# Microinversions in mammalian evolution

M. J. Chaisson, B. J. Raphael, and P. A. Pevzner

*PNAS* 2006;103;19824-19829; originally published online Dec 22, 2006;
doi:10.1073/pnas.0603984103

# Microinversions in mammalian evolution

**M. J. Chaisson*†, B. J. Raphael‡, and P. A. Pevzner§**

*Bioinformatics Program and §Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92093; and ‡Department of Computer Science, Center for Computational Molecular Biology, Brown University, Providence, RI 02906

Edited by Michael S. Waterman, University of Southern California, Los Angeles, CA, and approved October 10, 2006 (received for review May 15, 2006)

**We propose an approach for identifying microinversions across different species and show that microinversions provide a source of low-homoplasy evolutionary characters. These characters may be used as ''certificates'' to verify different branches in a phylogenetic tree, turning the challenging problem of phylogeny reconstruction into a relatively simple algorithmic problem. We estimate that there exist hundreds of thousands of microinversions in genomes of mammals from comparative sequencing projects, an untapped source of new phylogenetic characters.**

genome rearrangements | phylogenetics

Chromosomal inversions have been used as phylogenetic characters since Dobzhansky and Sturtevant in 1938. Recent comparisons of whole mammalian genomes have revealed a surprisingly large number of microinversions (1, 2). While the microinversions were first met with skepticism and were attributed to assembly errors and alignment artifacts, recent comparative study of human and chimpanzee genomes convincingly proved that microinversions are indeed widespread (3). We therefore decided to perform a fine-grained search for inversions across many mammals in the greater cystic fibrosis transmembrane conductance regulator (*CFTR*) region. This is a 1.8-megabase region on chromosome 7 in the human genome that encompasses the *CFTR* gene, and its many neighboring genes, that was sequenced for the ENCyclopedia Of DNA Elements (ENCODE) project (4, 5). We found that microinversions are frequent across all species and occur at roughly one microinversion per megabase per 66 million years of evolution. We show that microinversions have low homoplasy and thus provide ample characters for phylogenetic studies.

Our work follows in the steps of the pioneering work by Okada's group (6) and Lake's group (7) that demonstrated the power of repeat-based and deletion-based characters to resolve difficult phylogeny problems that the traditional point mutation analysis failed to resolve. The repeat-based and deletion-based approaches, although very successful, have some drawbacks as reviewed in ref. 8. However, Bashir *et al.* (9) and Kriegs *et al.* (10) recently demonstrated that many repeat-based characters may be extracted from genomic sequences to alleviate these drawbacks and to resolve some existing controversies. Our work reveals a source of low-homoplasy phylogenetic characters that complement these previous studies in two respects. First, microinversion homoplasy (if any) may be detected, and such characters can be simply deleted from further consideration without affecting the tree reconstruction algorithm. Second, microinversions may be identified as long as there is a detectable sequence similarity thus not necessarily limiting the comparison to close species as in the case of repeats and deletions. Indeed, Bourque *et al.* (11) documented many microinversions between human and chicken genomes, whereas Fischer *et al.* (12) found many microinversions between yeast genomes, which are molecularly as diverse as the genomes of the entire phylum of chordates.

While microinversions represent powerful evolutionary characters, their detection is far from simple. A naive approach is to detect reverse-strand local alignments between orthologous sequences. However, reverse-strand local alignments may also be caused by palindromes and inverted repeats (Fig. 1), ubiquitous genomic features that do not reflect any variations in the genomic architecture between two genomes, i.e., they may be detected within a single genome without a need to align to another genome. Reverse-strand alignments may also be detected in inverted transpositions (Fig. 1) and more complex interleaving rearrangement events. The computational challenge of distinguishing between microinversions and other genomic features is not widely appreciated, leading to an implicit assumption that whole-genome reverse-strand alignments retained in a net by the University of California Santa Cruz chaining and netting algorithms (2) provide a universal solution to the characterization of microrearrangements. Chaining and netting combine optimal ordered sequences of pairwise alignments to create a genome-scale alignment that allows for gaps and inversions [see supporting information (SI) *Appendix A* and Figs. 7 and 8 for additional details]. We developed a method, InvChecker, to find inversions in the *CFTR* region, and we applied it to the human and chimpanzee genomes to show that ≈80% of the 1,576 putative microinversions recently found (3) are repeat-induced artifacts. At the same time we uncovered 167 human–chimpanzee microinversions missed in ref. 3. These findings reveal some limitations of chaining and netting (2) as a microinversion detection tool in ref. 3. This comment is not a criticism of this method but rather is an indication that accurate validation, parameter setting, and postprocessing are necessary to extract microinversions from netted alignments. The chaining and netting algorithms (2) were carefully designed as a compromise between providing a simple and intuitive representation of rearrangements on one hand, and reflecting all complexities of the rearrangement process on the other hand. This representation, although extremely useful, does not attempt to model complex rearrangements (e.g., overlapping inversions) in full generality. We further illustrate the use of microinversions in evolutionary studies by reconstructing the phylogeny of 15 mammalian species by using sequences from the ENCODE project (4) and the phylogeny of 38 species partially sequenced as part of the National Institutes of Health Intramural Sequencing Center (NISC) Comparative Vertebrate Sequencing Program.

