Review TRENDS in Genetics Vol.20 No.12 December 2004

Mammalian phylogenomics comes of age

William J. Murphy[1,*], Pavel A. Pevzner[2] and Stephen J. O'Brien[3]

[1]Basic Research Laboratory, SAIC-Frederick, Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD 21702, USA
[2]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 21702, USA
[3]Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD 21702, USA

The relatively new field of phylogenomics is beginning to reveal the potential of genomic data for evolutionary studies. As the cost of whole genome sequencing falls, anticipation of complete genome sequences from divergent species, reflecting the major lineages of modern mammals, is no longer a distant dream. In this article, we describe how comparative genomic data from mammals is progressing to resolve long-standing phylogenetic controversies, to refine dogma on how chromosomes evolve and to guide annotation of human and other vertebrate genomes.

Once a topic of discussion that was limited to mammalogists and paleontologists, mammalian phylogeny and evolution is now the driving force behind modern comparative genomic analysis: investigating the details of mammalian genomes and how they evolved [1–3]. In the past few years, data from a range of research disciplines – molecular systematics, genome sequencing and comparative cytogenetics – have tackled the evolutionary relationships between humans and their mammalian kin. Together, these tools are now converging on a well-established phylogeny and timescale of mammalian species. Traditionally, mammalian phylogenies were defined by bones and other anatomical characters; however, recent compilations of datasets with tens of thousands of base pairs of DNA sequence, analysis of indels (insertions and deletions) in genes, sequences of repetitive element families and CHROMOSOME ASSOCIATIONS (see Glossary) are converging on a precisely documented history for the mammalian radiation. This new phylogeny of mammalian relationships is being used as a comparative framework in a variety of fields, from functional and adaptive evolution to selecting genomes to aid in the annotation of the human genome. In this article, we discuss the recent advances in resolving higher-level mammalian phylogeny by using independent lines of support from various types of genomic data, and also the role of phylogenetics in mammalian comparative genomic analysis.

Mammalian phylogeny: towards a modern classification

Extant mammals are classified into three major lineages: the egg-laying monotremes, marsupials and placental mammals. Historically, these subdivisions have not been disputed. However, the timing of divergence and the relationships among the three lineages, and among their constituent orders, have been the subject of intense taxonomic debates [4,5]. Beginning over a decade ago, the application of mitochondrial DNA sequencing [6], later combined with nuclear DNA sequencing, began to paint a surprisingly different picture of interordinal mammal phylogeny [7–10] and of the timescale during which these species evolved [11–12]. Most notably these differences were observed in the relationships among the eighteen established and defined orders of placental mammals. Using DNA markers traditional mammalian groups, such as insectivores (e.g. shrews, moles and tenrecs) and ungulates (e.g. horses, ruminants and elephants) were found to be PARAPHYLETIC. However, several of the placental mammal orders, with fossil origins dating back to the early Cenozoic era of Africa, were found to comprise a natural, or MONOPHYLETIC, group dubbed Afrotheria (Box 1) [10]. The subsequent compilation and phylogenetic analysis of large concatenations of nuclear genes revealed the existence of four superordinal clades of

Glossary

Chromosome association: segments that are syntenic to two or more human chromosomes (the standard reference genome for Zoo-FISH experiments) found on a single chromosome in another species. For example, segments of human chromosome 3 and 21 are found on a single chromosome in most mammalian species examined to date.
Microbiotheria: a South American marsupial order represented by only a single living species, *Dromiciops australis*, commonly known as the Monito del Monte or Colocolo.
Monophyletic: a group of species including an ancestor and all of its descendants.
Paenungulata: a superordinal group of placental mammals that includes elephants (Proboscidea), hyraxes (Hyracoidea) and manatees+dugongs (Sirenia).
Paraphyletic: a group of species that includes an ancestor and some but not all of its descendants.
Paucituberculata: a South American marsupial order that includes five extant species in a single family, Caenolestidae, commonly referred to as rat-opossums or shrew-opossums.
Synapomorphy: a shared derived character supporting the monophyly of two or more taxa.
Synteny: the condition of two or more loci being located on the same chromosome.
Vicariance: separation of a formerly contiguous taxon by a geographic or ecological barrier, resulting in the formation of two new taxa.

