Supplement A Independent, semi-independent, and weakly-independent rearrangements

Let P be a genome represented a graph on black and obverse edges. For any m black edges in P, we define an *m-break* (or *multi-break*) rearrangement as replacement of these edges with other m black edges forming a matching on the same vertices (see Alekseyev and Pevzner (2008a); Alekseyev (2008)).

Let $P_1, P_2, \ldots, P_k$ be genomes that evolved by some unknown multi-breaks following an unknown evolutionary tree (we allow any combination of 2-breaks, 3-breaks, and m-breaks for $m > 3$ in a single evolutionary scenario). Without loss of generality, we assume that there was at least one multi-break on every branch of the tree. One may classify an m-break as *independent* if it does not reuse breakpoints, i.e., creates exactly m new breakpoints. Similarly, we call the rearrangement scenario *independent* if all its multi-breaks are independent.

If all multi-breaks are independent, the following theorem applies (the proof follows the proof of a similar result in Chaisson et al. (2006)).

Theorem 1. *If genomes $P_1, P_2, \ldots, P_k$ are produced by independent multi-breaks, then both the correct evolutionary tree for these genomes and the ancestral genomes in all its branching nodes may be reconstructed in polynomial time.*

We now consider the case when genomes $P_1, P_2, \ldots, P_k$ are produced by *semi-independent* 2-breaks that may re-use breakpoints *within* single branches of the evolutionary tree (i.e., a semi-independent 2-break does not share breakpoints with any other 2-break on a different branch of T). We call the 2-break rearrangement scenario *semi-independent* if all its 2-breaks are semi-independent. Since any semi-independent 2-break scenario corresponds to an independent multi-break scenario (see Alekseyev and Pevzner (2008a)), Theorem 1 implies:

Theorem 2. *If genomes $P_1, P_2, \ldots, P_k$ are produced by semi-independent 2-breaks, then both the correct evolutionary tree for these genomes and the ancestral genomes in all its branching nodes may be reconstructed in polynomial time.*

MGRA optimally solves the MGRP problem in case of semi-independent 2-breaks and uses heuristics to move beyond the semi-independent assumption. In a new manuscript (Alekseyev and Pevzner, 2009b) we define the notion of *weakly independent* rearrangements that relaxes the semi-independent assumption by allowing breakpoint re-uses within selected *pairs of incident branches* in the phylogenetic tree (as opposed to a single branch in semi-independent scenarios). We demonstrate that the TCMGRP can be solved efficiently in the case of weakly independent scenarios.[1] Below we show that most 2-breaks in mammalian evolution are either independent, or semi-independent, or weakly independent resulting in reliable ancestral reconstructions. The theoretical analysis of weakly independent scenarios is not crucial for understanding MGRA and will be described elsewhere.

Supplement B Simultaneous T-consistent transformations

The problem of finding a shortest rearrangement scenario typically has many solutions. To characterize all genomes that may appear in shortest rearrangement scenarios between genomes A and B, we say that a genome Q is *located between* A and B if the rearrangement distances between these genomes satisfy the condition: $d(A, Q) + d(Q, B) = d(A, B)$. In the case of a phylogenetic tree T with known genomes at all internal nodes, we say that a genome Q is *located on a branch* (A, B) of the

[1] In particular, while the reversal median problem is NP-complete for arbitrary scenarios (Caprara, 1999a), one can efficiently reconstruct the 2-break median for 3 genomes in case of weakly independent scenarios.

phylogenetic tree T if it located between nodes (genomes) A and B. Similarly, a genome Q is *located on the tree* T if it is located on a branch of T. A transformation between two genomes located on the tree T is called *T-consistent* if every intermediate genome in this transformation is also located on T.

Let T be a tree with known genomes specified at every node. A tree T' is *homeomorphic* to the tree T if it is derived from T by adding extra internal nodes (of degree 2) within branches of T and specifying some genomes at these added nodes. For example, Fig. 3c represents a tree homeomorphic to the tree in Fig. 3a with two extra nodes added to the branch (Q_2, Q_3). A homeomorphic tree T' defines a *T-consistent rearrangement scenario* if genomes at every two adjacent nodes of T' differ from each other by a single rearrangement and the total number of rearrangements along each branch is minimal. We now reformulate the problem of finding the most parsimonious rearrangement scenario as the problem of finding a T-consistent rearrangement scenario with the minimum number of nodes in the homeomorphic tree T'.

If X is an arbitrary genome (*root*) in T' then the path $path(X, P_i)$ from the root X to every leaf genome P_i in T' corresponds to a series of rearrangements transforming X into P_i $(i = 1, 2, \ldots, k)$. A set of nodes C in T' is called a *cut* if each path $path(X, P_i)$ contains exactly one node from C (for $1 \le i \le k$). For example, the sets $\{X\}$ and $\{P_1, \ldots, P_k\}$ represent cuts in T' with minimum and maximum number of nodes correspondingly. Given a cut C and a leaf genome P_i, let v_i^C be a (single) node in C located on a path $path(X, P_i)$ and let P_i^C be the genome assigned to the node v_i^C. Therefore, every cut C defines k genomes $P_1^{C_i}, P_2^{C_i}, \ldots P_k^{C_i}$.

One can orient branches of T' in the direction from X to the leaves and define $next(v)$ as the set of children of an internal node v (the number of children equals the degree of v minus one). Given an internal node v in a cut C, we define a new cut $next_v(C)$ obtained from C by deleting a node v and adding the set of nodes $next(v)$. A *simultaneous T-consistent transformation* of the root genome X into the leaf genomes $P_1, \ldots, P_k$ is a series of cuts $\{X\} = C_0, C_1, \ldots, C_d = \{P_1, \ldots, P_k\}$ such that $C_{i+1} = next_{v_i}(C_i)$ for some node $v_i \in C_i$ $(0 \le i < d)$. It is easy to see that for every T-consistent transformation there exists a simultaneous T-consistent transformation. Below we give an equivalent definition of the simultaneous T-consistent transformation in terms of multiple breakpoint graphs that motivates MGRA algorithm attempting to find a shortest simultaneous T-consistent transformation.

Any subset of edges from the multi-edge (x, y) represents a *sub-multi-edge* (x, y) of the multi-color formed by the colors of the edges in this subset. Any simultaneous T-consistent transformation of X into $P_1, \ldots, P_k$ defines a transformation of the identity breakpoint graph $G(X, \ldots, X)$ into $G(P_1, P_2, \ldots, P_k)$ with a series of rearrangements applied to sub-multi-edges of T-consistent multicolors. Namely, we define the multiple breakpoint graph corresponding to the cut C_i as $G_i = G(P_1^{C_i}, P_2^{C_i}, \ldots, P_k^{C_i})$. It is easy to see that G_{i+1} is obtained from G_i by a single rearrangement ρ applied to all copies of some genome Q in $P_1^{C_i}, P_2^{C_i}, \ldots, P_k^{C_i}$. Alternatively, a transformation of G_i into G_{i+1} can be viewed as applying ρ to T-consistent sub-multi-edges in G_i (of the multicolor $\{P_i \mid P^{C_i} = Q\}$).

Supplement C Reconstructing reliable CARs

To reconstruct reliable adjacencies in the ancestor genome at a node of the phylogenetic tree, we select this node as a root node X. Then we start to eliminate breakpoints in the breakpoint graph $G(P_1, \ldots, P_k)$ with reliable 2-breaks (for whatever definition of reliability) and stop when no further reliable 2-breaks exist. In the resulting breakpoint graph (that may still have some breakpoints) the multi-edges of multicolors containing X (as a subset of colors) represent the reliable block adjacencies in the target ancestor genome, and we generate CARs based only on such adjacencies. Note that this approach can reconstruct CARs in only one ancestor genome at a time, and multiple runs (with different root nodes X) are needed to reconstruct CARs in multiple ancestor genomes.

Supplement D Comparison of various ancestral reconstructions for a component of the breakpoint graph representing the human chromosomes 7, 16, and 19

Below we focus 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. (2006a). The advantage of the breakpoint graph approach is that it enables a simple analysis of this controversy since the analysis is reduced to "genomes" with only 6 synteny blocks after equivalent transformations performed by MGRA Stage 1 (Fig. S13). Indeed, the genomes represented by this component can now be viewed as:

Macaque: $4, 5, (2, -6), (3, -1)$
Mouse: $2, 3, (4, 6), (1, -5)$
Dog: $1, (4, -6, -5), (2, -3)$

where the block 1 is located on the human chromosome 7, the blocks 2 and 3 are located on the human chromosome 16 and the blocks 4, 5, 6 are located on the human chromosome 19 (Fig. S11, top panel and Fig. S13). Ma et al. (2006a) proposed the ancestral architecture with 4 chromosomes $1, 5, (2, -3), (6, -4)$ (no associations between chromosomes 7, 16, and 19), while Froenicke et al. (2006) proposed $4, 5, (1, -3), (2, -6)$ (associations 16 + 7 and 16 + 19). It is easy to see that the cytogenetics reconstruction is less parsimonious than Ma et al. (2006a) reconstruction. We therefore argue that the criticism of Ma et al. (2006a) in Rocchi et al. (2006) regarding the missing association 7 + 16 is not fully justified since the whole genome data do not support this associations.[2]

MGRA Stage 2 also generates a solution that improves on the cytogenetics reconstruction (Froenicke et al., 2006) and proposes the ancestral association 7 + 19. While both our and Ma et al. (2006a) solutions are more parsimonious than the cytogenetics reconstruction, we are not claiming that these solutions are necessarily correct (the most parsimonious scenarios on 6 genomes are not necessarily the most parsimonious scenarios on 100+ genomes). The important thing is that MGRA Stage 1 reduces analysis of the 7/16/19 controversy to such a small example that all possible scenarios can be explored.

Supplement E Selecting fair multi-edges in MGRA Stage 2

As described in the main text, the order of selected fair multi-edges may affect the ancestral reconstructions at some nodes of the phylogenetic tree. Below we specify how MGRA Stage 2 selects such edges.

