Reconstructing contiguous regions of an ancestral genome

Jian Ma,[1,5,6] Louxin Zhang,[2] Bernard B. Suh,[3] Brian J. Raney,[3] Richard C. Burhans,[1] W. James Kent,[3] Mathieu Blanchette,[4] David Haussler,[3] and Webb Miller[1]

[1]Center for Comparative Genomics and Bioinformatics, Penn State University, University Park, Pennsylvania 16802, USA; [2]Department of Mathematics, National University of Singapore, Singapore 117543; [3]Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA; [4]School of Computer Science, McGill University, Montreal, Quebec H3A 2B4, Canada

This article analyzes mammalian genome rearrangements at higher resolution than has been published to date. We identify 3171 intervals, covering ~92% of the human genome, within which we find no rearrangements larger than 50 kilobases (kb) in the lineages leading to human, mouse, rat, and dog from their most recent common ancestor. Combining intervals that are adjacent in all contemporary species produces 1338 segments that may contain large insertions or deletions but that are free of chromosome fissions or fusions as well as inversions or translocations >50 kb in length. We describe a new method for predicting the ancestral order and orientation of those intervals from their observed adjacencies in modern species. We combine the results from this method with data from chromosome painting experiments to produce a map of an early mammalian genome that accounts for 96.8% of the available human genome sequence data. The precision is further increased by mapping inversions as small as 31 bp. Analysis of the predicted evolutionary breakpoints in the human lineage confirms certain published observations but disagrees with others. Although only a few mammalian genomes are currently sequenced to high precision, our theoretical analyses and computer simulations indicate that our results are reasonably accurate and that they will become highly accurate in the foreseeable future. Our methods were developed as part of a project to reconstruct the genome sequence of the last ancestor of human, dogs, and most other placental mammals.

[Supplemental material is available online at www.genome.org and http://www.bx.psu.edu/miller_lab/.]

Using computer simulations, we have shown (Blanchette et al. 2004) that the genome sequence of the so-called Boreoeutherian ancestor (Fig. 1) can be computationally predicted at high accuracy within most euchromatic intervals that are free of large-scale rearrangements, given adequate data from living mammals. For instance, when sequences from 20 appropriately chosen mammalian species are available, we expect that >98% of the reconstructed nucleotides will be identical to the corresponding ancestral base. Because all mammals have experienced large-scale genomic rearrangements since their last common ancestor, in order to determine regional correspondence we analyze these rearrangements to infer a partition of each genome into intervals where nucleotide-level reconstruction methods can be applied.

The regional correspondence between modern and ancestral chromosomes has been predicted with increasing accuracy by a number of groups using a variety of methods. Currently, the main experimental technique is chromosomal painting (for surveys, see Wienberg 2004 and Froenicke et al. 2006), in which fluorescently labeled chromosomes from one species are hybridized to chromosomes from another species. Although the requirement of optical visibility means that the cytogenetic approach can recognize only rearrangements with conserved segments longer than 4 Mb (Froenicke et al. 2006) and cannot identify intrachromosomal rearrangements (Wienberg 2004), the chromosomal painting approach has the advantage that data are available for over 80 mammals (50 primates). Alternatively, computational methods that attempt to identify orthologous genomic intervals have much higher resolution, potentially down to under a kilobase. However, only a handful of vertebrate genomes are currently sequenced with sufficient precision and completeness to be informative for such an analysis. Using a combination of these two approaches, Murphy et al. (2005) estimated the rearrangement rates in the lineages leading to human, mouse, rat, cat, cattle, dog, pig, and horse, and predicted that the Boreoeutherian ancestor had 24 chromosomes.

To predict large-scale relationships among modern and ancestral genomes with sufficient accuracy for our needs, we have devised new methods. Conserved genomic segments are identified directly from freely available data and analyzed by a new computer program, as described below. We also estimate the accuracy of our results, compare them with published analyses, and explore the biological properties of rearrangement sites. The computer software described herein and details of our predictions for human, mouse, rat, and dog are freely available at http://www.bx.psu.edu/miller_lab/.

[5]Present address: Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA.
[6]Corresponding author.
E-mail jianma@bx.psu.edu; fax (814) 863-6699.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5383506. Freely available online through the *Genome Research* Open Access option.

Results

Segmenting the genomes based on pair-wise alignments

To predict segments of the ancestral genome, we start with "nets" (Kent et al. 2003), downloaded from the UCSC Human Genome Browser (http://genome.ucsc.edu/) (Kent et al. 2002). A net is an alignment between putatively orthologous regions in two ge-

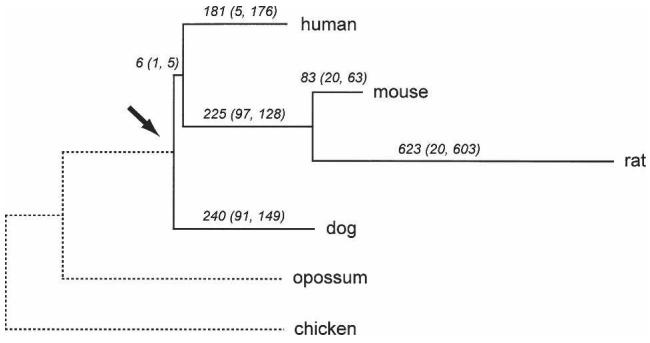

Figure 1. Position of the Boreoeutherian ancestor. Branch labels give the estimated number of chromosomal breaks from our study, also categorized as (interchromosomal, intrachromosomal). If conserved segments *i* and *j* are adjacent in the ancestral genome but not in the descendant genome, then we call the break interchromosomal if *i* and *j* are on different chromosomes in the descendant, and intrachromosomal otherwise. We suspect that many of the predicted intrachromosomal breaks in rat are assembly artifacts.

nomes, where it is predicted that no large-scale rearrangements occurred since the last common ancestor. We split nets if necessary to guarantee that they never contain an indel (insertion or deletion) of length exceeding a chosen threshold, e.g., 50 kb. Based on nets, we progressively construct sets of genomic intervals called, respectively, "orthology blocks," "conserved segments," and "contiguous ancestral regions" (abbreviated CARs). Each set contains pair-wise orthologous genomic intervals, one from each species under consideration, which for the current study means human (build hg18, March 2006) (Human Genome Sequencing Consortium 2001), mouse (build mm8, Feb. 2006) (Mouse Genome Sequencing Consortium 2002), rat (build rn3, June 2003) (Rat Genome Sequencing Consortium 2004), and dog (build canFam2, May 2005) (Lindblad-Toh et al. 2005). Since the intervals in a given set are orthologous, the set corresponds to a genomic interval in the last common ancestor of those species. We categorized the set according to restrictions on the kinds of large-scale evolutionary operations predicted to have happened in the lineages leading to the modern species, as summarized in Table 1.

