene losses following a Whole Genome Duplication (WGD). Cytogenetic results or manual reconstructions are not available for very ancient genomes.

The goal of the present work is to propose a general methodology for inferring ancestral genome segments and linkage groups, based on several comparative genomics principles grouped into a *physical mapping* framework. The framework we describe can be applied to any ancestral species following the speciation from the teleost

*To whom correspondence should be addressed.

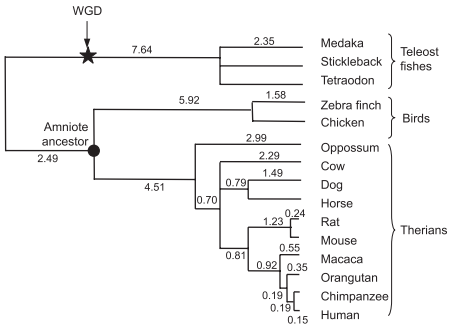

Fig. 1. Phylogeny of the 15 considered species: 12 amniotes (birds and therians) and 3 teleosts fishes. The whole genome duplication at the base of the teleost lineage is indicated by a star symbol.

fishes (tetrapods, sarcopterigians, for example), provided sequenced extant genomes from two branches descending from an ancestor are available. We apply it to the ancestral amniote genome. We study the validity and robustness of the method and results, and we argue that it is possible to evaluate an ancestral genome reconstruction based on the careful examination of the underlying principles of the method.

These principles are: (i) the detection of ancestral features (adjacencies, collinear or contiguous segments, syntenic groups) based on their conservation in the extant genomes; and (ii) the assembly of such features into ancestral chromosomal segments and linkage groups. Hence, we design the reconstruction process as a pipeline decomposed into several phases, each focused on the detection/assembly of a specific kind of ancestral genomic feature. This requires to define, for each such phase, an appropriate combinatorial representation of the corresponding assembled ancestral features, including PQ-trees to represent contiguous ancestral regions as in Bergeron *et al.* (2004); Chauve and Tannier (2008); Ma *et al.* (2006) and graphs to represent ancestral linkage groups (ALGs), as in Muffato *et al.* (2010). The final representation of an ancestral genome is a nested combination of such combinatorial objects.

We apply this pipeline to reconstruct the amniote ancestral genome at a 100 kb resolution, tailoring the method to account for the phylogeny and presence of a whole genome duplication at the base of the teleost lineage (Fig. 1). We obtain 39 ALGs, covering a significant part of the extant amniote genomes, which is a improvement on the two previous studies of this ancestral genome (Kohn *et al.*, 2006; Nakatani *et al.*, 2007). The syntenic associations between extant amniote chromosomes that we predict confirm three associations found in previous studies, but also include two new ones. We discuss the accuracy of the results produced by our method, and the robustness of the method to variations of data and parameters.

2 METHODS

We give a generic description of the different objects, and call $\mathcal{A}$ the ancestral species, identified by an internal node in a species phylogeny, of which we reconstruct the genome. Descendants of $\mathcal{A}$ are the *ingroup* species, and all others are the *outgroups*. An *informative pair* of species is composed of two extant species whose evolutionary path in the phylogenetic tree contains $\mathcal{A}$.

Ancestral blocks: we use orthologous genomic segments among ingroup genomes to derive *ancestral genomic blocks*. An ancestral block represents an oriented ancestral genome segment that evolved into a unique genomic segment in each extant genome, with limited internal rearrangements and point mutations, and that was not involved in a large-scale duplication. These ancestral blocks represent the basic bricks of the ancestral genome. They are constructed using Ensembl Compara multiple alignments (Paten *et al.*, 2008) as seeds. A block result from joining seeds collinear in all ingroup genomes, if it spans at least min_len = 100 kb in all ingroup genomes (see Supplementary Material).

Contiguous ancestral regions: we assemble ancestral blocks into *contiguous ancestral regions (CARs)* (Chauve and Tannier, 2008; Ma *et al.*, 2006). CARs are groups of ancestral blocks with partial ordering information. To this aim, we detect conserved *contiguity* signal in amniote extant genomes: maximum common intervals, conserved adjacencies and adjacencies inferred from reliable rearrangements.

Ancestral synteny graph and ALGs: next, we cluster the CARs by constructing the *ancestral synteny graph (ASG)*. CARs are linked by detecting a significant conserved *synteny* signal between pairs of blocks, including a *double synteny* signal when a WGD separates the compared extant species. ALGs are defined as connected components of the ASG, after edges with weaker support have been discarded.

2.1 Computing CARs

Concepts and methods to construct and represent CARs are extensively described in Chauve and Tannier (2008); Chauve *et al.* (2010), and we review them briefly here. We consider the *extremities* (head and tail) of the ancestral blocks that we call *ancestral markers*. Using the extremities of the blocks as markers is necessary to infer the ancestral orientation of each ancestral block.

Ancestral contiguous sets: for a given ancestral genome $\mathcal{A}$, an *ancestral contiguous set* (ACS) is a set of ancestral markers that are believed to be contiguous in $\mathcal{A}$, with no prior information on their order along the ancestral chromosome that contained them. An ACS with two markers is called an *adjacency*. We consider four types of ACS:

- *Block adjacencies*, defined as pair of extremities of a same ancestral block. The two markers composing a block adjacency are adjacent in all genomes.
- *Supported adjacencies*, defined as pairs of markers that are adjacent in both genomes of an informative pair.
- *Maximal common intervals* between both genomes of an informative pair. A maximal common interval between two genomes is a set of markers that are contiguous in both genomes (possibly with different orders) and is not included into another common interval between the same two genomes.
- *Reliable adjacencies*, defined as adjacencies that can be inferred from probable genome rearrangements (Chauve *et al.*, 2010; Zhao and Bourque, 2009).