Note that if a fair multi-edge is not $\vec{T}$-consistent then this multi-edge can only be affected by 2-breaks on adjacent multi-edges. Although two such 2-breaks are possible, ordering of these 2-breaks does not influence the final result (see Fig. 6, bottom panel). However, the situations when two fair multi-edges share a vertex (and both are $\vec{T}$-consistent) may be ambiguous since the final result of their processing may be affected by the order in which these edges are processed. MGRA Stage 2 starts by processing unambiguous fair edges first and selects the order of remaining fair edges according to the following heuristics.

For a fixed $\vec{T}$-consistent multicolor Q, an "ideal" 2-break of multicolor Q should satisfy two conditions: (i) it increases the number of cycles alternating between every pair of colors from $Q + \overline{Q}$ (i.e., one color from Q while the other is from $\overline{Q}$), and (ii) it does not decrease the number of cycles alternating between the other pairs of colors. It is easy to see that if a 2-break on a $\vec{T}$-consistent multicolor Q increases the number of connected components in the breakpoint graph then both these conditions are satisfied. For each $\vec{T}$-consistent multicolor Q, MGRA Stage 2 finds all 2-breaks on

[2]We are not claiming that the analysis of 7 + 16 association in Froenicke (2005); Robinson et al. (2006) is incorrect but instead argue that it is not supported by data used in Ma et al. (2006a).

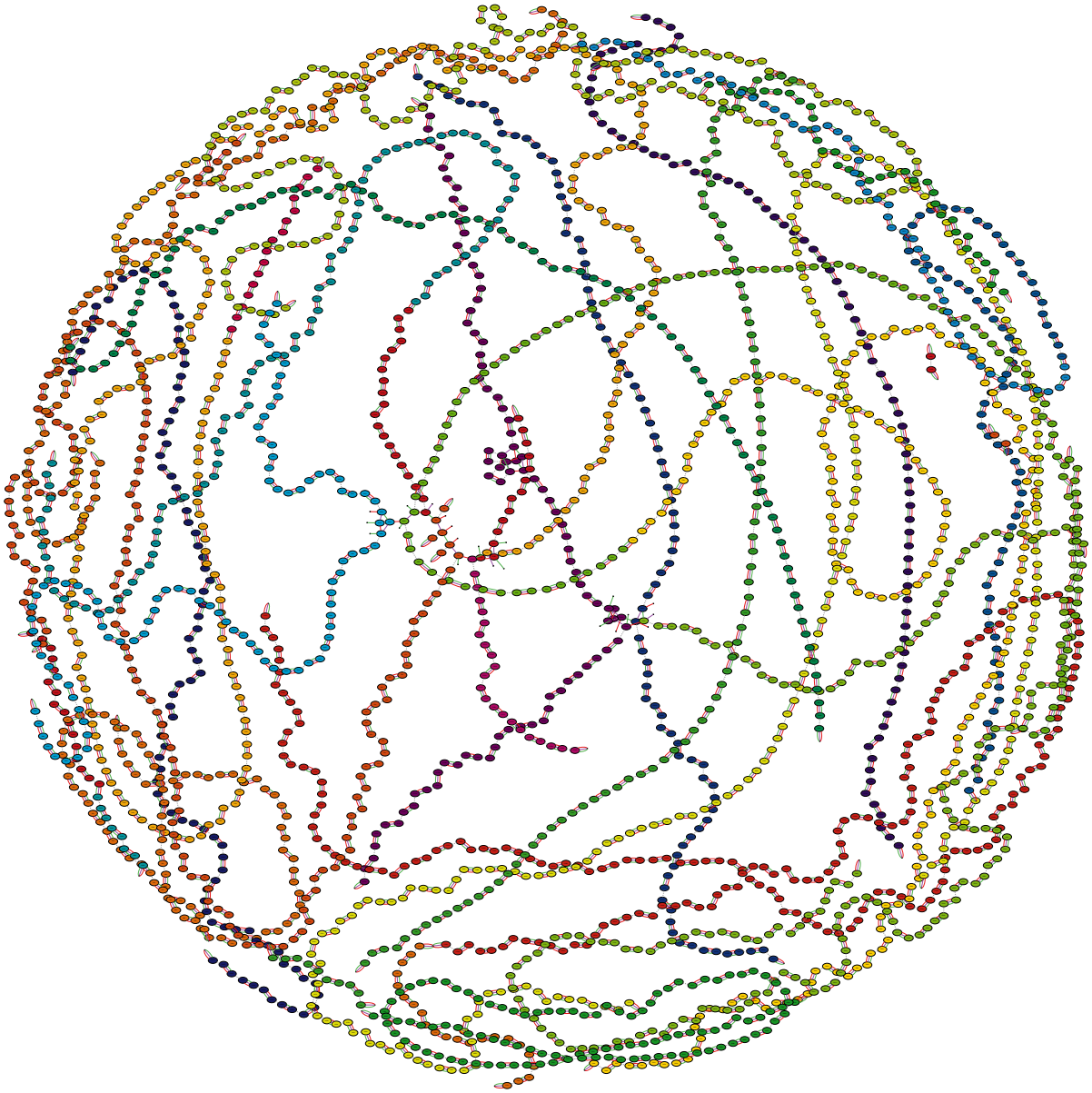

Figure S10: The breakpoint graph of Mouse (*red*), Dog (*green*), and macaQue (*violet*) genomes after MGRA Stage 1 with complete multi-edges and obverse edges shown (in contrast to Fig. 7). The obverse edges reveal many unicolored paths formed by alternating obverse edges and complete multi-edges. Vertices are labeled and colored similarly to Fig. 4.

multicolor Q that increase the number of connected components and perform them (in an arbitrary order).

Supplement F MGRA Stage 3: Processing complex breakpoints

If one attempts to find the positions and orientations of short synteny block in the ancestors, there are two possibilities. If the signs of the blocks are inferred correctly then the same micro-inversion happened independently on two different branches of the evolutionary tree. However, if the signs of the blocks are incorrect, manual re-examination of some blocks may be necessary. Recently, Ma et al. (2006a) and Chaisson et al. (2006) emphasized the difficulties in detecting micro-inversions and improved on previous work in detecting micro-inversions (Feuk et al., 2005). While these papers

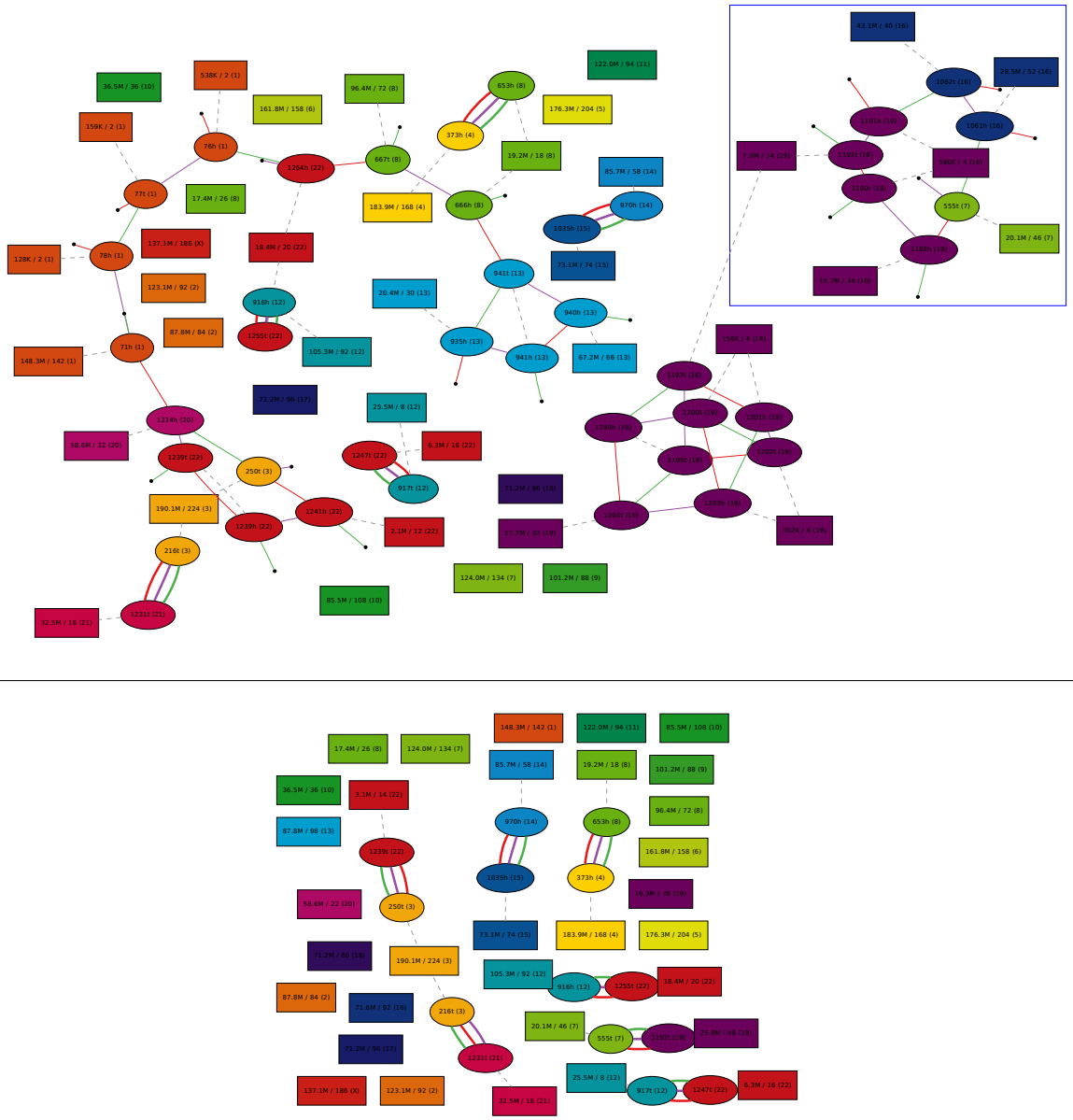

Figure S11: Compact representation of the breakpoint graph of the Mouse (*red*), Dog (*green*), and macaQue (*violet*) genomes after MGRA Stage 1 (top) and Stages 1-2 (bottom) (compare to Fig. S10). Every unicolored alternating path of obverse edges and complete multi-edges (with possible exception of the initial and terminal synteny blocks) is represented as a rectangular vertex labeled by the overall length and number of synteny blocks in this alternating path. The numbers in parentheses as well as vertex colors indicate the corresponding human chromosome. The isolated vertices of the total length shorter than 15 Mb are not shown. The observed edges are shown as dashed edges. The boxed selected component (top) is analyzed in Fig. S13.

resulted in two largely consistent sets of human-chimpanzee micro-inversions, there are still some differences between the sets/signs of human-chimpanzee micro-inversions generated by algorithms in Ma et al. (2006a) and Chaisson et al. (2006), indicating that some micro-inversions detected by these approaches may be unreliable. The micro-inversions detection becomes even more difficult when one moves from human and chimpanzee to more distant mammals.