## Results

**Identification of Microinversions.** We searched for microinversions in 15 genomes with the nearly finished greater *CFTR* region:¶ human, chimpanzee, baboon, macaque, marmoset, galago, rabbit, mouse, rat, cow, dog, platypus, opossum (*Monodelphis domestica*), and hedgehog (5). These sequences [May 2005 ENCODE release (13)] average 1.84 megabases in length. Sequences composed of multiple contigs were ordered and oriented according to alignments

Author contributions: M.J.C., B.J.R., and P.A.P. designed research; M.J.C. and P.A.P. performed research; M.J.C., B.J.R., and P.A.P. analyzed data; and M.J.C., B.J.R., and P.A.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS direct submission.

Abbreviations: ENCODE, Encyclopedia of DNA Elements; MGR, Multiple Genome Rearrangements.

†To whom correspondence should be addressed. E-mail: mchaisso@bioinf.ucsd.edu.

¶We limited our analysis to genomes with no more than 200,000 unfinished base pairs.

This article contains supporting information online at www.pnas.org/cgi/content/full/0603984103/DC1.

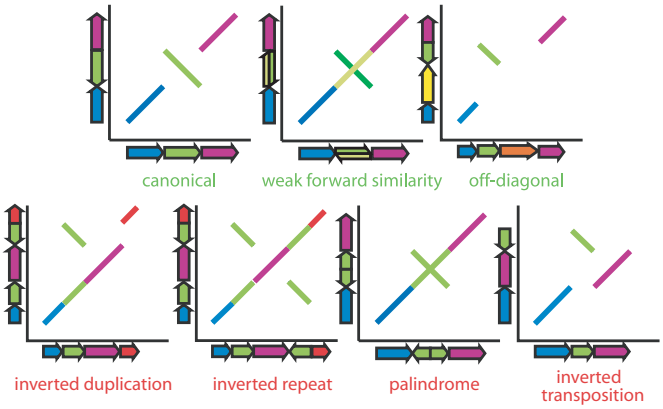


**Fig. 1.** Diagrammatic dot-plot of inversions (*Upper*) and genomic structures that are often misclassified as inversions (*Lower*). *Upper* shows alignments that are retained as inversions: canonical inversions, inversions with spurious forward similarities, and off-diagonal inversions that have been shifted by insertions/deletions or have highly diverged segments that elude similarity search. *Lower* represents inverted duplications, inverted repeats, palindromes, and inverted transpositions.

with the human sequence. Each sequence was repeat masked with Repeat Masker using both the RepBase and species-specific RepeatScout libraries (14, 15).

Although many putative inversions are detected by using chaining and netting, simple analysis of the netted files and reciprocal best hits as in ref. 3 has serious shortcomings. First, some more divergent inversions are missed in netted alignments because netted alignments have a tendency to favor the direct alignment even if the reverse alignment is more statistically significant, albeit miniscule. For example, two microinversions were found in the alignment of human and chimpanzee in the greater *CFTR* region that both remained undetected in whole-genome nets (one of them is shown in Fig. 2). Second, BLASTZ-based alignments that are processed by chaining and netting miss some "ancient" inverted segments. For example, while a segment in human may have a detectable similarity in mouse but not in rat, aligning the mouse segment against the rat genome may lead to detecting similarity between the human and rat sequences. Third, genomes typically have many palindromes and inverted repeats that may mimic microinversions, thus misleading the netting and chaining algorithms. An example of this is shown in Fig. 3, which represents an inverted repeat that is misclassified as a microinversion.

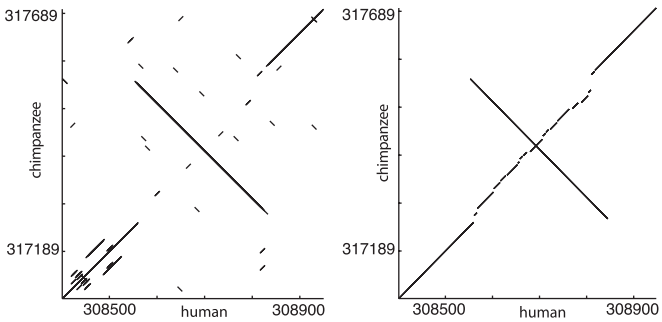


**Fig. 2.** The dot plot (*Left*) and the corresponding BLASTZ alignments (*Right*) of a 290-bp microinversion that is not detected by using netted alignments of human and chimpanzee (false negative). *Left* clearly shows the presence of a microinversion. However, the gap spanning the forward alignment is small enough to be closed by chain-and-extend alignment, although the score of the direct strand alignment is not statistically significant, particularly when compared with the much higher score of the reverse strand alignment.