Assembling ACS into CARs: an ordering of all the markers *satisfies* a given ACS S if the markers of S form a contiguous interval in this ordering. A set of ACS $\mathcal{S}$ is said to be in *conflict* if there is no total ordering of the ancestral markers which satisfies all ACS of $\mathcal{S}$. For example, if x, y and z are ancestral markers and $\{x,y\}$, $\{x,z\}$, $\{z,y\}$ are three ACS, then there is no total ordering of x, y and z such that the markers in each of the three pairs are contiguous. If a set of ACS is not in conflict, all the orderings which satisfy them can be represented by a structure called a PQ-tree (Booth and Lueker, 1976), widely used in physical mapping studies. A PQ-tree defines CARs, from which it is possible to order, at least partially, ancestral markers. If a set of ACS is in conflict, then each ACS is weighted according to its conservation in extant genomes (see Supplementary Material), and CARs are obtained from the subset of non-conflicting ACS with the highest summed weight.

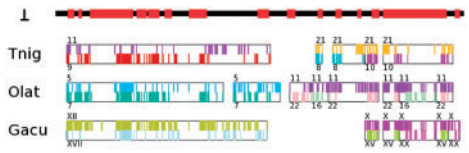

Fig. 2. Examples of double conserved syntenies mapped onto a segment of human chromosome 1 representing 18% of its length. The top row represents the human chromosome segment and the ancestral blocks it contains. The three bottom rows represent the DCS with tetraodon (Tnig), medaka (Olat), and stickleback (Gacu): each box is a DCS, with the top (respectively bottom) half representing the genes mapping on a teleost fish chromosomes, whose number appears above (respectively below) the box.

We ensure all block adjacencies are conserved by giving them a high weight that prevents them from being discarded [See Chauve and Tannier (2008) for a precise description].

2.2 Regrouping CARs into the ASG

The contiguity condition used to define ACSs and compute CARs is too stringent to define purely syntenic features of an ancestral genome. Indeed, some groups of markers may have been kept syntenic (on the same chromosome) during evolution, but spread out along this chromosome by genome rearrangements. Also, assembly errors, related to contigs ordering for example, can lead to an apparent loss of contiguity between group of syntenic markers. It is then necessary to relax the condition of contiguity in order to detect more subtle homology signal from the comparison of extant genomes composing an informative pair. We describe below two ways to detect such signal.

As the ancestral blocks we constructed are limited to the ingroups and we want to detect syntenies common to ingroups and outgroups as well, we use orthologous genes retrieved from Ensembl Compara [Vilella *et al.* (2009), see Supplementary Material] to detect the syntenic signal.

Ancestral synteny: for a given informative pair (A,B) in which no species is involved in a whole genome duplication, a *contiguous sets of autosomal markers (CSAM)* (Kumar *et al.*, 2001) is a set O of contiguous genes of A, such that all their orthologs are in a same chromosome in B. The set O of genes is called an *ancestral synteny* if its conservation in B is statistically significant according to a test described in Durand and Sankoff (2003) (see Supplementary Material for a detailed description of the tests and the controls for multiple testing).

To detect CSAMs, we use informative pairs where A is a bird genome and B is a therian genome, or conversely.

Double ancestral synteny: for a given informative pairs (A,B) for which a whole genome duplication happened in the lineage of B, we expect to see a set of genes that span one segment in A and their orthologs span two segments in B (Fig. 2).

The general principle defining double syntenies is relatively simple, but due to different patterns of genome rearrangements and gene losses following a WGD (Hufton and Panopoulou, 2009), it has been applied with different *ad hoc* implementations in vertebrates (Jaillon *et al.*, 2004; Pham and Pevzner, 2010; Sémon and Wolfe, 2007c), plants (de Peer, 2004) or yeasts (Dietrich *et al.*, 2004; Kellis *et al.*, 2004). We are not aware of any formal notion of double synteny. It is a contribution of the present method to propose such a formalization, which can be used widely in any clade of the living world where we know whole genome duplications happened (yeasts or plants). We use a generalization of the gene teams defined in Luc *et al.* (2003). An *$\alpha-\beta$-double team* is a set O of genes that span one segment S_A in A and two segments S_B^1 and S_B^2 in B, such that, along S_A (respectively S_B^1, S_B^2), there is no gap of size greater than α (respectively β). The set O of genes is called a *double conserved synteny* (DCS) if its conservations in A and B are statistically significant, and if there is no significantly clustered segment S_B^3 in B, disjoint from S_B^1 and S_B^2, which has orthologs in S_A.

This definition relies on two parameters (the gaps parameters α and β) which can vary to include a flexibility to noisy data as well as a high post-WDG rearrangement rate in B. We choose the parameters maximizing the coverage of A by DCS (due to forbidding the presence of a segment S_B^3 and the use of statistical tests, the coverage can decrease when α increases).

With our data, the parameters maximizing the coverage of extant amniote genomes by DCS are $\alpha=1$, accounting for possible isolated wrong gene annotation or retro-transpositions for example, and $\beta=\infty$, which matches the high rate of rearrangement after the teleost WGD (Hufton and Panopoulou, 2009; Sémon and Wolfe, 2007b). Supplementary Figure S3 shows the coverage of the chicken genome by DCS with different values of α and β.