Our methods for constructing orthology blocks and conserved segments are illustrated in Figure 2, while later sections and the Supplemental material explain the construction of CARs. Figure 2A shows human–mouse nets. Four mouse intervals are depicted, as ordered and oriented by the orthologous human segments. The second and third mouse intervals are actually adjacent (and appropriately orientated) on a mouse chromosome, and the intervening bases, if any, do not align to human; this is

Table 1. Types of orthologous-interval sets discussed in this article

Name	Species	From ancestor to descendants
Net	2	No large rearrangements or indels
Orthology block	N	No large rearrangements or indels
Conserved segment	N	No large rearrangements
Contiguous ancestral region (CAR)	N	Arbitrary rearrangements or indels

By "large rearrangement," we mean inversions and translocations involving genome segments exceeding some appropriate threshold in length (50 kb for the current study), or chromosome fissions or fusions.

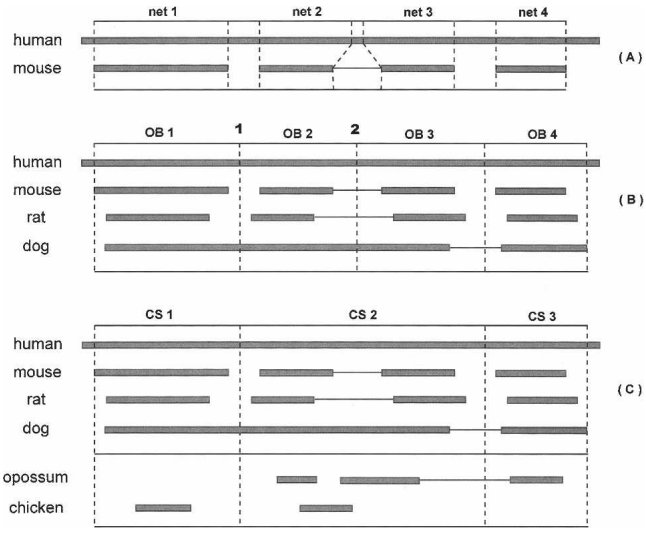

Figure 2. (*A*) Nets. Human is the reference species. The line between intervals indicates that a genomic interval of zero or more unaligned bases exists in the nonreference species between the adjacent intervals (see text). (*B*) Orthology blocks. (*C*) Conserved segments, including outgroup nets. The order and orientation of OB2 and OB3 are conserved in all four species, so we merge them into a conserved segment.

depicted by a thin line connecting the representations of those intervals.

Figure 2B shows the human–mouse, human–rat, and human–dog nets for a segment of the human sequence and illustrates the creation of orthology blocks. A dashed line between orthology blocks lies halfway between two intervals that are adjacent relative to human. For instance, the line at human position 2 in Figure 2B is midway between two mouse intervals. When the gap between two adjacent intervals in one species overlaps a gap relative to another species, as at position 1, we use the point halfway between the larger of the interval endpoints to the left and the smaller of the endpoints on the right. We discard all orthology blocks that cover <50 kb of human, because experiments showed that they tend to be unreliable (e.g., aligned seg-

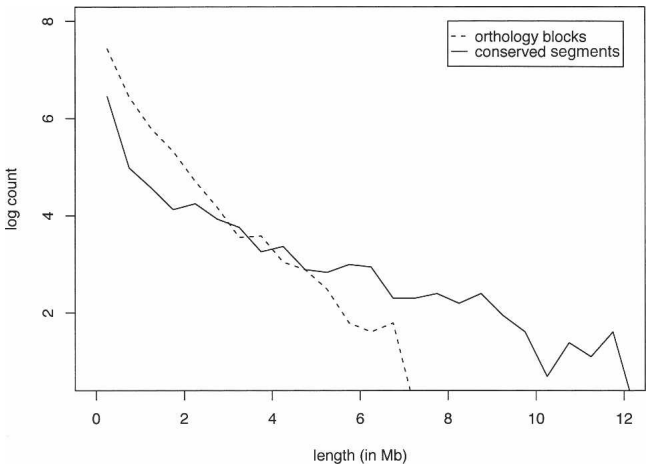

Figure 3. Length distribution of orthology blocks and conserved segments. Both orthology blocks and conserved segments are grouped into bins of 500 kb. Counts scaled by natural logarithm are plotted against lengths (in Mb).

ments are not always clearly orthologous). For this reason we describe our orthology blocks as having "50-kb resolution." Application of this process created 3171 genomic intervals, which include 92.36% of the available human genome sequence.

As illustrated in Figure 2C, we fuse runs of consecutive orthology blocks whenever the order and orientation of these blocks are conserved in each of the contemporary genomes. In terms of the convention used in Figure 2A, this means that for all nonhuman species, the boundary between the blocks is crossed either by a net or by a thin line. The results of the fusion process are independent of the order that fusions are performed. We call each resulting union of blocks a conserved segment. We found 1338 conserved segments, each containing an average of ~2.4 orthology blocks, which include 94.81% of the available human genome sequence. To help the process of inferring adjacencies between conserved segments in the ancestral sequence, we add nets from outgroup species to the conserved segments (Fig. 2C). The intervals in a conserved segment from an outgroup species are not required to be consecutive on the same chromosome.

Figure 3 shows the length distributions of orthology blocks and conserved segments across the whole genome.

