Breakpoint graphs and ancestral genome reconstructions

Max A. Alekseyev and Pavel A. Pevzner[1]

Department of Computer Science and Engineering, University of California at San Diego, La Jolla, California 92093-0404, USA

Recently completed whole-genome sequencing projects marked the transition from gene-based phylogenetic studies to phylogenomics analysis of entire genomes. We developed an algorithm MGRA for reconstructing ancestral genomes and used it to study the rearrangement history of seven mammalian genomes: human, chimpanzee, macaque, mouse, rat, dog, and opossum. MGRA relies on the notion of the multiple breakpoint graphs to overcome some limitations of the existing approaches to ancestral genome reconstructions. MGRA also generates the rearrangement-based characters guiding the phylogenetic tree reconstruction when the phylogeny is unknown.

[Supplemental material is available online at www.genome.org.]

The first attempts to reconstruct the genomic architecture of ancestral mammals predated the era of genomic sequencing and were based on cytogenetics approaches (Wienberg and Stanyon 1997). The rearrangement-based phylogenomic studies were pioneered by Sankoff and colleagues (Sankoff et al. 1992; Blanchette et al. 1997; Sankoff and Blanchette 1998) and were based on analyzing the *breakpoint distances*. Moret et al. (2001) further optimized this approach and developed a popular GRAPPA software for rearrangement analysis. MGR, another genome rearrangement tool (Bourque and Pevzner 2002), uses *genomic distances* instead of breakpoint distances for ancestral reconstructions. Since genomic distances lead to more accurate ancestral reconstructions (Moret et al. 2002; Tang and Moret 2003), GRAPPA has been modified for genomic distances as well. While MGR has been used in several phylogenomic studies (Bourque et al. 2005; Murphy et al. 2005; Bulazel et al. 2007; Pontius et al. 2007; Xia et al. 2007; Cardone et al. 2008; Deuve et al. 2008), both MGR and GRAPPA have limited ability to distinguish reliable from unreliable rearrangements and to address the "weak associations" problem in ancestral reconstructions (Bourque et al. 2004, 2005, 2006; Froenicke et al. 2006).

Recently, Ma et al. (2006) made an important step toward reliable reconstruction of the ancestral genomes. In contrast to MGR and GRAPPA (which analyze both reliable and unreliable rearrangements), they have chosen to focus on the reliable breakpoint reconstruction in the ancestral genomes and to avoid assignments in the case of weak associations (complex breakpoints). This proved to be a valuable approach since, as it turned out, most breakpoints in the ancestral mammalian genomes can be reliably reconstructed. However, there are some limitations (discussed in Rocchi et al. 2006) that this approach has to overcome to scale for large sets of genomes. First, while the Ma et al. (2006) inferCARs algorithm assumes that the phylogeny is known, it remains a subject of enduring debates even in the case of the primate–rodent–carnivore split (which is assumed to be resolved in Ma et al. 2006). With the increase in the number of species, the reliability of the phylogeny will become even a bigger concern, thus raising the question of devising an approach that does not assume a fixed phylogeny but instead uses rearrangements as new characters for constructing phylogenetic trees (see Chaisson et al. 2006). While MGR does not assume a fixed phylogeny, its heuristically derived weak associations are less reliable. The challenge then is to integrate the reliability of inferCARs with the flexibility of MGR. Another avenue to improve inferCARs algorithms is to find out how to deal with complex breakpoints that create gaps in reconstructions.

Note that the Ma et al. (2006) approach focuses on the reliable ancestor reconstruction rather than on the specific rearrangements that happened in the course of evolution. These are related but different problems that both can benefit from incorporating into a single computational framework. Indeed, Ma et al. (2006) consider individual breakpoints and do not distinguish between particular types of rearrangements that generated a breakpoint of interest. In reality, the reversals and translocations operate on pairs of dependent breakpoints rather than individual breakpoints. Some rearrangements (and synteny associations) cannot be inferred from the analysis of single breakpoints but become tractable via analyzing the breakpoint graph.[2] As a result, while MGR constructs provably optimal scenarios in the absence of breakpoint reuse, it is not clear whether the same result holds for inferCARs.

Recently, Zhao and Bourque (2007, 2009) developed the EMRAE algorithm, which reconstructs both reliable rearrangements and ancestors, thus addressing the shortcomings of both MGR (difficulty in distinguishing between reliable and putative rearrangement events) and inferCARs (ancestor reconstruction only). However, EMRAE (in contrast to MGR) does not attempt to reconstruct the phylogenetic tree and is limited to unichromosomal genomes. Below we address some limitations of MGR, EMRAE, and inferCARs by developing the Multiple Genome Rearrangements and Ancestors (MGRA) algorithm (available from http://www.cs.ucsd.edu/users/ppevzner/software.html). In particular: (1) MGRA constructs provably optimal scenarios even when there is some breakpoint reuse and when other tools do not guarantee optimality. (2) MGRA is suitable for ancestral reconstructions of multi-chromosomal genomes (in contrast to EMRAE). (3) MGRA is conceptually simpler and orders of

[1]Corresponding author.
E-mail ppevzner@cs.ucsd.edu; fax (858) 534-7029.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.082784.108.

[2] The breakpoint graphs represent a popular technique for the rearrangement analysis since they reveal pairs of breakpoints representing footprints of the rearrangement events. See chapter 10 of Pevzner (2000) for background information on genome rearrangements and breakpoint graphs.

magnitude faster than MGR. (4) MGRA is not limited to reconstructing ancestral genomes in the case of known phylogeny (like inferCARs and EMRAE). Instead, it can guide the rearrangement-based reconstruction of phylogenetic trees. (5) MGRA does not require prior information about the approximate lengths of the branches of the phylogenetic trees (in contrast to inferCARs).

To evaluate the performance of MGRA, we compared ancestral reconstructions generated by MGRA and inferCARs. Despite the fact that MGRA and inferCARs are very different algorithms, their reconstructions turned out to be remarkably similar (98.5% of synteny associations are identical). We further analyzed some differences between MGRA, inferCARs, and the cytogenetics approach.

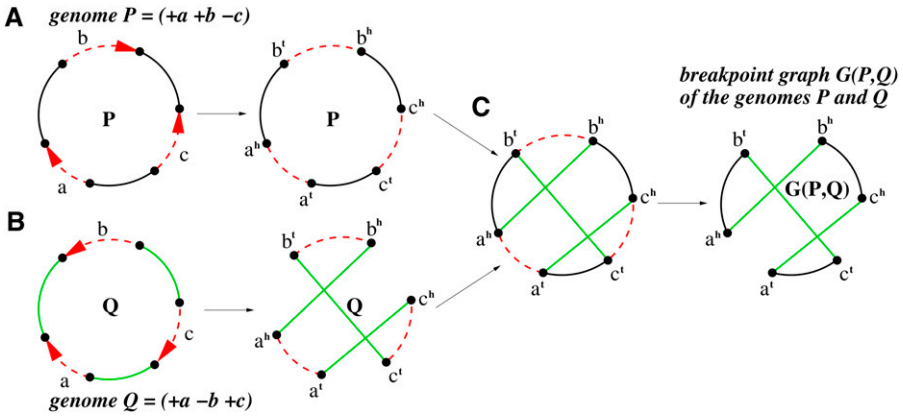

Figure 1. (*A*) Unichromosomal genome $P = (+a +b -c)$ represented as a black-obverse cycle. (*B*) Unichromosomal genome $Q = (+a -b +c)$ represented as a green-obverse cycle. (*C*) The breakpoint graph $G(P, Q)$ with and without obverse edges.

Methods

From pairwise to multiple breakpoint graphs

We start with analysis of rearrangements in circular genomes (i.e., genomes consisting of circular chromosomes) and later extend it to genomes with linear chromosomes. We assume that each genome is formed by the same set of synteny blocks, which are arranged differently in different genomes. We will find it convenient to represent a chromosome formed by synteny blocks $b_1, \ldots, b_n$ as a cycle with n directed labeled edges (corresponding to blocks) alternating with n undirected unlabeled edges (connecting adjacent blocks). The directions of the edges correspond to signs (strand) of the blocks. We label the tail and head of a directed edge b_i as b_i^t and b_i^h, respectively (Fig. 1) and represent a genome as a set of disjoint cycles (one for each chromosome). The edges in each cycle alternate between two colors: one color (e.g., "black") used for undirected edges and the other color (traditionally called "obverse") used for directed edges.

Let P be a genome represented as a collection of alternating black-obverse cycles (a cycle is alternating if the colors of its edges alternate). For any two black edges (x_1, x_2) and (y_1, y_2) in the genome (graph) P, we define a *2-break rearrangement* (first introduced as a *DCJ rearrangement* in Yancopoulos et al. 2005 and recently studied in Bergeron et al. 2006 and Lin and Moret 2008) as "replacement of these edges with either a pair of edges (x_1, y_1), (x_2, y_2), or a pair of edges (x_1, y_2), (x_2, y_1)" (Fig. 2A,B). In the case of circular genomes, 2-breaks correspond to the standard rearrangement operations of reversals, fissions, or fusions/translocations (Fig. 2).[3]

Let P and Q be genomes on the same set of blocks B. The (pairwise) breakpoint graph $G(P, Q)$ is simply the superposition of genomes (graphs) P and Q (Fig. 1C). Formally, the breakpoint graph $G(P, Q)$ is defined on the set of vertices $V = \{b^t, b^h \mid b \in B\}$ with edges of three colors: obverse (connecting vertices b^t and b^h), black (connecting adjacent blocks in P), and green (connecting adjacent blocks in Q). The black and green edges form the black–green alternating cycles that play an important role in analyzing rearrangements (Bafna and Pevzner 1996). From now on we will ignore the obverse edges in the breakpoint graph so that it becomes simply a collection of (black–green) cycles (Fig. 1).

The 2-break distance $d_2(P, Q)$ between genomes P and Q is defined as "the minimum number of 2-breaks required to transform one genome into the other." In contrast to the Genomic Distance Problem (Hannenhalli and Pevzner 1995; Tesler 2002a; Ozery-Flato and Shamir 2003) (for linear multi-chromosomal genomes), the 2-Break Distance Problem for circular multi-chromosomal genomes has a trivial solution (Yancopoulos et al. 2005; Alekseyev and Pevzner 2007): $d_2(P, Q) = b(P, Q) - c(P, Q)$, where $b(P, Q) = |B|$ is the number of synteny blocks in P and Q, and $c(P, Q)$ is the number of black–green cycles in $G(P, Q)$.

A linear genome is a collection of linear chromosomes represented as sequences of signed synteny blocks. Each linear chromosome on n blocks is represented as a path of n directed obverse edges (encoding blocks and their direction) alternating with $(n - 1)$ undirected black edges (connecting adjacent blocks). In addition, we introduce an extra vertex ∞ and connect it by an undirected (*irregular*) black edge with every vertex representing a chromosomal end (hence, the degree of vertex ∞ is twice the number of linear chromosomes). A "linear chromosome" is an alternating path of black and obverse edges, starting and ending at the vertex ∞, and a "linear genome" is a collection of such paths. The 2-breaks involving irregular edges model the rearrangements affecting the chromosome ends (Fig. 2C,D).

Analyzing reversals, translocations, fusions, and fissions in linear genomes poses additional algorithmic challenges as compared to analyzing 2-breaks in circular genomes. However, rearrangement scenarios in linear genomes are well approximated by 2-break scenarios in circular genomes (Alekseyev 2008). Hence, we use 2-breaks as a single substitute for reversals, translocations, fusions, and fissions, admitting that 2-breaks may violate linearity of the genomes by creating circular chromosomes.

While previous rearrangement studies (e.g., MGR) were limited to analyzing the pairwise breakpoint graphs, MGRA uses multiple breakpoint graphs (Caprara 1999b), which simplify the rearrangement analysis. Let $P_1, \ldots, P_k$ be genomes on the same set of synteny blocks B. Similarly to the pairwise breakpoint graph, the (multiple) breakpoint graph $G(P_1, \ldots, P_k)$ is simply the superposition of genomes (graphs) $P_1, \ldots, P_k$ on the same vertex set

[3] In this study, we use the term "reversal" (common in bioinformatics literature) instead of the term "inversion" (common in biology literature). For circular chromosomes, fusions and translocations are not distinguishable, that is, every fusion of circular chromosomes can be viewed as a translocation, and vice versa.

