Reconstruction and evolutionary history of eutherian chromosomes

Jaebum Kim[a,1], Marta Farré[b,1], Loretta Auvil[c], Boris Capitanu[c], Denis M. Larkin[b,2], Jian Ma[d,2], and Harris A. Lewin[e,2]

[a]Department of Biomedical Science and Engineering, Konkuk University, Seoul 05029, South Korea; [b]Comparative Biomedical Science Department, Royal Veterinary College, University of London, London, NW1 0TU, United Kingdom; [c]Illinois Informatics Institute, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [d]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213; and [e]Department of Evolution and Ecology, University of California, Davis, CA 95616

Contributed by Harris A. Lewin, May 11, 2017 (sent for review February 13, 2017; reviewed by William J. Murphy and Pavel A. Pevzner)

Whole-genome assemblies of 19 placental mammals and two outgroup species were used to reconstruct the order and orientation of syntenic fragments in chromosomes of the eutherian ancestor and six other descendant ancestors leading to human. For ancestral chromosome reconstructions, we developed an algorithm (DESCHRAMBLER) that probabilistically determines the adjacencies of syntenic fragments using chromosome-scale and fragmented genome assemblies. The reconstructed chromosomes of the eutherian, boreoeutherian, and euarchontoglires ancestor each included >80% of the entire length of the human genome, whereas reconstructed chromosomes of the most recent common ancestor of simians, catarrhini, great apes, and humans and chimpanzees included >90% of human genome sequence. These high-coverage reconstructions permitted reliable identification of chromosomal rearrangements over ∼105 My of eutherian evolution. Orangutan was found to have eight chromosomes that were completely conserved in homologous sequence order and orientation with the eutherian ancestor, the largest number for any species. Ruminant artiodactyls had the highest frequency of intrachromosomal rearrangements, and interchromosomal rearrangements dominated in murid rodents. A total of 162 chromosomal breakpoints in evolution of the eutherian ancestral genome to the human genome were identified; however, the rate of rearrangements was significantly lower (0.80/My) during the first ∼60 My of eutherian evolution, then increased to greater than 2.0/My along the five primate lineages studied. Our results significantly expand knowledge of eutherian genome evolution and will facilitate greater understanding of the role of chromosome rearrangements in adaptation, speciation, and the etiology of inherited and spontaneously occurring diseases.

chromosome evolution | ancestral genome reconstruction | genome rearrangements

Chromosome rearrangements are a hallmark of genome evolution and essential for understanding the mechanisms of speciation and adaptation (1). Determining chromosome rearrangements over evolutionary time scales has been a difficult problem, primarily because of the lack of high-quality, chromosome-scale genome assemblies that are necessary for reliable reconstruction of ancestral genomes. For closely related species with good map-anchored assemblies, such as human, chimpanzee, and rhesus, it is possible to infer most inversions, translocations, fusions, and fissions that occurred during evolution by simple observational comparisons (2). However, for sequence-based genome-wide comparisons that require resolving large numbers of rearrangements of varying scale, determining ancestral chromosomal states is challenging both methodologically and computationally because of the complexity of genomic events that have led to extant genome organizations, including duplications, deletions, and reuse of evolutionary breakpoint regions (EBRs) flanking regions of homologous synteny (3, 4).

A variety of methods have been used for resolving the evolutionary histories of mammalian chromosomes, with limited success and resolution. For example, chromosome painting by FISH (5–8) was used to predict ancestral karyotypes dating back ∼105 My to the ancestor of all eutherian (placental) mammals (9). Although yielding an outline of the basic reconstructed karyotypes, FISH-based methods do not have sufficient resolution to permit accurate identification of EBRs, homologous synteny blocks, and fine-scale rearrangements. Low-resolution methods also severely limit study of the relationship between chromosome rearrangements and structural variants, which are associated with adaptive evolution and the presence of EBRs (4, 10, 11). Thus, a distinct advantage of resolving EBRs at high resolution is that sequence features within them can be interrogated for genes that may be associated with lineage-specific phenotypes. This is an important motivation for creating finer-scale ancestral chromosome reconstructions (10, 12, 13).

Several algorithms have been developed to reconstruct the order and orientation of syntenic fragments (SFs) in common ancestors by using DNA sequence-level syntenic relationships among genomes of extant species. These methods use SFs constructed from whole-genome sequence alignments as input to infer the order and orientation of the SFs in a specific target ancestor. Different algorithmic approaches are used by the different reconstruction algorithms. For example, the multiple genome rearrangement (MGR) algorithm (14) uses a heuristic approach to reconstruct ancestral genomes by considering reversals (inversions), translocations, fusions, and fissions based on genome rearrangement distance. inferCARs (3) finds the most parsimonious scenario for the history of SF adjacencies and then greedily connects the adjacencies into contiguous ancestral regions (CARs).

Significance

Determining the order and orientation of conserved chromosome segments in the genomes of extant mammals is important for understanding speciation events, and the lineage-specific adaptations that have occurred during ∼200 My of mammalian evolution. In this paper, we describe the computational reconstruction of chromosome organization for seven ancestral genomes leading to human, including the ancestor of all placental mammals. The evolutionary history of chromosome rearrangements that occurred from the time of the eutherian ancestor until the human lineage is revealed in detail. Our results provide an evolutionary basis for comparison of genome organization of all eutherians, and for revealing the genomic origins of lineage-specific adaptations.

Author contributions: J.K., M.F., D.M.L., J.M., and H.A.L. designed research; J.K., M.F., L.A., B.C., D.M.L., J.M., and H.A.L. performed research; J.K., M.F., L.A., B.C., D.M.L., J.M., and H.A.L. analyzed data; and J.K., M.F., D.M.L., J.M., and H.A.L. wrote the paper.

Reviewers: W.J.M., Texas A&M University; and P.A.P., University of California, San Diego.

The authors declare no conflict of interest.

[1]J.K. and M.F. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: dmlarkin@gmail.com, jianma@cs.cmu.edu, or lewin@ucdavis.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1702012114/-/DCSupplemental.

The multiple genome rearrangements and ancestors (MGRA) algorithm (15) uses multiple breakpoint graphs based on SFs in descendant species to infer the ancestral order of SFs, and ANGES (ANcestral GEnomeS) (16) uses "consecutive one property" to cluster and order SFs in a target ancestor. However, these methods have been used to reconstruct just a small number of ancestral mammalian genomes, primarily because there are a very limited number of chromosome-scale whole-genome assemblies (4, 12). Furthermore, it has not been shown whether these existing algorithms for reconstructing chromosome organization are suitable for fragmented assemblies produced by next-generation sequencing technologies.

Examples of mammalian genome reconstructions reveal the limitations of earlier datasets. Murphy et al. (17) applied MGR to human, cat, cow, and mouse genome maps and assemblies to reconstruct the chromosome organization of the boreoeutherian ancestor, which lived ~97.5 Ma (9). Subsequently, the boreoeutherian, ferungulate, carnivore, and other ancestral genomes were reconstructed using MGR, combining physical maps and sequence information from eight species representing five mammalian orders (4). Twenty-three pairs of autosomes plus sex chromosomes were predicted for the boreoeutherian ancestor, but sequence coverage as measured against the human genome was only about 50% (4), resulting in limited definition and accuracy of both large-scale and fine-scale (<1.0 Mbp) chromosome rearrangements. In a later study (3), inferCARs was used to reconstruct continuous ancestral regions of the boreoeutherian ancestor that were generally consistent with chromosome painting results, but the reconstruction was limited and coarse because of the small number of descendant species used. In addition, there were studies using genes as markers to reconstruct the order and orientation of SFs in the boreoeutherian ancestor (e.g., ref. 18), but it is unclear how much gene-based reconstruction represents the ancestral reconstruction using whole-genome sequencing data. Therefore, although these recent results were an improvement over earlier work, missing information from other mammalian orders and use of low-resolution maps contributed to the reduced coverage, thus limiting the potential usefulness of the reconstructions for evolutionary and functional analysis.

Despite some recent improvements in reconstruction algorithms (3, 14–16), the field has been more or less stagnant for the past decade because of the paucity of new genome assemblies suitable for ancestral reconstructions. In this paper, we introduce a method, called DESCHRAMBLER, which uses SFs constructed from whole-genome comparisons of both high-quality chromosome-scale and fragmented assemblies. The method is an extension of the algorithm for reference-assisted chromosome assembly (RACA) (19), which implements a probabilistic framework to predict adjacencies of SFs in a target species. DESCHRAMBLER has the flexibility to handle chromosome-level and scaffold assemblies, and is scalable to accommodate a large number of descendant species. In the present study, we applied DESCHRAMBLER to sequenced genomes of 21 species that included representatives of 10 eutherian orders. Results reveal a detailed picture of chromosome rearrangements that occurred during ~105 My of eutherian evolution.