Corresponding author: William J. Murphy (wmurphy@cvm.tamu.edu).
* Present address: Department of Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843-4458, USA
Available online 30 September 2004

Box 1. Sorting out the afrotherians

Most recently, molecular cytogeneticists have been exploring genomes of species within Afrotheria, the much-debated mammalian superordinal clade with 80 million year old roots in Africa [10,18]. Although there has been no support from morphological data to substantiate this grouping of species, molecular datasets corroborate the existence of an endemic African clade of placental mammals. In 2003, two research groups published reciprocal chromosome painting results of afrotherian (elephant and aardvark) and human chromosomes [41–42]. Although these studies differ slightly, they highlight several important points. First, two shared chromosome associations (Figure I), link all afrotherians to the exclusion of other placental mammals: segments homologous by Zoo-FISH to human chromosome (HSA) 5+HSA21 on a single chromosome, and HSA1+19p [41–42]. Furthermore, Zoo-FISH analysis of species from the afrotherian orders Macroscelidea (elephant shrews) [43,66] and Afrosoricida (golden moles) [66] confirmed the presence of these two afrotherian-diagnostic chromosome associations. They also supported the monophyly of aardvarks, elephant shrews and afrosoricids by sharing an additional association: HSA2+8. Two additional associations, HSA3+20, and HSA10q+17 [39], potentially link elephant shrews and aardvarks to the exclusion of golden moles, although this is in contrast to phylogenies based on the largest available molecular datasets [13–17], which alternatively hypothesize macrosceledid+afrosoricid monophyly. An equally parsimonious interpretation of these data suggests that the 3+20 and 10q+17 associations could have been lost in the species of golden mole analyzed in Ref. [66]. Future Zoo-FISH analysis of additional afrosoricid taxa, such as tenrecs and additional golden mole species, plus sirenians and hyraxes, might help sort out the relationships among the six afrotherian orders.

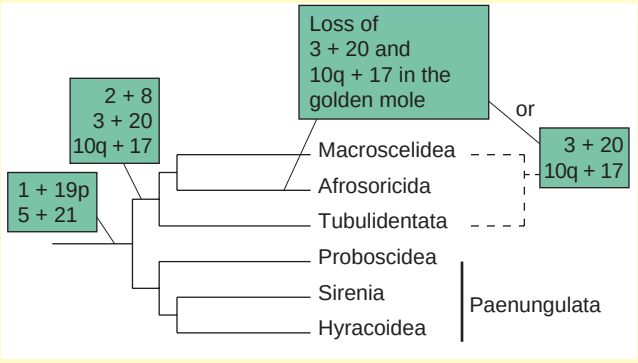

Figure I. Phylogenetic relationships within Afrotheria, based on the largest multi-gene-based molecular phylogenies [15,17]. Chromosome associations based on individual human chromosome homologous (numbered) segments diagnostic for each afrotherian branch are mapped onto the tree. All afrotherians are diagnosed by having a single chromosome in their genome that contains segments homologous to human chromosomes 1+19p, and a single chromsome that contains segments homologous to human chromosomes 5+21. At least one further association, HSA2+8, supports the monophyly of Tubulidentata (aardvark)+Afrosoricida (tenrecs+golden moles) +Macroscelidea (elephant shrews). Two further associations, HSA3+20 and HSA10q+17 are hypothesized to have arisen once in the ancestor of these three orders and been lost in the golden mole. Alternatively, these two associations might support the monophyly of elephant shrews and the aardvark to the exclusion of afrosoricids [66], in contradiction to current molecular-based phylogenies.

placental mammals (Figure 1) [13–17]. One of the most remarkable findings was that two of these clades, Afrotheria and Laurasiatheria, contained parallel adaptive radiations of similar forms that were originally thought to comprise monophyletic groups: namely the insectivores and ungulates. New sequence data also revealed the independent origins of exclusively aquatic and anteating species [13]. The early diversification of placental mammals appears to have been caused by the separation of African and South America during the late Cretaceous period ~100–105 million years ago (Mya) [15–18]. Most of the molecular-based superordinal relationships within the placental tree are stable, being supported by corroborating complex and rare genomic characters (Table 1). Some outstanding issues include the exact root of the placental tree, the relationships within the superordinal clade Laurasiatheria and resolving the trichotomy of elephants, sirenians and hyraxes (i.e. PAENUNGULATA) (Figure 1).