Predicting contiguous ancestral regions from modern adjacencies

Adjacencies of genomic content (e.g., genes) have been used as a binary character to infer phylogeny in a parsimony framework (for a survey, see Savva et al. 2003). However, in a different context, where the phylogeny is known, our objective is to predict the ancestral order and orientation based on adjacencies in modern genomes. Consider an end of a conserved segment that does not correspond to a human telomere or centromere. How can we identify the conserved segment that was adjacent in the ancestral genome? If the segment that is currently adjacent in human is identical to the one that is adjacent in dog (but a different segment is adjacent in mouse and rat), the most parsimonious assumption is that the first and second segments were adjacent in the ancestral genome (and that a disruption occurred in the rodent lineage at this genomic position).

If the same segment is adjacent to the chosen segment in human, mouse, and rat but not in dog, we need more information to confidently predict the ancestral configuration, since there is a chance that the dog adjacency is ancestral and that the breakage occurred on the short branch from the human–dog ancestor to the human–rodent ancestor (see Fig. 1). To help resolve these cases, we add outgroup information in the form of matches to opossum (build monDom4, Jan 2006, http://www.ncbi.nlm.nih.gov/; K. Linblad-Toh, pers. comm.) and chicken sequence (build galGal2, Feb 2004) (Hillier et al. 2004) to the conserved segments (Fig. 2C). If the outgroup information does not resolve the issue (by agreeing with either the human adjacency or the dog adjacency), we assume the more likely scenario, i.e., that the break occurred in the lineage leading to dog.

We have generalized these observations to develop a computational procedure for predicting the order and orientation of conserved segments (and hence of orthology blocks) in the ancestor, based on observed adjacency relationships in the modern genomes. In broadest outline, the method is analogous to Fitch's

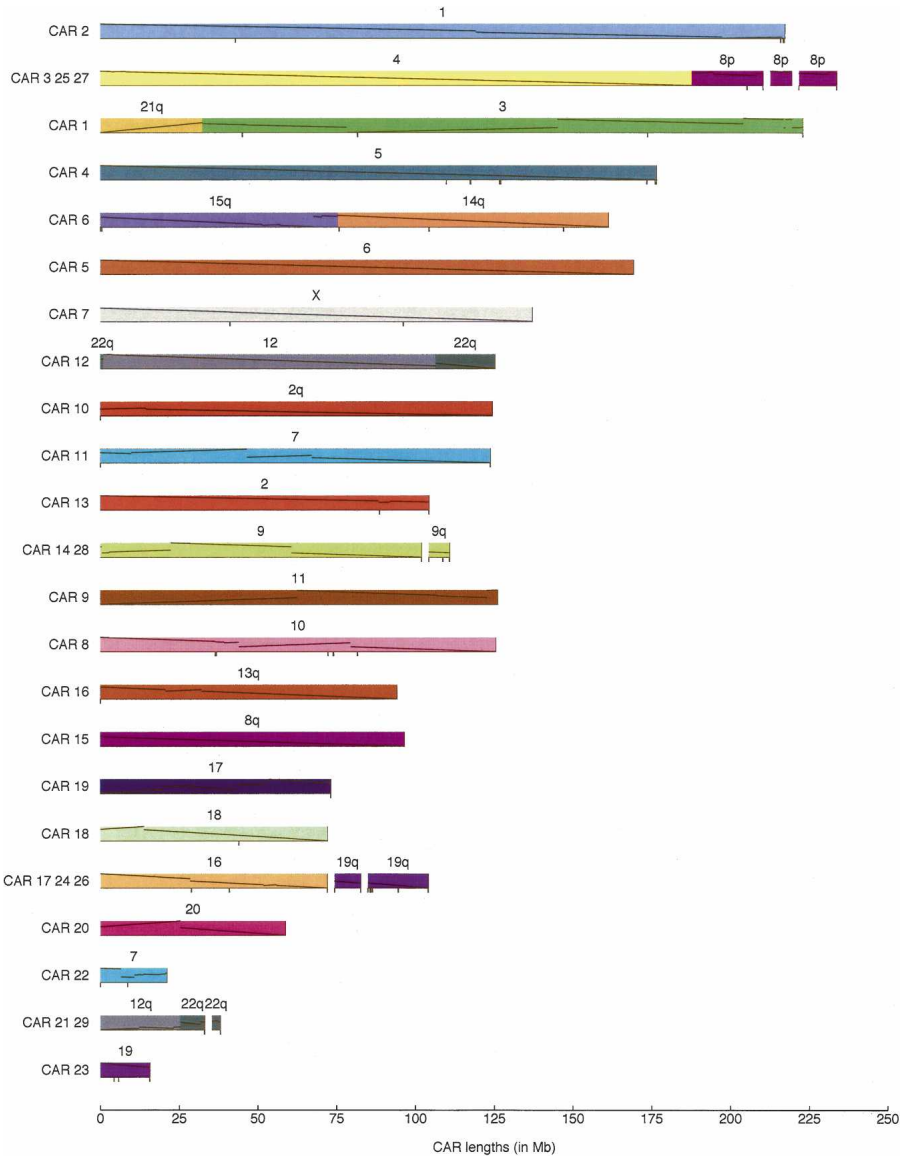

Figure 4. Map of the Boreoeutherian ancestral genome. For lengths of each CAR and corresponding parts in mouse, rat, and dog, see Table 2. Numbers *above* bars indicate the corresponding human chromosomes. Black tick marks *below* the bars indicate ambiguous joins (Fig. 7 in Methods; for details, see Supplemental material). Our predicted CARs are colored and ordered to facilitate comparison with Froenicke et al. (2006). Gaps between CARs are joins suggested by Froenicke et al. (2006). Diagonal lines within each block show the orientation and position in the human chromosome (Bourque et al. 2006).

parsimony method (Fitch 1971) for phylogenetic reconstruction, but adjacencies replace nucleotides as phylogenetic characters. Details of the method can be formulated precisely using concepts from graph theory (see Methods and a detailed example in the Supplemental material). We call each predicted ancestral run of segments a "contiguous ancestral region," abbreviated CAR. We found 29 CARs in total from the data we used (see Fig. 4 and Table 2). If we add the human sequence between conserved segments that are adjacent in both human and the ancestor (since a nucleotide-level reconstruction can include those intervals), 96.8% of the available human genome sequence is included. In Figure 4 we also use some chromosome painting results to combine CARs (leaving gaps in the figure) into our prediction of the genome structure of the Boreoeutherian ancestor. In Figure 4, black tick marks indicate joins with relatively weak support. For example, the leftmost tick mark on CAR 1 (which corresponds to human chr21 and chr3) shows a predicted ancestral adjacency between two conserved segments (conserved segments 238 and 239 in the Supplemental material) that are adjacent in human, mouse, and rat but not in dog and the outgroups.