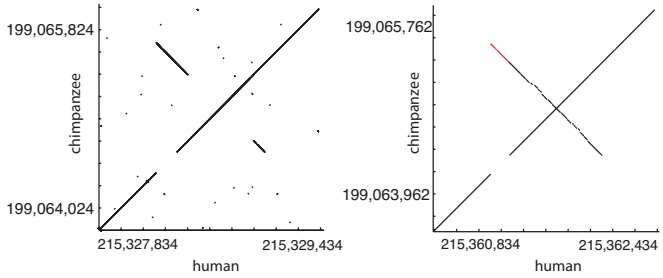


**Fig. 3.** The dot plot (*Left*) and the corresponding BLASTZ alignments (*Right*) of an inverted repeat that is misclassified as a putative microinversion by chaining and netting (false positive). The red segment in *Right* corresponds to an inversion taken from the list of inversions presented by ref. 13. The undetected forward alignment, directly below the red segment, is caused by a short run of masked nucleotides in the chimpanzee assembly. The same effect may be caused by some regions in inverted repeats/palindromes being more diverged than others, thus eluding alignment algorithms.

Fig. 1 shows seven genomic dot-plots, only the first three of which represent microinversions. The four others may be mistakenly classified as microinversions if one considers only reverse-strand alignments. On the other hand, the inversions shown in the second and third dot-plots are often missed in chaining and netting. Detecting inversions is not unlike the problem of finding orientations of highly diverged synteny blocks (see ref. 16), which requires a careful computational analysis. To address these complications, we developed a program, Inv-Checker, that analyzes artifacts shown in Fig. 1 by searching for inversions in all reverse-strand alignments, rather than those simply retained in a net (see *SI Appendix A*).

The pairwise representation of microinversions detected by our analysis hides the fact that there are insertions/deletions and alignment artifacts that affect each genome in a different way, making it difficult to rigorously define the term "inverted loci" (orthologous regions involved in inversions) across multiple species. The intuitive definition of inverted loci as a set of such orthologous regions across all species is somewhat imprecise because these regions may differ in length (because of deletions) and may include diverged nonalignable parts, making it difficult to construct their multiple alignment. Pairwise inverted regions between species $i$ and $j$ form a set of regions $S_{ij}$ in species $i$ and a set of regions $S_{ji}$ in species $j$. The union of all such regions over all pairs of species is denoted $\cup S_{ij}$ (the union is taken over all indices $i$ and $j$). This set represents the set of all regions in all genomes that were subjected to rearrangements (as demonstrated in SI Fig. 9). The exact endpoints of the regions in the sets $S_{ij}$ and $S_{ik}$ may vary widely (for $j \neq k$) and therefore the set $\cup S_{ij}$ provides a more accurate estimate of the span of the inversions than individual sets $S_{ij}$. We remark that although a microinversion between two species A and B may be easily detectable, the inverted region between species B and C may be too diverged to pass the alignment threshold. However, if one may align the corresponding regions in A and C, the inversion between B and C may be confirmed. To address this complication, for every continuous region in $\cup S_{ij}$ we use a more sensitive search to find all similar regions in other species, resulting in an extended set of species in which an inversion locus is detected. We iteratively use the extended set of regions to find divergent loci that were undetected in the pairwise comparisons until no new loci are discovered. This procedure allowed finding inversions in related yet divergent species and resulted in a 40% increase in the number of regions found as compared with the set $\cup S_{ij}$.

All loci are expected to be present in each species (in direct or reverse orientation) unless ($i$) the locus is in a gap in an

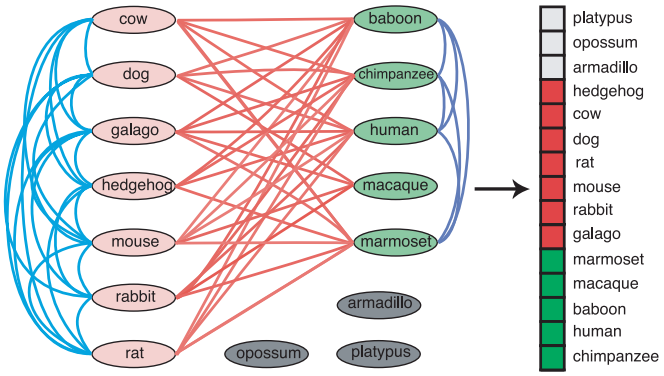


**Fig. 4.** The inversion graph (*Left*) and corresponding character vector (*Right*) for an inversion of length 1,300 bp created with edges assigned as the higher scoring alignments ($k = 2$). The graph is bipartite, indicating that there are no spuriously assigned orientations. Note that the graph is not complete; in particular, the inversion locus in macaque is partially deleted. It is counterintuitive that both macaque and human loci are aligned to mouse locus but not aligned to each other. It is explained by independent deletions in different regions in macaque and human that left no "common" sequence between macaque and human sequences in this locus. However, both partial sequences in macaque and human genomes are alignable to the mouse locus.

assembly, (*ii*) the locus was deleted in the course of evolution, or (*iii*) the locus is so diverged that it escapes the detection by sequence alignment. As a result, we find only 520 of the $68 \cdot 15 = 1{,}020$ possible regions in all inversion loci in 15 mammals.

We also detect a small number of regions that show evidence of overlapping microinversions such as $\overline{\text{ABCD}}\text{E} \rightarrow \text{A-}\overline{\text{D-C}}\text{-BE} \rightarrow$ ACD-BE in the human–baboon comparison. Although such microrearrangements are filtered out by InvChecker and are not considered as characters for phylogenetic reconstruction, they are perfectly suitable for phylogenetic analysis (unpublished work).