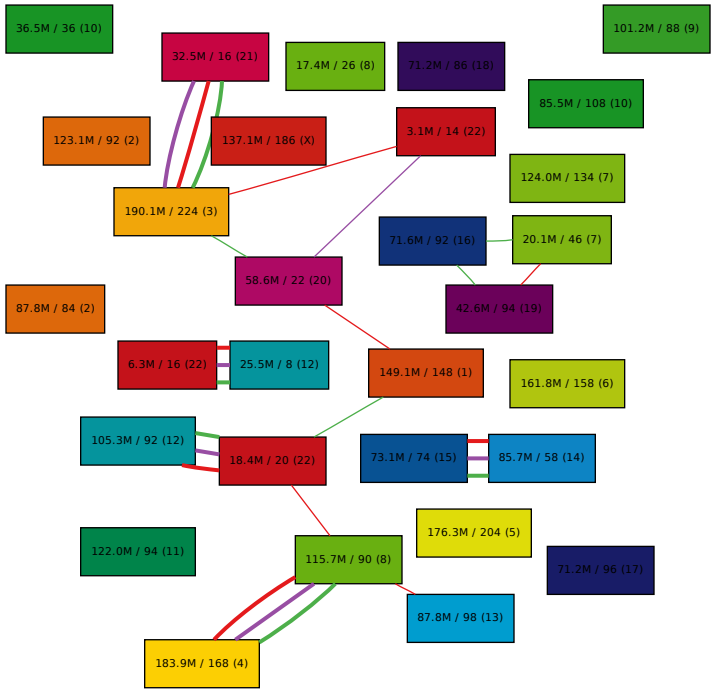

Figure S12: Compact representation of the unicolored connected components in Fig. S11 (top).

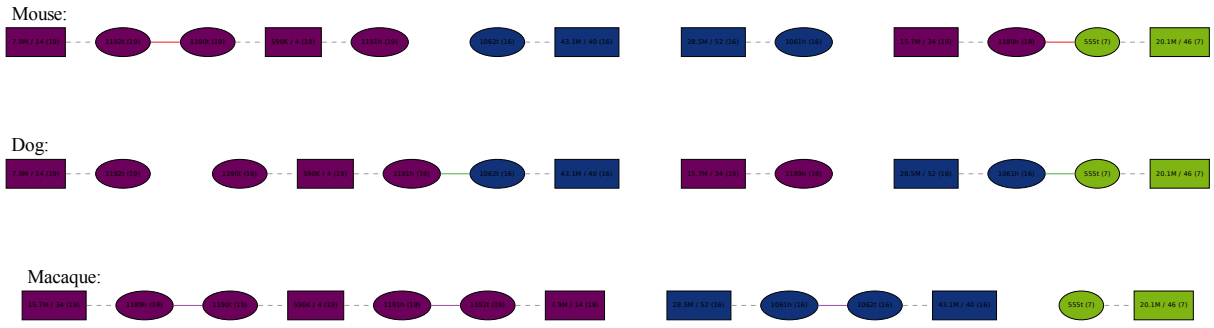

Figure S13: The regions of the Mouse, Dog, and Macaque genomes corresponding to the boxed component of the graph in Fig. S11, top panel.

The manual analysis of block 1300 (Jian Ma, personal communication) revealed that it indeed represents two independent micro-inversions in rat and macaque resulting in arrangements +1299, −1300, +1301 in rat and macaque as opposed to the arrangement +1299, +1300, +1301 in the human, chimpanzee, dog, and mouse genomes. It also found small aligned regions between blocks 1299 and 1300 (block 1299*a*) as well as between blocks 1300 and 1301 (block 1301*a*). While the regions 1299*a* and 1301*a* are too short to pass through any reasonable threshold on the synteny blocks size, they revealed the following arrangements: (+1299, −1300, −1299*a*, +1301*a*, +1301) in rat, (+1299, +1299*a*, −1301*a*, −1300, +1301) in macaque, and (+1299, +1299*a*, +1300, +1301*a*, +1301) in

human/chimpanzee/dog/mouse genomes.

We use similar arguments to process the remaining components of the breakpoint graph. For example, the simplest explanation for a component with two vertices 970t and 971t is a fission in dog that transforms the T-consistent split (mouse/rat/dog vs. human/chimpanzee/macaque) into an inconsistent split (mouse/rat vs. human/chimpanzee/macaque/dog). Note that the dog genome was subject to frequent fissions resulting in nearly doubling the number of chromosomes as compared to other five mammals. We remark that this processing at MGRA Stage 3 is viewed as less reliable and the resulting associations are not considered in the proposed ancestral reconstructions.

Supplement G The architecture of the ancestral X chromosome

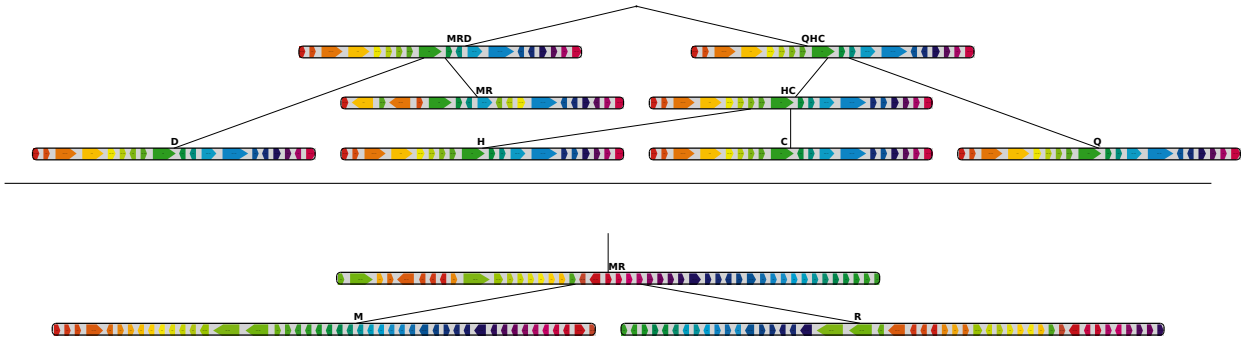

Figure S14: The architecture (up to micro-rearrangements) of 19 synteny blocks forming X chromosomes in the Dog, macaQue, Human, and Chimpanzee genomes as well as their common ancestral genomes (top panel). The Mouse and Rat X chromosomes (along with the MR ancestral X chromosome) are shown on a separate (bottom) panel since they display much higher fragmentation (46 synteny blocks).

Supplement H Rearrangements between the reconstructed ancestral genomes

Table S5 illustrates how the rearrangement distances between genomes at the leaves of the phylogenetic tree are being reduced while progressing through Stages 1 and 2 of MGRA.

	M	R	D	Q	H	C
M	0	438	436	392	395	406
R		0	739	689	696	707
D			0	283	284	292
Q				0	104	113
H					0	22
C						0

M	R	D	Q	H	C
0	37	90	95	95	97
	0	91	93	94	96
		0	43	40	39
			0	16	16
				0	7
					0

M	R	D	Q	H	C
0	4	3	9	5	7
	0	7	5	5	7
		0	9	5	5
			0	6	6
				0	4
					0

Table S5: The estimated pairwise genomic distances (based on the formula from Alekseyev (2008)) between the genomes before (*left table*) and after MGRA Stage 1 (*center table*) as well as after MGRA Stage 2 (*right table*).

Table S6 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 $\vec{T}$-consistent transformation found by MGRA is close to the most parsimonious.

It is not surprising that some of the 2-break distances in Tables S6 are smaller than the corresponding genomic distances. The explanation for this phenomenon is that 2-breaks have an "advantage" over the standard rearrangements in the presence of complex components (such as hurdles (Hannenhalli and Pevzner, 1999)) in linear genomes. Such components can be typically resolved with

	M	R	D	Q	H	C	MR	MRD	QHC	HC
M	0	499	450	407	409	421	81:81	285	354	404
R		0	800	749	753	765	436:384	637	701	748
D			0	291	295	304	380	173:170	241	290
Q				0	110	117	334	130	54:53	107
H					0	23	336	133	59	7:6
C						0	347	145	72	18:17
MR							0	212:213	281	331
MRD								0	76:76	128
QHC									0	54:53
HC										0

	M	R	O	D	Q	H	C	MR	MRO	MROD	QHC	HC
M	0	442	822	412	370	378	382	74:77	276	279	334	371
R		0	1107	713	665	674	676	382:341	579	581	631	665
O			0	714	675	682	682	761	587:586	591	637	673
D				0	245	253	256	351	156	150:148	210	246
Q					0	94	95	305	107	102	44:44	85
H						0	21	315	118	112	50	9:9
C							0	317	119	114	53	12:12
MR								0	212:215	215	269	306
MRO									0	7:9	69	109
MROD										0	63:65	103
QHC											0	41:40
HC												0

Table S6: The pairwise rearrangement distances between the Human, Mouse, Rat, Dog, Chimpanzee, and macaQue genomes (top table) as well as the Opossum genome (bottom table) and their ancestral genomes *MR*, *MRD*, *MRO*, *MROD*, *HC*, and *QHC* reconstructed by MGRA. Each cell contains a number x or a pair of numbers $x : y$ where x is the genomic distance (computed by GRIMM (Tesler, 2002b)) and y is the number of 2-breaks between the genomes in the $\vec{T}$-consistent transformation constructed by MGRA. The distances corresponding to the branches of the phylogenetic tree T are grayed.

smaller number of 2-breaks via temporary creation of circular chromosomes.