Results

Chromosome Reconstruction for Seven Eutherian Ancestors of *Homo sapiens*. The chromosome organizations of seven common ancestors in the lineage leading to human were reconstructed using genome assemblies of 19 extant eutherian species and two outgroup species, one a marsupial and one a bird (*SI Appendix*, Table S1). Genomes were selected on the basis of their availability in public databases, quality of genome assembly, and taxonomic order (*Materials and Methods*, Fig. 1, and *SI Appendix*, Table S1). The set of species contains representatives of 10 orders of eutherian mammals: primates (human, chimpanzee, orangutan, rhesus, and marmoset), rodentia (mouse, rat, and guinea pig), lagomorpha (pika), cetartiodactyla (cattle, goat, and pig), perissodactyla (white rhinoceros and horse), carnivora (dog), eulipotylpha (shrew), proboscidea (elephant), sirenia (manatee), and afrosoricida (tenrec), and two outgroup species to eutheria (opossum and chicken). Among the 21 genome assemblies, 14 were chromosome-level and the remaining 7 were assembled as sequence scaffolds with N50 ranging from 14.4–46.4 Mbp. The number of scaffolds in fragmented assemblies ranged from 2,352 (elephant) to 12,845 (shrew). Total sequenced genome size varied from 1 Gbp (chicken) to 3.5 Gbp (opossum) (*SI Appendix*, Table S1). For reconstruction of ancestral chromosomes, the human genome was used as the reference for alignments because of the relative quality of the assembly, and because we focused reconstructions on the evolution of lineages leading to human. Two resolutions (500- and 300-Kbp minimum breakpoint distance in the human genome) were selected to create the SFs that were used by the DESCHRAMBLER reconstruction algorithm as input. Herein, we made our interpretations on the basis of 300-Kbp resolution; results at 500 Kbp (*SI Appendix*, Tables S4–S6) were used for comparison to help resolve discrepancies with FISH data and to better understand differences in breakpoint rates along the different lineages.

The number of reconstructed ancestral chromosome fragments (RACFs) ranged from 30 in the common ancestor of great apes, to 35 in the common ancestor of human and chimpanzee (Table 1). The SFs of each ancestor were defined using only the descendant species from the corresponding ancestral node (with the rest as outgroup species). Therefore, SFs of the more ancient ancestors contained homologous genomic regions from a larger number of descendant species than the more recent ancestors. This accounts for the greater number of smaller SFs and the smaller total size of RACFs in more ancient ancestors. However, the difference in RACF sizes among ancestors was minimized by allowing missing coverage in SF definitions for a small number of descendant genomes (*Materials and Methods*). The RACFs of the simian, catarrhini, great apes, and common ancestor of human and chimpanzee cover more than 90% of the human genome, whereas the eutherian, boreoeutherian, and euarchontoglires RACFs each cover more than 80% of the human genome.

Comparison with Existing Ancestral Genome Reconstruction Algorithms. The performance of DESCHRAMBLER was first compared with three existing tools for ancestral chromosome reconstruction, ANGES (16), inferCARs (3), and MGRA (15), using simulation evaluation (*Materials and Methods*). The simulation data were created to allow missing sequences from some species' genomes in our evaluation. For predicted ancestral adjacencies, DESCHRAMBLER was superior to the three existing tools, with the agreement scores 99.66% for the boreoeutherian ancestor and 99.90% for the euarchontoglires ancestor (*SI Appendix*, Table S8). For the number of reconstructed ancestral chromosomes, DESCHRAMBLER's reconstruction was the closest to the true numbers in the simulation data (20.04 for boreoeutherian and 20.10 for euarchontoglires ancestors; the true numbers used in the simulation are 20 and 19 for boreoeutherian and euarchontoglires ancestors, respectively).

The reconstruction results for seven eutherian ancestral genomes were then compared using ANGES (16), inferCARs (3), MGRA (15), and DESCHRAMBLER. For a fair comparison, the same sets of SFs were used as input to the above three tools, and the predicted adjacencies of SFs in the seven target ancestors were compared. The number of RACFs obtained with DESCHRAMBLER ranged from 30 in the common ancestor of great apes to 35 in the common ancestor of human and chimpanzee (*SI Appendix*, Table S2). The other three tools produced larger numbers of RACFs for the eutherian ancestor, which are apparently because of the increased number of descendant species with scaffold assemblies having

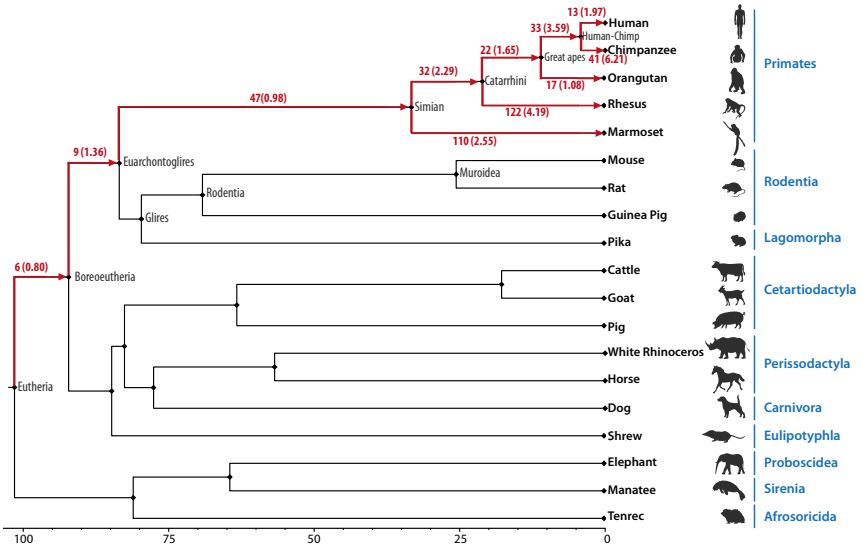

Fig. 1. Phylogenetic tree of descendant species and reconstructed ancestors. The numbers on branches from the eutherian ancestor to human are the numbers of breakpoints in RACFs, with breakpoint rates (the number of breakpoints per 1 My) in parentheses. The unit of time of branch lengths is 1 My. The details of the genome assemblies of descendant species and the classification of rearrangements are shown in *SI Appendix*, Tables S1 and S3, respectively.

unclear definition of chromosome ends (*SI Appendix*, Table S1). Other than for the eutherian ancestor, ANGES consistently produced the fewest RACFs, whereas MGRA produced very large numbers of RACFs, particularly for the most distant common ancestors to human. Comparison of predicted SF adjacencies among the four tools showed that the results obtained with DESCHRAMBLER were highly similar to those of ANGES and inferCARs (Jaccard similarity coefficient > 0.8) (*SI Appendix*, Fig. S2). Results from DESCHRAMBLER and inferCARs were the most similar for all of the seven reconstructed ancestral genomes, whereas the greatest discrepancies were found between MGRA and the other tools (*SI Appendix*, Fig. S2).

Comparison with FISH-Based Reconstructions of Ancestor Chromosomes. We compared the eutherian, boreoeutherian, and simian ancestral karyotypes determined by FISH (6, 8, 20) with those obtained using DESCHRAMBLER and three additional tools (see *Materials and Methods* for details). In this evaluation, we focused on interchromosomal rearrangements using human chromosomes as a reference. For example, there are seven fusions of human chromosomes found in the eutherian and boreoeutherian ancestors, and two fusions of human chromosomes in the simian ancestor (Table 2). DESCHRAMBLER agreed with FISH data in 12 of 16 cases, thus outperforming the other three tools. In three of four cases where FISH data and DESCHRAMBLER disagreed, DESCHRAMBLER partially predicted the interchromosomal rearrangements. For example, in the reconstructed chromosomes of the eutherian ancestor, the descendant homologs HSA8p and parts of HSA4 were predicted to be fused by DESCHRAMBLER, but joining of HSA8p to another segment of what is now HSA4q was not detected (*SI Appendix*, Table S6). Similarly, in the reconstructed chromosomes of the eutherian and boreoeutherian ancestors, the descendant homologs HSA12pq and HSA22q were predicted to be fused by DESCHRAMBLER, but joining to what is now HSA10p was not detected. However, in the eutherian and boreoeutherian ancestral genomes, the fusion of HSA10p to 12pq-22q is weakly supported in FISH-based reconstructions (6). ANGES was the next best performer with 11 agreed cases. MGRA produced the lowest agreement with the FISH-based reconstructions because of the highly fragmented nature of its RACFs in the three ancestors used in this evaluation (*SI Appendix*, Table S2).