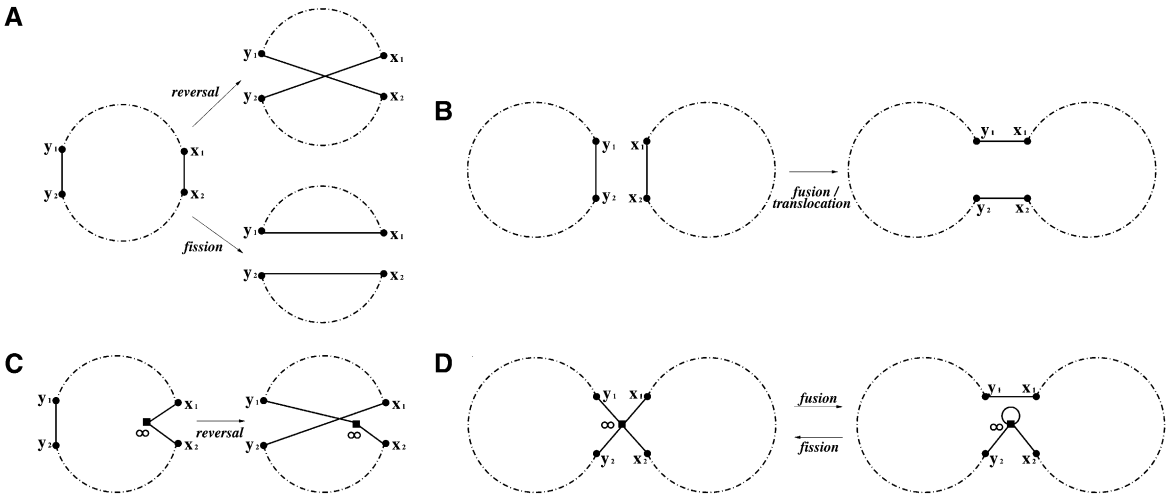

Figure 2. (*A*) A 2-break on edges (x_1, x_2) and (y_1, y_2) from the same chromosome corresponds to either a reversal, or a fission. (*B*) A 2-break on edges (x_1, x_2) and (y_1, y_2) from different chromosomes corresponds to a translocation/fusion. (*C*) A 2-break on edges (y_1, y_2) and (x_1, ∞) of a linear chromosome corresponds to a reversal affecting a chromosome end x_1 and creating a new chromosome end y_1. (*D*) A 2-break on edges (x_1, ∞) and (y_1, ∞) from different chromosomes models a fusion. Fissions can be modeled as 2-breaks operating on an irregular loop edge (∞, ∞) and an arbitrary regular edge in the genome.

$V = \{b^t, b^h \mid b \in B\} \cup \{\infty\}$ (Fig. 3A,B; Supplemental Fig. S20). Figure 4 shows the breakpoint graph on 1357 synteny blocks[4] of six mammalian genomes: *M* (mouse), *R* (rat), *D* (dog), *Q* (macaque), *H* (human), and *C* (chimpanzee).

A vertex in the breakpoint graph is *regular* if it is different from ∞. Similarly, an edge is *regular* if both its endpoints are regular, and *irregular* otherwise. The edges of $G(P_1, \ldots, P_k)$ are represented by undirected edges from the genomes $P_1, \ldots, P_k$ of k different colors (hence, the degree of each regular vertex is k). To simplify the notation, we will use $P_1, \ldots, P_k$ also to refer to the colors of edges in the multiple breakpoint graph, and denote the set of all colors $C = \{P_1, \ldots, P_k\}$. Furthermore, any non-empty subset of C is called a *multi-color.* All edges connecting vertices x and y in the (multiple) breakpoint graph form the *multi-edge* (x, y) of the multi-color represented by the colors of these edges [e.g., the multi-edge (e^h, f^h) in Fig. 3B has multi-color $\{P_3, P_4\}$ shown as red and yellow edges]. The number of multi-edges incident to a vertex (also equal to the number of adjacent vertices) is called the *multi-degree* (note that the multi-degree of a vertex may be smaller than its degree; e.g., the vertex e^h in Fig. 3B has degree 4 and multi-degree 3). Multi-edges correspond to adjacent synteny blocks that are conserved across multiple species and thus represent valuable phylogenetic characters (Sankoff and Blanchette 1998).

A breakpoint in the multiple breakpoint graph $G(P_1, P_2, \ldots, P_k)$ is a vertex of the multi-degree >1. A multiple breakpoint graph without breakpoints is an *identity breakpoint graph* $G(X, \ldots, X)$ *of some genome X*. Alternatively, the identity breakpoint graph can be characterized as a breakpoint graph consisting of *complete multi-edges* (i.e., multi-edges of the multi-color C) that correspond to the synteny blocks adjacencies in X.

Multiple genome rearrangement problem

The key observation in studies of pairwise genome rearrangements is that every 2-break transformation of a "black" genome P into a "green" genome Q corresponds to a transformation of the breakpoint graph $G(P, Q)$ into the identity breakpoint graph $G(Q, Q)$ (Supplemental Fig. S21) with 2-breaks on pairs of black edges (black 2-breaks). MGR (Bourque and Pevzner 2002) implicitly applies a similar observation and attempts to come up with rearrangements that bring the multiple breakpoint graph $G(P_1, P_2, \ldots, P_k)$ closer to the identity multiple breakpoint graph $G(P_i, P_i, \ldots, P_i)$ for i varying from 1 to k. However, this approach does not allow one to use the internal edges of the phylogenetic tree for finding reliable rearrangements. Below we formalize the Multiple Genome Rearrangement Problem in terms of multiple breakpoint graphs. The key element of MGRA is finding a shortest transformation of the multiple breakpoint graph $G(P_1, P_2, \ldots, P_k)$ into an arbitrary identity multiple breakpoint graph $G(X, X, \ldots, X)$ for some a priori unknown genome X. We first illustrate this concept with pairwise breakpoint graphs.

Let $G(P_1, P_2) \rightarrow G(X, X)$ be an m-step transformation of $G(P_1, P_2)$ into $G(X, X)$ by either black or green 2-breaks (in contrast to the standard breakpoint graph analysis based on black 2-breaks only).[5] It is easy to see that every such transformation corresponds to a transformation $P_1 \rightarrow X \rightarrow P_2$ that uses m black 2-breaks. Therefore, instead of searching for a shortest transformation $G(P_1, P_2) \rightarrow G(P_2, P_2)$, one can search for a shortest transformation of $G(P_1, P_2)$ into any identity breakpoint graph $G(X, X)$ without knowing X in advance.

In the case of $k \geq 2$ genomes $P_1, P_2, \ldots, P_k$, 2-breaks can be applied to multi-edges in the multiple breakpoint graph $G(P_1,$

[4] The detailed information about synteny blocks and assembly builds is provided in the Supplemental material. Out of 1360 synteny blocks (kindly provided by Jian Ma), three synteny blocks represent intermixed segments of the chromosome X and other chromosomes (the mouse chromosome 7 and the rat chromosomes 15 and 20). Since these blocks are short (16, 47, and 17 kb, respectively), we have discarded them to simplify the chromosome X analysis below. For better illustration of the breakpoint graphs, the vertex ∞ is shown in multiple copies as black dots, each connected by a single multi-edge to a regular vertex.

[5] Switching from black rearrangements to a mixture of black and green rearrangements is a simple but powerful paradigm that proved to be useful in previous studies (Bafna and Pevzner 1998; Tannier and Sagot 2004).

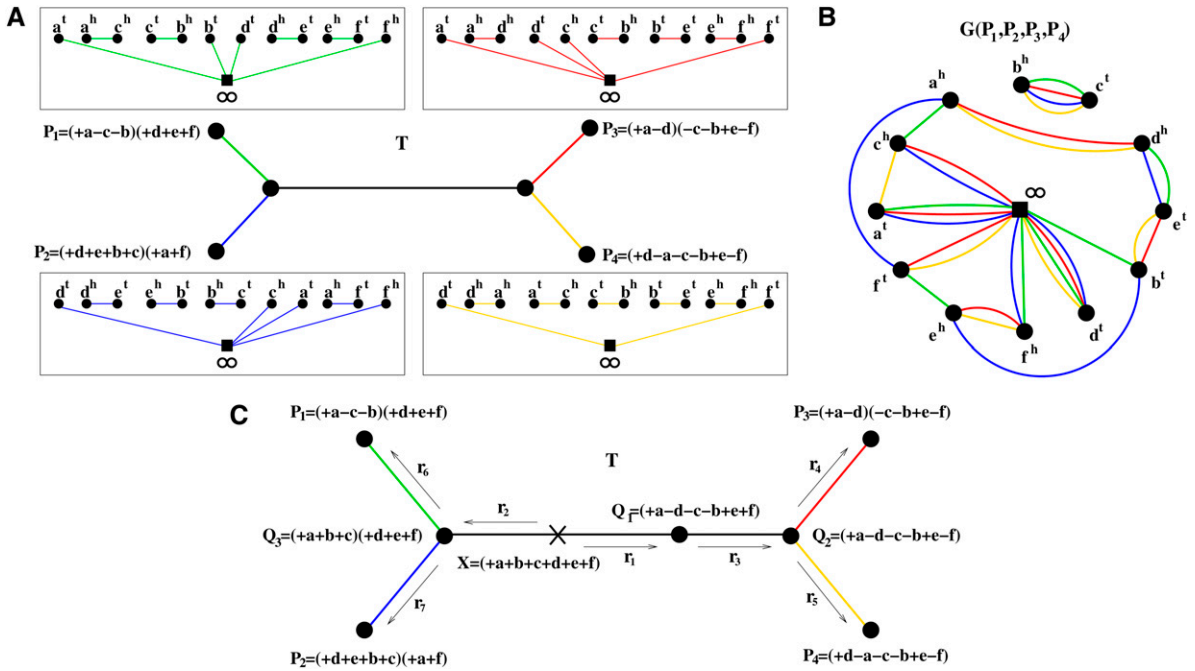

Figure 3. (*A*) A phylogenetic tree T with four linear genomes P_1, P_2, P_3, P_4 (represented as green, blue, red, and yellow graphs, respectively) at the leaves. The obverse edges are not shown. (*B*) The multiple breakpoint graph $G(P_1, P_2, P_3, P_4)$ is a superposition of graphs representing genomes P_1, P_2, P_3, P_4. The multi-degrees of regular vertices vary from 1 (e.g., vertex b^h) to 3 (e.g., vertex e^h). (*C*) The same phylogenetic tree T with all intermediate genomes specified and a genome X selected as a root. A T-consistent transformation of X into P_1, P_2, P_3, P_4 can viewed as a transformation of the quadruple (X, X, X, X) into the quadruple (P_1, P_2, P_3, P_4), where a rearrangement at each step is applied to some copies of the same genome in the quadruple. A particular such transformation takes the following steps: $(X, X, X, X) \xrightarrow{r_1} (X, X, Q_1, Q_1) \xrightarrow{r_2} (Q_3, Q_3, Q_1, Q_1) \xrightarrow{r_3} (Q_3, Q_3, Q_2, Q_2) \xrightarrow{r_4} (Q_3, Q_3, P_3, Q_2) \xrightarrow{r_5} (Q_3, Q_3, P_3, P_4) \xrightarrow{r_6} (P_1, Q_3, P_3, P_4) \xrightarrow{r_7} (P_1, P_2, P_3, P_4)$, where r_1 is a reversal in two copies of X; r_2 is a fission in two copies of X; r_3 is a reversal in both copies of Q_1; r_4 is a fission in one copy of Q_2; r_5 is a reversal in the other copy of Q_2; r_6 is a reversal in one copy of Q_3; and r_7 is a translocation in the other copy of Q_3.

$P_2, \ldots, P_k$) of as many as $(2^k - 2)$ different multi-colors formed by proper subsets of C. However, not every series of such 2-breaks makes sense in terms of ancestral genome reconstructions. A basic property of ancestral genome reconstructions is that 2-breaks on multi-edges of multi-color $Q \in C$ can be applied only when all genomes corresponding to colors in Q are merged into a single genome. We give an alternative definition of this property as follows: A transformation (series of 2-breaks) S of the multiple breakpoint graph $G(P_1, P_2, \ldots, P_k)$ is "strict" if for any 2-breaks operating on multi-edges of multi-colors $Q_1 \subseteq Q_2$, ρ_1 precedes ρ_2 in S. The Multiple Genome Rearrangement Problem (MGRP) is reformulated as follows:

Given genomes $P_1, \ldots, P_k$, find a shortest strict series of 2-breaks that transforms the breakpoint graph $G(P_1, \ldots, P_k)$ into an identity breakpoint graph.

Let T be a (unrooted) phylogenetic tree of the genomes $P_1, \ldots, P_k$ (Fig. 3A). The tree T consists of k leaf nodes (or simply leaves), $(k - 2)$ internal nodes, and $(2k - 3)$ branches connecting pairs of nodes, so that the degree of each leaf is 1, while the degree of each internal node is 3.