Table S7 shows the breakdown of intrachromosomal and interchromosomal rearrangements (generated by MGRA) between different branches of the phylogenetic tree. While the number of intrachromosomal 2-breaks is roughly twice larger than the the number of interchromosomal 2-breaks (on average), some branches (*D* + *MRQHC* and *MR* + *DQHC*) reveal an elevated number of interchromosomal rearrangements (approaching and even exceeding the number of intrachromosomal rearrangements).

In presence of the Opossum genome, MGRA assigns the following 2-breaks to the contested *MRO* + *DQHC* branch: three fissions (1547h with 710h, 1420h with 627h, and 1377h with 748t at MGRA Stage 2), five fusions (1548t with 1547h, 1668t with 1667h, 1531h with 1377h, 748t with 747h, and 957t with 924h at MGRA Stage 2), and one translocation (on edges (952t, 953t) and (951t, 952h) at MGRA Stage 3).

Supplement I Detailed comparison of MGRA and inferCARs reconstructions

To further compare these MGRA and inferCARs reconstructions we constructed the breakpoint graph $G(MRD_{\mathsf{MGRA}}, MRD_{\mathsf{CARs}})$ (Fig. S15). The non-trivial components of $G(MRD_{\mathsf{MGRA}}, MRD_{\mathsf{CARs}})$ are formed by 2 cycles (on 4 vertices each) and 19 paths (on 55 vertices), out of which 8 paths consist of single purple edges and represent various CARs (constructed by inferCARs) that were connected by MGRA. 5 out of 19 paths are purple-purple paths (representing CARs that are connected in MRD_{MGRA} and disconnected in MRD_{CARs}), 11 are cyan-cyan paths (representing CARs that are connected in MRD_{CARs} and disconnected in MRD_{MGRA}), and 3 are a purple-cyan path. One out of

Branch	# intrachromosomal 2-break	# interchromosomal 2-breaks	Total
M+RDQHC	53	28	81
R+MDQHC	294	90	384
D+MRQHC	92	78	170
Q+MRDHC	32	21	53
H+MRDQC	5	1	6
C+MRDQH	16	1	17
MR+DQHC	80	133	213
HC+MRDQ	40	13	53
MRD+QHC	55	21	76
Total	667	386	1053

Table S7: The statistics of the 2-break scenario reconstructed by MGRA between the Mouse, Rat, Dog, macaQue, Chimpanzee, and Human genomes. For each branch of the phylogenetic tree, it gives the number of intrachromosomal 2-breaks (reversals and intrachromosomal translocations) and the number of interchromosomal 2-breaks (fissions/fusions and interchromosomal translocations).

the two cycles as well as some paths in Fig. S15 represent different interpretations of micro-inversions (formed by synteny blocks that are located closely to each other in some genomes) by MGRA and inferCARs algorithms and do not affect the large-scale view of ancestral architectures.

Supplement J How stable are the ancestral reconstructions?

In order to test the stability of MGRA reconstructions with changing resolution (minimum size of the synteny blocks), we removed short synteny blocks from the original set of 1357 blocks for six genomes and compared the resulting reconstructions. While removing some synteny blocks unavoidably affects the ancestral reconstructions (e.g., some adjacencies may become "invisible"), it is important to verify that the number of changes is relatively small.

Note that removal of a synteny block may "enlarge" others by merging them (two blocks are merged as soon as they are adjacent in all 6 genomes). Therefore, we performed short block removal as iterative procedure that removes the shortest block (w.r.t. the human genome) and possibly merges all pairs of consistently adjacent blocks into longer blocks. The procedure stops when the length of the shortest blocks exceeds the specified threshold. We further reconstructed the Boreoeutherian ancestors using the genomes with all short blocks removed.

Removing synteny blocks may result in either loosing some ancestral adjacencies (e.g., breaking a single CAR into two CARs) or in introducing new adjacencies (as compared to the original ancestral reconstruction). To compare reconstructions on different sets of blocks we selected the set of blocks share between two reconstructions and computed the number of "missing" and "extra" adjacencies between two ancestral reconstructions. The results for minimal blocks thresholds of $100K$, $250K$, and $500K$ are shown in Tab. S8 that illustrates that MGRA reconstructions are rather stable. For example, removing all 168 blocks shorter than 100 Kb results (12% of all blocks) in reconstructions that retain 99.5% adjacencies compared to each other. Increasing the threshold to 250K results in removing 34% of all blocks but retains 98.5% of adjacencies.

MinBlockLength	#Blocks Left	#Adjacencies	Extra Adjacencies	Missing Adjacencies
$100K$	1189	1161	5	6
$250K$	903	871	11	17
$500K$	711	676	15	24

Table S8: Comparison of MGRA reconstructions on the original 1357 synteny blocks (for 6 genomes) with MGRA reconstruction on the reduced set of synteny blocks (blocks shorter than MinBlockLength threshold removed).

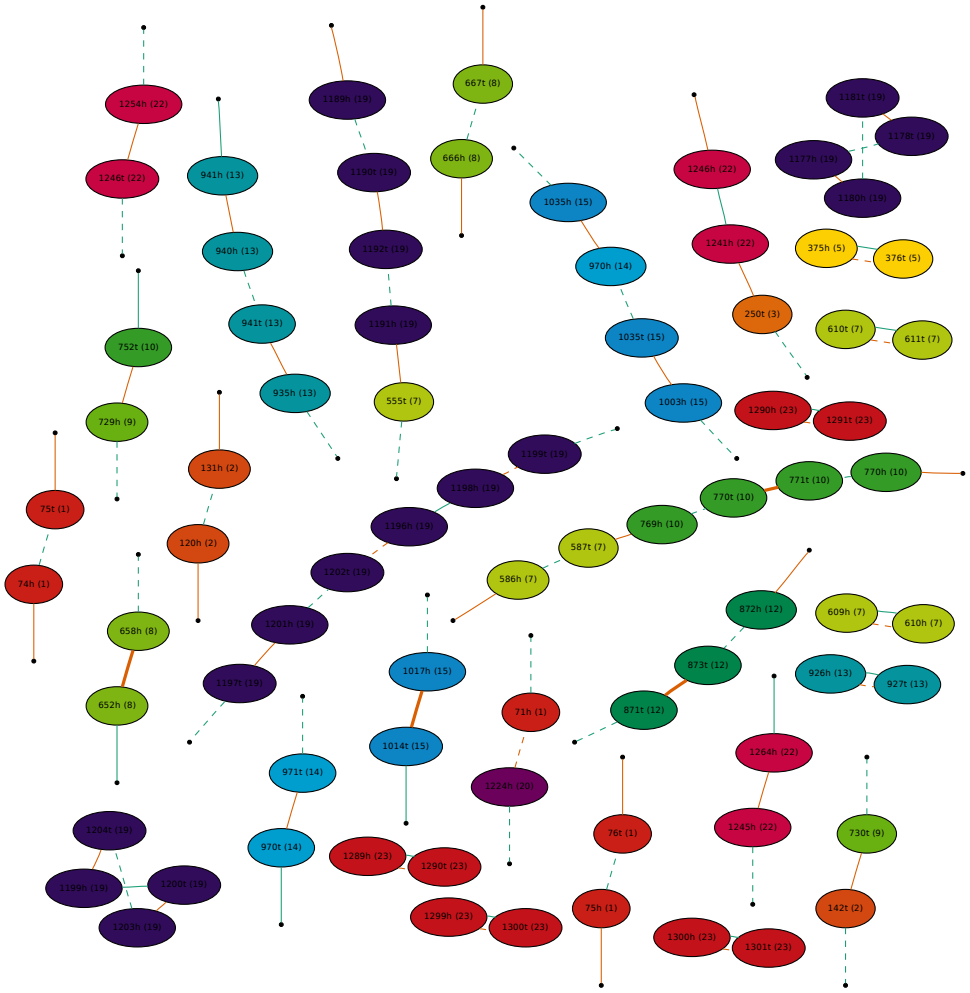

Figure S15: The breakpoint graph of the genomes MRD_{CARs} (*cyan*) and MRD_{MGRA} (*purple*). Bold purple edges represent reliable adjacencies obtained by MGRA Stage 1, while dashed purple edges (shown even if they are parts of complete multi-edges) represent adjacencies (between vertices incident to a split in $M/R/D$ colors in Fig. 7, bottom panel) viewed as less reliable. Dashed cyan and orange edges represent ambiguous joins made by inferCARs.

Supplement K CytoAncestor software

To bridge the gap between the cytogenetics and the rearrangement-based approaches we implemented CytoAncestor software, which follows the logic of the cytogenetics approach described in Kemkemer et al. (2006). The tests of CytoAncestor revealed that the cytogenetics approach does not scale well with increase in the number of synteny blocks. In particular, on Ma et al. (2006a) data CytoAncestor produces a Boreoeutherian ancestor that does not agree with the widely accepted cytogenetics reconstruction (Supplement K).[3] MGRA Stage 1, in contrast to CytoAncestor, produces a reconstruction that is largely consistent with the current view of the Boreoeutherian ancestor.

Kemkemer et al. (2006) recently applied the cytogenetics approach to *E-painting* data using semi-manual data analysis. We implemented their algorithm and applied it to the *Human*, *Mouse*, and

[3]The results improve when one limits attention to very large synteny blocks (e.g., larger than 3 Mb) indicating that further studies are needed to extend the cytogenetics approach to high-resolution data.