One large eutherian RACF produced by DESCHRAMBLER was not supported by FISH data. This RACF (see EUT1 in Dataset S1) joined what is now all of HSA4 and HSA13, and parts of HSA8 and HSA2. The organization of this large RACF partially agrees with the ancestral eutherian chromosome formed by what is now HSA8p and HSA4pq as predicted by chromosome painting (Table 2) (4). It is noteworthy that both eutherian ancestral adjacencies involving homologs of HSA8 and HSA2, and HSA2 and HSA13, have a high DESCHRAMBLER score (>0.999) and are spanned by one chromosome or scaffold in the Afrotherian and outgroup species. In addition, ANGES predicted the same ancestral configurations in the eutherian ancestor, whereas inferCARS split it into two RACFs (*SI Appendix*, Table S6). Therefore, there are multiple lines of evidence to support the EUT1 adjacencies in the eutherian ancestral genome, although there are discrepancies among the reconstruction methods and at different resolutions (*SI Appendix*, Table S6). Finally, the fusion of two ancestral chromosomes

Table 1. Statistics of the reconstructed ancestors (300-Kbp resolution)

Ancestor	No. of RACFs	Total size (Kbp)	Coverage (%)*	Maximum RACF (Kbp)	Minimum RACF (Kbp)	No. of SFs	Maximum SF (Kbp)	Minimum SF (Kbp)
Eutherian	32	2,467,725	81	386,409	523	2,404	8,322	523
Boreoeutherian	34	2,536,880	84	213,005	350	2,213	8,322	350
Euarchontoglires	33	2,671,496	88	221,686	317	1,646	13,092	317
Simian	33	2,752,920	91	226,255	1,079	618	40,460	1,079
Catarrhini	33	2,767,322	91	192,635	350	508	60,356	350
Great apes	30	2,784,232	92	193,721	355	301	96,716	355
Human–chimpanzee	35	2,809,400	93	194,693	325	174	110,079	325

*Percentage of sequence coverage against the human genome size (3,036,303,846 bp for autosomes and the X chromosome, including Ns).

Table 2. Comparisons of computationally reconstructed ancestral chromosomes with reconstructions made using Zoo-FISH or BAC-FISH

Ancestor	Interchromosomal event*	DESCHRAMBLER	ANGES	inferCARs	MGRA
Eutherian	4q-8p-4pq	−	−	−	−
	3–21	+	+	+	−
	14–15	−	+	−	−
	10p-12pq-22q	−	−	−	−
	16q-19q	+	+	+	−
	16p-7a	+	−	+	−
	12q-22q	+	+	+	+
Boreoeutherian	4–8p	+	+	+	+
	3–21	+	+	+	−
	14–15	+	+	+	−
	10p-12pq-22q	−	−	−	−
	16q-19q	+	+	−	−
	16p-7a	+	−	+	−
	12q-22q	+	+	+	+
Simian	3–21	+	+	+	−
	14–15	+	+	−	+
	Consistent	12	11	10	4
	Inconsistent	4	5	6	12

Chromosome fusions are indicated with a hyphen between chromosomes. A plus sign denotes that the fusion of chromosomes was detected in the ancestral genome. A minus sign denotes that the fusion of chromosomes was not detected in the ancestral genome. Sources of FISH data used for the comparisons: eutherian (8), boreoeutherian (6), and simian (20) reconstructed ancestral chromosome. "+" indicates that the adjacency detected by FISH was also detected by the algorithm. "−" indicates that the adjacency detected by FISH was not detected by the algorithm.

*Numbers represent human chromosome numbers, and "p" and "q" indicate the p-arm and q-arm, respectively. If a chromosome fragment does not perfectly match to the p- or q-arm, a letter is used based the order of the fragment on the chromosome.

homologous to HSA7 was predicted by bacterial artificial chromosome (BAC)-FISH to occur in the ancestral catarrhini genome (20), whereas DESCHRAMBLER placed it in the simian ancestor. High-confidence FISH-based chromosomal configurations in each ancestor were incorporated into the final reconstruction of ancestral genomes predicted by DESCHRAMBLER (*Materials and Methods* and Dataset S1; see also Dataset S2 for the number of bases and fraction of reconstructed ancestral, descendant, and outgroup genomes found in fully conserved eutherian ancestor chromosomes and those affected only by intrachromosomal rearrangements of eutherian ancestor chromosomes, and Dataset S3 for mapping between original scaffold identifiers of an extant species with a scaffold assembly and new identifiers used in the Evolution Highway comparative chromosome browser).

Evolutionary Breakpoints and Chromosome Rearrangements. At 300-Kbp resolution, we detected 162 chromosomal breakpoints that occurred during 105 My of mammalian evolution, from the eutherian ancestor's genome to the human genome (Fig. 1 and *SI Appendix*, Table S3). Six breakpoints occurred on the branch from eutheria to boreoeutheria, which correspond to three fissions, one inversion, and one complex rearrangement. There were nine breakpoints in the euarchontoglires ancestor's genome in comparison with the boreoeutherian ancestor's genome, resulting in one fusion, two fissions, three inversions, and two complex rearrangements. The number of rearrangements increased during evolution from the euarchontoglires ancestor to the more recent ancestors. Among them, the largest number of rearrangements ($n = 38$) occurred from the euarchontoglires ancestor to the simian ancestor, producing 47 evolutionary breakpoints. Mostly inversions and complex rearrangements were observed during the evolution of the eutherian ancestor to human, whereas fusions and fissions were less prevalent.

We next examined the number of chromosome breakpoints in terms of divergence time from common ancestors (*SI Appendix*, Tables S3–S5). At 300-Kbp resolution, the lowest breakage rate was 0.80/My, occurring from the eutherian ancestor to the boreoeutherian ancestor [false-discovery rate (FDR) $P < 0.05$]. The breakage rate was lower on the branch from the euarchontoglires ancestor to the simian ancestor (0.98/My, FDR $P < 0.05$), and higher on the branch from the common ancestor of great apes to the common ancestor of human and chimpanzee (3.59/My, FDR $P < 0.10$). During the evolution of primate ancestors to extant primate genomes, breakage rates in the lineages leading to rhesus and chimpanzee were significantly higher than along other branches (4.19/My, FDR $P < 0.05$, and 6.21/My, FDR $P < 0.05$, respectively) and was lower in the lineage leading to orangutan (1.08/My, FDR $P < 0.05$). We then compared the results obtained at 300-Kbp resolution with those obtained at 500-Kbp resolution (*SI Appendix*, Tables S4 and S5). Although breakage rates were consistently lower at 500-Kbp resolution, levels of statistical significance were consistent for all comparisons except for orangutan.

We then investigated possible causes of the differences in chromosome breakage rates at 300- and 500-Kbp resolution. The number of SFs below the 500- and 300-Kbp thresholds were compared by counting the number of SFs at 300-Kbp resolution corresponding to each branch and then correlating these results with the amount of breakpoint increase (*SI Appendix*, Fig. S3). There was a high linear correlation between the two measures in terms of both the absolute number and the fraction of small SFs. Thus, the increase in breakpoints was mostly attributed to smaller scale rearrangements between 300 and 500 Kbp because inversions and complex rearrangements were observed in higher numbers at 300-Kbp resolution (*SI Appendix*, Tables S3 and S4).

Evolutionary History of the Eutherian Ancestor's Genome. A complete summary of the evolutionary history of each reconstructed ancestral eutherian chromosome is presented in Fig. 2, *SI Appendix*, *Supplementary Text*, and Dataset S4. An integrated summary