Assembling ancestral syntenies into the ASG: if two blocks b_1 and b_2 are contained in a same genome segment of an ingroup which is involved either in an ancestral synteny or a double ancestral synteny, then b_1 and b_2 are believed to have been syntenic in the ancestral genome $\mathcal{A}$. Since, each block belongs to a CAR, then we get a set of synteny relationships between CARs. This naturally motivates the representation of a set of ancestral syntenies as a graph, whose vertices are the CARs, and where there is an edge between two vertices C_1 and C_2 if $C1$ (respectively C_2) contains a block b_1 (respectively b_2) such that b_1 and b_2 both belong to an ingroup genome segment that is involved in an ancestral or double ancestral synteny. We call this graph the *Ancestral Synteny Graph*; a connected component of this graph represents a set of CARs that are assumed to belong to a same proto-chromosome, and we call it an ALG (with no additional information on the ordering of the CARs along this proto-chromosome).

Due to the relaxed notion of conservation used to define the edges of the ASG (contiguity is enforced in a single genome) and the larger evolutionary distance between ingroups and outgroups, the risk of convergent evolution or inaccurate ancestral syntenies is higher. To account for this issue, edges can be examined individually and filtered, using phylogenetic support, DCS statistical support or alignment quality between blocks for example.

3 RESULTS

We implemented all the principles described above and analyzed a dataset containing 12 amniote genomes (human, chimpanzee, orangutan, macaca, mouse, rat, dog, cow, horse, opossum, zebra finch, chicken) and 3 teleost fish genomes (tetraodon, stickleback, medaka). The 12 amniote genomes were chosen because they were assembled and present in a multiple alignments of 16 amniote genomes available in the release 58 of the Ensembl database obtained using Pecan (Paten *et al.*, 2008). All data (multiple alignments of amniote genomes, gene contents and gene trees) were downloaded from the Ensembl 58 Compara database (Vilella *et al.*, 2009).

The phylogeny, including branch lengths, was taken from paleontological data (Benton and Donoghue, 2007) (Fig. 1). Branch lengths are defined to be the normalized lower bounds of speciation times given by paleontological evidences, in order to avoid any *a priori* hypotheses þ, on rearrangement rates.

We computed ancestral blocks with resolution min_len = 100 kb, and we obtained 773 ancestral blocks, together with their orientations in the amniote genomes, resulting in 1546 ancestral markers (two markers per block, one for each extremity).

3.1 Amniote contiguous ancestral regions and ALGs

We detected 1861 ACS: 773 block adjacencies, 457 supported adjacencies, 8 reliable adjacencies and 623 maximum common intervals. We discarded 10 of these ACS (2 supported adjacencies,

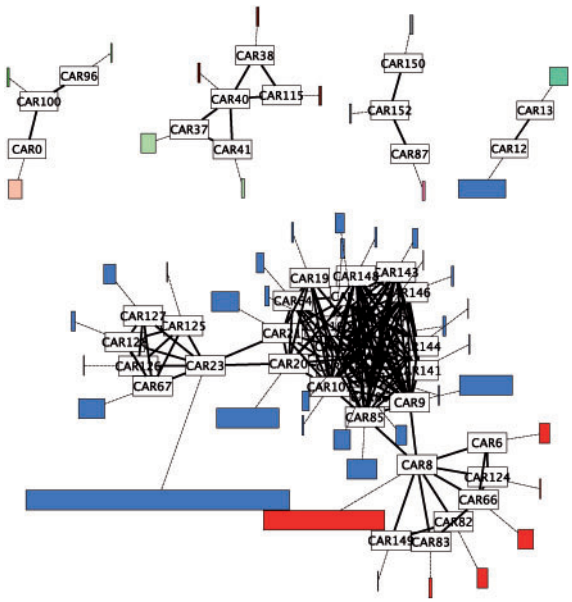

Fig. 3. Five connected components of the ASG. Vertices represented by white rectangles are CARs. The colors of the rectangles linked to the CARs by dashed lines represent the chicken chromosomes to which the corresponding blocks belong: beige (21), green (26), light green (18), brown (27), gray (22), pink (Z), blue (1), turquoise (24) and red (2).

3 reliable adjacencies and 5 maximum common intervals) to obtain a non-conflicting set of ACS that resulted in 164 CARs. The low number of reliable adjacencies, when compared with mammalian studies (Zhao and Bourque, 2009), is probably due to the large evolutionary distance between birds and therians and illustrates the less clear rearrangement signal that can be detected at this evolutionary scale.

We then computed the CSAMs, the DCSs and the ASG using the gene orthologies from Ensembl 58. We retained only edges of the ASG with a strong phylogenetic support: a pair of ancestral blocks defines an edge between two CARs in the the ASG if and only if both blocks belong to an amniote segment involved in an ancestral synteny or double ancestral synteny, and both blocks are syntenic in genomes of at least three clades of the four amniote clades we considered: birds, opossum, laurasiatheria (cow, horse, dog), euarchontoglire (primates and rodents). This resulted in 39 connected components representing putative amniote linkage groups, partially depicted on Figure 3 (See Supplementary Information for the whole ASG picture).

The largest connected components mainly consists of two or three very dense graphs linked through a few number of CARs.