Dog data (1357 synteny blocks from Ma et al. (2006a)). The goal of our analysis is to investigate whether CytoAncestor scales well when one moves from the cytogenetics resolution (typically 100-200 synteny blocks) to genomic resolution (1000+ synteny blocks).

We briefly describe the cytogenetics approach for the case of 3 genomes P_1, P_2, P_3 with p_1, p_2, p_3 chromosomes (see Kemkemer et al. (2006)). We use a *synteny-triple* (t_1, t_2, t_3) to describe a synteny block located on chromosome t_1 in P_1, chromosome t_2 in P_2, and chromosome t_3 in P_3. Clearly, there exist at most $p_1 \cdot p_2 \cdot p_3$ distinct synteny-triples. In reality the number of synteny-triples is much smaller than this maximum, and for 1357 synteny blocks in the *Human*, *Mouse*, and *Dog* genomes we have only 204 synteny-triples. The synteny-triples represent vertices in the *synteny graph* that are further connected by edges as described in Kemkemer et al. (2006) (Fig. S16). The connected components in the resulting graph represent the ancestral chromosomes and reveal the synteny associations. For example, the unicolored connected components representing human chromosomes 6, 9, 11, 17, 18, 20, and X all correspond to single chromosomes in the ancestor and are consistent with the now favored cytogenetics reconstruction. However, all other connected components disagree with the existing reconstruction (Froenicke et al., 2006). In particular, the giant multicolored component formed by human chromosomes 1 + 5 + 10 + 16 + 4 + 7 + 8 + 13 was never reported in previous cytogenetics studies and is likely to reflect the limitations of the cytogenetics approach when applied to a small number of species with many synteny blocks. We remark that with the same dataset, the rearrangement-based approaches inferCARs and MGRA produce ancestors that are largely consistent with the now favored cytogenetics reconstruction.

In an attempt to alleviate these shortcomings of CytoAncestor we limited our attention to long synteny blocks by excluding synteny-triples that cover less 300 Kb, 1 Mb, and 3 Mb from the dataset in Ma et al. (2006a) (Fig. S16). While the size of the giant component reduces, even for synteny-triples of size 3 Mb and longer (typical cytogenetics resolution), most of the resulting synteny associations remain unrealistic.

Supplement L Benchmarking MGRA on simulated data

We benchamarked MGRA on various simulated datasets with a fixed phylogenetic tree shown in Fig. 5 (for illustration purposes, we refer to the leaves of the tree as M, R, D, Q, H, and C). In addition, we evaluated MGRA's ability to reconstruct an unknown tree in case of short internal branches.

In the first "constant branch length" simulation, we fixed the number of rearrangements on each branch to the same number varying from 25 to 250 and generated the leaf genomes by performing rearrangements on a fixed MRD genome consisting of 20 chromosomes with 75 synteny blocks each. The total number of synteny blocks in this simulation is close to the number of synteny blocks for six mammalian genomes studied in Ma et al. (2006a). The leaf genomes were generated from the MRD genome by applying random 2-breaks (preserving linearity) along the branches of the tree. We further applied MGRA to the leaf genomes and compared the reconstructed MRD ancestral genome with the simulated one, counting the number of missing and incorrectly reconstructed adjacencies (Table S9).

Below we focus on the simulation with branch lenght 125 which results in a rather difficult ancestral reconstruction problem with high breakpoint re-use rate[4] of 1.5. We remark that 125 rearrangements on each branch imply $5 \cdot 125 = 625$ rearrangements between the simulated H ("human") and M ("mouse") nodes in Fig. 5, a rather large number of rearrangements (as compared to the number of rearrangements between the real human and mouse genomes). Note that 625 rearrangements break the lion share of adjacencies between 1500 synteny blocks in the simulated genomes, making the ancestral reconstruction difficult. Nevertheless, MGRA produced an error-

[4]Similarly to Pevzner and Tesler (2003b), we estimate the breakpoint re-use rate as two times the 2-break distance divided by the number of breakpoints between two genomes.

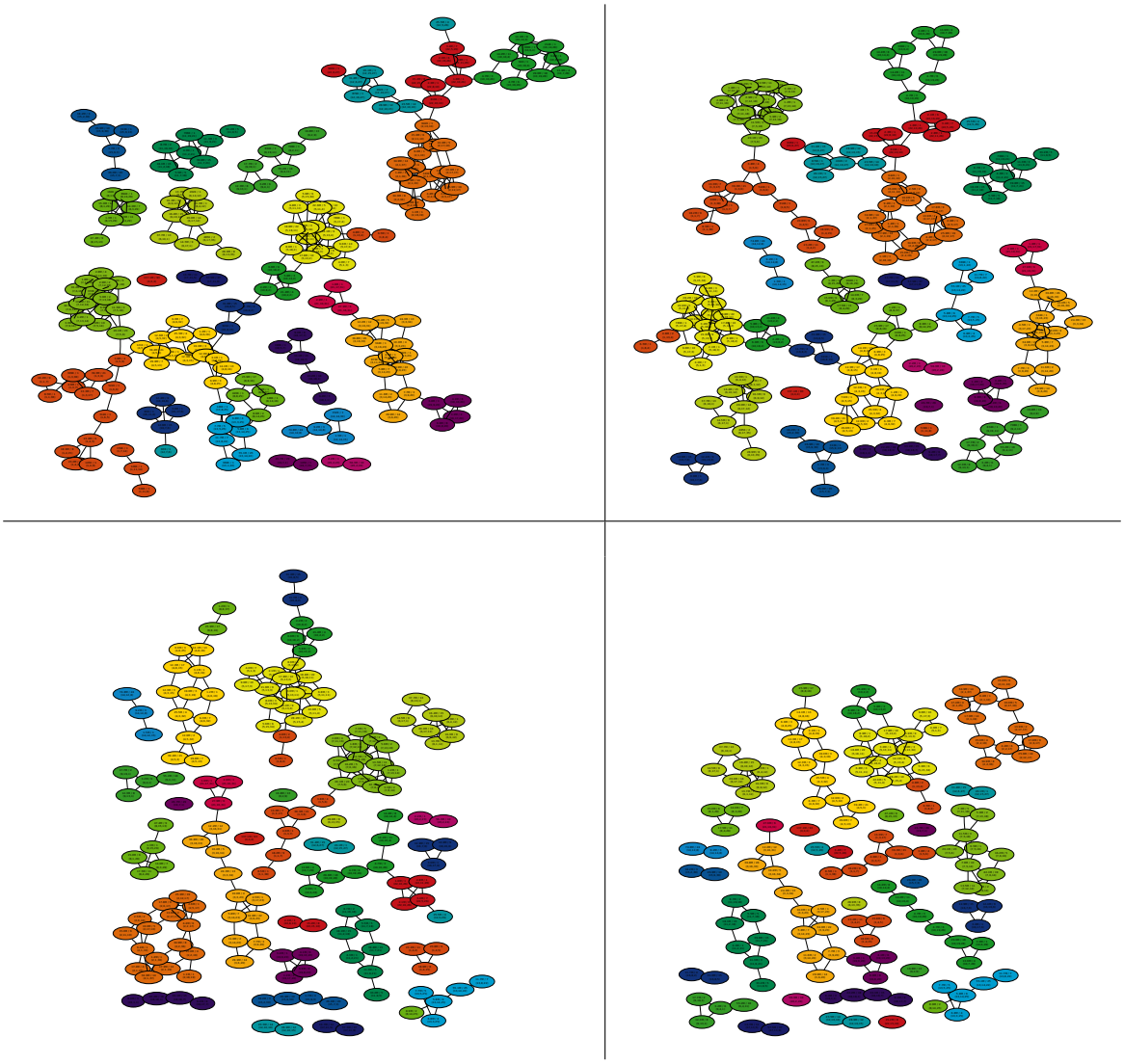

Figure S16: Chromosomal associations between the *Human*, *Mouse*, and *Dog* genomes on 1357 synteny blocks revealed by the cytogenetics approach for all synteny-triples (top left), and synteny-triples longer than 300 Kb (top right), 1 Mb (bottom left), and 3 Mb (bottom right). Each vertex corresponds to a synteny-triple (t_1, t_2, t_3) located on chromosomes t_1 in *Human*, t_2 in *Mouse*, and t_3 in *Dog*. Each vertex is also labeled with the total size and number of the synteny blocks corresponding to the synteny-triple (t_1, t_2, t_3). For example, a blue vertex labeled as $\frac{74.8M\ /\ 24}{(14,12,8)}$ describes 24 synteny blocks of the total size 74.8 Mb described by the synteny-triple $(14, 12, 8)$.

free reconstruction of the ancestral MRD genome in this case (with only 2 missing adjacencies). As expected, MGRA becomes less accurate and more fragmented when the genomes become extremely scrambled (e.g., 4% of adjacencies are incorrect and 9% of adjacencies are lost for the branch length 250).

Table S9 also shows the results of inferCARs reconstructions and illustrates that MGRA generates more accurate ancestral reconstructions for all choices of parameters. In particular, for simulation with the branch length 125, about 2% of adjacencies reconstructed by inferCARs are incorrect. While it is a relatively small proportion of incorrect adjacencies (for a rather difficult ancestral reconstruction problem with high breakpoint re-use), MGRA produced an error-free and less fragmented reconstruction in this case.