Removing a branch from T breaks it into two subtrees, each of which is induced by the set of its own leaves. A multi-color consisting of all colors (leaves) of either of these induced subtrees is called "T-consistent." Let G be the set of all T-consistent multi-colors. Note that if a multi-color Q is T-consistent, then its complement $\overline{Q} = C \setminus Q$ is also T-consistent. Therefore, there is a one-to-one correspondence between the pairs of complementary T-consistent multi-colors and the branches of T (Fig. 5).

When a phylogenetic tree is given, MGRA addresses a restricted version of MGRP where 2-breaks are applied only to multi-colors consistent with the phylogenetic tree.

The Tree-Consistent Multiple Genome Rearrangement Problem (TCMGRP) is as follows:

Given genomes $P_1, \ldots, P_k$ at the leaves of a phylogenetic tree T, find a shortest strict series of T-consistent 2-breaks, transforming the breakpoint graph $G(P_1, \ldots, P_k)$ into an identity breakpoint graph.

Note that MGRP and TCMGRP problems in the case of three unichromosomal genomes correspond to the median problem that is NP-complete (Caprara 1999a; Tannier et al. 2008). While existence of exact polynomial algorithms for solving MGRP and TCMGRP is unlikely, we describe a heuristic approach to "eliminating" breakpoints in $G(P_1, \ldots, P_k)$ that uses reliable rearrangements. In particular, MGRA optimally solves these problems in case of semi-independent rearrangement scenarios with some breakpoint reuses (see below).

We will find it convenient to fix a branch χ of the tree T and assume that this branch contains a root X (viewed as yet another node), the precise location of which is to be determined later. The choice of X defines directions "toward" X on all branches of the tree T (Fig. 5). We label every leaf node P_i of the directed tree T with the corresponding singleton multi-color $\{P_i\}$, and then recursively label each internal node with the union of the multi-colors of the starting nodes of all incoming branches (e.g., in Fig. 5, a common endpoint of branches coming from the leaf nodes M and R is labeled MR). The multi-colors forming node labels of the tree T are called

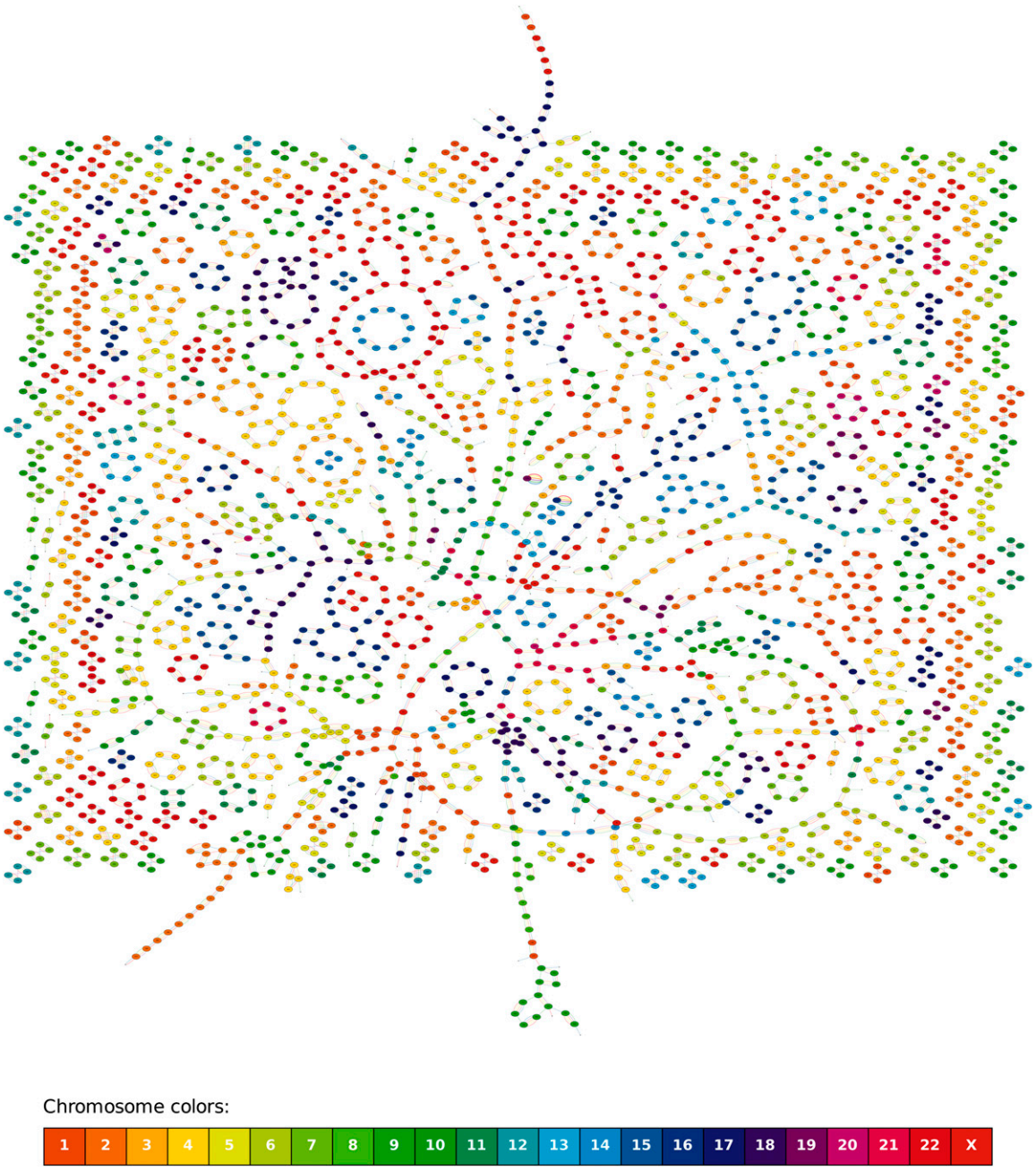

Figure 4. The breakpoint graph $G(M,R,D,Q,H,C)$ (obverse edges are not shown) of six mammalian genomes: mouse (red edges), rat (blue edges), dog (green edges), macaque (violet edges), human (orange edges), and chimpanzee (yellow edges). The graph has $1357 \times 2 = 2714$ vertices labeled as *n*t or *n*h (where *n* is a synteny block number) and colored in 23 colors representing chromosomes in the human genome.

"$\overrightarrow{T}$-consistent." Alternatively, $\overrightarrow{T}$-consistent multi-colors can be defined as T-consistent multi-colors whose induced subtrees do not contain χ. Note that exactly one of the multi-colors in each pair of complementary T-consistent multi-colors is $\overrightarrow{T}$-consistent and it labels the starting node of the corresponding directed branch in T (except for the multi-colors corresponding to the branch χ that both are $\overrightarrow{T}$-consistent).

MGRA transforms the genomes $P_1, \ldots, P_k$ into X along the directed branches of T, using 2-breaks on $\overrightarrow{T}$-consistent multi-colors ($\overrightarrow{T}$-consistent 2-breaks). In terms of breakpoint graphs, MGRA eliminates breakpoints in $G(P_1, P_2, \ldots, P_k)$ with $\overrightarrow{T}$-consistent 2-breaks and transforms it into the identity breakpoint graph $G(X, \ldots, X)$.[6] This transformation defines a *reverse transformation* of the genome X into the genomes $P_1, \ldots, P_k$ by $\overrightarrow{T}$-consistent 2-breaks (such as in Fig. 3C). MGRA keeps track of

[6] The use of $\overrightarrow{T}$-consistent 2-breaks here is motivated by an important property that every $\overrightarrow{T}$-consistent transformation can be turned into a strict $\overrightarrow{T}$-consistent transformation by changing the order of 2-breaks. Therefore, we do not directly address the strictness requirement in MGRA that first produces a $\overrightarrow{T}$-consistent transformation of the genomes $P_1, P_2, \ldots, P_k$ into the genome X and then reorders it into a strict transformation.

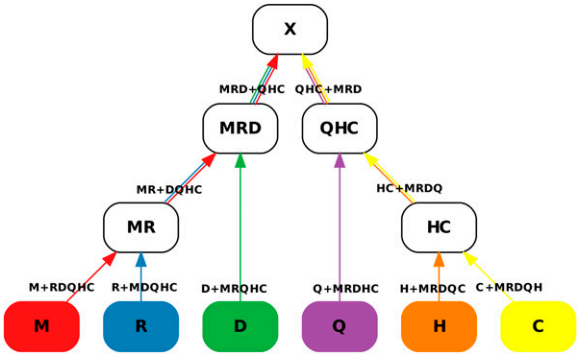

Figure 5. The phylogenetic tree T of six mammalian genomes: mouse (red), rat (blue), dog (green), macaque (violet), human (orange), and chimpanzee (yellow) with a root X on the $MRD + QHC$ branch. The branches are directed toward X and labeled with the corresponding pairs of complementary T-consistent multi-colors. The $\overrightarrow{T}$-consistent multi-color from each pair also labels the starting node of the corresponding directed branch. Note that the tree orientation may not necessarily correlate with the time scale, and the root genome X may not necessarily be a common ancestor of the leaf genomes.

rearrangements applied to the breakpoint graph $G(P_1, \ldots, P_k)$ during its transformation into an identity breakpoint graph $G(X, \ldots, X)$. The recorded rearrangements (in the reverse order) define a reverse transformation that passes through every internal node of the tree T and, thus, can be used to reconstruct the ancestral genomes at the internal nodes of T.

While initial steps in transformation of the breakpoint graph $G(P_1, \ldots, P_k)$ into an identity breakpoint graph usually correspond to reliable rearrangements, sooner or later one needs to use less reliable heuristic arguments in order to complete the transformation. However, sometimes it is preferable to stop after reaching a certain level of reliability even if the transformation is not complete (and the TCMGRP problem is not solved). In this case, we stop short of reconstructing the ancestral genomes since the transformation has not resulted in an identity breakpoint graph. In Supplement C, we describe an alternative method (not requiring solution of the TCMGRP problem) for reliable reconstruction of (parts of) ancestral genomes (similar to CARs from Ma et al. 2006) at internal nodes of the phylogenetic tree.

Results

MGRA algorithm

Supplement A introduces the notion of independent (no breakpoint reuses), semi-independent (breakpoint reuses may occur only within single branches of the phylogenetic tree), and weakly independent (breakpoint reuses are limited to adjacent branches of the phylogenetic tree) rearrangements. MGRA optimally solves the MGRP problem in case of semi-independent 2-breaks and uses heuristics to move beyond the semi-independent assumption. Below we show that most 2-breaks in mammalian evolution are either independent, semi-independent, or weakly independent, resulting in reliable ancestral reconstructions.

Cycles and paths in the breakpoint graph

Visual inspection of a rather complex breakpoint graph in Figure 4 (the giant component contains 630 vertices) reveals a large number of cycles and simple paths that are characteristic of independent and semi-independent rearrangement scenarios. MGRA uses the cycles/paths in the breakpoint graphs as a guide for finding reliable ancestor reconstructions.

We note that the immediate result of a 2-break performed along a branch $Q + \overline{Q}$ in the phylogenetic tree T is a cycle of four multi-edges whose multi-colors alternate between Q and $\overline{Q}$. All vertices in this cycle have multi-degree 2 and represent breakpoints that were not "reused." Even if one of these multi-edges is used in later rearrangements, the remaining three multi-edges still form an alternating path that serves as a footprint of the 2-break. This observation motivates a search for alternating paths and cycles in the breakpoint graphs. We introduce the following definitions to analyze such cycles/paths.

We define a *simple vertex* as a regular vertex of multi-degree 2 and a *simple multi-edge* as a multi-edge connecting two simple vertices. Simple multi-edges form *simple cycles/paths* in the breakpoint graphs, that is, cycles/paths in which multi-colors of consecutive multi-edges alternate between Q and $\overline{Q}$. Simple multi-edges/paths/cycles are called "good" if their multi-colors are T-consistent.

Table 1A describes the statistics of the breakpoint graph and illustrates how rearrangement analysis contributes to construction of phylogenetic trees. Indeed, all three internal branches (correct tree partitions) are supported by large numbers of good paths/cycles and good multi-edges (86 and 305 for $MR + DQHC$, 37 and 111 for $MRD + QHC$, 30 and 87 for $HC + MRDQ$). Each of 32 incorrect partitions (only eight of them are shown in Table 1A) have at most one simple path/cycle and at most six simple multi-edges, an order of magnitude smaller number than non-trivial correct partitions. This observation illustrates that reconstruction of the correct tree topology is a simple exercise in this case (see Chaisson et al. 2006). This and other statistics produced by MGRA (see below) may be used to determine the phylogenetic tree rather than to assume that it is given. In contrast to Cannarozzi et al. (2007), MGRA provides a large number of certificates supporting the tree topology in Figure 5. Below we show how MGRA reconstructs the ancestral genomes.