**Removing Conflicting Microinversions.** Ideally, each inversion locus yields a valid evolutionary character. However, in rare cases spurious alignments and overlapping inversions may produce ambiguous characters that need to be detected and removed before the tree reconstruction begins. We may remove ambiguous microinversions before tree reconstruction by using two methods: an *alignment-consistency test* that is based on the consistency of alignments within a single inversion loci, and a *four-gamete test* that is based on consistency of pairs of inversion loci.

The alignment-consistency test checks that the parity of a putative inversion is consistent across multiple species. For example, if a segment in species A is inverted relative to species B, and the same segment in B is inverted relative to a segment in C, then the segments in A and C should have the same orientation. To check alignment consistency within a given inversion locus, we construct an *inversion graph* for each inversion locus: vertices correspond to the inversion locus in each species; red edges connect vertices whose loci are in opposite orientation; and blue edges connect vertices whose loci are in the same orientation (Fig. 4). We determine the relative orientation of two inversion loci by comparing the local alignment scores of the inversion loci in the forward and reverse orientations to the (empirically determined) expected alignment score of two random sequences of the same length. The orientation is determined by the alignment that has a score $k$ times greater than the random alignment scores, where $k$ is a parameter. Lower values of $k$ allow one to analyze more divergent sequences, but they are more prone to errors. Furthermore, sequences that align with scores above $k$ in both orientation (as is the case with ambiguous loci boundaries such as partially deleted loci) are not assigned an edge. When all alignments for an inversion locus are consistent, then all cycles in its inversion graph should have an even number of red edges (in particular, red edges form a *bipartite graph*). Inversion loci that violate the "even number of red edges in a cycle" condition are discarded. We found that for $k = 2$ there are no cycles violating the even number of red edges condition.

The inversion graphs that do not violate the even number of red edges in a cycle condition (e.g., all inversion graphs for $k = 2$) may be used to derive evolutionary characters. The vertices of such graphs may be partitioned into two disjoint sets such that every path between vertices from two sets has an odd number of red edges (loci in one set are inverted as compared with loci in another set). We arbitrarily assign "direct" orientation to all loci in the first set and reverse orientation to all loci in the second set. Orientation is encoded in character vectors by assigning an orientation 1 to species on one side of the graph, and 0 on the other. We also assign ? to species outside the connected component (Fig. 4 *Right*). Combining all inversion loci results in an $n \times m$ matrix $C$ (with 0s, 1s, and ?s) for $m$ character vectors and $n$ species, shown in Fig. 5 *Upper*. The condition $C_{ij} = ?$ implies that the inversion locus $j$ is unresolved in species $i$. Note that the partition of each column into 0s and 1s is arbitrary and may be switched (i.e., the characters are *undirected*). The increase in the number of unresolved inversion loci with evolutionary time is attributed to difficulties in validating such inversions, incomplete coverage of *CFTR* regions for some species, and the stringent parameters we use in this study (see *Discussion*).

Next, we apply the four-gamete test to pairs of inversion loci. We assume that the set of microinversions is homoplasy-free. Therefore, the microinversions form a *perfect phylogeny* and all pairs of characters must satisfy the compatibility or four-gamete *test* (17, 18): no $n \times 2$ submatrix of $C$ formed by a pair of columns has the rows 00, 01, 10, and 11. Four-gamete violations may arise either from spurious alignments or from inversions that are not homoplasy-free. While in general violating characters may be detected and removed by using the Maximal Conflict Removal technique from ref. 9, our original dataset of 68 inversions contained only one violation. Manual inspection of this violation revealed that it is caused by a misclassification of a rather diverged inverted duplication (fourth diagram in Fig. 1). This misclassified microinversion was removed, resulting in 67 characters. *SI Appendix B* and SI Fig. 10 describe the distribution of these microinversions along the human genome.

While microinversions rarely reuse breakpoints, there is a difference between large-scale rearrangements and microinversions when it comes to breakpoint reuse. Sequenced genomes revealed large breakpoint reuse imposed by different (large-scale) rearrangements that happen at the same rearrangement hot-spots (19, 20). However, when one claims that two genome-scale rearrangements use the same breakpoint it does not mean that the breaks occur at exactly the same nucleotide but rather that their exact breakpoint positions are indistinguishable at the genomic scale. The situation is very different for microinversions, where even closely located breakpoints may be distinguished due to the smaller scale of the aligned regions. Because we found very limited microinversion breakpoint reuse with this higher level of resolution we postulate that repeated microinversions with exactly the same pairs of breakpoints (that would paraphyletically replicate microreversions and create homoplasy) are unlikely. The microinversions that share only one of their breakpoints do not represent a problem because they may be detected [by breakpoint graph (21) analysis] and discarded from further consideration.

**Reconstructing Phylogenetic Trees.** We first consider the case when all inversion loci are fully resolved (i.e., matrix $C$ has no ? signs).