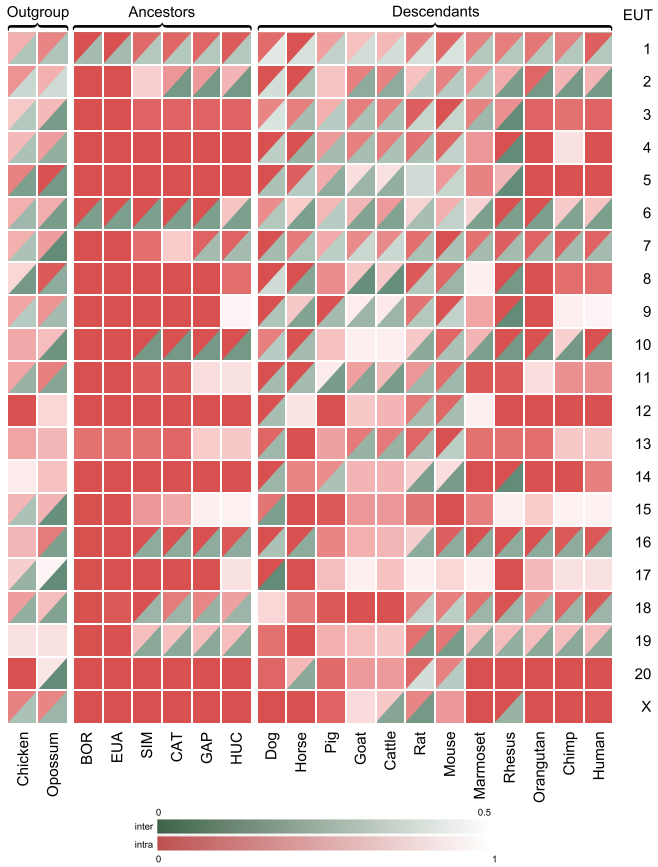

Fig. 2. Summary visualization of rearrangements of ancestral eutherian chromosomes in chromosomes of reconstructed descendant ancestors, and extant descendant and outgroup species. Solid red-brown blocks indicate eutherian chromosomes that were maintained as a single synteny block, with shades of the color indicating the fraction of the chromosome affected by intrachromosomal rearrangements (lightest shade is most affected). Split blocks demarcate eutherian chromosomes that were also affected by interchromosomal rearrangements: that is, fissions and translocations. Shades of green in split blocks indicate the fraction of an ancestral chromosome affected by translocations or fissions (lightest shade is most affected), and the shades of red-brown indicate the fraction of eutherian chromosomes affected by intrachromosomal rearrangements measured and summed for all SFs. The heatmap shows the color shades used to represent different fractions of outgroup, descendant ancestors' and extant species chromosomes affected by interchromosomal (shades of green) or intrachromosomal (shades of brown-red) rearrangements. Because of undefined positions of ancestral centromeres, the intrachromosomal rearrangements are measured relative to the prevailing orientation of SFs within each outgroup or descendant chromosome and therefore the fraction of intrachromosomal rearrangements cannot exceed 50%. As it follows from the heatmap, dark shades indicate high level of conservation with the ancestral chromosome and light shades of the same color indicate high level of rearrangements. BOR, boreoeutherian ancestor; CAT, catarrhini ancestor; EUA, euarchontoglires ancestor; EUT, eutherian ancestor; GAP, great apes ancestor; HUC, human–chimp ancestor; SIM, simian ancestor.

of results with emphasis on chromosome rearrangements in the lineage leading to human is presented below.

Comparative analysis of reconstructed chromosomes of the eutherian ancestor revealed that a majority were highly stable in both the boreoeutherian and euarchontoglires ancestral genomes (Fig. 2 and Dataset S4). The exceptions to this pattern were the descendant homologs of EUT1 and EUT6, which were separated by fission into three and two chromosomes, respectively, in the boreoeutherian ancestor's genome. Another exception was the descendant homolog of EUT13, which gained an ∼10-Mbp inversion in the boreoeutherian ancestor. The descendant homolog of EUT18 gained large inversions in the euarchontoglires ancestor's genome but was maintained as a single chromosome (Dataset S4).

In the reconstructed simian ancestor's genome, 15 of 21 eutherian ancestor chromosomes were conserved as a single chromosome, of which 5 underwent intrachromosomal rearrangements (Fig. 2). Among the 15 conserved full-chromosome syntenies, 13 were conserved as single chromosomes or chromosome blocks within larger chromosomes in human, chimpanzee, and orangutan, the largest number for any extant species. Two descendant homologs of eutherian ancestor chromosomes with synteny conserved in the simian ancestor's genome underwent interchromosomal rearrangements later in the primate lineage; EUT2 (a fission in the catarrhini ancestor) and EUT7 (a fission in the ancestor of great apes) (Dataset S4). In comparison, 12 eutherian ancestor chromosomes have homologs in pig with completely conserved synteny, the greatest number for any extant nonprimate species in our analysis; however, 11 of these underwent intrachromosomal rearrangements. The species with the fewest conserved chromosomes relative to the eutherian ancestor was mouse, with three.

No additional rearrangements in evolutionary stable eutherian ancestor chromosomes (i.e., those without internal rearrangements) were introduced in the reconstructed catarrhini ancestor genome compared with the simian ancestor. However, three descendant homologous chromosomes of the eutherian ancestor (EUT8, EUT9, and EUT17) underwent lineage-specific complex rearrangements in the human–chimpanzee ancestor (Fig. 2 and Dataset S4). We found six eutherian ancestral chromosomes (EUT4, EUT5, EUT12, EUT14, EUT20, and EUTX) that had no interchromosomal or intrachromosomal rearrangements during ∼98.4 My of evolution until the common ancestor of human and chimpanzee (Fig. 2 and Dataset S4). Among all extant species studied, orangutan was found to have the largest number of chromosomes ($n = 8$) that were completely conserved in SF order and orientation compared with homologs in the eutherian ancestor. In the human lineage, the descendant homolog of EUT14 underwent a large (∼12-Mbp) inversion (Fig. 3), whereas in chimpanzee its structure follows the ancestral eutherian configuration.

The largest number of intrachromosomal rearrangements in the primate lineage occurred in the evolution of EUT15 (Fig. 3), with the majority of these events dating to the simian ancestor, and additional rearrangements occurring later in the catarrhini and in the human–chimpanzee ancestor's genomes. Both the human and chimpanzee genomes exhibit additional rearrangements in the descendant homologs of EUT15 (HSA17 and PTR17, respectively). In contrast, EUT15 was found completely conserved in the mouse and horse genomes (Fig. 3), whereas the cattle and goat genomes contained just one large inversion in their descendant homologs of EUT15.

Although EUTX was highly conserved among primates, artiodactyl species had significant numbers of X chromosome inversions, whereas the order and orientation of EUTX SFs in horse (a perissodactyl) were conserved. There are small inversions and interchromosomal rearrangements observed in the X chromosomes of murid rodents, dog (a carnivore), cattle, and other lineages, but assembly errors cannot be ruled out as causing at least some of these apparent rearrangements.

Overall, 537.5 Mbp of the reconstructed eutherian ancestor's genome (21.8% of total eutherian genome size) lack both interchromosomal and intrachromosomal rearrangements, and an additional 798.5 Mbp (32.4% of total genome size) of the eutherian ancestor chromosomes had intrachromosomal but no detectable interchromosomal events during evolution to the human genome (Dataset S2). The remaining 45.8% was found in reconstructed eutherian chromosomes that underwent intrachromosomal and interchromosomal rearrangements. This compares to 3.8% and 2.6% maximum eutherian ancestor genome coverage observed for

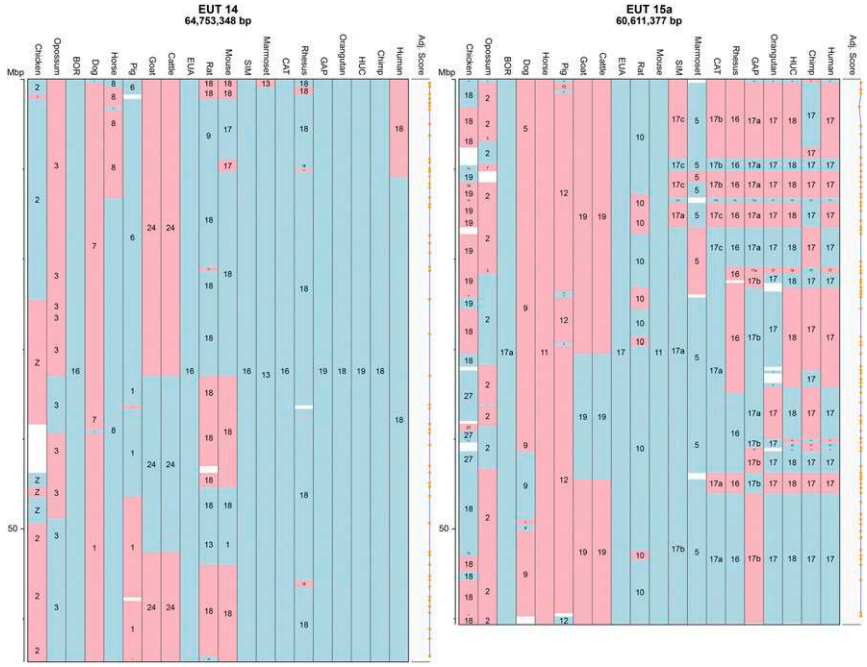

Fig. 3. Two examples of eutherian ancestor chromosomes with dramatically different evolutionary histories in the primate lineage. Order and orientation of SFs overlaid on the reconstructed eutherian ancestor chromosomes are visualized using the Evolution Highway comparative chromosome browser (eh-demo.ncsa.illinois.edu/ancestors/). The eutherian chromosome number and its total length are given at the top of each ideogram. Only the main fragment of EUT15 (EUT15a) is shown for this comparison. Blue and pink colors represent orientation of blocks relative to the reference, with blue indicating the same orientation, and pink indicating the opposite orientation. Pink does not always indicate an inversion because the orientation of RACFs is randomly chosen during the reconstruction. Also, as in the case of dog for EUT14, numbering of nucleotides may begin from the opposite end of the chromosome. The number within each block represents a chromosome of a reconstructed ancestor (Dataset S1) or an extant species; a letter indicates a fragment of the chromosome. Adjacency scores computed with DESCHRAMBLER are shown in the right-most tracks. Letter codes of reconstructed ancestors are the same as given in the legend of Fig. 2. Only extant species with full chromosome-scale assemblies are shown. BOR, boreoeutherian ancestor; CAT, catarrhini ancestor; EUA, euarchontoglires ancestor; GAP, great apes ancestor; HUC, human–chimp ancestor; SIM, simian ancestor.

chromosomes with no interchromosomal or intrachromosomal rearrangements, and 36.5% and 7.0% maximum coverage for intrachromosomal-only rearrangements in artiodactyl and murid genomes, respectively (Dataset S2). Thus, compared with the reconstructed eutherian genome, the primate lineage tends to have a larger fraction of genomes in unrearranged SFs compared with other eutherian lineages.