A finer analysis of the contribution of each type of signal we considered (adjacencies, common intervals, ancestral syntenies and double syntenies) shows several interesting properties. First CARs are mostly defined by common intervals (common intervals only define 165 CARs). The impact of adjacencies is then more in ordering blocks within CARs. Second, most connected components of the ASG are due to edges induced by CSAMs: discarding the DCS would result in 44 connected components. However, all syntenic associations are due the very few edges (24) edges induced uniquely by DCS. Details are provided in Supplementary Material.

Table 1. Number of chicken protein-coding genes covered by the amniote ancestor reconstructed in three different studies

Kohn *et al.* (2006)	Nakatani *et al.* (2007)	Present method
3283	5883	10759/13996

The gene coverage for the two previous studies were calculated from Supplementary Table S1 in Kohn *et al.* (2006) and Supplementary Table S3 in Nakatani *et al.* (2007). For the present method, we show both the numbers of genes covered by the CARs and by the ALGs, showing the contribution of the integrative appraoch.

Table 2. Inferred syntenic associations between chicken chromosomes in three different studies

Kohn *et al.* (2006)	Nakatani *et al.* (2007)	Present method
18-19-27		18-27
1-24		1-24
21-23-26-32	26-32	21-26
17-Z	10-13-17-Z	
2-9-16	2-9	
3-14	1-3-14-18, 3-14	
5-10, 4-22, 8-28		
	1-7, 3-5	
		1-2, Z-22

Numbers refer to chicken chromosomes, and association between numbers means that segments from these chromosome are assumed to descend from a single ancestral amniote chromosome.

3.2 Structure of the amniote ancestral genome

We say a region of an extant genome is covered by an amniote ALG if all blocks contained in this region belong to the ALG. The ancestral groups we inferred cover 972 Mb (94%) of the chicken genome and 2540 Mb (83%) of the human genome. In terms of genes, they cover 13996 (91%) of the chicken genes, and 15628 (76%) of the human genes. This coverage of the chicken and human genomes is much higher than what was attained by previous studies (Kohn *et al.*, 2006; Nakatani *et al.*, 2007) (Table 1 and Supplemental Material). The study of Muffato *et al.* (2010), while attaining a similar coverage, gives a much more fragmented configuration (606 ancestral segments with >1 gene for the amniote genome) as it uses more genomes, but only adjacencies to infer ancestral features.

To push further the comparison with the previous proposals for the ancestral amniote genome (Kohn *et al.*, 2006; Nakatani *et al.*, 2007), we use features of the reconstructions called *syntenic associations*. A syntenic association is the presence in a proto-chromosome (a single component of the ancestral synteny graph), of blocks from two different chromosomes of an amniote extant genome chosen as reference. In Table 2, we report all syntenic associations we detected on the chicken genome, as well as the ones reported by two previous studies [Muffato *et al.* (2010) do not propose any syntenic association].

3.3 Robustness

Our method relies on few parameters. Among them, the ones that define the DCS are optimized to get a maximum coverage of amniote genomes (Supplementary Fig. S3). Their variation is then part of the

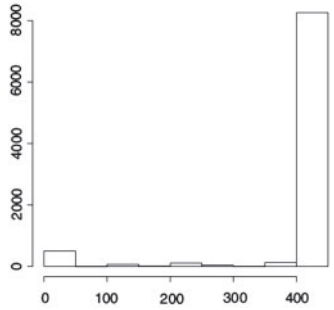

Fig. 4. For min_*len* = 100 kb, frequency of the pairs of genes being in two different ALGs, in function of the number of situations over the 440 runs. One bar of heigth y at position x ($0 \leq x \leq 440$) means that y pairs of genes were found in different ALGs in the amniote ancestral genome in x situations over 440.

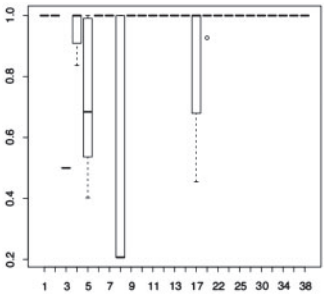

Fig. 5. Stability of the ALGS presented in the Section 3, that is, using min_len = 100 kb, and all available genomes. On the x-axis are the 39 ALGs (in an arbitrary order). On the y-axis is plotted, for each ALG A, the distribution of the number of runs (over 440) in which two genes belonging to A are in a same ALG.

construction process, so we did not include them in the robustness analyses.

Instead we made the set of sampled species and the resolution parameter min_len (the minimum length of a block) vary. We considered 11 variations of the set of amniote genomes obtained by either retaining all the 12 genomes or discarding 1 of the 10 non-reference genomes (all genomes except human and chicken). Four variations of the set of teleost fish genomes were also considered by either retaining all the three genomes or discarding one of them. The parameter min_len varied from 50 to 500 kb by steps of 50 kb, leading to 10 variations. In total, we performed $11 \times 4 \times 10 = 440$ different experiments.

The number of ALGs is quite stable: it ranges from 34 to 47, with most situations resulting in 36–39 ALGs (Supplementary Fig. S5). Interestingly, Supplementary Figure S4 shows the stability of the number of CARs, which is much less visible. Again, this shows the interest of the integrative method, since each feature can be unstable while the whole is not (see a discussion on this in the Supplementary Material). We then wanted to know how much these ALGs from different runs looked alike. For this, we randomly selected 100 000 pairs of chicken genes, and filtered those covered by blocks in all 440 experiments. For each such pair of genes, we recorded the number of situations over the 440 ones in which they belong to two different ALGs (Fig. 4). We observe two predominant modes at 0-49 and 400-440. Because of the small probability to randomly select two genes that are in the same ALG, the 0-49 mode is obviously smaller, but this shows a relative constance in the choices to group genome parts or not. Indeed, a perfectly robust method would show only values 0 and 440, whereas a non-robust one would have a dispersed signal.