In order to estimate the breakages on each lineage, we also reconstructed the intermediate ancestral genomes, i.e., the rodent ancestor and human–rodent ancestor. Boreoeutherian adjacencies were propagated to the intermediate ancestors. See the Methods section for the inference algorithm.

Identification of small inversions

Within each conserved segment, we identified in-place inversions that are too small to create a new orthology block. Because we currently lack a good outgroup species (such as elephant or armadillo), it was frequently difficult to confidently predict ancestral orientation. In ambiguous cases, we assumed that human is in the ancestral orientation relative to the immediately flanking regions, in part because the human assembly is more accurate than the others. However, this means that inversions in the human lineage are currently underestimated.

This method assigns 856 inversions to human, 3210 to mouse, 3067 to rat, and 4924 to dog. Among the 4924 inversions assigned to the dog lineage, only 703 were confirmed by an outgroup (e.g., opossum agreed with human); many of the remaining 4221 will be resolved by better outgroup data, e.g., from elephant. Figure 5 shows the length distributions of observed inversions assigned to each species. Among human inversions, 29 were assigned to the short branch leading from the Boreoeutherian ancestor to the human–rodent ancestor. The shortest inversions we found are 87 bp (in human), 31 bp (in mouse), 36 bp (in rat), and 34 bp (in dog). Figure 6 gives a detailed map of CAR 16, including small inversions. The detailed coordinates of these inversions and the tools for identifying them are freely available from http://www.bx.psu.edu/miller_lab/.

Properties of the breakpoints

Of 1309 pairs of conserved intervals that were predicted to be adjacent in the Boreoeutherian ancestor, 149 (11%) were separated by events in at least two independent lineages (12 were separated in three lineages). When we omit rat (because of the potential assembly problems indicated in Fig. 1), we find breakpoint reuse for 57 of 742 (8%). The ratio of breakpoint reuse we found is lower than what was reported in Murphy et al. (2005). One reason is that we use higher resolution to partition the ge-

Table 2. Number of conserved segments involved in each of 29 CARs

CAR	Bases covered (Mb)	No. of conserved segments	Human	Mouse	Rat	Dog
1	225.16	111	21 3	10 17 16 3 9 14 11 6	20 11 2 8 16 9 15 4	31 33 23 34 20
2	219.22	98	1	4 6 3 5 1 13 8 11 7	5 4 2 14 13 17 19 10 1	5 2 9 15 6 17 7 38 4 14 8 16
3	208.27	96	4 8	5 6 3 8	14 4 2 19 16	3 13 6 15 32 19 25 16
4	175.49	96	5	13 15 17 1 18 11	1 17 2 9 18 10	34 4 2 3 11
5	166.60	76	6	13 17 14 1 9 4 10	17 20 15 9 8 5 1	35 12 1
6	161.57	68	15 14	2 9 7 14 12	3 8 1 15 6	30 13 3 15 8
7	140.61	90	X	X	X 15	X
8	128.75	70	10	13 2 18 8 6 14 10 19 7	17 19 4 16 15 20 1	2 4 28 26
9	126.79	50	11	9 7 2 19	8 7 1 3	5 21 18
10	125.76	45	2	1 18 2	9 18 13 3	19 36 37 25
11	124.12	66	7	11 6 13 9 5 12	14 4 17 8 6	16 18 14
12	123.57	53	22 12	6 16 15 10 8	4 11 7 19	27 3 10 15
13	106.48	49	2	12 5 17 11 6 2 1 10	6 14 4 3 9 20	17 10
14	101.40	40	9	13 4 19 2	17 5 1 3	1 11 9
15	97.05	38	8	16 1 4 3 13 15	11 5 2 7	29 13
16	93.57	54	13	14 5 3 8 1	15 12 2 16 9	25 22
17	75.69	41	16	11 17 16 7 8	10 1 19	6 15 2 5
18	73.14	35	18	18 17 5 1	18 9 13	1 7
19	72.91	47	17	11	10	9 5
20	58.45	11	20	2	3	23 24
21	33.58	16	12 22	5 11 10	12 19 14 20	26
22	26.21	24	7	5	12	6
23	22.27	22	19	7	1	1
24	19.40	15	19	10 8 17 9	7 12 8 19 16	20
25	11.58	3	8	8 14	16 15	25
26	8.08	5	19	7	1	1
27	6.84	9	8	8	16	37 16
28	6.25	4	9	13	17	1
29	2.71	6	22	16 10	11 20	26

This table also shows how each contiguous ancestral region has been scattered among chromosomes in human, mouse, rat, and dog.