MGRA Stage I: Processing good cycles and paths

Alternating cycles represent well-studied objects in the case of the pairwise breakpoint graphs. Every such cycle of length $2m$ is formed by $(m - 1)$ 2-breaks (Alekseyev and Pevzner 2008) in each most parsimonious scenario.[7] Therefore, there is little difference between alternating cycles in the pairwise breakpoint graphs and good cycles in the multiple breakpoint graphs: indeed, the good cycles with alternating multi-colors Q and $\overline{Q}$ in the breakpoint graph model the rearrangements separating the sets of the genomes Q and $\overline{Q}$ exactly in the same way as in the pairwise genome comparison. We therefore argue that such alternating cycles (and the corresponding rearrangements) can be reliably assigned to the branch $Q + \overline{Q}$ in the phylogenetic tree T. This operation generalizes the notion of "good rearrangements" in MGR by extending them from cycles alternating multi-colors P_i and $\overline{P}_i = \{P_1, \ldots, P_{i-1}, P_{i+1}, \ldots, P_k\}$ to cycles alternating any complementary T-consistent multi-colors. While MGR attempts to find rearrangements bringing P_i closer to all genomes from $\overline{P}_i$ (i.e., rearrangements on the leaf branches of the phylogenetic tree), MGRA processes reliable rearrangements on all (both leaf and internal) branches of the phylogenetic tree (c.f. Zhao and Bourque 2007).

[7] While this representation is not unique, all these representations are equivalent (i.e., they produce the same final result). Figure 6B illustrates transformation of a simple cycle on six vertices into three complete multi-edges with two 2-breaks.

Table 1A. The statistics of the breakpoint graph of the mouse, rat, dog, macaque, human, and chimpanzee genomes

Multi-colors	Multi-edges	Simple vertices	Simple multi-edges	Simple paths
R + MDQHC	1173	1080	1036	235
MR + DQHC	487	376	305	86
D + MRQHC	473	368	310	105
M + RDQHC	223	145	118	45
MRD + QHC	208	135	111	37
Q + MRDHC	162	130	120	33
HC + MRDQ	140	104	87	30
C + MRDQH	45	32	26	11
H + MRDQC	15	8	6	3
QC + MRDH	9	6	6	1
MRQ + DHC	8	1	0	0
MD + RQHC	8	1	0	0
QH + MRDC	7	2	1	1
RQ + MDHC	7	4	4	1
DC + MRQH	6	4	4	1
DQ + MRHC	5	0	0	0
∅ + MRDQHC	2	0	0	0
MRC + DQH	1	0	0	0

Table 1B. The statistics of the breakpoint graph of the mouse, dog, macaque, and opossum genomes after MGRA Stages 1 and 2 on confident branches

Multi-colors	Multi-edges	Simple vertices	Simple multi-edges	Simple paths
∅ + MDQO	1693	0	0	0
O + MDQ	45	4	0	0
M + DQO	42	4	0	0
Q + MDO	35	0	0	0
D + MQO	26	0	0	0
MO + DQ	26	5	2	1
MD + QO	19	7	4	1
MQ + DO	12	3	0	0

For every pair of complementary multi-colors, we show the number of multi-edges of these multi-colors, the number of simple vertices that are incident to such multi-edges, the number of simple multi-edges, and the number of simple paths and cycles. The *T*-consistent multi-colors are shown in bold. (A) Only 18 out of 32 possible multi-colors are shown (the remaining 14 multi-colors have zero corresponding multi-edges). (B) See Supplemental Figure S18, bottom.

Similarly, good paths can be also assigned to branches of the phylogenetic tree by transforming them into good cycles first. Consider a good path $x_1, x_2, \ldots, x_m$ consisting of $(m - 1)$ multi-edges with *T*-consistent multi-colors alternating between a multi-color *Q* of the multi-edge (x_1, x_2) and its complement $\overline{Q}$. We extend this path by vertices x_0 and x_{m+1} incident to its first and last vertices, respectively, resulting in the path $p = (x_0, x_1, x_2, \ldots, x_{m+1})$. If the first and the last multi-edges in this path have the same $\overrightarrow{T}$-consistent multi-color, we perform a 2-break over the multi-edges (x_0, x_1) and (x_m, x_{m+1}) to transform *p* into a good cycle $c = (x_1, x_2, \ldots, x_m)$ and a multi-edge (x_0, x_{m+1}) (Fig. 6A,C).[8] If the first or/and last multi-edges of *p* are of non-$\overrightarrow{T}$-consistent multi-color, we remove it/them to obtain a path flanked by a $\overrightarrow{T}$-consistent multi-color that is processed (if it is longer than one edge) as above. Note that processing good cycles/paths in the breakpoint graph can create new good cycles/paths. We therefore process the good cycles/paths in an iterative fashion until no more good cycles/paths remain.[9]

Figure 7 (top panel) shows the breakpoint graph after processing (i.e., removing) good cycles/paths and illustrates that it is significantly simplified as compared to Figure 4. The size of the giant component is reduced from 630 to 193 vertices, and the overall number of vertices (not counting vertices incident to complete multi-edges) is reduced 10-fold from 2712 to 253. Figure 7 (top panel) illustrates how MGRA improves upon MGR: MGR is able to reduce the same graph only threefold to 924 vertices (414 vertices in the giant component) before it runs out of good rearrangements. While MGRA Stage 1 greatly reduces the rearrangement distance between the analyzed genomes, Supplemental Table S5 (center panel) illustrates that it still does not reveal the ancestral genomes *MR*, *MRD*, *HC*, and *QHC*. Moreover, it is not clear how to derive these ancestors based on a rather complex topology of the breakpoint graph in Figure 7 (top panel). MGRA Stage 2 introduces the notion of *fair cycles/paths* that allows one to reveal the rearrangements that violate the semi-independence assumption and to further simplify the graph in Figure 7 (top panel).

The results of MGRA Stage 1 already reveal valuable insights about the ancestral genomes (even without MGRA Stage 2). To simplify the analysis of the Boreoeutherian ancestral reconstruction[10] by MGRA Stage 1, we restrict the set of genomes to single representatives of rodents (mouse), carnivores (dog), and primates (macaque). The resulting breakpoint graph (with obverse edges shown) reveals many long unicolored paths formed by alternating obverse edges and complete multi-edges (Supplemental Fig. S10). Such paths represent parts of different human chromosomes in the reconstructed ancestor genome. We compress every such path into a single rectangular vertex as shown in Supplemental Figure S11 (top panel), resulting in a rather small graph. We further show the chromosomal associations present in this graph in Supplemental Figure S12. We emphasize that MGRA Stage 1 reveals some subtle but reliable adjacencies that other ancestral reconstruction algorithms may miss. In particular, it reveals two adjacencies that are absent in any of the extant genomes and many adjacencies that are present in only one of the extant genomes.

The compressed breakpoint graph reveals only 5 complete multi-edges connecting vertices of different colors: 12 + 22, 12 + 22, 3 + 21, 4 + 8, and 14 + 15. These are exactly the same five adjacencies $12a + 22a$, $12b + 22b$, 3 + 21, $4a + 8p$, 14 + 15 revealed in Ma et al. (2006). It also reveals the CARs corresponding to the human chromosomes 2, 2, 5, 6, 7, 8, 9, 10, 10, 11, 17, 18, and X (represented as isolated boxes in Supplemental Fig. S12), exactly the same as the ancestral chromosomes revealed by previous cytogenetics analysis (Froenicke et al. 2006) (2q, 2pq, 5, 6, 7a, 8q, 9, 10q, 11, 17, 18, X) with a single exception: The second segment from chromosome 10 is identified as an isolated chromosome by us and is tentatively assigned as $10p + 12a + 22a$ by

[8] In the special case $x_0 = x_{m+1} = \infty$, and the flanking edges are of the same $\overrightarrow{T}$-consistent multi-color; we perform a fusion 2-break as shown in Figure 6D. In the case of $m = 1$ (i.e., when *p* contains a single simple multi-edge), *c* represents a complete multi-edge rather than a cycle (Fig. 6C) and does not require further processing.

[9] One can prove that the topology of the resulting graph does not depend on the order in which good cycles/paths are processed.

[10] We use the MRD node of the phylogenetic tree in Figure 5 to approximate the Boreoeutherian ancestor. While this study focuses on the Boreoeutherian ancestor, MGRA reconstructs ancestral genomes for every node of the phylogenetic tree. We emphasize that while reconstruction starts with selection of the root branch (as in Fig. 5), the choice of this branch and the exact location of the root *X* on this branch are rather arbitrary and not correlated with a specific ancestral genome of interest (in contrast to the alternative "root-driven" approach described in Supplement C). As described in the Reconstructing Ancestral Genomes section, the ancestral genomes are defined by the reverse transformation from the (whatever) root genome *X* to the leaf genomes. Ideally, different choices of the root branch and locations of the root *X* itself will result in the same set of ancestral genomes.

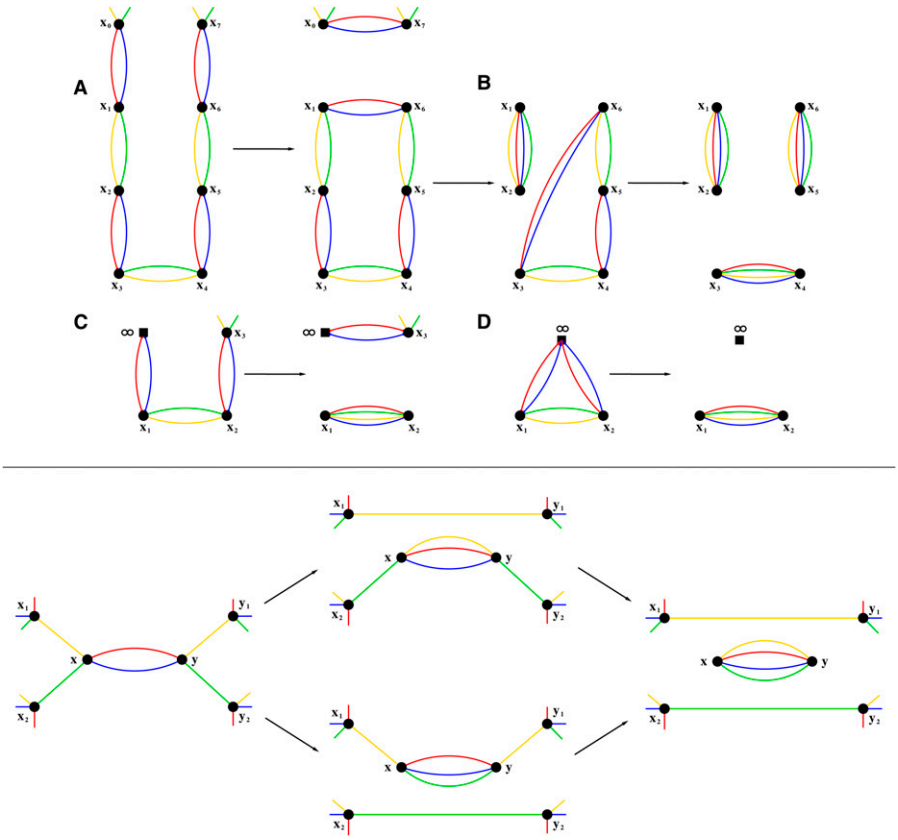

Figure 6. (*Top* panel) Processing good paths using a $\vec{T}$-consistent red–blue multi-color. (*A*) A good path on vertices $x_1, x_2, \ldots, x_6$ is transformed into a cycle on the same vertices by extending it into $x_0, x_1, x_2, \ldots, x_6, x_7$ and performing a 2-break on the multi-edges (x_0, x_1) and (x_6, x_7). (*B*) Transformation of a good cycle on 6 vertices into complete multi-edges with a 2-break on the multi-edges (x_1, x_2), (x_3, x_4) followed by a 2-break on the multi-edges (x_1, x_4), (x_5, x_6). (*C*) A 2-break on an irregular edge corresponds to a reversal involving chromosome ends. (*D*) A 2-break on two irregular edges corresponds to a fusion. (*Bottom* panel) Two ways of transforming a fair edge (x, y) into a good edge: (*top*) by a 2-break on yellow edges or (*bottom*) by a 2-break on green edges. In either case, the follow-up processing of the generated simple path results in the same graph with the complete multi-edge (x, y).

Froenicke et al. (2006). However, Froenicke et al. (2006) acknowledged that the association of 10p and 12a is only weakly supported (indicated by a question mark in Froenicke et al. 2006).[11] Our analysis also rules out the associations 1 + 22, 5 + 19, 2 + 18, 1 + 10, and 20 + 2 suggested in Murphy et al. (2005) as weak associations and later criticized by Froenicke et al. (2006) as unreliable. Supplement D further focuses on the connected component of the breakpoint graph representing the human chromosomes 7, 16, and 19, where the cytogenetics approach disagrees with Ma et al. (2006).