We then evaluated the robustness of the result we present, that is, the ALGs computed with all genomes and a resolution of min_len=100 kb. Again, we sampled a high number of pairs of genes belonging to blocks in all 440 runs. Then for each particular ALG, we recorded the number of runs in which both genes still belong to a same ALG. The results are shown in Figure 5.

In this figure, each column represent the stability of an ALG A. A bold dot marks the mean of the distribution of the number of runs (over 440) in which two genes belonging to A are in a same ALG. Among the 39 ALGs, 33 have a perfect stability, meaning that genes in this ALG are always together than ALG in the other situations, three having a good stability, with a mean close to 1. Three ALGs show a bad stability, with a support between 0.2 and 0.7. This analysis put a doubt in the structure of these three ALGs at the 100 kb resolution. The number of low supported ALGs decreases with the resolution. In Supplementary Figure S6, the same analysis is reproduced for min_len = 500 kb. All 32 ALGs in this case have a support $> 90\%$, and 30 of them have 100%. This suggests there is a trade-off between the resolution and the support of the results.

4 DISCUSSION

We described a general methodology for inferring ancestral genomes, that relies on a hierarchical process: computing ancestral blocks, computing CARs and computing the ASG and ALGs. We implemented this general approach using stringent methods for each phase, tailored to account for the specificity of the evolution of amniote genomes, and we applied it to infer large ancestral amniote linkage groups at a 100 kb resolution, from whole-genome alignments and gene families. This resulted in the first proposal of an ancestral amniote genome that covers large parts of its extant descendants.

4.1 Comparison with previous ancestral amniote genomes

Kohn *et al.* (2006) and Nakatani *et al.* (2007) proposed significantly more amniote syntenic associations than we do. One explanation is the stringency of the method we use to detect double ancestral syntenies. Indeed, due to the stability of bird genomes and the absence of reptile genomes in the considered dataset, the only syntenic associations that can be detected are supported by double ancestral syntenies. We find that the associations 18-27 and Z-22 are the only ones to be supported by a pair of blocks that belong to more than one DCS. This illustrates the fact that combinatorial traces of synteny conservation at a large evolutionary distance are quite degraded and not widespread, even when several amniote genomes are considered. Even if we cannot definitively argue that the syntenic associations described in the previous ancestral genome proposals are wrong, we believe that they rely on a weaker synteny conservation signal as illustrated by the case of the putative syntenic association 17-Z (see discussion below). We even found no evidence that the medaka genome is as stable as claimed by Nakatani *et al.* (2007) (for example, Fig. 2 does not show more conserved DCS signal in the medaka than in the two other teleost fishes).

Table 3. Syntenic associations of chicken chromosomes, compared with the syntenic associations of human chromosomes

Chicken syntenic associations	Human syntenic association
18-27, 1-24, 21-26, 1-2, Z-22	1-2-3-7-10-12-13-15-21-22-X 1-6, 14-15, 7-16-17, 5-6-8-18 5-9-18, 2-8, 2-3, 4-8, 12-22 16-19, 2-20, 3-9

We now discuss a few examples of syntenic associations that were reported by both (Kohn *et al.*, 2006; Nakatani *et al.*, 2007) but not by our method, as they illustrate the interest of methodological discussions to analyze such discrepancies. The association between chicken chromosomes 26 and 32 was out of reach for us as no ancestral block did belong to chicken chromosome 32. The signal we observe for an association between chicken chromosomes 2 and 9 does not support clearly the fact it was present in the amniote ancestor: it can be related to two ancestral blocks, contiguous in all placental mammalian genomes but not in the opossum genome, and that do not belong to a DCS, which indicates more likely an ancestral boreoeutherian feature. Finally, the case of the association 17-Z is interesting. We detect a faint signal for it based on three blocks present on a single DCS in the macaca genome and on the same chromosomes in the human, chimpanzee and orangutan genomes, but which does not satisfy the phylogenetic conservation criterion we use to retain edges of the ASG. There are also two blocks contiguous in all placental mammalians and that also belong to a genome segment that shows a DCS signal that does not satisfy the stringent definition we used. So we leave open the possibility that the association 17-Z is present in the amniote ancestral genome. The signal related to associations 2-9 and 3-14, as well as for associations Z-22 and 1-2, which are specific to our study, is described in Supplementary Material.

4.2 Insights on the evolution of bird genomes

It can be remarked in Table 3 that the ancestral amniote linkage groups we inferred are remarkably close to the bird chromosomes, and somehow more distant from the mammalian ones.

It has already been remarked that bird genomes show an incredibly conserved degree of synteny (Backström *et al.*, 2008; Consortium, 2004; Ellegren, 2010; Griffin *et al.*, 2007, 2008; Hansson *et al.*, 2010; Skinner *et al.*, 2009; Stapley *et al.*, 2008). It has been claimed that the avian ancestral karyotype is only one fission away from the chicken one, concerning chicken chromosome 4. Our analysis, that splits chicken chromosome 4 into three ancestral linkage groups, including two small and one large segments (1.7, 5 and 41 Mb, respectively), does not contradict this hypothesis. Griffin *et al.* (2007) place the appearance of micro-chromosomes after the divergence with mammals, since no mammal seems to carry some, but the fact that most chicken micro-chromosomes each belong to a single ALG, leaves open the hypothesis of ancestral micro-chromosomes: apart from the 18-27, 21-26 and Z-22 fissions, which concern at least one micro-chromosome each, all micro-chromosomes could be ancestral. Even the groups 18-27 and 21-26 can be considered as ancestral micro-chromosomes.