Unassigned RACFs. DESCHRAMBLER produced two small chromosomal fragments, Un29 (1 Mbp) and Un30 (0.5 Mbp) that were not joined to any reconstructed chromosomes in the eutherian ancestor genome (Dataset S4). These fragments must have been produced by multiple independent rearrangements [i.e., reuse breakpoints (11)] in several mammalian clades. It is likely that in the lineage leading to primates these fragments were adjacent and located at the telomeric region of the EUT1 homolog. In the simian and later in the catarrhini ancestral genomes, several inversions separated Un29 and Un30, which are found about 10-Mbp apart on HSA1. Thus, independent chromosomal rearrangements apparently reorganized these fragments in artiodactyl, rodent, and perissodactyl lineages, indicating that these RACFs are bounded by highly dynamic intervals in eutherian chromosomes.

Discussion

Chromosomes of seven ancestral genomes along the 98.4 My lineage, from the ancestor of all placental mammals to the common ancestor of humans and chimpanzees, were reconstructed using the DESCHRAMBLER algorithm. Seven of the extant species had subchromosomal, scaffold-level assemblies that were effectively used by DESCHRAMBLER to reconstruct ancestral chromosome fragments and to identify lineage-specific chromosome breakpoints. The reconstructions were made using genomes of extant species from 10 of 19 orders of eutherian mammals representing the Laurasiatheria, Afrotheria, and Euarchontoglires superorders. Although Xenarthra was not represented, species from three superorders permitted reconstruction of the eutherian, boreoeutherian, and euarchontoglires ancestor's chromosomes at high resolution compared with the earlier FISH-based reconstructions (6, 8, 20). The ancestral reconstructions far surpassed the quality of previous map and sequence-based reconstructions in terms of the number of descendant species included, coverage of ancestor genomes relative to the human genome, and the number of ancestors in the evolutionary path to the human genome (3, 4), thus providing novel insights into eutherian and primate genome evolution.

The choice of a reference genome is critical for the completeness of chromosome reconstructions because the reference is used as a backbone to find orthologous chromosomal regions in different species using whole-genome sequence alignment, and to construct SFs that are shared between species. It is noteworthy that our reconstruction algorithm itself does not bias toward any descendant genome, but the reference genome has an impact on the SFs that are used for the reconstruction. The human genome was used as a reference because it is considered to have the highest quality assembly among the mammals, and because all ancestors targeted for genome reconstruction were ancestral to human. In addition, assembly quality is also important for overall accuracy and completeness of the SFs. To reduce the complications in reconstruction introduced by extensively fragmented genome assemblies and misalignments, we selected species with assemblies that have N50 scaffold size > 14 Mbp and that could be aligned against more than 80% of the reference

human genome. Because we only used one reference genome in the present work for defining SFs, it is possible that some ancestral sequences that are not present in the human genome were omitted in the reconstructions. It would be useful to develop SF construction methods that consider multiple reference genomes, similar to what has been done for bacterial genomes (21). In addition, recent developments in long-read sequencing technologies (22), genome scaffolding (23–25), and comparative and integrative mapping (19, 26) produce higher quality assemblies that approach whole chromosomes. These methods are now cost-effective relative to creating high-density BAC maps, linkage maps, and radiation hybrid maps (12), and will be useful for providing higher-quality SFs that may greatly facilitate the understanding of chromosome evolution using ancestral genome reconstruction methods.

For ancestral genome reconstruction, DESCHRAMBLER takes into account clade-specific or species-specific insertions and deletions. If the SFs are constructed by requiring orthologous chromosomal regions from all descendant species, the genome of their common ancestor would not be well covered, especially when the genomes of the descendant species are highly diverged or the assemblies are incomplete. To address this issue, SFs were created without the above constraint of the inclusion of all orthologous genomic regions. Instead, all possible SFs were first created with a different number of genomic regions of descendant species, and then candidate SFs for each target ancestor were chosen by a parsimony algorithm based on the presence and absence of orthologous genomic regions in each descendant species. To take advantage of these new SFs, the reconstruction algorithm must be able to use them. Most existing algorithms, such as ANGES, inferCARs, and MGRA, were developed using the assumption of strict constraint on orthologous regions in SFs that orthologous regions from all descendant species must exist in an SF. However, DESCHRAMBLER is more flexible in using SFs when some of the species have deletions of genomic regions or there is missing data. This is one of the reasons why DESCHRAMBLER outperformed other existing tools in the reconstruction of the oldest (eutherian) ancestor.

After incorporating high-confidence FISH-based chromosomal configurations in each ancestor, we deduced an ancestral eutherian karyotype having 2n = 44 chromosomes (assuming a separate Y chromosome). This number is lower than FISH-based inferences of 2n = 46 (5, 6, 8, 27, 28), and is because of the reconstructed EUT1 (ascendant homolog of HSA13, HSA2, HSA4, and HSA8) and EUT6 (partially homologous to HSA7 and HSA10). Our results are in agreement with previous studies that used FISH-based and sequenced-based methods to deduce the ancestral boreoeutherian karyotype to have 2n = 46 chromosomes (3, 5, 6, 27, 28). We also deduced an ancestral catarrhini karyotype of 2n = 46, an ancestral great apes karyotype of 2n = 48, and 2n = 48 for the human–chimpanzee ancestor, which all agree with results from chromosome painting and BAC-FISH experiments (20, 28).

The major differences with FISH-based ancestral karyotype reconstructions for eutherian and boreoetherian karyotypes likely result from the incomplete set of mammalian orders included in our reconstruction dataset. For example, the lack of Xenarthra could cause DESCHRAMBLER to put a higher weight on the adjacencies observed in tenrec and outgroup genomes to reconstruct EUT1. Taking in to account the atypical outgroup mammalian opossum karyotype with 2n = 12, we cannot exclude the possibility that some of the adjacencies reconstructed by DESCHRAMBLER were introduced because of recurrent rearrangements formed in some ingroup and outgroup genomes. On the other hand, well-established ancestral adjacencies that were missed by DESCHRAMBLER (e.g., the HSA14-HSA15 fusion in eutherian RACFs) could result from inclusion of some highly rearranged ingroup genomes. Future ancestral karyotype reconstructions built with DESCHRAMBLER will highly benefit from inclusion of representative species of the nine mammalian orders not included in this work and additional sampling within previously studied taxa. The advantage of additional sampling is demonstrated by the results obtained with primate genomes.

In the simian ancestor (the ancestor of Old World and New World monkeys), we reconstructed an ancestral karyotype with 2n = 46 chromosomes. This number is lower than obtained with FISH-based methods, which inferred 2n = 48 (5, 28) or 2n = 50 (20). The main differences are SIM7 (homolog to HSA7) and SIM10 (homolog to HSA10), where DESCHRAMBLER created one ancestral chromosome for each chromosome, whereas FISH data consistently supported reconstruction of HSA7 and HSA10 each into two fragments (5, 20, 28). In summary, the diploid numbers of ancestor genomes deduced by DESCHRAMBLER were very similar to the results of previous reconstructions. Additional high-quality genome assemblies will help to resolve remaining discrepancies.

We have demonstrated that each eutherian chromosome has a unique evolutionary history in the different mammalian lineages, and that many ancestral eutherian chromosomes were stable in descendant lineages, with relatively few large-scale rearrangements in the ancestral genomes leading to human. Among the primate species included in the analysis, more than 100 putative breakpoints were detected during evolution from the simian ancestor to marmoset, and from the catarrhini ancestor to rhesus (*SI Appendix*, Table S5), thus indicating an accelerated rate of evolution in these nonhuman primates during the past 43 My (see below). Although the time from the great ape ancestor to the common ancestor of human and chimpanzee has a relatively short branch length (9.2 My), there were 14 inversions and 10 complex rearrangements (i.e., a combination of inversions and putative transpositions) assigned to that branch, which also gives the highest breakpoint rate on that particular lineage. For comparison, we looked at the breakpoint rates from these ancestral nodes along the lineages to other nonhuman descendant species. We found that the branch from the great ape ancestor to orangutan has the lowest breakpoint rate (1.08/My) compared with other branches (*SI Appendix*, Table S5), and the result was consistent when we used 500 Kbp as the SF resolution. This finding suggests an overall higher chromosomal rearrangement rate on the branch from the great ape ancestor to the ancestor of human and chimpanzee, but a much slower rate from the great ape ancestor to orangutan. In addition, our results refined the previously reported comparison between the orangutan genome and human–chimpanzee ancestor (29), where 40 rearrangement events were identified at 100-Kbp resolution. Regardless of varying rates of rearrangements within different primate lineages, comparison with other mammalian orders included in this work indicates that the primate ancestor and several descendant species' genomes contain the largest fraction of descendant homologs of eutherian ancestor chromosomes either totally conserved or affected by intrachromosomal rearrangements only. This finding suggests that the small insectivorous and scansorial common ancestor of all existing placental mammals (30) had chromosome structures highly resembling those of some contemporary primates (e.g., orangutan and human).