MGRA Stage 2: Processing fair cycles and paths

Figure 7 (top panel) reveals many pairs of vertices of multi-degree 3 connected by a multi-edge. Each such multi-edge (x, y) corresponds to six vertices x, x_1, x_2, y, y_1, y_2 and five multi-edges (x, y), (x, x_1), (x, x_2), (y, y_1), (y, y_2) (including cases with $x_i = \infty$, $y_i = \infty$, or $x_i = y_j$ for some $1 \le i, j \le 2$). A multi-edge (x, y) is called *composite* if edges (x, x_1) and (y, y_1) have the same multi-color Q_1 and edges (x, x_2) and (y, y_2) have the same multi-color Q_2. A composite multi-edge is called *fair* if Q_1 and Q_2 represent T-consistent multi-colors (Fig. 6, bottom panel). Table 2 shows the statistics of composite multi-edges (depending on pairs of complementary multi-colors $Q_1 + \overline{Q}_1$ and $Q_2 + \overline{Q}_2$) and reveals that (1) most composite multi-edges are fair and (2) while some types of composite multi-edges are common [e.g., (*M*+, *R*+), (*M*+, *MR*+), (*R*+, *MR*+), (*MR*+, *D*+), (*D*+, *QHC*+), (*MR*+, *QHC*+)], others [e.g., (*Q*+, *R*+)] are either rare or absent. Table 2 illustrates the extremely biased statistics of composite multi-edges: The branches $Q_1 + \overline{Q}_1$ and $Q_2 + \overline{Q}_2$ corresponding to the multi-colors Q_1 and Q_2 of a composite multi-edge are likely adjacent in the phylogenetic tree (compare to the weakly independent rearrangements). Table 2 provides yet another illustration of utility of MGRA for deriving phylogenetic trees. Indeed, it reveals valuable information about the topology of the phylogenetic tree (incident edges) that can be combined with information (valid partitions) in Table 1A,B to infer the trees.

Every fair multi-edge (x, y) can be transformed into a good multi-edge by a 2-break (fair 2-break) either on multi-edges (x, x_1) and (y, y_1) (of multi-color Q_1) or on multi-edges (x, x_2) and (y, y_2) (of multi-color Q_2) (Fig. 6, bottom panel). In the former case, (x, y) is transformed into a good multi-edge of color $\bar{Q}_2$, while in the latter case, it is transformed into a good multi-edge of color $\bar{Q}_1$. The resulting good paths (formed by fair 2-breaks) can be further processed as described in MGRA Stage 1. An important observation is that the final result of processing a fair multi-edge does not depend on whether we start with a 2-break on Q_1 or Q_2 multi-color (see Fig. 6, bottom panel). A cycle/path in the breakpoint graph is called fair if (1) all its edges are either good or fair, and (2) it can be transformed into a good cycle/path by some fair 2-breaks.

MGRA Stage 2 detects fair paths/cycles, transforms them into good paths/cycles by fair 2-breaks, and further processes the resulting good paths/cycles as in MGRA Stage 1. In some cases, fair paths in Stage 2 should be chosen with caution since the choice of fair paths may influence ancestral reconstructions in some nodes (see Supplement E). Figure 7 (bottom panel) shows the breakpoint graph after processing fair cycles/paths and illustrates that it becomes so small that it now can be analyzed in a step-by-case fashion by brute-force analysis of every connected component.

Reconstructing ancestral genomes

After removing vertex ∞, the breakpoint graph (after MGRA Stage 2) consists of only nine connected components (Fig. 7, bottom panel). Five out of nine components contain vertices corresponding to both start and end of the same synteny blocks 80, 610, 795, 1290, and 1300. This is surprising since generally

[11] We are not claiming that this association does not exist since it may be present in some of 100+ genomes with available cytogenetics data. However, there is no support for this association in the six mammalian genomes. We remark that Ma et al. (2006) also did not find support for this association.

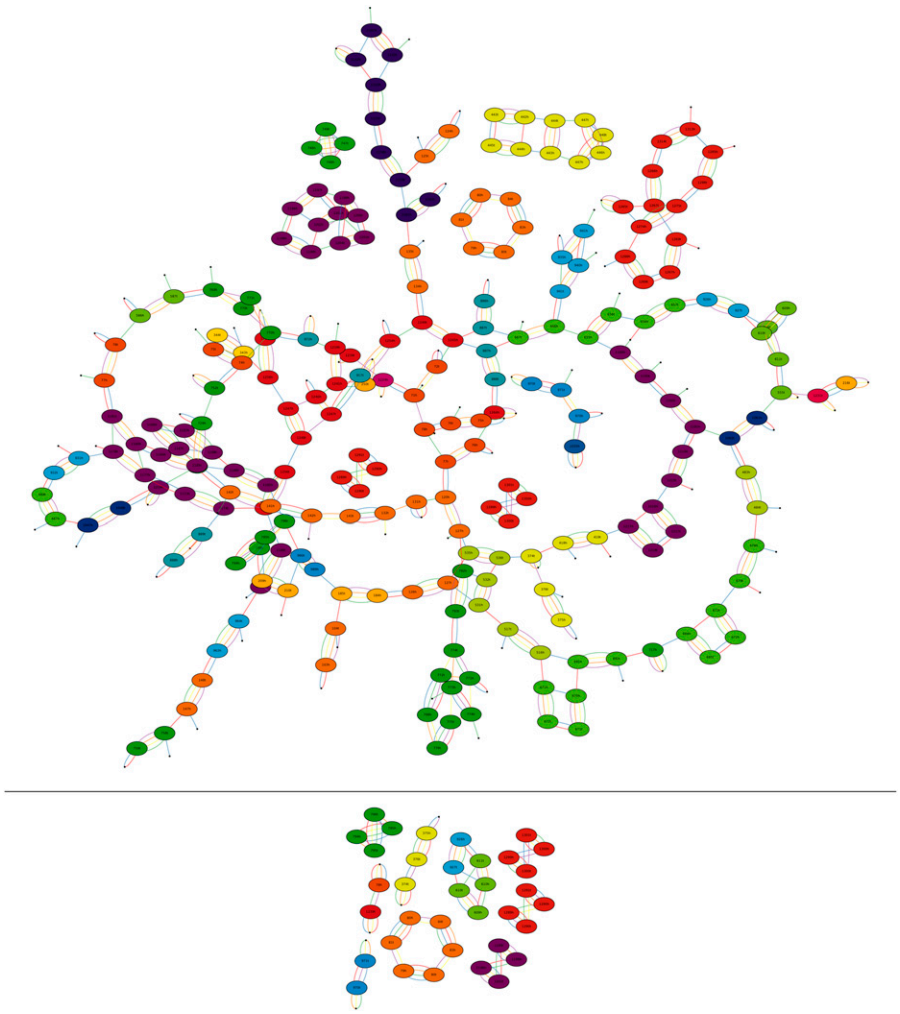

Figure 7. The breakpoint graph $G(M,R,D,Q,H,C)$ (the complete multi-edges are not shown) after MGRA Stage 1 (*top* panel) and after MGRA Stages 1–2 (*bottom* panel). The edge colors represent mouse (red), rat (blue), dog (green), macaque (violet), human (orange), and chimpanzee (yellow) genomes. Vertices are labeled and colored similarly to Figure 4.

the start and end of a synteny block are not expected to be present in the same (small) connected component unless this block was subject to a "micro-inversion" (Chaisson et al. 2006). Indeed, blocks 80, 610, 795, 1290, and 1300 turned out to be short (all under 500 kb), with blocks 610, 1290, and 1300 even shorter than 100 kb (block 1300 is 91 kb in human, 41 kb in mouse, and only 10 kb in the dog genome) that is near the threshold of 50 kb used in Ma et al. (2006) for generating reliable synteny blocks.

The simplest way to deal with such short blocks is to simply remove them from the set of input synteny blocks (Supplement J). Such removal will not significantly affect the architecture of the ancestral genomes (indeed, these blocks are well below the resolution of the cytogenetics approaches) while at the same time resolving five out of nine remaining components in the graph. Supplement F describes a different approach that attempts to find the positions and orientations of such short synteny blocks in the ancestors by processing complex breakpoints (MGRA Stage 3). We remark that processing at MGRA Stage 3 is viewed as less reliable, and the resulting associations are not considered in the proposed ancestral reconstructions (see below).

Recall that a strict T-consistent rearrangement scenario uniquely defines ancestral genomes at all internal nodes of the phylogenetic tree T. However, because of the use of 2-breaks instead of reversals/translocations/fissions/fusions, the ancestral genomes initially obtained by MGRA may contain (a small number of) circular chromosomes. Whenever possible, MGRA linearizes them by rearranging 2-breaks in the transformation. While circular chromosomes may occasionally appear in the initial rearrangement scenario obtained by MGRA, their appearance is a result of either 2-breaks applied in the "wrong" order (that can be avoided by reordering the 2-breaks [see Pevzner 2000], or a "shortcut" in processing hurdles that can be remedied by introducing additional 2-breaks [Hannenhalli and Pevzner 1999]). MGRA eliminates possible circular chromosomes in the reconstructed genomes at the post-processing stage. We emphasize that the outcome of MGRA is the set of ancestral (linear) genomes, while the 2-break rearrangement scenario produced by MGRA is considered only as a starting point for constructing the reversals/translocations/fusions/fissions scenario. An optimal linear rearrangement scenario can be found by applying GRIMM to the ancestral genomes reconstructed by MGRA.

Supplemental Figure S14 illustrates the results of ancestral genome reconstruction for chromosome X for six mammalian genomes. Supplement H shows the pairwise rearrangement distances between the ancestral and leaf genomes, following the strict T-consistent transformation constructed by MGRA and compares them to the genomic distances computed by GRIMM (Tesler 2002b). The differences between these distances are rather small, suggesting that the $\overrightarrow{T}$-consistent transformation found by MGRA is close to the most parsimonious.

Benchmarking MGRA

Benchmarking of the ancestral genome reconstruction algorithms may be challenging since the architecture of ancestral genomes is not known. While MGR, GRAPPA, inferCARs, and MGRA showed excellent performance on simulated data sets, these benchmarks were mainly designed for rearrangements generated according to the Random Breakage Model (RBM). Since MGRA improves on MGR and is guaranteed to produce optimal solutions for semi-independent scenarios, it is bound to provide even better results than MGR on such benchmarks. Supplement L compares MGRA and inferCARs on simulated data and illustrates that MGRA generates more accurate ancestral reconstructions for all choices of parameters. However, analyzing all these tools on simulated data may generate over-optimistic results since RBM does not reflect the realities of mammalian evolution (Bailey et al. 2004; van der Wind et al. 2004; Zhao et al. 2004; Murphy et al. 2005; Webber and Ponting 2005; Hinsch and Hannenhalli 2006; Ruiz-Herrera et al. 2006; Yue and Haaf 2006; Caceres et al. 2007; Gordon et al. 2007;

Table 2. The statistics of composite multi-edges (nonzero counts only) in the breakpoint graph *G*(*M, R, D, Q, H, C*) after MGRA Stage 1

	M+	R+	MR+	D+	MRD+	Q+	HC+	H+	C+	*DQ*+	*QH*+	*RD*+
M+	—	19	11	3								
R+	19	—	21	8	2	3	4					2
MR+	11	21	—	19	7		2		1			
D+	3	8	19	—	11		2		1			2
MRD+		2	7	11	—		4	1	2		2	
Q+		3				—	4		1			
HC+		4	2	2	4	4	—		1			
H+					1			—				
C+				1	2	1	1		—			
DQ+			1							—		
QH+					2						—	
RD+		2		2								—

Each pair of complementary multi-colors is denoted by one of its representative multi-colors (e.g., *M*+ represents the complementary multi-colors *M* + *RDQHC*). The bold row/column labels correspond to *T*-consistent (pairs of) multi-colors. The shaded entries correspond to pairs of adjacent branches in the phylogenetic tree *T* and account for 87% of all composite multi-edges.

Kikuta et al. 2007; Mehan et al. 2007). We therefore decided to analyze the differences between MGRA and inferCARs reconstructions and to further track evidence for each such difference in a case-by-case fashion.