4.3 Methodological issues, accuracy and validation

A fundamental question regarding any ancestral genome reconstruction method is its accuracy, and the confidence we may have in the results it produces. As such methods aim at inferring an information that has been lost and cannot be recovered, they cannot be assessed on real data. Validation may be obtained by

- comparing the results to well-established ones, as in mammalian data studied by cytogeneticians (Wienberg, 2004), or yeast data curated by experts (Gordon *et al.*, 2009);
- comparing to results obtained, *in silico*, on synthetic data generated by simulating an evolutionary process;
- studying and validating the underlying principles of the method and their implementation;
- studying the robustness of the choices made by the method to small parameter changes or data perturbations.

First, we can remark that, when applied to reconstruct the ancestral boreoeutherian genome, whose structure is widely agreed upon by bioinformaticians and cytogeneticians, the introduction of the ASG does not bring additional information here (i.e. all connected components are single CARs; results not shown). So our method defines this ancestor uniquely in terms of CARs, and obtains results similar to previous reconstructions (Chauve and Tannier, 2008; Ma *et al.*, 2006).

We think that using simulated data is not relevant presently: relevant simulations would require to follow a realistic model of genome structural evolution (here in terms of genome rearrangements), and such a model is currently not available. Crucial elements of this model—for example, rearrangement rates, ratio of different types of events, distribution of breakpoints along genomes, breakpoints re-use, rearrangement and duplication lengths, fate of duplicated genes following a WGD—are still not well understood. Accurate ancestral genomes inferred using model-free methods (methods that do not rely on a genome rearrangement model) will be key to improve our understanding of the evolution of vertebrate genomes and participate to develop better models, that can then be used to refine inference methods and develop realistic simulated datasets.

We then propose here to evaluate our results by highlighting important properties of the method itself, that complement the extensive robustness analysis described in the previous section.

All ancestral features are supported: if two blocks are found adjacent or syntenic in the ancestor, it is immediately possible to recover the signal that defines this feature (extant or reliable adjacency, genome segment defining a common interval, synteny, double synteny). Hence, every such feature can be assessed individually by looking at the precise genome segments, alignments and genes that support it. Although this is not a validation of the method we propose, this provides the appropriate ground for discussing the results it produces, either in the light of evolutionary arguments (such as possible convergent evolution, for example) or from a data processing point of view (alignment quality, orthology detection, DCS boundaries, to cite a few).

This is not the case, for example, of rearrangement techniques (Alekseyev and Pevzner, 2009), where an ancestral adjacency can be inferred while never present in the data, its presence being due to a parsimony criterion in a simplified model of evolution. While

this does not rule out that parsimony methods can provide accurate results, the lack of ways to assess ancestral feature is a problem our method aims at addressing.

Optimization is almost absent: any optimization step based on a simplified evolutionary model (Alekseyev and Pevzner, 2009) or on clustering methods (Muffato *et al.*, 2010) results in arbitrary and uncontrolled choices, and is subject to the pitfall of the multiplicity of optimal (and slightly suboptimal) solutions (Eriksen, 2007). Here we try to reduce the optimization step as much as possible, by being very stringent on the retrieved signals and thus limiting the conflicts. And indeed we discard only 0.5% of the inferred information to construct the CARs, which is the only optimization step we use.

Stringent implementations: we chose to implement each phase in a stringent way, using established comparative genomics concepts. First, we defined ancestral blocks using reference genome alignments from Ensembl that were conserved in *all* amniote genomes (universal seeds) and then joined seeds into blocks if collinear again in all amniote genomes. Hence, we followed a more conservative approach than in most synteny/orthology blocks models, that allow local rearrangements (Chauve and Tannier, 2008; Ma *et al.*, 2006; Pham and Pevzner, 2010).

Next, we inferred CARs mostly from sets of blocks that are contiguous in pairs of informative extant genomes: common intervals and conserved adjacencies; the information added from reliable rearrangements had very little impact. It is important to note here that common intervals are associated with an implicit statistical significance. Indeed, as shown in Xu *et al.* (2008), one can expect to find in a random gene order two common intervals, both of length two. Hence, larger common intervals are statistically significant. Together with the fact that most CARs are defined primarily by common intervals, this shows that the signal defining CARs is significant. Relying on such stringent conserved characters resulted in a set of ACSs that was almost conflict free, which suggests a low rate of false positives.

Regarding the edges of the ASG, to reduce the impact of possible false positives, we relied on statistical tests of synteny conservation, and corrected the whole procedure for multiple testing. Moreover, to add one additional layer of safety, we discarded from the ASG all edges that were not supported in extant genomes of at least three of the four main clades we considered. In conclusion, all ancestral features are defined in terms of statistically significant combinatorial structures.

4.4 Future research

The computational reconstruction of ancestral genomes is still a recent research topic, especially when it comes to ancient genomes with faint traces of synteny conservation. We discuss below possible ways to improve the different stages of the method we introduced.

Ancestral blocks: our work relies strongly on the set ancestral blocks, that are the basic bricks of the reconstruction process: both CARs and edges of the ASG are defined in terms of these blocks. Hence, their accuracy is central to ensure the accuracy of the inferred ancestral genome, and false positives (i.e. blocks that do not descend from a single ancestral genomic segment) can lead to wrong ALGs or syntenic associations. This calls for new research on orthology blocks, their desirable properties and methods to detect them or assess their accuracy. In particular, the traditional approach that relies on universal, or almost universal, seeds (DNA alignment or orthologous genes) to define such blocks needs to be revisited to cope with the increasing number of sequenced vertebrate genomes, as illustrated recently in Pham and Pevzner (2010).