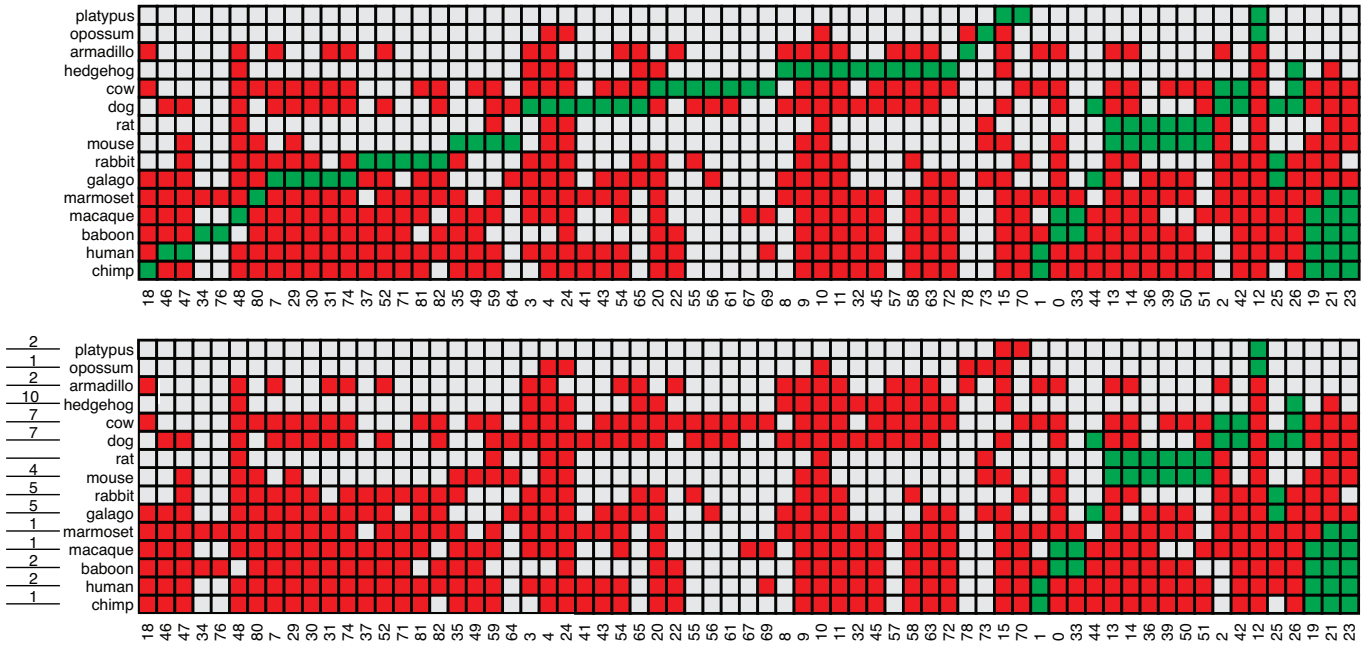


**Fig. 5.** The character matrix for 67 microinversions in 15 species (*Upper*) and the matrix after performing the first 49 good inversions (*Lower*). Each column represents an orthologous inversion locus. Red and green cells represent inversion loci in opposite orientation, and gray cells correspond to ? signs (unknown orientation). Columns with a single green cell are inversions unique to a species. The number of inversions performed on each species is shown to the left of *Lower*.

Let $\pi_1, \ldots \pi_n$ be *n signed genomes*[‖] that evolved by some (possibly overlapping) unknown inversions according to an unknown evolutionary tree. Without loss of generality, we assume that there was at least one inversion on every branch of the tree as zero length edges may be contracted. Every inversion creates up to two breakpoints, pairs of adjacent orthologous sequences that are out of order.

One may classify an inversion as *independent* if it creates exactly two new breakpoints (i.e., increases the number of breakpoints by 2) and does not reuse breakpoints (16). We call the rearrangement process *independent* if all its inversions are independent, up to the resolution of the boundaries of each inversion.

If all inversions were resolved (no ? signs in matrix $C$) and did not reuse breakpoints, the following variation of the perfect phylogeny theorem (22) would immediately resolve the problem of reconstructing inversion-based phylogenetic tree.

**Theorem.** *If n genomes of length m are produced by independent inversions, then both the correct evolutionary tree (up to the zero-length edges) for these genomes and the ancestral architectures of all its branching vertices may be reconstructed in polynomial time*.

For the sake of completeness, we give the outline of the proof. Let $\pi_1, \ldots \pi_n$ be $n$ genomes that evolved according to an (unknown) evolutionary tree $T$ and let $b(\pi_i, \pi_j)$ denote the number of breakpoints between genomes $\pi_i$ and $\pi_j$. Because every rearrangement creates exactly two breakpoints, the tree path between leaves $\pi_i$ and $\pi_j$ accounts for $b(\pi_i, \pi_j)/2$ rearrangements. The inversion distance between these genomes is at least $b(\pi_i, \pi_j)/2$, implying that the tree $T$ is additive (23). Because the (unknown) tree $T$ is additive and because the distances between its leaves are known, Zaretskii's theorem (24) implies that it may be uniquely reconstructed in linear time. An observation that the *median* (25) of every three genomes $\pi_i$, $\pi_j$, and $\pi_t$ evolved by independent rearrangements is unique and may be reconstructed in linear time implies that the permutations corresponding to all branching vertices in the tree $T$ may be uniquely reconstructed.