Like the relationships among placental mammals, the details of marsupial relationships have been equally difficult to resolve. Seven orders found in the Neotropics, Australia and New Guinea, appear to have diversified in a more recent yet rapid burst compared with placental mammals, sometime around the Cretaceous–Tertiary boundary [19,20]. Although early morphological classifications split the American and Australasian forms into two separate, monophyletic groups [21,22], molecular studies revealed that two American marsupial orders (PAUCITUBERCULATA and the monotypic MICROBIOTHERIA) are more closely related to the Australian orders than to the American opossums [19,23]. The sister-group relationship between Microbiotheria and the Australian orders, if correct, would suggest VICARIANCE (i.e. the separation of South America from Australia via Antarctica ~52–65 Mya) as the cause of this divergence.

Using bayesian approaches that relax the assumption of a molecular clock (which does not hold across mammalian lineages; [16,18]), a well-defined timescale of all extant mammalian lineages is now emerging (Figure 1). These include an early divergence of the Prototheria (monotremes) from the Theria (marsupials and placentals) nearly 240 Mya, followed by a split between the

Table 1. Molecular synapomorphies diagnostic for the four major placental superordinal clades[a]

Diagnostic criteria	Refs
Afrotheria	
9 bp deletion in *BRCA1*	[13]
237–246 bp deletion in *APOB*	[17]
Unique family of SINEs (AfroSINES)	[35]
Two chromosome associations (HSA1+19q, HSA5+21)	[41–43,66]
Xenarthra	
9 bp deletion in *CRYAA*	[83]
Euarchontoglires	
54 bp deletion in *ATXN1*	[84]
6 bp deletion in *PRNP*	[84]
3 bp deletion in *TNF*	[85]
Three mariner transposon insertions in *CFTR* genomic region	[33]
SINE element phylogeny (dog, mouse and human)	[34]
Laurasiatheria	
10 bp deletion in *PLCB4* 3′UTR	[14,15,86]

[a]Abbreviations: *APOB*, apolipoprotein B; *ATXN1*, ataxin 1; *BRCA1*, breast cancer 1, early onset; *CRYAA*, crystallin, alpha A; *CFTR*, cystic fibrosis transmembrane conductance regulator; *PRNP*, prion protein; SINEs, short interspersed nuclear elements; *TNF*, tumor necrosis factor; UTR, untranslated region.