MGRA and inferCARs produce highly consistent ancestral reconstructions. For illustration purposes, we have chosen to focus on the reconstruction of the *MRD* ancestral genome (Fig. 5), remarking that the results for the other ancestor genomes are similar. As an input to inferCARs, we provided six mammalian genomes and the same phylogenetic tree as used in Ma et al. (2006). The *MRD* genomes reconstructed by MGRA and inferCARs consist of 25 and 30 chromosomes (CARs), respectively.[12] However, MGRA does not consider associations obtained at Stage 3 (Fig. 7, bottom panel) as reliable. Most of these associations correspond to micro-inversions and thus do not significantly affect the ancestral reconstructions.

Comparison of two inferCARs reconstructions and using MGRA to improve inferCARs ancestral reconstructions

We start by comparing inferCARs with itself on two inputs: the original six mammalian genomes *M*, *R*, *D*, *Q*, *H*, *C* and the genomes *M*′, *R*′, *D*′, *Q*′, *H*′, *C*′ produced by MGRA Stage 1 (Fig. 7, top panel). We denote the reconstructed *MRD* genomes as MRD_{CARs} and MRD'_{CARs}, respectively.

Since MGRA Stage 1 processes only good cycles/paths that are unambiguously present in every optimal rearrangement scenario, one can safely assume that any optimal ancestral reconstruction should include the rearrangements performed at Stage 1. Therefore, running inferCARs on *M*, *R*, *D*, *Q*, *H*, *C* genomes should ideally produce the same results as running inferCARs on the "equivalent" *M*′, *R*′, *D*′, *Q*′, *H*′, *C*′ genomes. However, since inferCARs makes some greedy decisions and does not claim optimality, it does not guarantee to produce the same results on *M*, *R*, *D*, *Q*, *H*, *C* as compared with *M*′, *R*′, *D*′, *Q*′, *H*′, *C*′. Any such inconsistency would point to either somewhat less reliable CARs reconstructed by inferCARs or to reliable adjacencies missed by inferCARs. Therefore, inferCARs reconstructions can be potentially improved if MGRA Stage 1 runs before inferCARs as a preprocessing step.

Comparison of the reconstructed genomes MRD_{CARs} and MRD'_{CARs} indicates that while they share the overwhelming majority (99.0%) of reconstructed adjacencies, there are 13 adjacencies present in MRD_{CARs} but absent in MRD'_{CARs} and 13 adjacencies absent in MRD_{CARs} but present in MRD_{CARs} (out of the 1325 reconstructed adjacencies). Figure 8 (top) displays the breakpoint graph between the corresponding MRD_{CARs} and MRD_{CARs} reconstructions and reveals that the MRD'_{CARs} reconstruction is arguably more reliable than the MRD_{CARs} reconstruction. Indeed, Figure 8 (top) reveals that while most of the adjacencies (12 out of 13) present in MRD_{CARs} but not in MRD'_{CARs} correspond to "ambiguous joins" (in terms of Ma et al. 2006), MRD'_{CARs} contains four reliable adjacencies (i.e., resolved by MGRA Stage 1) that are nevertheless absent in MRD_{CARs}.

To resolve the conflicts between inferCARs results on equivalent inputs, we analyze each of these adjacencies [(658*h*, 652*h*), (871*t*, 873*t*), (770*t*, 771*t*), and (1014*t*, 1017*h*)] in a case-by-case fashion. For example, in the case of the (658*h*, 652*h*) adjacency, inferCARs failed to connect them, since the vertices 658*h* and 652*h* represent breakpoints of multi-degree 3 (Fig. 8, bottom panels), and it is not immediately clear how to process them using "local" rules employed by inferCARs. inferCARs turns 658*h* into a CAR end in *MRD*, although it is not a chromosome end in any of the six genomes. The breakpoint graph provides a clear view of connection between 658*h* and 652*h* by revealing good paths connecting them (see Fig. 8, bottom panels).

Comparison of inferCARs and MGRA reconstructions

Supplemental Figure S19 displays the breakpoint graph of the three *MRD* reconstructions: MRD_{MGRA}, MRD_{CARs}, and MRD'_{CARs}, and illustrates that the number of differences between MRD_{CARs} and MRD'_{CARs} (we consider the latter reconstruction to be more reliable) is comparable to the number of differences between MRD_{MGRA} and MRD'_{CARs}. Indeed, MRD'_{CARs} differs from MRD_{CARs} by 30 adjacencies and differs from MRD_{MGRA} by 39 adjacencies. Since the large-scale architecture of MRD_{CARs} was shown to be largely consistent with previous cytogenetics reconstructions (Ma et al. 2006) and since MRD'_{CARs} (that is arguably even more reliable than MRD_{CARs}) and MRD_{MGRA} share at least 98.5% of all adjacencies, all these reconstructions can be viewed as largely consistent with the cytogenetics-based reconstructions. Remarkably, most differences between MGRA and inferCARs reconstructions are represented by ambiguous joins that MGRA labels as less reliable anyway (shown as dashed edges). In particular, inferCARs reports eight less reliable adjacencies as unambiguous (complete multi-edges with dashed purple edges in Supplemental Fig. S19). However, most of them correspond to micro-inversions and have minor effects on the large-scale ancestral architectures (see Supplement I for detailed comparison of MGRA and inferCARs reconstructions). Table 3 shows the genomic distances from MRD_{MGRA} and MRD_{CARs} to each of the six leaf genomes and illustrates that MGRA results in a slightly more parsimonious scenario as compared to inferCARs (the total distance is 1503 for MGRA and 1518 for inferCARs).

The primate–rodent–carnivore split in mammalian evolution

Knowledge of the correct phylogeny is an important prerequisite for many comparative genomics approaches (Blanchette and

[12] inferCARs reconstructions slightly differ from those reported in Ma et al. (2006) since we use the synteny blocks from the latest builds of mammalian genomes (provided by Jian Ma, University of California, Santa Cruz). Similar to Ma et al. (2006) and Kemkemer et al. (2006), we ignore very short CARs blocks in both inferCARs and MGRA reconstructions to simplify the analysis (see Supplemental Table S14).

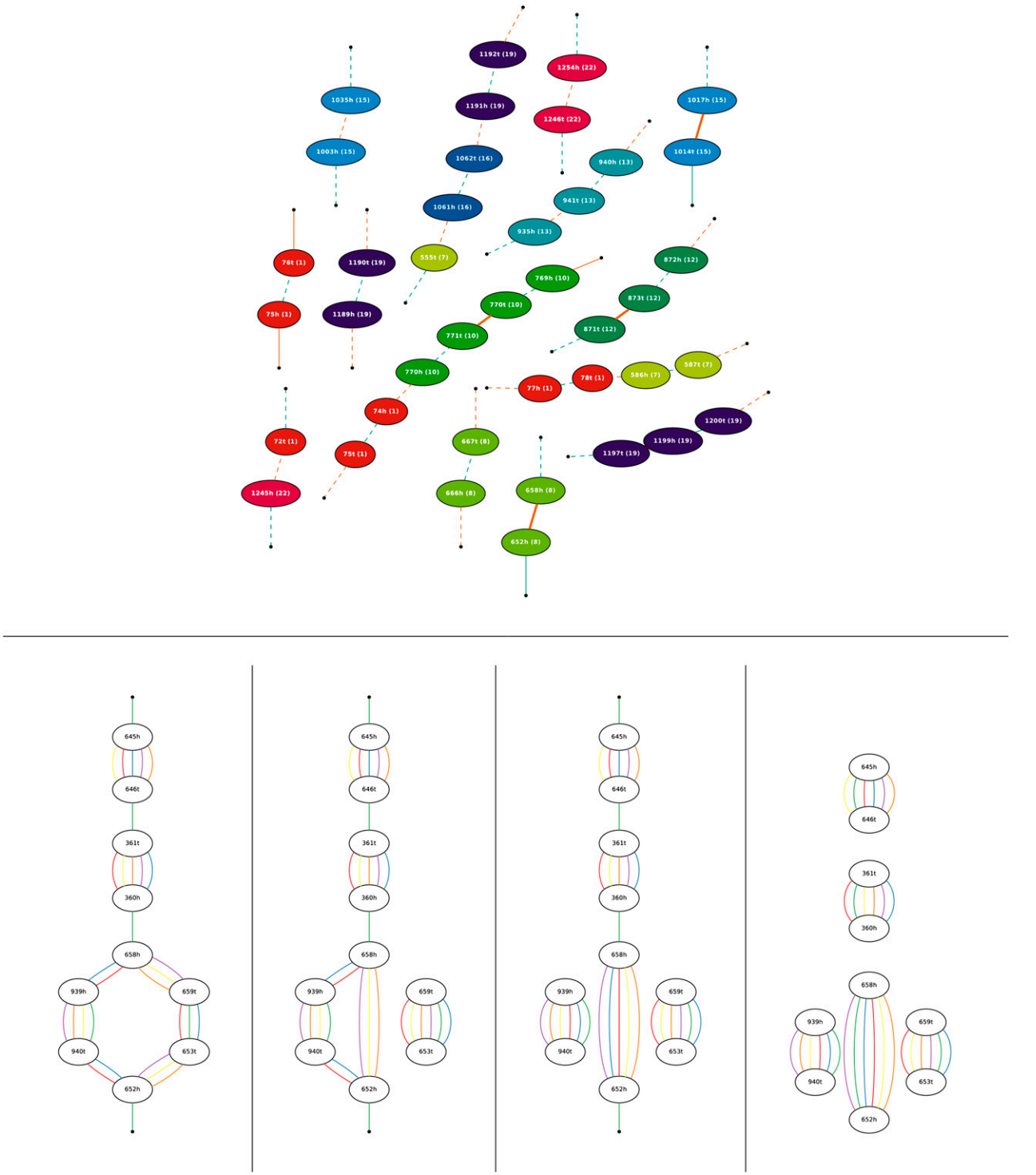

Figure 8. (*Top* panel) The breakpoint graph of the genomes MRD_{CARs} (cyan) and MRD'_{CARs} (orange) reconstructed by inferCARs (common adjacencies are not shown). (Bold edges) Reliable adjacencies (resolved by MGRA Stage 1); (dashed edges) "ambiguous joins" (see Ma et al. 2006) made by inferCARs. Vertex colors are coded as in Figure 4. (*Bottom* panel) A most parsimonious transformation of one connected component (containing vertices 658*h* and 652*h*) of the breakpoint graph $G(M,R,D,Q,H,C)$ from Figure 4. (First panel from *left*) The initial component is transformed with (second panel) a 2-break in primates, (third panel) a 2-break in rodents, and (fourth panel) two 2-breaks in dog resulting from processing of a good D + *MRQCH* path.

Tompa 2002; Kellis et al. 2003). However, even the basic features of the mammalian phylogeny (e.g., the primate–rodent–carnivore split) remain controversial (Fig. 9). While the morphology studies support the primate–rodent clade (Shoshani and McKenna 1998), the early molecular studies supported the primate–carnivore clade (Graur 1993; Janke et al. 1994; Kumar and Hedges 1998; Reyes

Table 3. The genomic distances between the MRD reconstructions MRD_{CARs} and MRD_{MGRA} and the genomes *M, R, D, Q, H, C*

	M	*R*	*D*	*Q*	*H*	*C*	Total
MRD_{CARs}	303	656	180	124	122	133	1518
MRD_{MGRA}	285	637	173	130	133	145	1503

et al. 2000). Although starting from Murphy et al. (2001), the phylogeny based on the primate–rodent clade (Madsen et al. 2001; Poux et al. 2002; Amrine-Madsen et al. 2003; Thomas et al. 2003; Reyes et al. 2004) has become widely accepted, the question is far from being settled: recent studies provided arguments against the primate–rodent clade (Arnason et al. 2002; Misawa and Janke 2003; Jorgensen et al. 2005; Cannarozzi et al. 2007; Huerta-Cepas et al. 2007; Huttley et al. 2007; Niimura and Nei 2007). Below we analyze some rearrangement-based characters supporting both the primate–carnivore and (to a smaller extent) the primate–rodent clade. Similarly to other approaches, the rearrangement analysis reveals some pros and cons for each alternative but does not definitely resolve the long-stranding controversy.

Chaisson et al. (2006) made an attempt to analyze mammalian phylogeny using micro-rearrangements in the CFTR region representing 0.06% of mammalian genomes. However, the small size of this region and ambiguities in revealing micro-rearrangements between distant mammals made it difficult to find micro-rearrangements that can certify the deep branches of the mammalian phylogenetic tree. Cannarozzi et al. (2007) made an attempt to analyze large-scale rearrangements (as opposed to micro-rearrangements) for reconstructing the mammalian evolutionary history. Their approach, while promising, left many questions unanswered. In particular, Cannarozzi et al. (2007) discussed only reversals and ignored translocations, fusions, and fissions. Also, they computed the reversal distances using a (unpublished) greedy algorithm. Since breakpoint reuse is prominent (Pevzner and Tesler 2003) in mammalian evolution, greedy approaches are unlikely to provide an adequate rearrangement scenario. Finally, Cannarozzi et al. (2007) used the "distance-based" rather than "character-based" methods for computing the phylogenetic tree. It is well known that the performance of the distance-based methods deteriorates in the case of the large breakpoint reuse typical for mammalian genomes.