Doubly conserved syntenies: the present work illustrates the interesting potential of DCS to infer syntenic information from genomes having evolved through a WGD (already noticed by several previous works), but it also shows that the lesser constraints on the detected signal requires great care in using such signal to infer ancestral syntenic features. In particular, it should be remarked that a DCS in an amniote might not represent an ancestral genome segment but the union of several ancestral syntenic genome segments. This is the reason why we chose not to use DCSs as ancestral contiguous sets but as synteny signals only. In our previous study (Ouangraoua *et al.*, 2009), using DCSs as ACSs for the amniote ancestor reconstruction led to a high ratio of discarded ACSs (14%), most of the discarded ACSs being linked to DCSs. The problem of detecting high confidence DCS is of importance beyond the case of amniote genomes, as WGD happened in yeasts genome evolution (Sémon and Wolfe, 2007a), or, at a larger scale, in plant genome evolution (Salse *et al.*, 2008, 2009a, b).

The ASG: the combinatorial properties of the ASG deserve deeper investigations. A close look at the graph we obtain for the ancestral amniote genome shows both highly connected sets of CARs, indicating clear groups of syntenic CARs, and weakly connected structures such as bridges, illustrating the faint traces of evolution supporting some syntenic associations. Among the natural questions on the ASG, it remains to see if CARs grouped in dense subgraphs can be ordered. i.e. if they can be arranged in linear structures representing larger ancestral chromosomal segments. This problem could be attacked for example by weighting the edges of the ASG with estimated genomic length representing the distance between blocks along an ancestral chromosome. From a theoretical point of view, this becomes related to the graph bandwidth problem.

Amniote genomes evolution: our dataset is unbalanced, with only two birds. It will be interesting to see the additional information that can be obtained from the turkey, lizard and xenopus genomes. Our method can handle genomes in scaffold form, but they were excluded from our study because they were not present in the Ensembl multiple alignments we used to compute ancestral blocks.

Our reconstruction of ancestral amniote linkage groups suggests that this branch is slowly evolving in structure since the amniote-therians divergence. To refine this hypothesis, as well as to better understand the structure of amniote breakpoint regions (Larkin, 2010), we would need to compute genomic distances and evolutionary scenarios along the different lineage. The availability of a high-resolution ancestral amniote genome and of efficient methods for sampling rearrangement scenarios (Miklos and Tannier, 2010), adapted to handle fragmented and partially ordered ancestral genomes, should allow such studies.

Funding: Agence Nationale pour la Recherche (ANR-06-BLAN-0045 to A.O.); Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and by a France-Canada Travel Grant (to C.C.); Agence Nationale pour la recherche (ANR-08-GENM-036-01 and ANR-08-EMER-011-03 to E.T.).

Conflict of Interest: none declared.

REFERENCES

Alekseyev,M. and Pevzner,P. (2009) Breakpoint graphs and ancestral genome reconstruction. *Genome Res.*, **19**, 943–957.

Backström,N. *et al.* (2008) A gene-based genetic linkage map of the collared flycatcher (*Ficedula albicollis*) reveals extensive synteny and gene-order conservation during 100 million years of avian evolution. *Genetics*, **179**, 1479–1495.

Benton,M.J. and Donoghue,P.C.J. (2007) Paleontological evidence to date the tree of life. *Mol. Biol. Evol.*, **24**, 26–53.

Bergeron,A. *et al.* (2004) Reconstructing ancestral gene orders using conserved intervals. *Lect. Notes Comput. Sci.*, **3240**, 14–25.

Bertrand,D. *et al.* (2010) Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss. *Lect. Notes Comput. Sci.*, **6293**, 78–89.

Booth,K. and Lueker,G. (1976) Testing for the consecutive ones property, interval graphs and graph planarity using PQ-tree algorithms. *J. Comput. Syst. Sci.*, **13**, 335–379.

Bourque,G. *et al.* (2004) Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse and rat genomes. *Genome Res.*, **14**, 507–516.

Catchen,J. *et al.* (2008) Inferring ancestral gene order. In Keith,J. (ed.) *Bioinformatics, Volume I: Data, analysis, and Evolution*. Vol. 452. Humana Press, Springer, New York, NY, pp. 365–383.

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.

Chauve,C. *et al.* (2010) Yeast ancestral genome reconstructions: the possibilities of computational methods II. *J. Comput. Biol.*, **17**, 1097–1112.

Consortium,I.C.G.S. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.

de Peer,Y.V. (2004) Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.*, **5**, 752–763.

Dietrich,F. *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.

Durand,D. and Sankoff,D. (2003) Tests for gene clustering. *J. Comput. Biol.*, **10**, 453–482.

Ellegren,H. (2010) Evolutionay stasis: the stable chromosomes of birds. *Trends Ecol. Evol.*, **25**, 283–291.

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

Faraut,T. (2008) Addressing chromosome evolution in the whole-genome sequence era. *Chromosome Res.*, **16**, 5–16.

Ferguson-Smith,M. and Trifonov,V. (2007) Mammalian karyotype evolution. *Nat. Rev. Genet.*, **8**, 950–962.