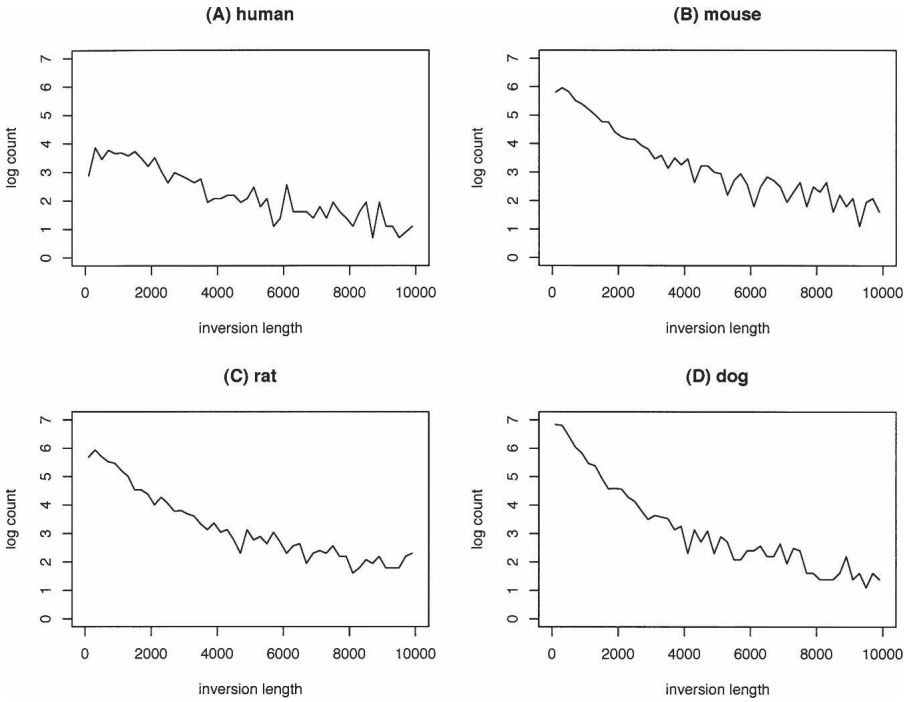

Figure 5. Length distribution of predicted inversions in human (*A*), mouse (*B*), rat (*C*), and dog (*D*). Inversions of lengths >10 kb are not represented in the plots. Inversions lengths are grouped into bins of 250 bp.

nomes. Another reason is that we only use four Boreoeutherian descendant species in our study. Some of the breakpoints identified using dog and rodent might be reused in other Boreoeutherian descendants. Simulations, described below, suggest that this frequency of breakpoint reuse is approximately what one would expect if breakage was equally likely for every genomic position, but a careful analysis is beyond the scope of this study. See Peng et al. (2006), Sankoff (2006), and references therein for an introduction to the long-standing debate about validity of the uniform breakage model.

We inspected intervals around breakpoints in the human sequence, looking for properties that might help explain why breaks occur at some positions but not others. We used 50-kb intervals centered on the end of a conserved segment where the adjacent segments in the ancestor and human differ. Breaks that occurred only in the human lineage were treated separately from those that were reused in another species, giving 16.97 Mb of human-specific breakpoint regions and 11.96 Mb around reused breakpoints. For small inversions in human, we used 1-kb intervals centered on each endpoint, covering 1.60 Mb.

Our observations are summarized in Table 3. GC content around breakpoints is slightly higher than the genome average, but not as elevated as reported for dog chromosomal breaks by Webber and Ponting (2005). The breakpoint regions are substantially enriched for RefSeq genes, consistent with what Murphy et al. (2005) observed in larger (~1 Mb) regions around breakpoints. The density of SINEs is also much higher than average. Finally, we observed that a large amount of DNA (41.72%) in human-specific breakpoint regions is in human segmental duplications (Bailey et al. 2001).

Theoretical analysis

We wanted to assess what proportion of the ancestral adjacencies inferred by our method might be correct. We started with a theoretical analysis, which necessitated some rather severe assumptions. Following the method of Sankoff and Blanchette (1999), we imagine a genome π with n elements that evolves through a series of rearrangements. An entry of π can be thought of as one of the $2n$ numbers $1, -1, 2, -2, \ldots, n, -n$, where the sign designates the element's orientation. No assumptions are made about the relative frequencies of inversions, translocations, fusions, and fissions. Instead, if f precedes g in the genome $\pi = \ldots fg \ldots$ at a particular time, then g may be changed to h (other than f, $-f$ or g), with each of the $2n - 3$ possibilities equally likely. The probabilities that the successor of f is changed in time t along a specific branch are pivotal parameters of the model. We estimated those parameters using data observed in real genomes. A surprisingly complex computation (see the Supplemental material) then estimated that ~90.18% of the adjacencies we predict in the Boreoeutherian ancestor are correct.

Mathematical tractability of this model is purchased at a high cost in realism. For instance, the assumption that all replacements are equally likely contradicts the observations that inversions are more frequent than are other rearrangements and that short inversions are more common than are long ones.

Simulations

We used computer simulations to inject more realism into the analysis, employing a realistic evolutionary tree with branch lengths based on substitution frequencies. Starting with a hypothetical human–chicken ancestor having 6000 orthology blocks and 25 chromosomes, we simulated inversions, translocations, fusions, and fissions along each branch. Rearrangements were distributed as 90% inversions, 5% translocations, 3.75% fusions, and 1.25% fissions. We modeled lengths of inverted blocks with a γ distribution, with shape and scale parameters $\alpha = 0.7$ and

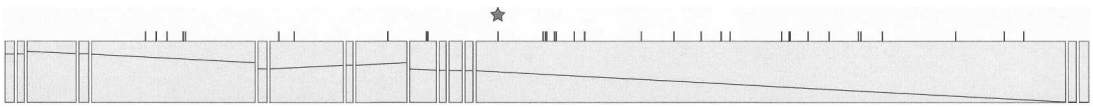

Figure 6. Detailed map of human chromosome 13q onto CAR 16. There are16 large-scale rearrangement-free pieces. Vertical lines *above* each piece indicate positions of small in-place inversions. The star indicates an inversion (around hg18.chr13:57,380,591–57,383,765) that happened on the branch from the Boreoeutherian ancestor to the human–rodent ancestor.

Table 3. Genomic content of breakpoint regions

	Human-specific breakpoints	Reused breakpoints	Human short inversions	Genome-wide average
GC content (%)	43.58	42.81	39.82	40.91
Segmental duplication (%)	41.72	17.89	3.94	5.24
Gene density (genes/Mb)	20.45	21.99	—	8.58
Repeats density (%)	52.88	54.30	44.23	48.58
SINE	18.58	16.60	14.30	13.60
LINE	19.97	23.04	16.87	21.32
LTR	9.15	9.92	7.53	8.61
DNA	2.48	2.74	3.56	3.01
Others	2.70	2.00	1.97	2.04

Gene density is not given for inversions because we used regions too short to give meaningful results. Repeats were identified with RepeatMasker, and segmental duplications were obtained from UCSC Genome Browser segmental duplication track.

scale $\theta = 500$, respectively. Since we required inversion lengths not to exceed 50 blocks, we truncated the γ distribution at 50 (probabilities for shorter lengths are renormalized). Simulation parameters, especially the rates of rearrangements, were further tuned according to our observed data. For example, we also allowed each branch to have its own adjustment parameter for each operation, to account for the differences among branches. The simulated genomes produced by this approach are consistent with actual mammalian genomes in terms of number of conserved segments, number of breakpoints, chromosome count, etc. Some important features of the simulated data are compared with what were seen in real data in Table 4.