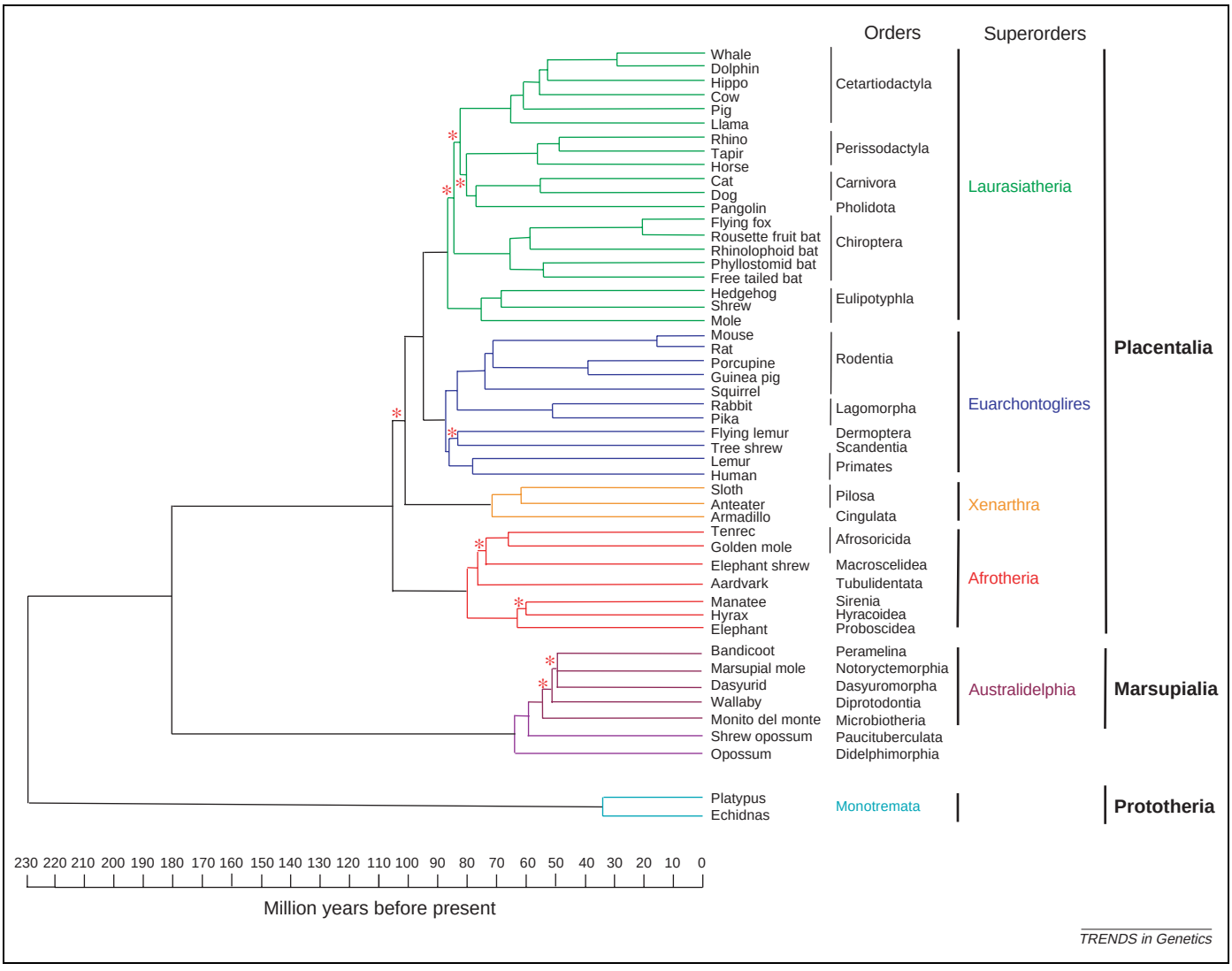

Figure 1. The emerging phylogeny and timescale of mammalian orders defined by DNA sequence data. The tree is a consensus of phylogenetic and divergence time estimation results from Refs. [15,18–20,24,26]. The red asterisks indicate nodes for which there is some ambiguity. It is hoped that new genomic data from additional mammals will eventually sort out these ambiguous nodes in the mammalian tree and aid in the annotation of the human genome.

marsupials and placental mammals ~175–190 Mya [24]. Therian monophyly has been controversial because mitochondrial genomes and the nuclear 18S rRNA gene favor the Marsupionta hypothesis (monotremes+marsupials) [25,26]. More recently, however, nuclear protein encoding genes have confirmed the traditional therian view uniting marsupials and placentals to the exclusion of monotremes [27,28]. Although controversy remains, such as the precise relationship between primates and rodents [29,30], support from large-scale genomic sequencing is reaffirming the current molecular view of mammalian relationships shown in Figure 1.

Signatures of ancestry from large-scale sequencing projects

The human genome sequence was completed in March 2003. The mouse, rat, dog and chimp genomes each have moderate to deep coverage (4.5–8X), and deep coverage shotgun sequencing of four other mammalian genomes are underway (rhesus macaque, cow, opossum and platypus). Plans to expand whole genome sequence (WGS) to include other mammal species (elephant, armadillo, tenrec, rabbit, cat, shrew, guinea pig, hedgehog and orangutan) are now underway (http://www.nhgri.nih.gov/12511858) [31].

A snapshot of the potential of large genome comparisons has been offered by a recent study of portions of the genomes of multiple, phylogenetically divergent vertebrates [32]. Two recent studies from megabase sequencing to WGS have been similarly illuminating. Thomas *et al.* [33] analyzed ~1 Mb of comparative sequence data from the cystic fibrosis transmembrane conductance regulator (*CFTR*) locus region; they found several mutational events in coding exons, plus the insertion of MLT1A0 elements, which confirmed the sister-taxon relationship of primates and rodents in a superordinal clade excluding artiodactyls and carnivores. Using a slightly different approach, Kirkness *et al.* [34] compared the 1.5X genome sequence from the domestic dog with the human and mouse genome assemblies. Specifically, these authors compared different repeat classes that are common to the three genomes with a deduced ancestral mammalian repeat sequence. As shown in Afrotheria and in other groups of mammals [35],

repetitive elements can provide powerful evidence of phylogeny. When human, dog and mouse repeat families were compared [34], it was affirmed that the mouse had an accelerated substitution rate compared with dog and human. In a phylogenetic context, however, short interspersed nuclear elements (SINEs) supported the grouping of primates and rodents, consistent with their inclusion in the superorder Euarchontoglires (Figure 1).