To evaluate effect of more complex rearrangements (e.g., transpositions) on MGRA performance,

Branch length	H-M breakpoint re-use	Conserved adjacencies	Reconstructed adjacencies					
			Correct		Missing		Incorrect	
			MGRA	inferCARs	MGRA	inferCARs	MGRA	inferCARs
25	1.14	1093	1480	1472	0	8	0	6
50	1.18	806	1480	1464	0	16	0	9
75	1.32	608	1479	1461	1	19	0	9
100	1.38	432	1478	1446	2	34	0	15
125	1.50	317	1478	1425	2	55	0	28
150	1.58	233	1463	1412	17	68	8	39
175	1.70	175	1460	1373	20	107	14	56
200	1.78	130	1448	1342	32	138	20	86
225	1.83	113	1429	1305	51	175	39	118
250	1.89	81	1343	1255	137	225	62	162

Table S9: Reconstruction of the *MRD* ancestor of six simulated genomes with the phylogenetic tree shown in Fig. 5 where all branches have the same length. The genomes were generated from a fixed MRD genome on 20 chromosomes and 1500 synteny blocks (with 1480 adjacencies), applying random 2-breaks (preserving linearity) along the branches of the tree. The second column refers to the breakpoint re-use rate between simulated genomes corresponding to *H* ("human") and *M* ("mouse") nodes. Some of the pairs of adjacent synteny blocks in the *MRD* genome remain adjacent in all six generated leaf genomes. The number of such conserved adjacencies is shown in the third column (hence, the effective number of the synteny blocks in each simulation is 1480 minus the number of conserved adjacencies). The columns from forth to ninth give the statistics of reconstructed adjacencies by classifying them into correct, missing, and incorrect (for both MGRA and inferCARs).

we further added 3-breaks (happening with the probability 0.1 at every step) to the set of simulated rearrangements. Table S10 illustrates that adding 3-breaks has only minor effect on MGRA performance. Again, MGRA improves on inferCARs in the case when both 2-breaks and 3-breaks are included in the simulations.

Branch length	H-M breakpoint re-use	Conserved adjacencies	Reconstructed adjacencies					
			Correct		Missing		Incorrect	
			MGRA	inferCARs	MGRA	inferCARs	MGRA	inferCARs
25	1.20	1092	1478	1467	2	13	0	11
50	1.23	804	1480	1466	0	14	0	9
75	1.36	587	1480	1449	0	31	0	17
100	1.42	429	1478	1430	2	50	0	29
125	1.49	315	1474	1417	6	63	6	33
150	1.63	233	1453	1387	27	93	15	53
175	1.75	186	1461	1375	19	105	11	58
200	1.80	116	1462	1350	18	130	14	77
225	1.85	105	1382	1304	98	176	68	121
250	1.90	61	1376	1256	104	224	60	148

Table S10: Simulations similar to those in Table S9, where rearrangements in addition to 2-breaks include 3-breaks occuring with the probability 0.1.

In the second "variable branch length" simulation, we selected the number of rearrangements on each branch of the tree according to the values from Table S6(top) to better reflect various rates of rearrangements in different mammalian lineages. To model breakpoint re-use, we varied the number of initial synteny blocks from 1000 to 2000 (the smaller is the number of blocks, the large is the breakpoint re-use). The benchmarking results for this simulation are shown in Table S11 (for MRD genome).

Table S11 illustrates that for 1400 synteny blocks (roughly the number of synteny blocks identified in Ma et al. (2006a)), MGRA is error-free but rather fragmentary with 55 missing adjacencies. We

Number of blocks	H-M breakpoint re-use	Conserved adjacencies	Reconstructed adjacencies					
			Correct		Missing		Incorrect	
			MGRA	inferCARs	MGRA	inferCARs	MGRA	inferCARs
1000	1.52	99	893	909	87	71	21	55
1100	1.50	123	1021	1015	59	65	17	36
1200	1.47	188	1120	1114	60	66	9	36
1300	1.33	242	1228	1227	52	53	2	34
1400	1.32	282	1325	1338	55	42	0	26
1500	1.32	346	1446	1441	24	39	0	25
1600	1.32	381	1553	1532	27	48	2	29
1700	1.32	451	1652	1629	28	51	0	34
1800	1.27	532	1740	1739	40	41	0	27
1900	1.28	592	1862	1842	18	38	0	20
2000	1.25	622	1954	1938	26	42	0	32

Table S11: Reconstruction of the MRD ancestor of six simulated genomes with the phylogenetic tree shown in Fig. 5, where the length of branches is the same as in Table S6(top) and the number of synteny blocks varies from 1000 to 2000. Compare to Table S9.

remark that such significant fragmentation was not observed when MGRA was applied to real data. A possible explanation is that many rearrangements in real scenarios are actually micro-rearrangements that are typically easier to analyze (unless they lead to breakpoint re-use). Our simulation does not model micro-rearrangements, thus making reconstruction of simulated genomes in this case somewhat more difficult than reconstruction of real genomes. We remark that while inferCARs generated slightly less fragmented reconstruction than MGRA in this case, it generated 26 incorrect adjacencies.

In the third simulation we investigated whether MGRA is capable of revealing the short branches of the phylogenetic tree in the "blind mode" when the tree is not known in advance. The goal is to evaluate whether the phylogenetic characters generated by MGRA (such as in Tables 1, S13, and 4) may be misleading in the case of very short branches. TSuch short branches typically incurred very few rearrangements that may be difficult to reconstruct due to a variety of factors (e.g., breakpoint re-use or long branch attraction).

We simulated H, M, D, and O genomes and preserved the rearrangement distances between the Human, Mouse, Dog, and Opossum genomes shown in Table S6(bottom). We considered a tree on 4 leaves and two internal nodes X_{HD} and X_{MO} with branch distance $d(H, X_{HD}) = 110$, $d(D, X_{HD}) = 140$, $d(M, X_{MO}) = 260$, $d(O, X_{MO}) = 560$ and the varying length of the internal edge between X_{HD} and X_{MO}) (from 0 to 50). We further simulated rearrangements on 5 branches of the resulting tree according to the specified rearrangement distances. We performed 4 simulations (for 1000, 1250, 1500, and 1750 synteny blocks) resulting in various breakpoint re-use rates. Table S12 illustrates that MGRA reveals the correct topology in nearly all cases with the exception of the cases when the length of the internal branch is close to zero.

Supplement M Paths in the breakpoint graph and the primate–rodent–carnivore split

We analyzed the paths in the breakpoint graph in Fig. S18(top) with the goal to find a path that may support or reject the primate–carnivore split. Since the branch corresponding to this split is relatively short ($\approx$ 7 million years), we do not expect to find many rearrangements supporting either the primate–carnivore, or the primate–rodent splits. Alternating MRO-DQHC paths would represent a strong supporting evidence for the primate-carnivore split, while alternating DO-MRDQHC paths would represent a strong supporting evidence for the primate-rodent split. Not surprisingly, neither the original breakpoint graph, nor the breakpoint graph after applying MGRA Stage 1 contains such paths, an indication that the branch corresponding to the split is indeed short. Fig. S17 enlarges a

MO + *DH* length	*MO* + *DH* multi-edges/paths	*MH* + *DO* multi-edges/paths	*MD* + *HO* multi-edges/paths
0	0 / 0	0 / 0	4 / 1
5	20 / 5	0 / 0	0 / 0
10	40 / 10	0 / 0	0 / 0
15	54 / 13	0 / 0	0 / 0
20	76 / 19	0 / 0	0 / 0
25	92 / 23	0 / 0	0 / 0
30	112 / 27	3 / 1	0 / 0
35	132 / 33	0 / 0	4 / 1
40	136 / 34	0 / 0	0 / 0
45	172 / 43	0 / 0	0 / 0
50	170 / 42	4 / 1	0 / 0
0	4 / 1	0 / 0	0 / 0
5	16 / 4	7 / 2	0 / 0
10	24 / 6	0 / 0	0 / 0
15	48 / 12	0 / 0	0 / 0
20	76 / 19	0 / 0	0 / 0
25	96 / 24	0 / 0	0 / 0
30	114 / 29	0 / 0	0 / 0
35	129 / 32	4 / 1	0 / 0
40	122 / 30	0 / 0	3 / 1
45	140 / 34	0 / 0	0 / 0
50	130 / 32	0 / 0	0 / 0
0	0 / 0	0 / 0	0 / 0
5	20 / 5	0 / 0	3 / 1
10	24 / 6	0 / 0	0 / 0
15	52 / 13	0 / 0	0 / 0
20	63 / 16	0 / 0	3 / 1
25	70 / 18	4 / 1	0 / 0
30	88 / 22	0 / 0	4 / 1
35	96 / 24	3 / 1	0 / 0
40	98 / 24	4 / 1	0 / 0
45	108 / 21	0 / 0	4 / 1
50	148 / 36	3 / 1	0 / 0
0	0 / 0	4 / 1	8 / 2
5	16 / 4	0 / 0	9 / 2
10	20 / 5	3 / 1	0 / 0
15	31 / 8	0 / 0	0 / 0
20	56 / 14	4 / 1	0 / 0
25	55 / 13	7 / 2	0 / 0
30	82 / 21	0 / 0	6 / 2
35	58 / 14	9 / 3	0 / 0
40	65 / 15	0 / 0	0 / 0
45	72 / 18	0 / 0	0 / 0
50	105 / 25	0 / 0	3 / 1

Table S12: The statistics of the breakpoint graph of the simulated M, H, D, and O genomes (using the $((H, D)(M, O))$ tree topology with 5 branches) on 1750 (first table), 1500 (second table), 1250 (third table), and 1000 (fourth table) synteny blocks. The tables represent the statistics after MGRA Stages 1-2 run on the confident leaf branches. The length of the branch separating M and O from H and D varied from 0 to 50. Compare to Tables 1(bottom), S13, and 4.

path of alternating O and MR edges in the breakpoint graph in Fig. S18(top) that groups opossums with rodents and represents the best supporting evidence for the primate–carnivore split (most vertices on this path represent chromosome endpoints in D, Q, H, and C genomes). We emphasize that while this path is hard to explain under the assumption of primate-rodent split, it does not

represent a "proof" of the primate-carnivore split since complex breakpoint re-uses combined with difficulties in finding the synteny blocks between distant mammals may skew the statistics of paths in the breakpoint graph.