For the species shown in Figure 1, we repeated the simulation 50 times, in each case running our program for inferring CARs on the resulting data set and comparing the predicted adjacencies with the known (simulated) ones. For determining the success rate, we considered only the ancestral joins that were broken in at least one lineage, since the unbroken joins will be found by essentially any procedure.

When using human, mouse, rat, and dog, with opossum and chicken as outgroups (Fig. 1), the frequency of correctly predicted adjacencies was 98.96% (SD = 0.39) for the Boreoeutherian ancestor, 98.37% (SD = 0.55) for the human–rodent ancestor, and 97.07% (SD = 1.01) for the mouse–rat ancestor. Note that as for inference of nucleotides (Blanchette et al. 2004), the prediction accuracy is also higher for the Boreoeutherian ancestor than for some younger ancient genomes.

We also reconstructed the Boreoeutherian ancestor without using opossum and chicken; the accuracy decreased to 97.40% (SD = 0.69). If we retain the outgroups but leave out rat, the accuracy drops to 98.29% (SD = 0.67). However, if chimp, cow, and macaque are included in the reconstruction, the simulation indicates that joins in the Boreoeutherian ancestor are computed with 99.34% (SD = 0.29) accuracy.

Comparison with other reconstructions

A comparison with other reconstructions identifies which part of the ancestral genome can be confidently reconstructed and highlights regions where further investigation is needed. Our predicted CARs agree well with predictions from chromosome painting with respect to interchromosomal operations (cf. Fig. 1 in Froenicke et al. 2006 and Fig. 4). Of eight strongly supported interchromosomal breaks in human predicted by Froenicke et al. our reconstruction agrees with five (see Table 5). In the prediction of Murphy et al. (2005), joins of Hsa16q/Hsa19q and Hsa16p/Hsa7 were reconstructed with weak support, using data from cat and pig, and cat and cow, respectively, all of which were unavailable for computing CARs. Also, in both cases, the species (cat and pig or cat and cow) come from the same superorder, so they did not provide strong evidence as to the Boreoeutherian ancestral state. Therefore, more species are needed. Moreover, the endpoints of the relevant conserved segments for these two joins are all connected with ambiguity in our prediction (see Fig. 4), suggesting other possible scenarios which are very likely to be joins of Hsa16q/Hsa19q and Hsa16p/Hsa7. The join of Hsa16p/Hsa7 corresponds to the centromere of Hsa16 (the first weakly supported join on CAR17) and the left side of CAR22 in the picture. We think the small sample size of Boreoeutherian descendant genomes in the current study is likely responsible for the failure to recover these adjacencies.

The predictions of Murphy et al. (2005) which were mainly based on the MGR algorithm (Bourque and Pevzner 2002), are more or less similar to Froenicke et al.'s (2006) results. However, five of Murphy's putative weakly supported interchromosomal breaks (Hsa1/Hsa22, Hsa5/Hsa19, Hsa2/Hsa18, Hsa1/Hsa10, Hsa2/Hsa20) are not supported by chromosome painting data. Our program made none of these joins, which is in agreement with Froenicke et al. To further examine the differences between reconstruction algorithms, we ran our CAR-building program on Murphy et al.'s data set with 307 conserved segments from their Supplemental materials. Table 6 compares our predicted ancestral joins with theirs.

Discussion

A number of additional mammals are already being sequenced "at low redundancy" for the purpose of identifying human regions under negative selection for substitutions (Margulies et al. 2005). The resulting sequence data are extremely useful for predicting ancestral nucleotides. However, they lack the long-range contiguity needed for accurate identification of large-scale evolutionary events. We look forward to the day when a high-accuracy assembly of, say, elephant or armadillo provides us with an ideal outgroup for large-scale reconstruction of the Boreoeutherian ancestral sequence.

Table 4. Comparison between our simulated data and real data

	No. of conserved segments	Breakpoint distance			% of breakpoint reuse
		H-M	H-R	H-D	
Real data	1338	564.5	1059	452	11.38
Simulated data	1375.64 (39.87)	559.02 (30.31)	1038.39 (37.94)	434.72 (25.24)	11.74 (0.92)

Simulated statistics are the average from 50 simulated data sets. The standard deviations of numbers in simulated data are in parentheses.

Table 5. Comparison with Froenicke et al. (2006) and Murphy et al. (2005)

Comparison	Froenicke et al. (2006)	Murphy et al. (2005)	Our method	Included in Figure 4	Comments
No. of species	>80	8	4		We also used two outgroups
Coverage of human genome	–	48%	96%		
Resolution	4Mb	120Kb	50Kb		
Identify inter/intra	only inter	both	both		
Hsa4a/Hsa8p	+	+ (weak)	+ (strong)	+	Join (379, −653). Supported by mouse, rat, dog, chicken
Hsa4b/Hsa8p	+	+ (weak)	–	–	
Hsa21/Hsa3	+	+ (strong)	+ (strong)	+	Join (−1212, 229). Supported by mouse, rat, dog, chicken
Hsa15/Hsa14	+	+ (weak)	+ (strong)	+	Join (−994, 968). Supported by mouse, rat, dog, opossum
Hsa10p/Hsa12a	+ (weak)	–	–	–	
Hsa12a/Hsa22a	+	–	+ (strong)	+	Join (909, 1239). Supported by mouse, rat, dog, opossum, chicken
Hsa12b/Hsa22b	+	+ (weak)	+ (strong)	+	Join (−1231, 910). Supported by mouse, rat, dog
Hsa16q/Hsa19q	+	+ (weak)	–	+	
Hsa7b/Hsa16p	+	+ (weak)	–	–	
Hsa1/Hsa22a	–	+ (weak)	–	–	
Hsa5/Hsa19p	–	+ (weak)	–	–	
Hsa2pq/Hsa18	–	+ (weak)	–	–	
Hsa1q/Hsa10q	–	+ (weak)	–	–	
Hsa20/Hsa2	–	+ (weak)	–	–	

The naming convention of human chromosomes' regions follows Figure 1 in Froenicke et al. 2006. "+" indicates the method that made that join, −, otherwise. If the join was made by our method, we also list the corresponding conserved segment numbers and show which species support that join.