The breakpoint rate in the lineage leading to chimpanzee was almost threefold higher than in the lineage leading to human at 300-Kbp resolution (6.21/My and 1.97/My, respectively), and more than fourfold greater at 500-Kbp resolution (*SI Appendix*, Tables S3–S5). These results indicate true differences in the rate of chromosome evolution in the lineages leading to humans and chimpanzees. Interestingly, the number and the rate of breakpoints in orangutan chromosomes remained constant for the two breakpoint resolutions, indicating few if any rearrangements that are in the 300- to 500-Kbp range in this species. On the basis of

the above analyses we recommend that ≥300-Kbp resolution be used to analyze chromosomal rearrangements that affect the synteny and order of homologous sequences to avoid most false breakpoints introduced by assembly errors, as well as segmental duplications and copy number variants. However, the use of multiple breakpoint resolutions can be advantageous when the goal is to draw more accurate and comprehensive conclusions from many descendant species to reveal the interplay between large-scale rearrangements and finer-resolution genomic changes (including duplications). Therefore, there should be additional efforts to enhance reconstruction algorithms to effectively aggregate results at different resolutions of breakpoint intervals.

The analysis of chromosome evolutionary breakpoint rates yielded results that are generally consistent with Murphy et al. (2), who found slow rates of chromosome evolution in mammals before the K–P boundary, which corresponds to the massive extinction event that led to the disappearance of the dinosaurs (except for birds) and the eventual rise of mammals. We also found an accelerated rate of chromosome rearrangements in primate ancestors, specifically along the branch leading to the common ancestor of humans and chimpanzees. The significance of these findings is unclear, but might be related to differences in genomic architecture, repetitive elements, and changes in the environment that are known to cause chromosome rearrangements (11). Assembly errors may also cause an increase in the apparent rate of rearrangements, and these must be excluded before drawing conclusions. One way to approach this problem is to compare breakpoint rates at different resolutions. Fewer breakpoints are expected at lower resolution, but the relative differences in rates should be stable. Consistent with this expectation, we found a linear correlation between the number of SFs <500 but >300 Kbp and the number of breakpoint differences at 300 and 500 Kbp (*SI Appendix*, Fig. S3). From additional analysis, we also observed that the small SFs contributed to creating rearrangements involving inversions and other complex rearrangements (*SI Appendix*, Tables S3 and S4). Breakpoints generated by rearrangements of these smaller SFs are either the footprint of bona fide structural rearrangements, or they may be artifacts produced by misassembled sequences. For example, previous studies revealed problems in the rheMac2 assembly version of the rhesus genome (31–33), which is one of the species showing a large discrepancy of the number and the rate of breakpoints at the two resolutions. Even though we used a more recent version of the rhesus genome (rheMac3), it is not clear whether all of the assembly problems in the previous version were completely fixed.

The reconstructed events of chromosome evolution in multiple ancestral genomes leading to human permitted assignment of breakpoints to different branches in the phylogeny. Such information can be useful for further analysis of the potential functional roles of chromosomal rearrangements in eutherian evolution. Earlier work reported an association between evolutionary breakpoints and gene functions that may contribute to lineage- and species-specific phenotypes (11, 34). More recently, such association analysis has been extended to understanding the relationship between chromosome rearrangements and noncoding function elements of the genome, such as open chromatin regions (18). In the present study, we found two small RACFs of the eutherian ancestor (Un29 and Un30) that were not assigned to specific ancestral chromosomes because of the fact that these two fragments were flanked by breakpoint regions with independent reuse in different eutherian lineages. If we examine the gene content within these EBRs using the human genome as a reference, we find them to contain multiple paralogs of zinc finger and olfactory receptor genes, which have been found previously to be enriched within EBRs (11, 35), are associated with adaptive evolution (36, 37), and may promote rearrangements by nonallelic homologous recombination (e.g., ref. 38). Specifically, the fragment Un29 is flanked by zinc finger genes ZNF678 and pseudogene ZNF847P at one end, and three histone genes (HIST3H3, HIST3H2BB, HIST3H2A) at the other. Among the other 17 genes found within Un29 are several gene family members, including *WNT3A* and *WNT9A*. It has been shown that small changes in expression of WNT genes can result in a radical alteration of body plan (39). In the human genome, Un30 is flanked by three zinc finger genes (*ZNF670*, *ZNF669*, *ZNF124*), one additional zinc finger gene (*ZNF496*), and three olfactory receptor genes (*OR2B11*, *OR2W5*, *OR2C3*). Because chromosome rearrangements are known to affect regulation of gene expression (40), these data suggest that reuse of evolutionary breakpoint sites near this fragment in multiple clades could be a contributor to producing new variation in gene content and gene expression. With additional mammalian genomes being sequenced, our genome reconstruction approach has the potential to provide the foundation for a more comprehensive evolutionary analysis to improve understanding of the relationship between genome rearrangements, functional elements (both coding and noncoding), and adaptive traits.

Reconstruction of the chromosomes of seven descendant genomes, from the eutherian ancestor to human, is an excellent example of what can be achieved by applying similar analysis to other clades. The recent advances in long-read technology and scaffolding techniques will enable more rapid production of assemblies that are suitable for accurate identification of lineage-specific breakpoints, which are the basis for high-quality ancestral chromosome reconstructions. Thus, in the near future, it will be possible to reconstruct genomes at the key nodes of all mammalian lineages, and to explore the nature of chromosome rearrangements that occurred during more recent radiations. As previously shown, karyotypes, physical maps, and whole-genome sequences with precise locations of centromeres and telomeres also add important information for understanding chromosome evolution, and for understanding the relationship between chromosome rearrangements, EBRs, cancers, and inherited human diseases (2, 41). Together with improved tools for aligning, comparing, and visualizing large numbers of genomes, these new chromosome-scale assemblies will offer unparalleled opportunities to study the mechanisms and consequences of chromosome rearrangements that have occurred during mammalian evolution. With efforts such as those to sequence 10,000 vertebrate genomes (42), it will be possible to extend reconstructions deeper into evolutionary time, and thus provide a more detailed picture of chromosome evolution in other vertebrate classes. Ultimately, it should prove possible to determine the ancestral eukaryote chromosome organization, and to create a new chromosome nomenclature system that is based on evolutionary principles.

Materials and Methods

Data. The pairwise genome sequence alignments (chains and nets) among 21 genome assemblies using the human genome as reference were downloaded from the UCSC Genome Browser (43) or directly constructed by using an alignment pipeline based on lastz (44) with the chain/net utilities from the UCSC Genome Browser. The genomes used were: human (*Homo sapiens*, GRCh37/hg19), chimpanzee (*Pan troglodytes*, CSAC 2.1.4/panTro4), orangutan (*Pongo pygmaeus abelii*, WUGSC 2.0.2/ponAbe2), rhesus (*Macaca mulatta*, BGI CR_1.0/rheMac3), marmoset (*Callithrix jacchus*, WUGSC 3.2/calJac3), mouse (*Mus musculus*, GRCm38/mm10), rat (*Rattus norvegicus*, RGSC 5.0/rn5), guinea pig (*Cavia porcellus*, Broad/cavPor3), pika (*Ochotona princeps*, OchPri3.0/ochPri3), cattle (*Bos taurus*, Baylor Btau_4.6.1/bosTau7), goat (*Capra hircus*, CHIR_1.0/capHir1), pig (*Sus scrofa*, SGSC Sscrofa 10.2/susScr3), white rhinoceros (*Ceratotherium simum*, CerSimSim1.0/cerSim1), horse (*Equus caballus*, Broad/equCab2), dog (*Canis lupus familiaris*, Broad CanFam3.1/canFam3), shrew (*Sorex araneus*, Broad/sorAra2), elephant (*Loxodonta africana*, Broad/loxAfr3), manatee (*Trichechus manatus latirostris*, Broad v1.0/triMan1), tenrec (*Echinops telfairi*, Broad/echTel2), opossum (*Monodelphis domestica*, Broad/monDom5), and chicken (*Gallus gallus*, ICGSC Gallus_gallus-4.0/galGal4). The tree topology of these 21 species was based on the tree used to align 45 vertebrate genomes with human in the

UCSC Genome Browser, and branch lengths were estimated based on TimeTree (9).