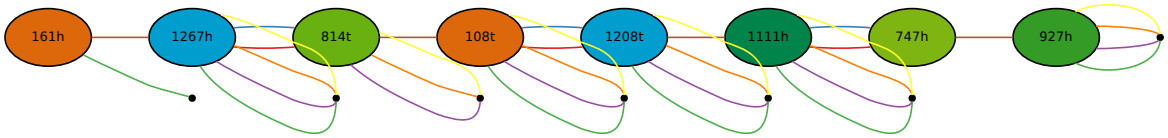

Figure S17: A path of the alternating *MR* and *O* edges from the breakpoint graph shown in Fig. S18(top). The vertices forming this path represent mostly chromosome endpoints in the other genomes.

Multicolors	Multi-edges	Simple vertices	Simple miltiedges	Simple paths+cycles
O + MRDQHC	561 + 738 = 1299	1120	559 + 434 = 993	125 + 92 = 217
R + MDQHCO	442 + 557 = 999	884	442 + 391 = 833	51 + 148 = 199
MR + DQHCO	226 + 177 = 403	288	104 + 126 = 230	44 + 24 = 68
D + MRQHCO	135 + 241 = 376	270	135 + 86 = 221	49 + 30 = 79
M + RDQHCO	138 + 64 = 202	128	39 + 64 = 103	25 + 16 = 41
QHC + MRDO	49 + 104 = 153	81	34 + 27 = 61	13 + 11 = 24
Q + MRDHCO	46 + 80 = 126	92	46 + 33 = 79	13 + 13 = 26
HC + MRDQO	38 + 66 = 104	70	32 + 23 = 55	12 + 8 = 20
C + MRDQHO	12 + 25 = 37	24	12 + 6 = 18	6 + 3 = 9
H + MRDQCO	9 + 18 = 27	18	9 + 6 = 15	3 + 2 = 5
MRO + DQHC	4 + 46 = 50	1	0 + 0 = 0	0 + 0 = 0
RO + MDQHC	31 + 2 = 33	1	0 + 0 = 0	0 + 0 = 0
DO + MRQHC	21 + 11 = 32	0	0 + 0 = 0	0 + 0 = 0
MRD + QHCO	15 + 7 = 22	1	0 + 0 = 0	0 + 0 = 0
HCO + MRDQ	5 + 2 = 7	0	0 + 0 = 0	0 + 0 = 0
DHC + MRQO	2 + 4 = 6	0	0 + 0 = 0	0 + 0 = 0
MO + RDQHC	0 + 5 = 5	0	0 + 0 = 0	0 + 0 = 0
DQO + MRHC	1 + 3 = 4	0	0 + 0 = 0	0 + 0 = 0
QC + MRDHO	0 + 3 = 3	0	0 + 0 = 0	0 + 0 = 0
QH + MRDCO	0 + 3 = 3	0	0 + 0 = 0	0 + 0 = 0
MRQ + DHCO	2 + 1 = 3	0	0 + 0 = 0	0 + 0 = 0
RDO + MQHC	1 + 1 = 2	0	0 + 0 = 0	0 + 0 = 0
DQC + MRHO	0 + 1 = 1	0	0 + 0 = 0	0 + 0 = 0
MRC + DQHO	0 + 1 = 1	0	0 + 0 = 0	0 + 0 = 0
DQH + MRCO	0 + 1 = 1	0	0 + 0 = 0	0 + 0 = 0
RQC + MDHO	0 + 1 = 1	0	0 + 0 = 0	0 + 0 = 0

Table S13: The statistics of the breakpoint graph for the Mouse, Rat, Dog, macaQue, Human, Chimpanzee, and Opossum genomes. For every pair of complementary multicolors, we show the number of multi-edges of these multicolors, 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 confident *T*-consistent multicolors are shown in bold.

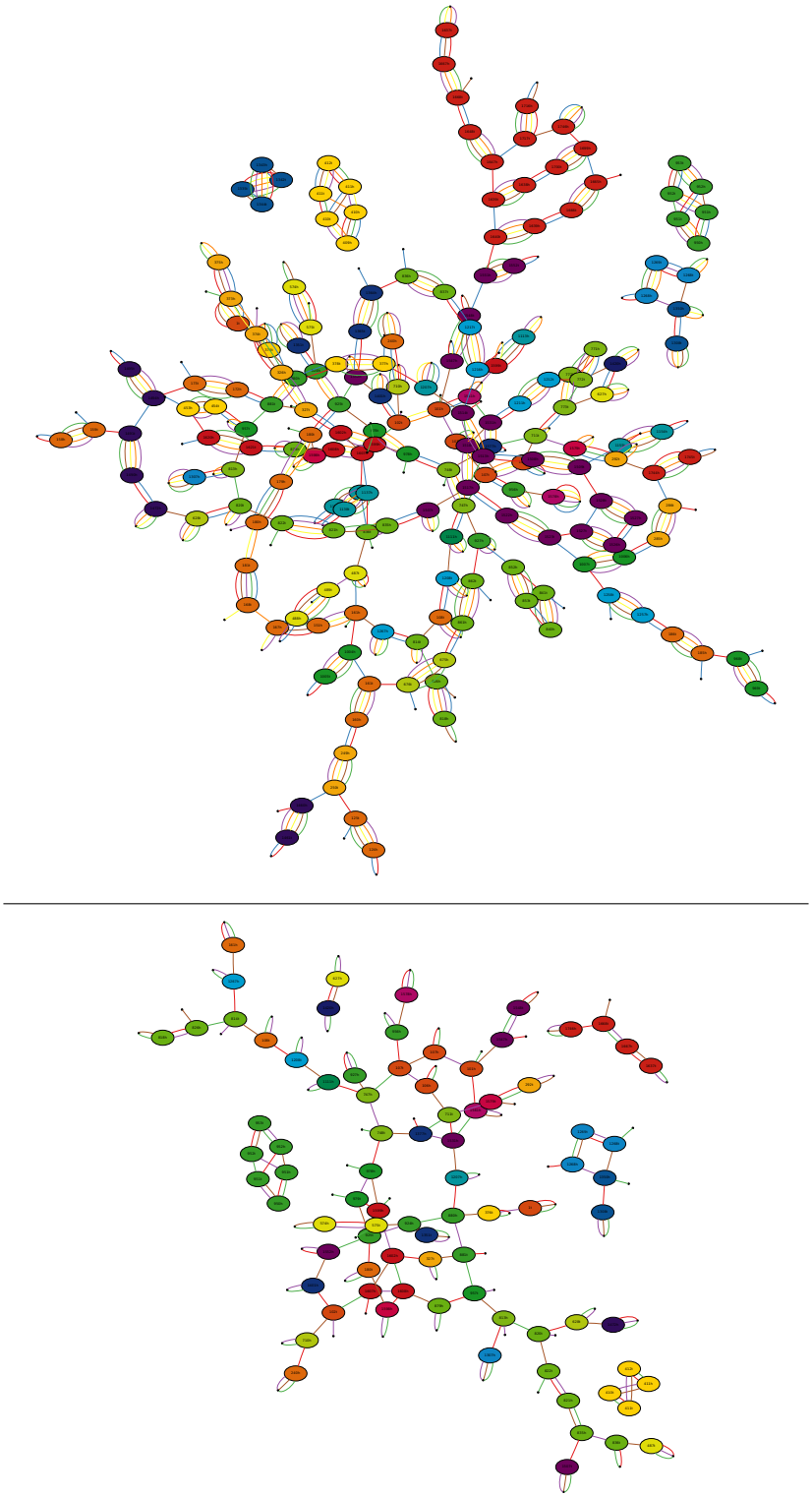

Figure S18: **Top panel:** The breakpoint graph of the Mouse (*red*), Rat (*blue*), Dog (*green*), macaQue (*violet*), Human (*orange*), Chimpanzee (*yellow*), and Opossum (*brown*) genomes after MGRA Stages 1-2 on the confident branches. Compare to Fig. 7(bottom). Restricting this graph to 4 genomes M,D,Q,O and running MGRA on this smaller graph using only 4 confident branches $M + DQO$, $D + MQO$, $Q + MDO$, and $O + MDQ$ results in the breakpoint graph $G(M, D, Q, O)$ shown at the bottom panel. **Bottom panel:** The breakpoint graph $G(M, D, Q, O)$ of the Mouse, Dog, macaQue, and Opossum genomes after MGRA Stages 1-2 on the confident branches.

Supplement N Additional Tables and Figures

CAR	Length
+1239	1.0 Mb
+72 +73 +74	11.3 Mb
+75	7.0 Mb
+76	0.5 Mb
+77 +78	0.3 Mb

Table S14: The list of short CARs (shorter than 15 Mb w.r.t. the Human genome) in MGRA reconstruction of the Boreoeutherian ancestral genome.