Our computational methods need further refinement. In particular, we anticipate improvements in the handling of large duplications and deletions. Additional progress may be possible through the modeling of other evolutionary events, such as gene conversion or expansion/contraction of short tandem repeats caused by strand slippage. Moreover, advances are needed to facilitate simultaneous reconstruction of ancestral genomes at all internal nodes of the phylogenetic tree.

An accurate conceptual model of large-scale evolutionary events will be critical for successful reconstruction of ancestral genomes. The discrepancy between our theoretical analysis of reconstruction accuracy and the results from simulation underline the need for more work in this area.

A number of challenges remain before the genome sequences of mammalian ancestors can be accurately predicted at nucleotide resolution. However, we believe that this goal is an appropriate focus for our twin aims of understanding the evolutionary history of every position in the human genome and of providing an Internet resource that optimally organizes and presents the ever-expanding wealth of mammalian and vertebrate sequence data in the context of the ancestral sequence they have in common.

Table 6. Results of running our CAR-building program on data from Murphy et al. (2005)

	Total number of joins based on Murphy et al.'s data	Two methods agree
All types of joins	338	287
All non-endpoint joins	276	255
Strongly supported non-endpoint joins	246	242
Weakly supported non-endpoint joins	30	13
All endpoint joins	62	32
Strongly supported endpoint joins	37	29
Weakly supported endpoint joins	25	3
Non-human-consecutive joins	9	7

Endpoint joins have two cases: (1) the join between the first conserved segment in a chromosome and the beginning of the chromosome; (2) the join between the last conserved segment in a chromosome and the end of the chromosome. Non-human-consecutive joins refer to two consecutive elements in human that are not consecutive in the ancestor, indicating a breakpoint in human.

Methods

Inferring CARs

Given information about adjacencies between conserved segments in each modern species, our goal is to infer segment order in the ancestral genome. To get a clean and precise statement of the problem, we formalize it using graph theory. We develop an algorithm that identifies a most parsimonious scenario for the history of each individual adjacency, although the whole-genome prediction is not guaranteed to optimize traditional measures like the number of breakpoints. We introduce weights to the graph edges to model the reliability of each adjacency.

By encoding each genomic element as a number, we view a chromosome as a signed permutation where the sign corresponds to the orientation (strand) of the encoded element. For example, $\pi = 1\text{–}4\ -3\ 5\ 2$ denotes a chromosome of five genomic elements, in which elements 1, 5, 2 are on the positive strand and 3, 4 are on the negative strand. Currently, we do not allow duplications, i.e., two elements with the same absolute value. Each element π_i in a chromosome $\pi = \pi_1\ \pi_2 \ldots \pi_{i-1}\ \pi_i\ \pi_{i+1} \ldots \pi_n$ has a unique predecessor π_{i-1} and successor π_{i+1},; we define that π_1 has predecessor 0 and π_n has successor 0, where 0 is a special symbol for the left and right end of each chromosome. Inferring CARs in the ancestral genome is equivalent to finding the unambiguous predecessor and successor of each element in the ancestral genome.

Recall that the Fitch algorithm infers the ancestral DNA sequence at the root of a phylogenetic tree from the DNA sequences at the leaves, site-by-site in two stages (Felsenstein 2004). For each site, in a bottom-up fashion, it first determines a

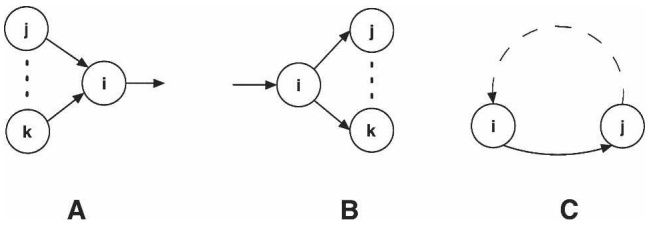

Figure 7. Three potential ambiguous cases. (*A*) *i* has several possible predecessors; (*B*) *i* has several possible successors; (*C*) *i* forms a cycle with *j*.

set C_v of candidate nucleotides at each internal node v according to the following rule: If v is a leaf, C_v just contains its nucleotide label; otherwise, if v has children u, w, then C_v equals $C_u \cup C_w$ or $C_u \cap C_w$ depending on whether C_u and C_w are disjoint or not. Then, in a top-down fashion, it assigns a nucleotide b_v from C_v to v according the following rule: Let v' be the parent of v; if the nucleotide $b_{v'}$ assigned to v' belongs to C_v, then, $b_v = b_{v'}$. Otherwise, set b_v to be any nucleotide in C_v. Although nucleotide assignment in the second stage is not unique, any assignment gives an evolutionary history with the minimum number of substitution events.

To infer the ancestral predecessor of an element x, our algorithm does exactly the same thing as described above. In a bottom-up fashion, for each node u, we compute a set $P_u(i)$ of the "predecessors" for each element i according to the following rule:

If u is a leaf, then $P_u(i)$ consists of the unique predecessor in the genome associated with u. Otherwise, assume u has children v and w; then, $P_u(i)$ is equal to $P_v(i) \cup P_w(i)$ or $P_v(i) \cap P_w(i)$ depending on whether $P_v(i)$ and $P_w(i)$ are disjoint or not.

The above inference method outputs a "predecessor graph" at each internal node. In this graph there is an arc from each element in $P_u(i)$ to element i. In general, such a predecessor graph may not be a union of vertex-disjoint paths. This is because two different series of evolutionary events may transform different genomes into the same one, and hence, we often do not have enough information to determine the true predecessor and successor relationships in the ancestor. However, the vast majority of the ancestral adjacencies are likely to be present as edges in the inferred graph.