The above theorem does not impose any restrictions on the reconstructed tree (such as parsimony) and assumes only that the evolutionary process consists of independent events. This is a reasonable assumption because microinversions rarely reuse breakpoints.

It is straightforward to show by following the arguments used in the proof of the above theorem, that in case of independent evolution the Multiple Genome Rearrangements (MGR) algorithm (25) reconstructs the correct evolutionary tree. MGR constructs an evolutionary tree while seeking to minimize the number of inversions. However, MGR assumes that all inversions are resolved. Because microinversions are often unresolved for distant species we developed an MGR-like heuristic that is directed toward data with unresolved characters (the ? signs in matrix $C$).

Note that a removal of an edge from a phylogenetic tree partitions the tree into two subtrees. Our goal is to reconstruct a tree and assign every character to an edge in the tree in such a way that if this edge is removed then all 1s are in one subtree, whereas all 0s are in another subtree (see ref. 26).

Intuitively, our algorithm attempts to move "back in time," undoing microinversions, i.e., performing inversions of inverted loci that bring the existing species closer to the ancestral mammalian genome. This is achieved by evaluating all possible inversions for each genome, and identifying good inversions that bring a genome closer to the ancestral genome. Of course, the ancestral genome is unknown and therefore it is unclear how to find good inversions. However, Bourque and Pevzner (25) argued that an inversion which brings a particular genome closer to all other genomes is likely to be a good inversion. If this is correct, then we do not need the ancestral genome to find good inversions. We then continue performing good inversions in all genomes and iterate until some of the genomes (e.g., $A$ and $B$) do not have any loci in different orientations (converge to their

[‖]See Pevzner (23) for a background on genome rearrangements.

most common ancestor). After $A$ and $B$ become identical there are no longer good inversions in $A$ and $B$ (because any inversion in $A$ will make it more distant from $B$) and we merge $A$ and $B$ into a single genome (thus enabling good inversions at the next iteration) and iterate. Of course, this approach works well only for "nearly perfect" characters, and we argue that it is the case for microinversions.

Therefore, our MGR-like algorithm is very simple: look for good inversions in all genomes and perform them (if there are any) until some of the genomes become identical, merge identical genomes, and iterate. For example, in Fig. 5 *Upper* there is one good inversion in chimpanzee (corresponding to the green cell in the first column), two good inversions in human (green cells in the second and third columns), two in baboon, one in macaque, etc. We "reverse" all 49 good inversions (Fig. 5 *Lower*) so that some species become identical. For example, human and chimpanzee, macaque and baboon, mouse and rat, etc. become identical in Fig. 5 *Lower*. The difficulty, however, is that, because some inversions are unresolved, there is a danger that some inversions may appear to be good whereas in fact they are not, depending on the value (0 or 1) assigned to one of the ? signs. Another danger is that some genomes may appear identical (after performing some good inversions) whereas in fact they are not if the ? signs are replaced by 0 or 1. Armadillo/hedgehog and platypus/opossum represent an extreme case of such potentially incorrect merges because they have a single shared inversion locus. We address the uncertainty caused by ? signs with a greedy heuristic: we postpone merging species in any iteration if they have less than $p$ percent resolved characters in common (where $p$ is a threshold). For $p = 90\%$, the merging of platypus/opossum and armadillo/hedgehog will be postponed despite the fact that they represent "valid" merges. The number of remaining characters decreases as species are merged, and so merges that are postponed in an early stage are performed later.

Because our character matrices include ? characters it is possible that there are species that are pairwise, but not transitively, equivalent. Consider a simple example of species A, B, and C, with three characters in the matrix:

$$\begin{array}{cccc} A & 1 & 1 & ? \\ B & 1 & ? & 0 \\ C & ? & 0 & 0 \end{array} \qquad [1]$$

In this example A $\sim$ B, and B $\sim$ C, but A $\not\sim$ C. Sequences that are highly divergent or that contain many gaps may be missing characters that create such inconsistencies. To avoid artifacts caused by unresolved characters we merge the largest set of transitively equivalent species for which there are no inconsistencies. While this greedy heuristic is important in cases when there are a limited number of microinversions, it may not be necessary when more sequences are available.

After the first round of good inversions, our greedy heuristic merges human and chimp, macaque and baboon, galago and rabbit, mouse and rat, and cow and dog. Afterward we are left with 18 characters that represent "earlier" microinversions (SI Fig. 11*a*). Again, there exists 1 good inversion in the human + chimpanzee ancestor (first row), 2 good inversions in the macaque + baboon ancestor (second and third rows), 6 good inversions in the mouse + rat ancestor, and 2 good inversions in the cow + dog ancestor (SI Fig. 11*a*). The further progression of the algorithm is shown in SI Fig. 11 *c–f* (only four iterations are required to build the phylogeny). The 4 "dotted" edges in SI Fig. 11 do not correspond to any microinversions (zero-length edges that have to be contracted) and thus represent the same genomic architecture. Representing this ancestral architecture as a single vertex results in the final tree shown in the left of Fig. 6. The methods used to generate this dataset are described in *SI Appendix C*. The currently accepted phylogeny on the same species constructed from accordant subsets from refs. 9, 27, and 28 is presented in the right of Fig. 6.