Froenicke,L. (2005) Origins of primate chromosomes - as delineated by zoo-fish and alignments of human and mouse draft genome sequences. *Cytogenet. Genome Res.*, **108**, 122–138.

Froenicke,L. *et al.* (2003) Towards the delineation of the ancestral eutherian genome organization: comparative genome maps of human and the African elephant (*Loxodonta africana*) generated by chromosome painting. *Proc. R. Soc. London*, **270**, 1331–1340.

Gordon,J.L. *et al.* (2009) Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome. *PLoS Genet.*, **5**, e1000485.

Griffin,D. *et al.* (2007) The evolution of the avian genome as revealed by comparative molecular cytogenetics. *Cytogenet. Genome Res.*, **117**, 64–77.

Griffin,D. *et al.* (2008) Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution. *BMC Genomics*, **9**, 168.

Hansson,B. *et al.* (2010). Avian genome evolution: insights from a linkage map of the blue tit (*Cyanistes caeruleus*). *Heredity*, **104**, 67–78.

Hufton,A. and Panopoulou,G. (2009) Polyploidy and genome restructuring: a variety of outcomes. *Curr. Opin. Genet. Dev.*, **19**, 600–606.

Jaillon,O. *et al.* (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyoty pe. *Nature*, **431**, 946–957.

Kellis,M. *et al.* (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, **428**, 617–624.

Kemkemer,C. *et al.* (2009) Gene synteny comparison between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. *BMC Evol. Biol.*, **9**, 84.

Kohn,M. *et al.* (2006) Reconstruction of a 450My-old ancestral vertebrate protokaryotype. *Trends Genet.*, **22**, 203–210.

Kumar,S. *et al.* (2001) Determination of the number of conserved chromosomal segments between species. *Genetics*, **157**, 1387–1395.

Larkin,D. (2010) Role of chromosomal rearrangements and conserved chromosome regions in amniote evolution. *Mol. Genet. Microbiol. Virol.*, **25**, 1–7.

Luc,N. *et al.* (2003) Gene teams: a new formalization of gene clusters for comparative genomics. *Comput. Biol. Chem.*, **27**, 59–67.

Ma,J. *et al.* (2006) Reconstructing contiguous regions of an ancestral genome. *Genome Res.*, **16**, 1557–1565.

Mikkelsen,T.S. *et al.* (2007) Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature*, **447**, 167–177.

Miklos,I. and Tannier,E. (2010) Bayesian sampling of genome rearrangement scenarios via double cut and join. *Bioinformatics,* **26**, 3012–3019.

Muffato,M. and Roest-Crollius,H. (2008) Paleogenomics, or the recovery of lost genomes from the mist of times. *BioEssays*, **30**, 122–134.

Muffato,M. *et al.* (2010) Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, **26**, 1119–1121.

Murphy,W. *et al.* (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, **309**, 613–617.

Nakatani,Y. *et al.* (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.*, **17**, 1254–1265.

Naruse,K. *et al.* (2004) A medaka gene map: The trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res.*, **14**, 820–828.

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

Paten,B. *et al.* (2008) Enredo and pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, **18**, 1814–1828.

Pham,S.K. and Pevzner,P.A. (2010) DRIMM-Synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics*, **26**, 2509–2516.

Putnam,N.H. *et al.* (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, **317**, 86–94.

Putnam,N. *et al.* (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064–1071.

Rascol,V.L. *et al.* (2007) Ancestral animal genomes reconstruction. *Curr. Opin. Immunol.*, **19**, 542–546.

Richard,F. *et al.* (2003) Reconstruction of the ancestral karyotype of eutherian mammals. *Chromosome Res.*, **11**, 605–618.

Salse,J. *et al.* (2008) Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell*, **20**, 11–24.

Salse,J. *et al.* (2009a) Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Brief. Bioinformatics*, **10**, 619–630.

Salse,J. *et al.* (2009b) Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc. Natl Acad. Sci. USA*, **106**, 14908–14913.

Sémon,M. and Wolfe,K.H. (2007a) Consequences of genome duplication. *Curr. Opin. Genet. Dev.*, **17**, 505–512.

Sémon,M. and Wolfe,K.H. (2007b) Rearrangement rate following the whole-genome duplication in teleosts. *Mol. Biol. Evol.*, **24**, 860–867.

Sémon,M. and Wolfe,K.H. (2007c) Reciprocal gene loss between tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet.*, **23**, 108–112.

Skinner,B. *et al.* (2009) Comparative genomics in chicken and pekin duck using fish mapping and microarray analysis. *BMC Genomics*, **10**, 357.

Stapley,J. *et al.* (2008) A linkage map of the zebra finch *Taeniopygia guttata* provides new insights into avian genome evolution. *Genetics*, **179**, 651–657.

Vilella,A.J. *et al.* (2009) Ensembl compara genetrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.

Wienberg,J. (2004) The evolution of eutherian chromosomes. *Curr. Opin. Genet. Dev.*, **14**, 657–666.

Woods,I. *et al.* (2005) The zebrafish gene map defines ancestral vertebrates chromosomes. *Genome Res.*, **15**, 1307–1314.

Xu,W. *et al.* (2008) Poisson adjacency distributions in genome comparison: multichromosomal, circular, signed and unsigned cases. *Bioinformatics*, **24**, i146–i152.

Yang,F. *et al.* (2003) Reciprocal chromsome painting among human, aardvark, and elephant (superorder afrotheria) reveals the likely eutherian ancestral karyotype. *Proc. Natl Acad. Sci. USA*, **100**, 1062–1066.

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