We treat outgroups in a consistent manner. We first infer the predecessor set $P_R(i)$ in the common ancestor R of all the species, including outgroups. Then we propagate $P_R(i)$ down the tree until we reach the target ancestor T. During the propagation process, if O and A are ancestor and descendant on one branch, respectively, for each element i, we adjust the inferred predecessor set $P_A(i)$ of i at the mammalian ancestor A as follows: If $P_O(i)$ and $P_A(i)$ share common elements, we just take them as the predecessor set of i at A; otherwise, $P_A(i)$ is unchanged. In our reconstruction, we first infer the predecessor set of human–chicken common ancestor. Then we used it to adjust the human–opossum common ancestor. And finally, we used human-opossum ancestor to adjust the human–dog common ancestor.

Similarly, we infer a "successor graph" at each internal node in the given phylogeny. In a successor graph, there is an arc from element i to each element in its inferred successor set $S(i)$ at each internal node.

The predecessor and successor graphs are the same at each leaf, but those at an internal node are generally different, although they typically have many common parts. We find the consistent parts by taking the intersection graph G of the predecessor and successor graphs. In the intersection graph, an element could still be involved in three kinds of ambiguities, as depicted in Figure 7. If none of these ambiguous cases is present, the graph itself forms the set of paths that covers all the elements and provides the reconstructed ancestral genome structure.

When ambiguity exists, we need to resolve it and choose appropriate directed edges to form CARs. We assign a weight to each of the directed edges in the intersection graph using the following approach. If the edge (i, j) is neither one of the incoming edges of case (a) nor one of the outgoing edges of case (b), we set $w_\alpha(i, j) = 1$. Otherwise, the weight $w_\alpha(i, j)$ is recursively determined from the leaves by

$$w_\alpha(i,j) = \frac{D_L \cdot w_R(i,j) + D_R \cdot w_L(i,j)}{D_L + D_R}$$

where D_L and D_R are the branch lengths to the left child L and right child R branching from ancestral node α. Here, $w_L(i, j)$ and $w_R(i, j)$ are the edge weights to left child and right child. Therefore, in the graph, all the edge weights are between 0 and 1.

On a leaf genome, if (i, j) is present in the predecessor graph, we set $w(i, j) = 1$; otherwise $w(i, j) = 1$. This weight can be determined by a post-order traversal. Note that if an edge (i, j) is involved in ambiguous case (a) or (b), then $w(i, j) < 1$. The greater the weight of an edge (i, j), the more likely it is that i and j should be joined. The underlying assumption is that rearrangement is more likely to happen on longer branches.

Our purpose is to have a set of paths that cover all the nodes in the directed graph and at the same time maximize the total edge weights in the paths (we allow degenerate paths where there is only one node in the path). We propose a greedy heuristic approach. We first sort all the edges by weight and start to add edges to the vertex-disjoint paths representing the CARs. When an edge (i, j) is being added, we make sure that it will not cause ambiguous cases. If it will, we discard that edge and choose the next available one. The process is performed until no edge can be added. In the graph resulting from this step, all the edges that are not involved in ambiguous cases (a) and (b) will be retained. For ambiguous case (c) where a cycle is formed, we can discard any one edge to break the cycle. In fact, we can easily prove that if there is a cycle, the weight of each edge in that cycle is 1.

When we are adding elements into an existing path, particular care is needed to avoid putting j and $-j$ in the same CAR. In addition, we add both (i, j) and its symmetric version, $(-j, -i)$. For each path found by this approach, a symmetric path in the opposite orientation is also found, since we have nodes for both i and $-i$. The two paths correspond to the same CAR, and we choose one of them.

A detailed example of the algorithm can be found in the Supplemental materials.

Acknowledgments

We thank the Broad Institute for making the opossum assembly publicly available, Francesca Chiaromonte for several helpful suggestions, Elliott H. Margulies for providing the phylogenetic tree used in ENCODE project, and Guillaume Bourque for the MGR program. We especially would like to thank the anonymous reviewers for their comments on the initial draft of this paper. J.M., R.C.B., and W.M. were supported by NIH grant HG02238. L.X.Z. was supported by ARF grant 146-000-068-112. W.J.K, B.J.R, and D.H. were supported by NHGRI grant 1P41HG02371 and NCI contract 22XS013A, and D.H. was supported additionally by the Howard Hughes Medical Institute.

References

Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., and Eichler, E.E. 2001. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11:** 1005–1017.

Blanchette, M., Green, E.D., Miller, W., and Haussler, D. 2004. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.* **14:** 2412–2423.

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

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

Felsenstein, J. 2004. *Inferring phylogenies*. Sinauer, Sunderland, MA.

Fitch, W.M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* **20:** 406–416.

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

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695–716.

Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci.* **100:** 11484–11489.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas III, E.J., Zody, M.C., et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803–819.

Margulies, E.H., Vinson, J.P., Miller, W., Jaffe, D.B., Lindblad-Toh, K., Chang, J.L., Green, E.D., Lander, E.S., Mullikin, J.C., and Clamp, M. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl. Acad. Sci.* **102:** 4795–4800.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Murphy, W.J., Larkin, D.M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Gatzke, L., et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309:** 613–617.

Peng, Q., Pevzner, P.A., and Tesler, G. 2006. The fragile breakage versus random breakage models of chromosome evolution. *PLoS Comp. Biol.* **2:** e14.

Rat Genome Sequencing Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.

Sankoff, D. 2006. The signal in the genomes. *PLoS Comput. Biol.* **2:** e35.

Sankoff, D. and Blanchette, M. 1999. Probability models for genome rearrangement and linear invariants for phylogenetic inference. *RECOMB '99: Proceedings of the third annual international conference on computational molecular biology,* pp. 302–309. ACM Press, New York.

Savva, G., Dicks, J., and Roberts, I.N. 2003. Current approaches to whole genome phylogenetic analysis. *Brief. Bioinform.* **4:** 63–74.

Webber, C. and Ponting, C.P. 2005. Hotspots of mutation and breakage in dog and human chromosomes. *Genome Res.* **15:** 1787–1797.

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

Received April 7, 2006; accepted in revised form June 22, 2006.