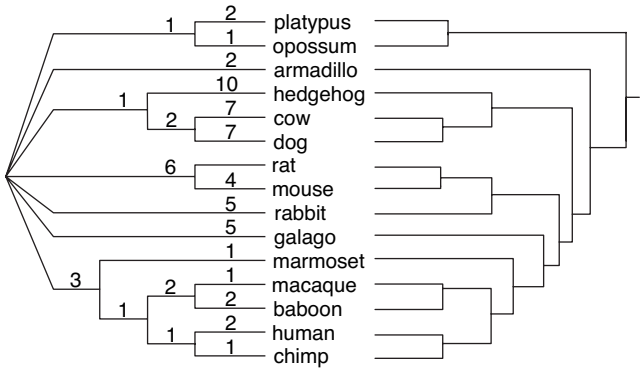


**Fig. 6.** The reconstructed tree (left), and corresponding canonical mammalian phylogeny (right). Vertices connected by dotted edges in SI Fig. 11 (no corresponding microinversions) are contracted into a single vertex.

Although a number of edges in the reconstructed tree remain unresolved, our analysis provides a proof of principle that microinversions represent valuable characters for phylogeny reconstruction. For example, the mitochondrial data analysis in ref. 29 places the hedgehog close to the root of the placentals, whereas others argue against this placement (30). Our result supports grouping of hedgehog with cow and dog, a result that is supported by most recent studies.

**Microinversions in Human and Chimpanzee Lineages.** The problem of discovering microinversions requires a careful analysis even when comparing the human and chimpanzee genomes, where very high sequence similarity suggests that microinversion breakpoints should be easy to detect. We analyzed the 1,460 putative microinversions reported in ref. 3 that are shorter than 15 kb, by running InvChecker on each inverted locus and 60 kb of flanking sequence. Only 293 putative microinversions were classified as inversions by InvChecker, whereas 1,005 inversions were classified as artifacts. The remaining 162 putative microinversions represent ambiguous genomic architectures that InvChecker is unable to call either way (an example is shown in SI Fig. 12). A large fraction of these artifacts are palindrome-like structures. Feuk *et al.* (3) experimentally validated some selected microinversions and confirmed that they indeed represent inverted sequences. Because a large portion of inversions in ref. 3 represent artifacts, the questions arises how these artifacts can possibly be experimentally validated. Of the 19 experimentally validated inversions from ref. 3 that were shorter than 15 kb, InvChecker classified all of them as inversions except one (of length 4,331 bp on chromosome 7) that turned out to be an inverted duplication. This finding suggests that the selection of inversions in ref. 3 for experimental validation had a bias for selecting canonical inversions (like the first inversion in Fig. 1). This bias is likely to be a consequence of the difficult repetitive nature of some breakpoint regions that makes PCR-based validation difficult.

We manually analyzed the genomic microarchitecture of these inversions by using genomic dot-plots of each inversion and flanking sequences. Whereas a large number of microinversions are flanked by inverted repeats (34%), many are flanked by insertions (31%) (*SI Appendix D*). This observation suggests that although inversions are thought to arise from nonhomologous recombination of inverted repeats, local genomic architecture is subject to rearrangement during repair of an insertion (or deletion).

**Reconstruction on Additional Sequence.** We applied InvChecker to two additional data sets: sequences of the same set of species across the 14 ENCODE manually selected target regions (ENCODE), and sequences of 38 species released as part of the National Institutes of Health Intramural Sequencing Center Comparative Vertebrate Sequencing project (CVSP). The ENCODE data set yielded 236 inversion loci that produce a phylogeny shown in the right of Fig. 6. The additional regions added the resolution of galago as well as further support for previously defined clades. The CVSP data set yielded 122 characters that correspond to internal edges on the phylogeny, and produced a phylogeny shown in *SI Appendix E*. On average, roughly 3 megabases are sequenced in each species. Our phylogeny supports a Rodentia topology that differs from recent nuclear gene-based phylogenies (31) and is discussed in *SI Appendix F*. This data set also highlights the rare but necessary need to be able to detect overlapping microinversions, as we found an instance of this in dog and ferret (discussed in *SI Appendix G*).

## Discussion

We have presented a method to discover microinversions and to use them as evolutionary characters to reconstruct phylogeny. The reconstructed tree has a number of unresolved branches but otherwise is in agreement with the currently accepted phylogeny.

High sequence divergence in noncoding regions makes it more difficult to find microinversions at the deep branches of mammalian evolution. In particular, we did not detect ancient inversions shared by a human/Rodentia (Euarchontogilres) ancestor versus dog/cow/hedgehog (Laurasiatheria) ancestor. The locus inverted in Laurasiatheria was deleted from Rodentia. This is consistent with findings of Bashir *et al.* (9), who found only two repeats in the *CFTR* greater region separating Euarchontogilres from Laurasiatheria, compared with 112 that resolve primates. We were able to reconstruct the primate phylogeny with rather short branches, thus indicating that our limited capacity in finding "ancient microrearrangements" may reflect limitations of our inversion detection tool and the choice of stringent parameters rather than the shortage of ancient microinversions. Indeed, throughout this study we use the same BLASTZ alignment scoring matrix and stringent $k = 2$ threshold for detecting microinversions, i.e., we apply the same threshold to human–chimpanzee comparison (high similarity) as to human–platypus comparison (low similarity). Making a variable threshold (depending on divergence of the species) will likely lead to significant decrease in the number of unresolved characters and discovery of ancient microinversions that remain undetected under the stringent threshold. Improving our algorithm to detect ancient microinversions remains an important goal; we believe it may be addressed by using ancestral sequence reconstructions (32).