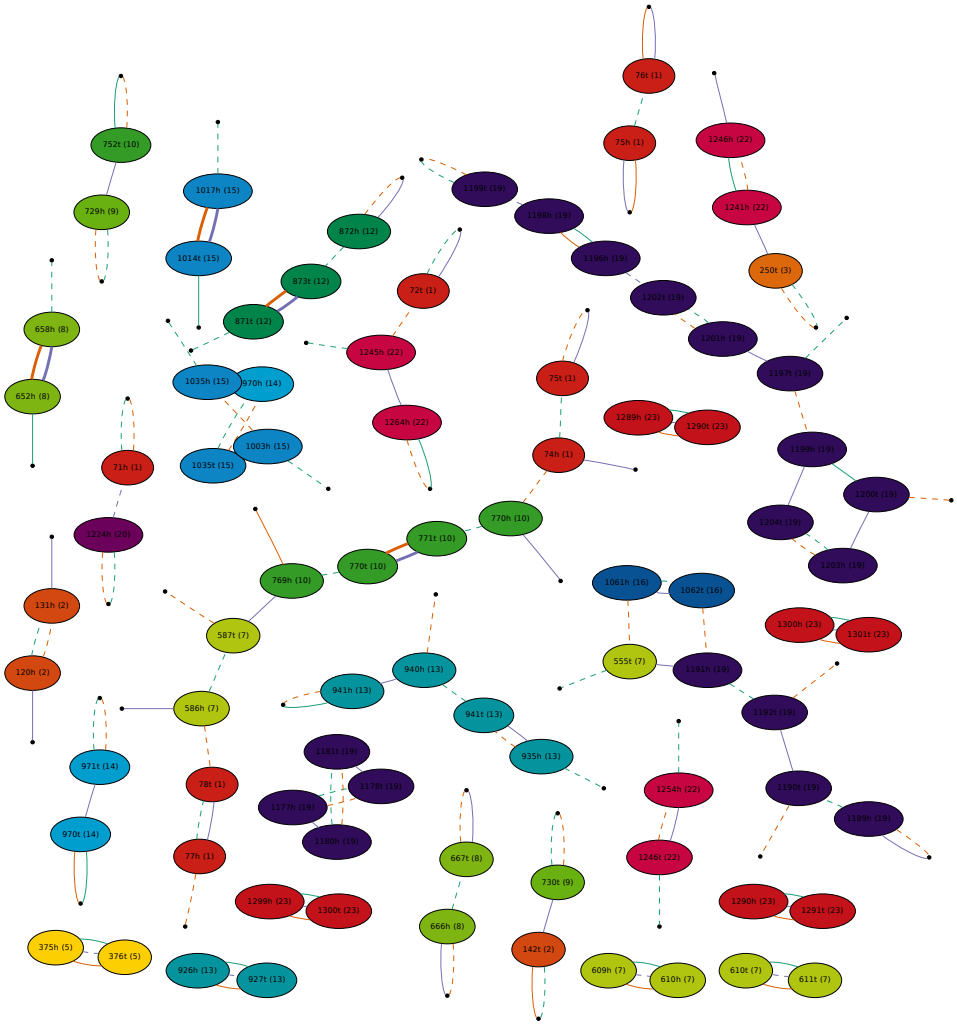

Figure S19: The breakpoint graph of the genomes MRD_{CARs} (*cyan*) and MRD'_{CARs} (*orange*) reconstructed by inferCARs as well as MRD_{MGRA} (*purple*) reconstructed by MGRA. Bold purple edges represent reliable adjacencies obtained by MGRA Stage 1, while dashed purple edges (shown even if they are parts of complete multi-edges) represent adjacencies (between vertices incident to a split in $M/R/D$ colors in Fig. 7, bottom panel) viewed as less reliable. Dashed cyan and orange edges represent ambiguous joins made by inferCARs.

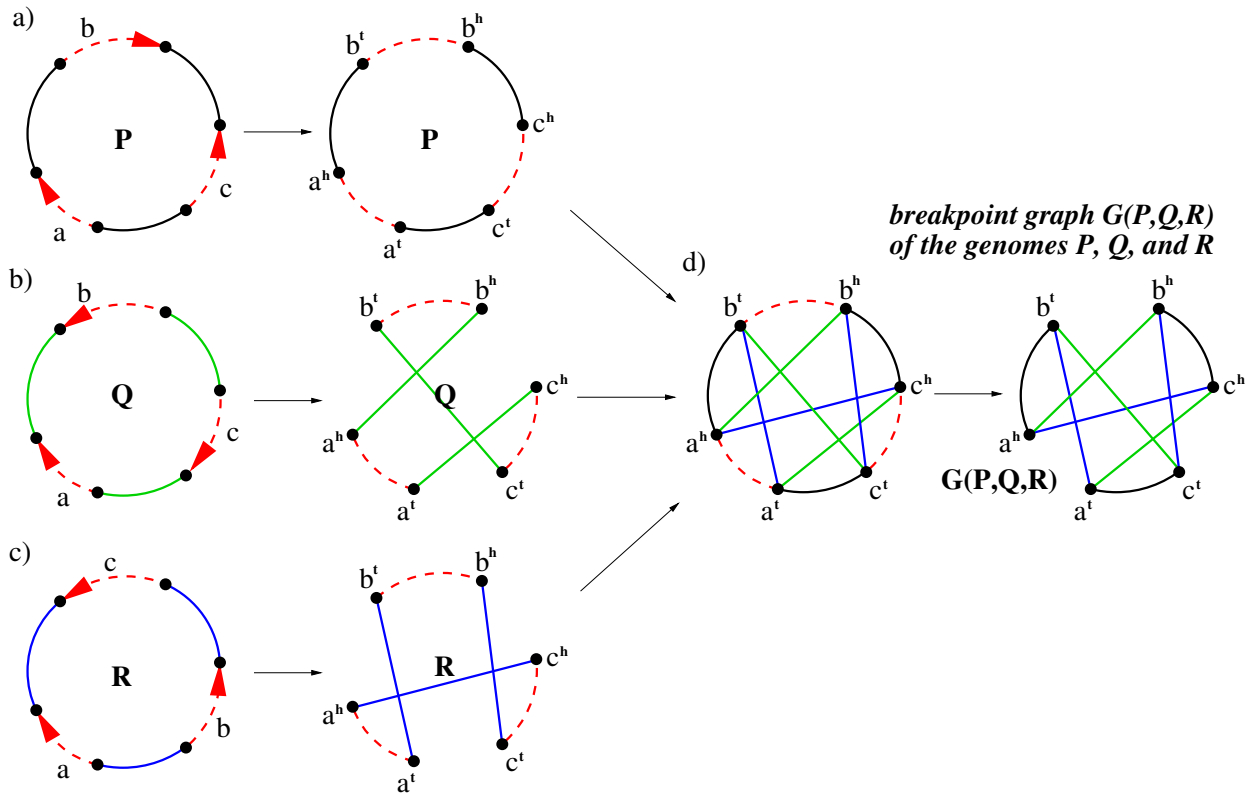

Figure S20: a) Unichromosomal genome $P = (+a+b-c)$ represented as a black-obverse cycle. b) Unichromosomal genome $Q = (+a-b+c)$ represented as a green-obverse cycle. b) Unichromosomal genome $R = (+a-c-b)$ represented as a blue-obverse cycle. d) The (multiple) breakpoint graph $G(P, Q, R)$ with and without obverse edges (compare to Fig. 1).

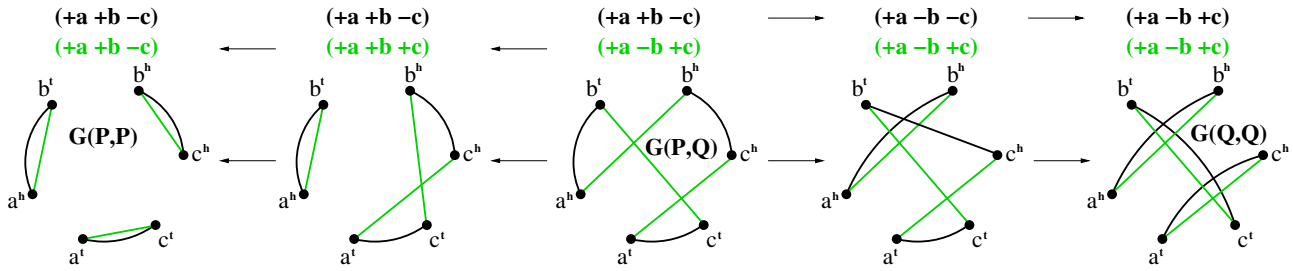

Figure S21: Transformation of the breakpoint graph $G(P, Q)$ of the "black" genome $P = (+a+b-c)$ and "green" genome $Q = (+a-b+c)$ (see Fig. 1) into the identity breakpoint graphs $G(P, P)$ (with "green" 2-breaks) and $G(Q, Q)$ (with "black" 2-breaks).

References

Alekseyev, M. A., 2008. Multi-Break Rearrangements and Breakpoint Re-uses: from Circular to Linear Genomes. *Journal of Computational Biology*, **15**(8):1117–1131.

Alekseyev, M. A. and Pevzner, P. A., 2008a. Multi-Break Rearrangements and Chromosomal Evolution. *Theoretical Computer Science*, **395**(2-3):193–202.

Alekseyev, M. A. and Pevzner, P. A., 2009b. Breakpoint Re-uses and Efficient Ancestral Genome Reconstruction. *Mimeo*, (in preparation).

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*, pages 84–93, New York, NY, USA. ACM.

Chaisson, M. J., Raphael, B. J., and Pevzner, P. A., 2006. Microinversions in mammalian evolution. *Proceedings of the National Academy of Sciences*, **103**(52):19824–19829.

Feuk, L., MacDonald, J. R., Tang, T., Carson, A. R., Li, M., Rao, G., Khaja, R., and Scherer, S. W., 2005. Discovery of Human Inversion Polymorphisms by Comparative Analysis of Human and Chimpanzee DNA Sequence Assemblies. *PLoS Genetics*, **1**:e56.

Froenicke, 2005. Origins of primate chromosomes – as delineated by zoo-fish and alignments of human and mouse draft genome sequences. *Cytogenetic and Genome Research*, **108**(1-3):122–138.

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**(3):306–310.

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

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 Research*, **14**(8):899–907.

Ma, J., Zhang, L., Suh, B. B., Raney, B. J., Burhans, R. C., Kent, J. W., Blanchette, M., Haussler, D., and Miller, W., 2006a. Reconstructing contiguous regions of an ancestral genome. *Genome Research*, **16**(12):1557–1565.

Pevzner, P. A. and Tesler, G., 2003b. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences*, **100**:7672–7677.

Robinson, T. J., Ruiz-Herrera, A., and Froenicke, L., 2006. Dissecting the mammalian genome – new insights into chromosomal evolution. *Trends in Genetics*, **22**(6):297–301.

Rocchi, M., Archidiacono, N., and Stanyon, R., 2006. Ancestral genomes reconstruction: An integrated, multi-disciplinary approach is needed. *Genome Res.*, **16**(12):1441–1444.

Tesler, G., 2002b. GRIMM: genome rearrangements web server . *Bioinformatics*, **18**(3):492–493.