These two examples illustrate the power and the potential of genomic sequence data to confirm (or reject) molecular phylogenetic relationships. Studies examining nuclear protein encoding genes have also revealed several rare deletions that provide strong cladistic characters in support of several lineages of placental mammals (Table 1, Box 1).

Thus, the recognition of multiple phylogenetic characters derived from these different methodologies provide support for the four superorder hierarchy established from likelihood and bayesian-based phylogenies (Figure 1), and reject any alternative arrangements among these major branches (Table 1). Nevertheless, fuzzy nodes persist in the mammalian tree, such as the 65 million year old elephant–manatee–hyrax trichotomy (Paenungulata) and the exact position of the placental tree root (Figure 1). It is our expectation that genomic data, when available from diverse ordinal lineages, will at last untangle the branches of the mammalian tree.

Genome rearrangements as evolutionary characters

Molecular evolutionary studies have traditionally focused on nucleotide or amino acid substitutions in individual genes (or groups of genes) rather than entire genomes. An alternative approach is to infer the evolutionary history of entire genomes. Because large-scale karyotypic evolution proceeds at a much slower pace than nucleotide evolution, chromosomal rearrangements should provide rare, yet powerful, footprints of evolutionary ancestry. Comparative cytogenetics studies offer the ability to demonstrate whole chromosome homologies among the genomes of distantly related mammalian orders. Chromosome painting [or Zoo-fluorescence *in situ* hybridization (Zoo-FISH)], using chromosome-specific probes obtained from flow-sorted chromosomes, has traditionally been the most efficient and productive approach for examining synteny conservation across many mammalian orders (Box 2). To date, >60 mammalian species have been examined using one-way (human on mammal) or reciprocal (mammal on human) chromosome painting, including several focused studies within different mammalian groups: primates [36,37], bears [38], canids [39] and mustelids [40]. What

Box 2. Methods for defining genomic characters

Several methods are available that identify genomic characters that can be used to infer evolutionary relationships.

Chromosome painting: also known as Zoo-fluorescence *in situ* hybridization (Zoo-FISH) utilizes DNA from individual flow sorted chromosomes from one species, which are then hybridized to metaphase chromosomes of a different species (Figure I).

- Advantages: quick and powerful for rapidly detecting whole chromosome homologies across mammalian orders.
- Disadvantages: breaks down over large evolutionary distances (placental mammals versus marsupials); can not identify most inversions; lacks resolution of blocks <4 Mb in size.

Single-copy FISH: fluorescence *in situ* hybridization of DNA from large insert clones (i.e. BACs, YACs, cosmids and fosmids) can be used to order loci in different species and can thus identify inversions between taxa. Standard FISH usually provides resolution of markers spaced several Mb apart, although specialized techniques such as fiber-FISH can provide kilobase level resolution [67].

- Advantages: good for intrachromosomal rearrangements within mammalian orders [68].
- Disadvantages: not as effective across orders because of insufficient non-coding sequence similarity in large insert clones.

Radiation hybrid (RH) mapping: a somatic cell hybrid method where the chromosomes from a particular species are fragmented with irradiation, then fused to a recipient hamster cell line deficient for a selectable marker. DNAs from 90–100 hybrid clones, which are isolated and expanded in selectable media, are tested using PCR-based methods for the presence or absence of specific loci in each cell line. Because each clone contains a random assortment of chromosome fragments from the donor genome, the distance between two loci in the genome can be estimated from the coretention frequency of both markers in the panel.

- Advantages: excellent for fine resolution ordered mapping in any vertebrate species. Mapping resolution can be adjusted by varying the irradiation dosage.
- Disadvantages: moderately time consuming and costly to make panel and genotype markers.

Genome sequencing: provides the highest resolution (base-pair level) for genome comparisons. Targeted sequencing across multiple species enables comparisons of gene content, orientation, repetitive element insertions, indels and so on, providing a wealth of evolutionary characters [33]. An assembled whole genome sequence enables comparison of conserved synteny and gene order across species, facilitating determination of genome rearrangement scenarios.