The main criterion for choosing assemblies was to have the maximum representation of mammalian orders. We also used a cut-off of scaffold N50 >14 Mbp for fragmented assemblies, and a minimum of 80% coverage of pair-wise alignment to the human genome. These thresholds were established to: (*i*) maximize the coverage of the reconstructed ancestral karyotype, (*ii*) minimize the number of RACFs obtained, and (*iii*) reduce the chances of EBRs being found in between scaffolds. The cut-off date for a genome assembly to be included in the analysis was May 2014. At that time, the next assembly with the largest N50 after manatee (*SI Appendix*, Table S1) was for the hedgehog (N50 = 3.3 Mbp), which we found was too fragmented to produce reliable reconstruction results. Thus, the scaffold N50 threshold was chosen empirically.

Ancestral Genome Reconstruction Algorithm. We developed a method, called DESCHRAMBLER, to reconstruct the order and orientation of SFs in eutherian ancestral genomes. The workflow of the method is shown in *SI Appendix*, Fig. S1. The algorithm starts with the construction of SFs. Using a chromosome evolution model-based probabilistic framework, DESCHRAMBLER computes the probabilities of pairs of SFs being adjacent in a target ancestor based on the order and orientation of SFs in descendant as well as outgroup species. The SFs and their degree of adjacency in the target ancestor are next represented as a graph, which is used to estimate the most likely paths of SFs. The paths represent the order and orientation of SFs in the target ancestor. DESCHRAMBLER does not generate the nucleotide sequence of a target ancestor as other reconstruction tools, such as ANGES (16), MGRA (15), and inferCARs (3). Breakpoint regions, which are genomic regions flanking SFs, are also not the part of reconstruction results. However, they can be easily extracted by using the coordinates of terminal nucleotides of SFs of individual species. Details of each step are presented below.

Construction of SFs. For each ingroup species, genomic blocks, which are matched to the nets of pairwise alignments with a reference, were mapped on reference genome sequences. The nets of length greater than a given threshold (resolution) were used, and colinear genomic blocks were merged together. After finishing this step for every ingroup species, the reference genome sequences together with the mapped genomic blocks of the other species were split at the boundaries where there were breaks in genomes of at least one species. Then aligned genomic blocks of outgroup species were added to each fragment, resulting in SFs. Not all SFs have genomic blocks from all ingroup species, and therefore not all SFs were used in reconstruction. The SFs were used in reconstruction if the genomic blocks in the SF were predicted to share a common ancestral block in a target ancestor by using a parsimony algorithm that minimizes the number of state changes in intermediate ancestors to account for the presence and absence of blocks in extant species. By convention, we use the term "syntenic fragment" rather than "homologous synteny block" throughout this paper because the former differentiates the use of fragmented assemblies from the chromosome-scale assemblies used in previous studies (4, 11).

Computation of SF Adjacency Probabilities in a Target Ancestor. Given input SFs, their order and orientation in each ingroup and outgroup species are collected, which are used as the SF adjacency information in extant species. The probabilities of pairs of these SFs being adjacent in a target ancestor are computed from their adjacencies in extant species based on the probabilistic framework used in the RACA algorithm (19). The basic idea of the probabilistic framework is to calculate the posterior probability of pairs of SFs b_i and b_j being adjacent in the target ancestor by multiplying two posterior probabilities: b_i precedes b_j, and b_j succeeds b_i. The two posterior probabilities were calculated by using the Felsenstein's algorithm for likelihood (45) and the extended Jukes-Cantor model for breakpoints (46). More details can be found in Kim et al. (19).

Prediction of the Order and Orientation of SFs in a Target Ancestor. The probabilities of SF adjacencies in a target ancestor are used to construct a SF graph $G(V, E)$, which is an undirected graph with a set of vertices V representing SFs, and a set of edges E connecting vertices whenever there is an adjacency probability between two vertices. Each SF is expressed by using two vertices representing the head and tail of a SF. This is required because one SF can be connected to either the head or tail of another SF. Each edge has a weight representing the probability of adjacency between two connected vertices, and the head and tail vertices of the same SF always have the highest probability, 1.0. From the constructed SF graph, a greedy algorithm is used to predict the order and orientation of SFs in the target ancestor by incrementally merging two adjacent SFs according to the descending order of their edge weights, which is followed by the construction of lists of adjacent SFs. All SF adjacencies with a probability >0 were used in the reconstruction for seven eutherian ancestors.

Refinement of Predicted SF Adjacencies. Weak SF adjacencies, which are (*i*) supported by just one ingroup species without any support from outgroup species or (*ii*) not supported by any ingroup species, are split. Then among the collection of lists of adjacent SFs, any two lists $L_1(a_1, \ldots, a_n)$ and $L_2(b_1, \ldots, b_m)$, where the adjacency between two SFs a_n and b_1 has a weight and is unambiguously supported by the parsimony algorithm by considering their adjacencies in descendant species, are merged to create a new list of adjacent SFs $L_{12}(a_1, \ldots, a_n, b_1, \ldots, b_m)$. This process repeats until no newer list of SFs is created. We note that L_1 and L_2 can be merged by four different ways (L_1 L_2, L_1 -L_2, -L_1 L_2, and -L_1 -L_2, where the "-" symbol represents a reversal of a list). Therefore, if there is more than one way to meet the above criteria, the one with the maximum adjacency weight is chosen. We note that DESCHRAMBLER and RACA (19) are similar in the sense that they calculate and use the probabilities of SF adjacencies to order and orient SFs. However, the target of prediction is different (an ancestor for DESCHRAMBLER, and an extant species for RACA). In addition, only DESCHRAMBLER has the refinement step described above, and can handle SFs with missing sequences from some species' genomes.

Many of the RACFs initially reconstructed using DESCHRAMBLER (and with the other tools) are subfragments of chromosomes. For example, the number of RACFs in each of the seven ancestral genome reconstructions is larger than 30 (Table 1), whereas the estimated number of chromosomes of those ancestors is 23 or 24 (5, 6, 8, 20, 28, 47). Chromosome fragmentation is caused primarily by large repetitive regions around centromeres and other regions of chromosomes that are difficult to bridge in assemblies that do not have an underlying genetic or physical map. The final step of the reconstruction to chromosome level was the semiautomated reordering of RACFs of each ancestor on the basis of their ancestral configuration predicted from FISH data. To accomplish this, we collected reconstructed karyotypes of ancestral genomes predicted by FISH experiments from the literature and used those as a standard (5, 6, 8, 20, 28, 47). The final reorganization of RACFs into chromosomes was done by ordering RACFs based on the correspondence of FISH-based definitions of interchromosomal events. Orientation of these RACFs was done to minimize differences with human chromosome orientation (Dataset S1).

Identification of Evolutionary Breakpoints and Chromosome Rearrangements. Analysis at 300- and 500-Kbp resolutions can identify breakpoints caused by translocations, inversions, fissions, fusions, deletions, insertions, and transpositions involving SFs of size above these thresholds. Apparent rearrangements involving SFs at higher resolution are possible with DESCHRAMBLER, but at resolutions less than 300 Kbp, presence or absence of breakage in synteny can be affected by assembly errors, alignment artifacts, segmental duplications, and copy number variants, leading to an overestimation of the number of chromosome rearrangements. Thus, these algorithmic thresholds yield a conservative definition of evolutionary breakpoints that capture most of the true chromosomal rearrangements that have occurred during evolution (see below).

Reconstructed ancestral genomes obtained using DESCHRAMBLER are more fragmented than what has been known in part because of scaffold assemblies of descendant species where the exact tips of chromosomes are not known, and in part because of ambiguous cases resulting from insufficient evidence of adjacency. Therefore, RACFs created by DESCHRAMBLER were first reorganized by referring to FISH-based reconstruction results (5, 6, 8, 20, 28, 47), which show large-scale organization of ancestral chromosomes. Then the reorganized RACFs of parent and child ancestors on each branch in a phylogenetic tree were compared with infer the history of the changes of RACFs from the parent to the child ancestor. This process was repeated for branches from the eutherian ancestor to human, and different types of chromosome rearrangements, such as fissions, fusions, inversions, and complex rearrangements (i.e., a combination of inversions and putative transpositions) were identified.

The reconstructed chromosomes of each ancestor were visualized using the Evolution Highway browser (eh-demo.ncsa.illinois.edu/ancestors/).

Comparison of Chromosome Rearrangement Rates. Rates of chromosome rearrangement (EBRs/My) were calculated using the number of EBRs detected for each phylogenetic branch divided by the estimated length of each branch (in My) of the tree (4). Only the ancestor rates and the rates on the branches leading to humans and other primates were included in the analysis. The

primate lineage was chosen for comparison of rearrangement rates because there is a very high-quality reference sequence (human) and it has the greatest number of represented species with chromosome-scale genome assemblies. We estimated rates of chromosome rearrangement at 300- and 500-Kbp resolution of SFs. The *t* statistics for each branch were obtained by calculating the difference between the rearrangement rate on the branch and the mean rate across all of the branches and then normalizing for the SE. *P* values were corrected by FDR using the p.adjust function from the R package (https://www.R-project.org).