Lunter (2007) criticized Cannarozzi et al. (2007) and wrote in April 2007: "It appears unjustified to continue to consider the phylogeny of primates, rodents, and canines as contentious." Huttley et al. (2007) wrote in May 2007: "We have demonstrated with very high confidence that the rodents diverged before carnivores and primates." (See Niimura and Nei 2007 and Huerta-Cepas et al. 2007 for other recent studies supporting the primate–carnivore clade.) We therefore argue that the rearrangement-based study of the primate–rodent–carnivore controversy is timely.

To analyze the primate–rodent–carnivore controversy, we added the opossum genome (Mikkelsen et al. 2007) to our rearrangement analysis.[13] However, while the phylogenetic tree of the previously considered six mammalian genomes is well established, the position of the opossum genome in this tree is being debated (Fig. 9). Supplemental Table S13 presents the statistics of the breakpoint graph and reveals simple edges supporting the debated tree topologies. Among the non-confident branches, the *MRO* + *DQHC* branch (corresponding the primate–carnivore clade) is supported by 50 multi-edges, while the *DO* + *MRQHC* branch (corresponding to the primate–rodent clade) is supported by 32 multi-edges. We emphasize that only four out of 50 multi-edges supporting the *MRO* + *DQHC* branch represent *MRO* multi-edges, resulting in a very small number of simple *MRO* + *DQHC* vertices (one simple vertex for the *MRO* + *DQHC* branch as compared to zero simple vertices for the *DO* + *MRQHC* branch).

To further address the uncertainty with the opossum branch, we applied MGRA only to the non-controversial parts of the tree with the goal to find characters supporting each of two currently debated tree topologies. The debated tree topologies share (non-controversial) *HC* + *MRDOQ*, *QHC* + *MRDO*, and *MR* + *DOQHC* branches (as well as seven leaf branches corresponding to single genomes). We refer to these branches as "confident" and consider only good and fair paths that correspond to confident branches in MGRA analysis. To further compare the support for the primate–carnivore and the primate–rodent clades, we run MGRA to simplify this breakpoint graph. MGRA Stages 1–2 result in the breakpoint graph (Supplemental Fig. S18, top) that encodes rearrangements during mammalian radiation.

While running MGRA on all seven genomes was important for simplifying the initial breakpoint graph of seven genomes, it hardly makes sense to analyze all these genomes in the complex graph in Supplemental Figure S18 (top). Indeed, we are not interested in subtle inconsistencies between mouse–rat and human–chimpanzee–macaque genomic architectures revealed by this graph. We therefore select single representatives of the primate (macaque–human–chimpanzee ancestor), rodent (mouse–rat ancestor), and carnivore (dog) as well as the outgroup (opossum genome) to simplify the analysis (see Supplemental Fig. S18, top, for a similar analysis with the representatives corresponding to extant macaque, mouse, dog, and opossum genomes).

Table 1B allows one to analyze features supporting the primate–carnivore clade (26 multi-edges of *MO* + *DQ* multi-colors) and the primate–rodent (12 multi-edges of *MQ* + *DO* multi-colors). While the rearrangement-based support for the primate–carnivore clade is more significant than for the primate–rodent clade (26 vs. 12 multi-edges), one cannot exclude a possibility that some complex breakpoint reuse events skewed the statistics in Table 1B in favor of the primate–carnivore clade (see Table 4A–C). Since the elephant genome provides a better (less diverged) outgroup than the opossum genome, there is a hope that the completion of the elephant sequencing project may eventually lead to the resolution of the primate–rodent–carnivore controversy.

We re-run MGRA on the set of seven genomes, assuming the primate–carnivore topology.[14] The resulting rearrangement distances as well as 2-breaks assigned to the *MRO* + *DQHC* branch (supporting the carnivore–primate split) are given in Supplement H.

Discussion

Recently, Froenicke et al. (2006) expressed a concern about some differences between the rearrangement-based and cytogenetics-based approaches to ancestral genome reconstruction. The problem is that some important insights developed by the cytogenetics community still did not find their way into the genome rearrangement tools like MGR, GRAPPA, inferCARs, and EMRAE. While MGRA started as an attempt to close this gap, we quickly realized that the problem of merging the cytogenetics-based and rearrangement-based approaches is far from being simple. First,

[13] Adding the seventh genome increases the number of the synteny blocks to 1746 (by ~30%) but reduces the coverage of the genomes by the synteny blocks from 89% to 79%.

[14] In contrast, Ma et al. (2006) assumed the primate–rodent topology.

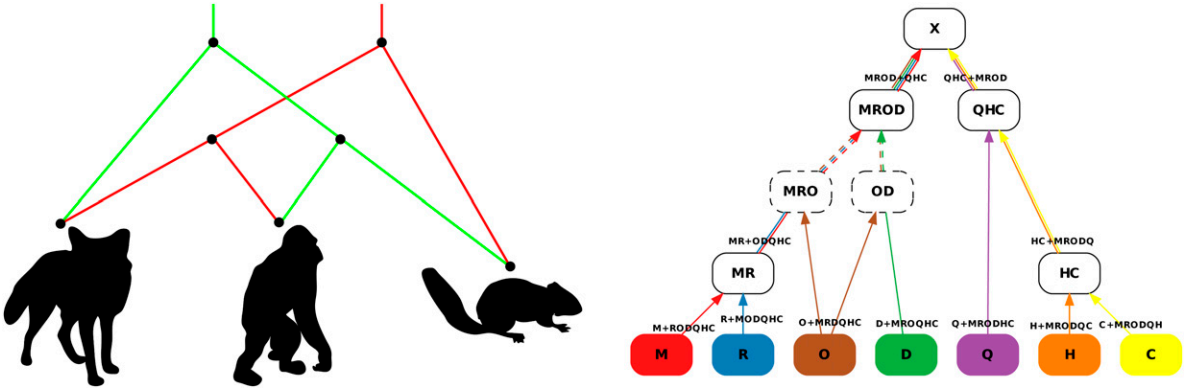

Figure 9. (*Left* panel) The primate–rodent–carnivore controversy: an alternative between the primate–rodent (green tree) and the primate–carnivore clades (red tree). (*Right* panel) The phylogenetic tree *T* of seven mammalian genomes: mouse (red), rat (blue), dog (green), macaque (violet), human (orange), chimpanzee (yellow), and opossum (brown). Since the opossum branch is subject to a controversy, the dashed branches represent possible variations, while the solid branches are confident and do not depend on the opossum branch.

there is still no cytogenetics-based software that can be automatically applied to genome-scale data sets to enable an unbiased comparison of two approaches on the same data set. Second, it is not clear how well the cytogenetics approach scales with increase in the resolution, for example, with 1000+ synteny blocks from Ma et al. (2006).

Despite the low resolution of the cytogenetics data, the cytogenetics-based ancestral reconstructions are very accurate as there are relatively few discrepancies between the cytogenetics-based and the recent genomics-based high-resolution reconstructions (Bourque et al. 2005; Murphy et al. 2005; Ma et al. 2006). Moreover, the discrepancies are usually attributed to

Table 4. The statistics of the breakpoint graph of the ancestral rodent (mouse–rat ancestor), ancestral primate (macaque–human–chimpanzee ancestor), carnivore (dog), and opossum genomes

Multi-colors	Multi-edges	Simple vertices	Simple multi-edges	Simple paths + cycles
		A. Before running MGRA		
∅ + RPCO	0 + 627 = 627	0	0 + 0 = 0	0 + 0 = 0
O + RPC	617 + 740 = 1357	1230	613 + 523 = 1136	93 + 109 = 202
R + PCO	280 + 173 = 453	346	121 + 173 = 294	52 + 28 = 80
C + RPO	142 + 246 = 388	284	142 + 96 = 238	46 + 33 = 79
P + RCO	49 + 124 = 173	98	49 + 31 = 80	18 + 13 = 31
RO + PC	5 + 46 = 51	1	0 + 0 = 0	0 + 0 = 0
RP + CO	32 + 11 = 43	1	0 + 0 = 0	0 + 0 = 0
RC +PO	16 + 8 = 24	2	0 + 0 = 0	0 + 0 = 0
		B. After MGRA Stage 1		
∅ + RPCO	0 + 1712 = 1712	0	0 + 0 = 0	0 + 0 = 0
O + RPC	10 + 26 = 36	10	0 + 0 = 0	0 + 0 = 0
R + PCO	22 + 5 = 27	5	0 + 0 = 0	0 + 0 = 0
C + RPO	0 + 18 = 18	0	0 + 0 = 0	0 + 0 = 0
P + RCO	0 + 16 = 16	0	0 + 0 = 0	0 + 0 = 0
RO + PC	4 + 16 = 20	6	2 + 1 = 3	2 + 0 = 2
RP + CO	5 + 4 = 9	4	0 + 0 = 0	0 + 0 = 0
RC + PO	5 + 4 = 9	4	1 + 0 = 1	1 + 0 = 1
		C. After MGRA Stage 2		
∅ + RPCO	0 + 1743 = 1743	0	0 + 0 = 0	0 + 0 = 0
O + RPC	0 + 2 = 2	0	0 + 0 = 0	0 + 0 = 0
C + RPO	0 + 2 = 2	0	0 + 0 = 0	0 + 0 = 0
P + RCO	0 + 1 = 1	0	0 + 0 = 0	0 + 0 = 0
RO + PC	9 + 10 = 19	12	3 + 3 = 6	5 + 0 = 5
RC + PO	8 + 8 = 16	10	2 + 3 = 5	3 + 1 = 4
RP + CO	7 + 8 = 15	10	4 + 2 = 6	4 + 0 = 4

The ancestral rodent and primate ancestors were reconstructed using MGRA and were used as the genomes in the leaves of the phylogenetic tree for four species: rodent, primate, dog, and opossum. The statistics are shown for before running MGRA, after MGRA Stage 1, and after MGRA Stage 2 run on the leaf branches (i.e., without assuming any particular topology of the phylogenetic tree). The *T*-consistent multi-colors are shown in bold. Compare to Table 1B and Supplemental Table S13.

some arbitrary assignments of the genomics-based MGR algorithm (Froenicke et al. 2006) rather than errors in the cytogenetics analysis. Indeed, MGR was developed for finding the most parsimonious scenario rather than finding which rearrangements in this scenario are less reliable than others. The discrepancies between MGR and the cytogenetics-based reconstructions are likely to be a reflection of the "strength in numbers" principle rather than shortcomings of the genomics-based approaches: while the cytogenetics reconstructions are based on more than 100 known cytogenetics maps, there are still only seven completed mammalian genomic sequences suitable for the rearrangement analysis. However, even with a small increase in the number of the genomes from three to four (as in Bourque et al. 2004, 2005) to six to seven (as in Murphy et al. 2005; Ma et al. 2006), there are very few discrepancies between the cytogenetics-based and the genomics-based approaches (Rocchi et al. 2006). Despite a recent debate (Bourque et al. 2006; Froenicke et al. 2006), the cytogenetics-based and genomics-based approaches are converging and benefiting from the higher resolution of the genomics-based approaches. However, the key condition for such convergence is the availability of algorithms that improve on the existing heuristics for separating between strong and weak associations. We addressed this challenge by devising the MGRA algorithm, which remedies some limitations of the previous approaches to ancestral reconstructions. Similarly to the algorithms recently proposed by Ma et al. (2008) and Chauve and Tannier (2008) (published after this paper was submitted), MGRA focuses on accurate rather than the most parsimonious ancestral reconstructions.

Acknowledgments

We thank Jian Ma for providing us with the synteny blocks for mammalian genomes from the latest builds and for numerous thoughtful discussions. We thank Bill Murphy for a discussion about the primate–rodent–carnivore controversy, and Guillaume Bourque and Glenn Tesler for discussions on the algorithmic aspects of this project. We also thank Lutz Froenicke and Claus Kemkemer for useful comments on the cytogenetics approach and CytoAncestor. This work was supported by the Howard Hughes Professor Award.

References

Alekseyev, M.A. 2008. Multi-break rearrangements and breakpoint reuses: From circular to linear genomes. *J. Comput. Biol.* **15:** 1117–1131.

Alekseyev, M.A. and Pevzner, P.A. 2007. Whole genome duplications, multi-break rearrangements, and genome halving theorem. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 665–679. SIAM, Philadelphia.

Alekseyev, M.A. and Pevzner, P.A. 2008. Multi-break rearrangements and chromosomal evolution. *Theor. Comput. Sci.* **395:** 193–202.