We also remark that the assemblies of most of the 15 genomes we consider are incomplete and are currently represented by multiple contigs. These contigs are mapped to the human genome to produce ordered sequences. This procedure imposes a "human order" on unfinished sequences and prevents us from detecting inverted sequences in some species even if they exist. Such sequences may be discovered upon completion of the National Institutes of Health Intramural Sequencing Center project. In the future, ongoing genome sequencing projects, even low-coverage sequencing, will enable microinversion-based phylogenomics.

**Note.** Jian Ma and Jim Kent (personal communication) have developed an approach that finds microinversions by postprocessing of netted alignments and have arrived at a similar estimate of the number of microinversions between human and chimpanzee (≈500).

We are grateful to Ali Bashir, Vineet Bafna, Bill Murphy, Stephen Scherer, and Glenn Tesler for many useful comments. We are grateful to Jian Ma and Jim Kent for the comparative analysis of human–chimpanzee microinversions found in this paper and many useful comments. Some of the used sequence data were generated by the National Institutes of Health Intramural Sequencing Center (www.nisc.nih.gov). B.J.R. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

1. Pevzner PA, Tesler G (2003) *Genome Res* 13:37–45.
2. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D (2003) *Proc Natl Acad Sci USA* 100:11484–11489.
3. Feuk L, MacDonald JR, Tang T, Carson AR, Li M, Rao G, Khaja R, Scherer SW (2005) *PLoS Genet* 1:e56.
4. ENCODE Project Consortium (2004) *Science* 306:636–640.
5. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, *et al.* (2003) *Nature* 424:788–793.
6. Nikaido M, Rooney AP, Okada N, (1999) *Proc Natl Acad Sci USA* 96:10261–10266.
7. Rivera MC, Lake JA (1992) *Science* 257:74–76.
8. Hillis DM (1999) *Proc Natl Acad Sci USA* 96:9979–9981.
9. Bashir A, Chun Y, Price A, Bafna V (2005) *Genome Res* 15:998–1006.
10. Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J (2006) *PLoS Biol* 4:e91.
11. Bourque G, Zdobnov EM, Bork P, Pevzner PA, Tesler G (2004) *Genome Res* 15:98–110.
12. Fischer G, Rocha EP, Brunet F, Vergassola M, Dujon B (2006) *PLoS Genet* 2:e32.
13. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, *et al.* (2003) *Nucleic Acids Res* 31:51–54.
14. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) *Cytogenet Genome Res* 110:462–467.
15. Price A, Jones NC, Pevzner PA (2005) *Bioinformatics* 21(Suppl 1):i351–i358.
16. Pevzner PA, Tesler G (2003) in *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*, eds Vingron M, Istrail S, Pevzner P, Waterman M (Association for Computing Machinery, New York), pp 247–256.
17. Hudson RR, Kaplan NL (1985) *Genetics* 111:147–164.
18. Felsenstein J (2003) *Inferring Phylogenies* (Sinauer, Sunderland, MA).
19. Pevzner PA, Tesler G (2003) *Proc Natl Acad Sci USA* 100:7672–7677.
20. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L, *et al.* (2005) *Science* 309:613–617.
21. Bafna V, Pevzner PA (1996) *SIAM J Comput* 25:272–289.
22. Buneman P, Hodson F (1971) in *Mathematics in the Archaeological and Historical Sciences*, eds Hodson, FR, Kendall DG, Tautu P. (Edinburgh Univ Press, Edinburgh), pp 387–395.
23. Pevzner PA (2000) *Computational Molecular Biology, an Algorithmic Approach* (MIT Press, Cambridge, MA).
24. Zaretskii KA (1965) *Uspekhi Mat Nauk* 20:90–92.
25. Bourque G, Pevzner PA (2002) *Genome Res* 12:26–36.
26. Steel M (1992) *J Class* 9:91–116.
27. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, deJong WW, Springer MS (2001) *Science* 294:2348–2351.
28. Reyes A, Gissi C, Catzeflis F, Nevo E, Pesole G, Saccone C (2004) *Mol Biol Evol* 21:397–403.
29. Arnason U, Adegoke JA, Bodin K, Born EW, Esa YB, Gullberg A, Nilsson M, Short RV, Xu X, Janke A (2002) *Proc Natl Acad Sci USA* 99:8151–8156.
30. Waddell PJ, Shelley S (2003) *Mol Phylogenet Evol* 28:197–224.
31. Murphy WJ, Pevzner PA, O'Brien SJ (2004) *Trends Genet* 20:631–639.
32. Blanchette M, Green ED, Miller W, Haussler D (2004) *Genome Res* 14:2412–2423.