Simulation-Based Evaluation. We evaluated the reconstruction methods using simulated datasets. Simulation data were obtained from Ma et al. (3), which consists of 50 datasets of seven ingroup species (human, chimpanzee, rhesus, mouse, rat, cattle, and dog) and two outgroup species (opossum and chicken). The simulation data were further modified to allow missing sequences from some species' genomes based on the rate of missing sequences estimated from real data used in our analysis (chimpanzee: 1.3%, rhesus: 2.0%, mouse: 0.9%, rat: 2.3%, cattle: 5.4%, dog: >0.4%). Each simulated dataset contains about 2,496 SFs. Three existing tools, ANGES (16), inferCARs (3), and MGRA (15), together with DESCHRAMBLER, were run by targeting two ancestors (boreoeutherian and euarchontoglires) and their performance was compared in terms of the agreement of predicted SF adjacencies with true adjacencies and the number of reconstructed ancestral chromosomes.

Comparison with Existing Tools. The reconstructed ancestors of DESCHRAMBLER were compared with results from three existing tools, ANGES (16), inferCARs (3), and MGRA (15). For fair comparison, the four tools were used to predict the adjacencies of the same set of SFs for ancestors, and the similarities and differences of their predicted adjacencies were measured by using the Jaccard index, which is calculated by the number of common adjacencies divided by the union of adjacencies between two sets of adjacencies predicted by two different tools. InferCARs was run with default parameters, and MGRA was run with three as the number of stages value along with other default parameters. The parameters used for ANGES are shown in *SI Appendix*, Table S7. For fair comparison the original reconstruction results obtained using DESCHRAMBLER, not the modified results based on the FISH data, were used.

Evaluation Using FISH Data. Interchromosomal rearrangements of human chromosomes referenced to computationally reconstructed ancestor chromosomes were identified and compared with reconstructions made using chromosome painting. The FISH-based reconstructions for the eutherian (8), boreoeutherian (6), and simian (20) ancestors were compiled from the literature.

Availability of Software and Datasets. The source code of DESCHRAMBLER and link to input and output files are available at https://github.com/jkimlab/DESCHRAMBLER.

ACKNOWLEDGMENTS. This work was supported by Ministry of Science, ICT & Future Planning of Korea Grant 2014M3C9A3063544 (to J.K.); Ministry of Education of Korea Grant 2016R1D1A1B03930209 (to J.K.); Rural Development Administration of Korea Grant PJ01040605 (to J.K.); Biotechnology and Biological Sciences Research Council Grants BB/K008226/1 and BB/J010170/1 (to D.M.L.); the Robert and Rosabel Osborne Endowment (to H.A.L.); National Institutes of Health Grant HG007352 (to J.M.); and National Science Foundation Grants 1054309 and 1262575 (to J.M.). This work was conducted as a contribution to the G10K Project.

1. White MJD (1969) Chromosomal rearrangements and speciation in animals. *Annu Rev Genet* 3:75–98.
2. Murphy WJ, et al. (2005) A rhesus macaque radiation hybrid map and comparative analysis with the human genome. *Genomics* 86:383–395.
3. Ma J, et al. (2006) Reconstructing contiguous regions of an ancestral genome. *Genome Res* 16:1557–1565.
4. Murphy WJ, et al. (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309:613–617.
5. Ferguson-Smith MA, Trifonov V (2007) Mammalian karyotype evolution. *Nat Rev Genet* 8:950–962.
6. Froenicke L (2005) Origins of primate chromosomes—As delineated by Zoo-FISH and alignments of human and mouse draft genome sequences. *Cytogenet Genome Res* 108:122–138.
7. Frönicke L, Wienberg J, Stone G, Adams L, Stanyon R (2003) Towards the delineation of the ancestral eutherian genome organization: Comparative genome maps of human and the African elephant (*Loxodonta africana*) generated by chromosome painting. *Proc Biol Sci* 270:1331–1340.
8. Ruiz-Herrera A, Farré M, Robinson TJ (2012) Molecular cytogenetic and genomic insights into chromosomal evolution. *Heredity (Edinb)* 108:28–36.
9. Murphy WJ, Eizirik E (2009) Placental mammals (Eutheria). *The Timetree of Life*, eds Hedges SB, Kumar S (Oxford Univ Press, New York), pp 471–474.
10. Elsik CG, et al.; Bovine Genome Sequencing and Analysis Consortium (2009) The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science* 324:522–528.
11. Larkin DM, et al. (2009) Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res* 19:770–777.
12. Lewin HA, Larkin DM, Pontius J, O'Brien SJ (2009) Every genome sequence needs a good map. *Genome Res* 19:1925–1928.
13. Groenen MA, et al. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491:393–398.
14. Bourque G, Pevzner PA (2002) Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Res* 12:26–36.
15. Alekseyev MA, Pevzner PA (2009) Breakpoint graphs and ancestral genome reconstructions. *Genome Res* 19:943–957.
16. Jones BR, Rajaraman A, Tannier E, Chauve C (2012) ANGES: Reconstructing ANcestral GEnomeS maps. *Bioinformatics* 28:2388–2390.
17. Murphy WJ, Bourque G, Tesler G, Pevzner P, O'Brien SJ (2003) Reconstructing the genomic architecture of mammalian ancestors using multispecies comparative maps. *Hum Genomics* 1:30–40.
18. Berthelot C, Muffato M, Abecassis J, Roest Crollius H (2015) The 3D organization of chromatin explains evolutionary fragile genomic regions. *Cell Reports* 10:1913–1924.
19. Kim J, et al. (2013) Reference-assisted chromosome assembly. *Proc Natl Acad Sci USA* 110:1785–1790.
20. Stanyon R, et al. (2008) Primate chromosome evolution: Ancestral karyotypes, marker order and neocentromeres. *Chromosome Res* 16:17–39.
21. Kolmogorov M, Raney B, Paten B, Pham S (2014) Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* 30:i302–i309.
22. Huddleston J, et al. (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* 24:688–696.
23. Mostovoy Y, et al. (2016) A hybrid approach for de novo human genome sequence assembly and phasing. *Nat Methods* 13:587–590.
24. Putnam NH, et al. (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* 26:342–350.
25. Seo JS, et al. (2016) De novo assembly and phasing of a Korean human genome. *Nature* 538:243–247.
26. Damas J, et al. (2017) Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Res* 27:875–884.
27. Froenicke L, et al. (2006) Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes? *Genome Res* 16:306–310.
28. Wienberg J (2004) The evolution of eutherian chromosomes. *Curr Opin Genet Dev* 14: 657–666.
29. Locke DP, et al. (2011) Comparative and demographic analysis of orang-utan genomes. *Nature* 469:529–533.
30. O'Leary MA, et al. (2013) The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* 339:662–667.
31. Zhang X, Goodsell J, Norgren RB, Jr (2012) Limitations of the rhesus macaque draft genome assembly and annotation. *BMC Genomics* 13:206.
32. Zimin AV, et al. (2014) A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biol Direct* 9:20.
33. Roberto R, Misceo D, D'Addabbo P, Archidiacono N, Rocchi M (2008) Refinement of macaque synteny arrangement with respect to the official rheMac2 macaque sequence assembly. *Chromosome Res* 16:977–985.
34. Farré M, et al. (2016) Novel insights into chromosome evolution in birds, archosaurs, and reptiles. *Genome Biol Evol* 8:2442–2451.
35. Rudd MK, et al.; NISC Comparative Sequencing Program (2009) Comparative sequence analysis of primate subtelomeres originating from a chromosome fission event. *Genome Res* 19:33–41.
36. Emerson RO, Thomas JH (2009) Adaptive evolution in zinc finger transcription factors. *PLoS Genet* 5:e1000325.
37. Hayden S, et al. (2010) Ecological adaptation determines functional mammalian olfactory subgenomes. *Genome Res* 20:1–9.
38. Ou Z, et al. (2011) Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome Res* 21:33–46.
39. Duffy DJ (2011) Modulation of Wnt signaling: A route to speciation? *Commun Integr Biol* 4:59–61.
40. Harewood L, Fraser P (2014) The impact of chromosomal rearrangements on regulation of gene expression. *Hum Mol Genet* 23:R76–R82.
41. Mitelman F, Mertens F, Johansson B (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nat Genet* 15:417–474.
42. Koepfli KP, Paten B, O'Brien SJ; Genome 10K Community of Scientists (2015) The Genome 10K Project: A way forward. *Annu Rev Anim Biosci* 3:57–111.
43. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res* 12:996–1006.
44. Harris R (2007) Improved pairwise alignment of genomic DNA. PhD dissertation (Pennsylvania State University, University Park, PA).
45. Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* 17:368–376.
46. Sankoff D, Blanchette M (1999) Probability models for genome rearrangements and linear invariants for phylogenetic inference. *Proceedings of the Third International Conference on Computational Molecular Biology (RECOMB99)* (Association for Computing Machinery, New York), pp 302–309.
47. Müller S, Wienberg J (2001) "Bar-coding" primate chromosomes: Molecular cytogenetic screening for the ancestral hominoid karyotype. *Hum Genet* 109:85–94.