Amrine-Madsen, H., Koepfli, K.-P., Wayne, R.K., and Springer, M.S. 2003. A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. *Mol. Phylogenet. Evol.* **28:** 225–240.

Arnason, U., Adegoke, J.A., Bodin, K., Born, E.W., Esa, Y.B., Gullberg, A., Nilsson, M., Short, R.V., Xu, X., Janke, A., et al. 2002. Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc. Natl. Acad. Sci.* **99:** 8151–8156.

Bafna, V. and Pevzner, P.A. 1996. Genome rearrangement and sorting by reversals. *SIAM J. Comput.* **25:** 272–289.

Bafna, V. and Pevzner, P.A. 1998. Sorting permutations by transpositions. *SIAM J. Discrete Math.* **11:** 224–240.

Bailey, J., Baertsch, R., Kent, W., Haussler, D., and Eichler, E. 2004. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5:** R23. doi: 10.1186/gb-2004-5-4-r23.

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

Blanchette, M. and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12:** 739–748.

Blanchette, M., Bourque, G., and Sankoff, D. 1997. Breakpoint phylogenies. *Genome Inform. Ser. Workshop Genome Inform.* **8:** 25–34.

Bourque, G. and Pevzner, P.A. 2002. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Res.* **12:** 26–36.

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

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

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

Bulazel, K., Ferreri, G., Eldridge, M., and O'Neill, R. 2007. Species-specific shifts in centromere sequence composition are coincident with breakpoint reuse in karyotypically divergent lineages. *Genome Biol.* **8:** R170. doi: 10.1186/gb-2007-8-8-r170.

Caceres, M., Sullivan, R.T., and Thomas, J.W. 2007. A recurrent inversion on the eutherian X chromosome. *Proc. Natl. Acad. Sci.* **104:** 18571–18576.

Cannarozzi, G., Schneider, A., and Gonnet, G. 2007. A phylogenomic study of human, dog, and mouse. *PLoS Comput. Biol.* **3:** e2. doi: 10.1371/journal.pcbi.0030002.

Caprara, A. 1999a. Formulations and hardness of multiple sorting by reversals. In *RECOMB '99: Proceedings of the Third Annual International Conference on Computational Molecular Biology,* pp. 84–93. ACM, New York.

Caprara, A. 1999b. On the tightness of the alternating-cycle lower bound for sorting by reversals. *J. Comb. Optim.* **3:** 149–182.

Cardone, M., Jiang, Z., D'Addabbo, P., Archidiacono, N., Rocchi, M., Eichler, E., and Ventura, M. 2008. Hominoid chromosomal rearrangements on 17q map to complex regions of segmental duplication. *Genome Biol.* **9:** R28. doi: 10.1186/gb-2008-9-2-r28.

Chaisson, M.J., Raphael, B.J., and Pevzner, P.A. 2006. Microinversions in mammalian evolution. *Proc. Natl. Acad. Sci.* **103:** 19824–19829.

Chauve, C. and Tannier, E. 2008. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput. Biol.* **4:** e1000234. doi: 10.1371/journal.pcbi.1000234.

Deuve, J., Bennett, N., Britton-Davidian, J., and Robinson, T. 2008. Chromosomal phylogeny and evolution of the African mole-rats (Bathyergidae). *Chromosome Res.* **16:** 57–74.

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

Gordon, L., Yang, S., Tran-Gyamfi, M., Baggott, D., Christensen, M., Hamilton, A., Crooijmans, R., Groenen, M., Lucas, S., Ovcharenko, I., et al. 2007. Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions. *Genome Res.* **17:** 1603–1613.

Graur, D. 1993. Towards a molecular resolution of the ordinal phylogeny of the eutherian mammals. *FEBS Lett.* **325:** 152–159.

Hannenhalli, S. and Pevzner, P. 1995. Transforming men into mouse (polynomial algorithm for genomic distance problem). In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science,* pp. 581–592.

Hannenhalli, S. and Pevzner, P. 1999. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *JACM* **46:** 1–27.

Hinsch, H. and Hannenhalli, S. 2006. Recurring genomic breaks in independent lineages support genomic fragility. *BMC Evol. Biol.* **6:** 90. doi: 10.1186/1471-2148-6-90.

Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldon, T. 2007. The human phylome. *Genome Biol.* **8:** R109. doi: 10.1186/gb-2007-8-6-r109.

Huttley, G.A., Wakefield, M.J., and Easteal, S. 2007. Rates of genome evolution and branching order from whole genome analysis. *Mol. Biol. Evol.* **24:** 1722–1730.

Janke, A., Feldmaier-Fuchs, G., Thomas, W.K., von Haeseler, A., and Paabo, S. 1994. The marsupial mitochondrial genome and the evolution of placental mammals. *Genetics* **137:** 243–256.

Jorgensen, F., Hobolth, A., Hornshoj, H., Bendixen, C., Fredholm, M., and Schierup, M. 2005. Comparative analysis of protein coding sequences from human, mouse and the domesticated pig. *BMC Biol.* **3:** 2. doi: 10.1186/1741-7007-3-2.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.
Kemkemer, C., Kohn, M., Kehrer-Sawatzki, H., Minich, P., Högel, J., Froenicke, L., and Hameister, H. 2006. Reconstruction of the ancestral ferungulate karyotype by electronic chromosome painting (E-painting). *Chromosome Res.* **14:** 899–907.
Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A.Z., Engstrom, P.G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., et al. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17:** 545–555.
Kumar, S. and Hedges, S.B. 1998. A molecular timescale for vertebrate evolution. *Nature* **392:** 917–920.
Lin, Y. and Moret, B.M. 2008. Estimating true evolutionary distances under the DCJ model. *Bioinformatics* **24:** i114–i122.
Lunter, G. 2007. Dog as an outgroup to human and mouse. *PLoS Comput. Biol.* **3:** e74. doi: 10.1371/journal.pcbi.0030074.
Ma, J., Zhang, L., Suh, B.B., Raney, B.J., Burhans, R.C., Kent, J.W., Blanchette, M., Haussler, D., and Miller, W. 2006. Reconstructing contiguous regions of an ancestral genome. *Genome Res.* **16:** 1557–1565.
Ma, J., Ratan, A., Raney, B.J., Suh, B.B., Miller, W., and Haussler, D. 2008. The infinite sites model of genome evolution. *Proc. Natl. Acad. Sci.* **105:** 14254–14261.
Madsen, O., Scally, M., Douady, C.J., Kao, D.J., DeBry, R.W., Adkins, R., Amrine, H.M., Stanhope, M.J., de Jong, W.W., Springer, M.S., et al. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409:** 610–614.
Mehan, M.R., Almonte, M., Slaten, E., Freimer, N.B., Rao, P.N., and Ophoff, R.A. 2007. Analysis of segmental duplications reveals a distinct pattern of continuation-of-synteny between human and mouse genomes. *Hum. Genet.* **121:** 93–100.
Mikkelsen, T.S., Wakefield, M.J., Aken, B., Amemiya, C.T., Chang, J.L., Duke, S., Garber, M., Gentles, A.J., Goodstadt, L., Heger, A., et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447:** 167–177.
Misawa, K. and Janke, A. 2003. Revisiting the Glires concept—Phylogenetic analysis of nuclear sequences. *Mol. Phylogenet. Evol.* **28:** 320–327.
Moret, B., Wyman, S., Bader, D.A., Warnow, T., and Yan, M. 2001. A new implementation and detailed study of breakpoint analysis. In *Pacific Symposium on Biocomputing,* pp. 583–594. Hawaii.
Moret, B.M.E., Siepel, A.C., Tang, J., and Liu, T. 2002. Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. *Lect. Notes Comput. Sci.* **2452:** 521–536.
Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A., and O'Brien, S.J. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* **409:** 614–618.
Murphy, W.J., Larkin, D.M., van der Wind, A.E., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Gatzke, L., et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative map. *Science* **309:** 613–617.
Niimura, Y. and Nei, M. 2007. Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS One* **2:** e708. doi: 10.1371/journal.pone.0000708.
Ozery-Flato, M. and Shamir, R. 2003. Two notes on genome rearrangement. *J. Bioinform. Comput. Biol.* **1:** 71–94.
Pevzner, P.A. 2000. *Computational molecular biology: An algorithmic approach*. MIT Press, Cambridge, MA.
Pevzner, P.A. and Tesler, G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci.* **100:** 7672–7677.
Pontius, J.U., Mullikin, J.C., Smith, D.R., Lindblad-Toh, K., Gnerre, S., Clamp, M., Chang, J., Stephens, R., Neelam, B., Volfovsky, N., et al. 2007. Initial sequence and comparative analysis of the cat genome. *Genome Res.* **17:** 1675–1689.
Poux, C., van Rheede, T., Madsen, O., and de Jong, W.W. 2002. Sequence gaps join mice and men: Phylogenetic evidence from deletions in two proteins. *Mol. Biol. Evol.* **19:** 2035–2037.
Reyes, A., Gissi, C., Pesole, G., Catzeflis, F.M., and Saccone, C. 2000. Where do rodents fit? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*. *Mol. Biol. Evol.* **17:** 979–983.
Reyes, A., Gissi, C., Catzeflis, F., Nevo, E., Pesole, G., and Saccone, C. 2004. Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods. *Mol. Biol. Evol.* **21:** 397–403.
Rocchi, M., Archidiacono, N., and Stanyon, R. 2006. Ancestral genomes reconstruction: An integrated, multi-disciplinary approach is needed. *Genome Res.* **16:** 1441–1444.
Ruiz-Herrera, A., Castresana, J., and Robinson, T.J. 2006. Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biol.* **7:** R115. doi: 10.1186/gb-2006-7-12-r115.
Sankoff, D. and Blanchette, M. 1998. Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.* **5:** 555–570.
Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F., and Cedergren, R. 1992. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci.* **89:** 6575–6579.
Shoshani, J. and McKenna, M.C. 1998. Higher taxonomic relationships among extant mammals based on morphology, with selected comparisons of results from molecular data. *Mol. Phylogenet. Evol.* **9:** 572–584.
Tang, J. and Moret, B.M. 2003. Scaling up accurate phylogenetic reconstruction from gene-order data. *Bioinformatics* **19:** i305–i312.
Tannier, E. and Sagot, M.-F. 2004. Sorting by reversals in subquadratic time. *Lect. Notes Comput. Sci.* **3109:** 1–13.
Tannier, E., Zheng, C., and Sankoff, D. 2008. Multichromosomal genome median and halving problems. *Lect. Notes Bioinformatics* **5251:** 1–13.
Tesler, G. 2002a. Efficient algorithms for multichromosomal genome rearrangements. *J. Comput. Syst. Sci.* **65:** 587–609.
Tesler, G. 2002b. GRIMM: Genome rearrangements web server. *Bioinformatics* **18:** 492–493.
Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.
van der Wind, A.E., Kata, S.R., Band, M.R., Rebeiz, M., Larkin, D.M., Everts, R.E., Green, C.A., Liu, L., Natarajan, S., Goldammer, T., et al. 2004. A 1463 gene cattle–human comparative map with anchor points defined by human genome sequence coordinates. *Genome Res.* **14:** 1424–1437.
Webber, C. and Ponting, C.P. 2005. Hotspots of mutation and breakage in dog and human chromosomes. *Genome Res.* **15:** 1787–1797.
Wienberg, J. and Stanyon, R. 1997. Comparative painting of mammalian chromosomes. *Curr. Opin. Genet. Dev.* **7:** 784–791.
Xia, A., Sharakhova, M., and Sharakhov, I. 2007. Reconstructing an inversion history in the *Anopheles gambiae* complex. *Lect. Notes Bioinformatics* **4751:** 136–148.
Yancopoulos, S., Attie, O., and Friedberg, R. 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* **21:** 3340–3346.
Yue, Y. and Haaf, T. 2006. 7E olfactory receptor gene clusters and evolutionary chromosome rearrangements. *Cytogenet. Genome Res.* **112:** 6–10.
Zhao, H. and Bourque, G. 2007. Recovering true rearrangement events on phylogenetic trees. *Lect. Notes Bioinformatics* **4751:** 149–161.
Zhao, H. and Bourque, G. 2009. Recovering genome rearrangements in the mammalian phylogeny. *Genome Res.* (this issue). doi: 10.1101/gr.086009.108.
Zhao, S., Shetty, J., Hou, L., Delcher, A., Zhu, B., Osoegawa, K., de Jong, P., Nierman, W.C., Strausberg, R.L., Fraser, C.M., et al. 2004. Human, mouse, and rat genome large-scale rearrangements: Stability versus speciation. *Genome Res.* **14:** 1851–1860.

Received June 30, 2008; accepted in revised form January 22, 2009.