- Advantages: the ultimate method for fine resolution ordered mapping in any species.
- Disadvantages: expensive and prohibitive for species without extraordinary biomedical utility or application to human genome annotation, which drives the current selection of species choices [31]. High coverage is necessary for full genome alignment and assembly, required for phylogenomic comparisons.

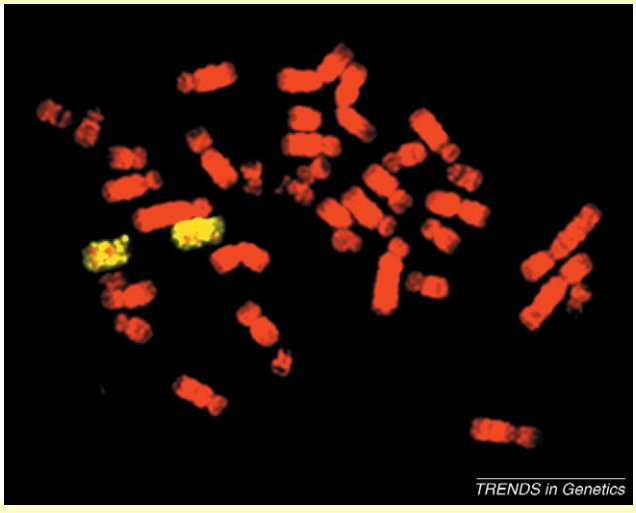

Figure I. Fluorescence *in situ* hybridization of a flow-sorted human chromosome 11 paint on a metaphase chromosome spread of a cat, illustrating the syntenic conversion of this entire chromosome in both species.

has emerged from these studies has been a remarkable confirmation of phylogenetic groups based on molecular sequence data and precise depictions of the mode and tempo of chromosomal change in different mammalian lineages.

Numerous recent reconstructions of the ancestral placental karyotype exist [41–46] and nearly all now agree that it probably consisted of 24 pairs of chromosomes with homology to the following human chromosomes: 1, 2p, 2q, 3+21, 4+8p, 5, 6, 7 partial, 7+16, 8q, 9, 10p, 10q, 11, 12p-qdis+22, 12qter+22, 13, 14+15, 16p+19q, 17, 18, 19p, 20, X and Y. The exact number and configuration of these major syntenic blocks in the ancestor of all mammals will require synteny comparisons of placentals, marsupials and monotremes. Aside from the X chromosome, autosomal genetic divergence across the three major mammalian lineages currently precludes synteny comparisons using Zoo-FISH. Future single locus gene mapping and genome sequencing efforts in marsupial and monotreme taxa [23] should clarify the ancestral genome of all mammals.

Comparative gene maps of representative species in multiple mammalian orders are sufficiently resolved to inform the reconstruction of genome organization. Although Zoo-FISH provides good delineation of conserved synteny blocks between species, estimating the precise boundaries of inversions within synteny blocks and the historical distribution of breaks across chromosomes requires knowledge of gene order that is not provided by this technique. Radiation hybrid maps (Box 2) have the requisite power to resolve megabase-level rearrangements that account for the majority of rearrangements within genomes [45,47,48]. Recently, radiation hybrid mapping, teamed with targeted reciprocal Zoo-FISH analysis of 15 mammalian species from six different orders, revealed that human chromosome 1 (HSA1) was intact in the ancestral placental mammal [45]. This was based on molecular cytogenetic confirmation of a single intact chromosome homologous to human chromosome 1 in species representing each of the four major placental clades, and confirmation that different evolutionary fission and/or translocation events in the ancestral HSA1 homologue occurred in mammalian species carrying two or more chromosomes syntenic to HSA1 [41–43,45]. Analyses of the distribution of HSA1 breakpoints over time revealed a biased distribution across the chromosome, supporting the idea that certain regions of mammalian genomes can be more prone to breakage than others [45].

Similarly, high-resolution comparative maps based on bacterial artificial chromosome (BAC)-end derived markers [48] recently made the detailed comparisons of cattle, human and mouse chromosomes possible. This enabled the precise identification of shared segment boundaries that were conserved throughout evolutionary time, and the identification of lineage-specific rearrangements. In addition, comparative radiation hybrid maps have also been employed to reveal extreme conservation of gene order. The X chromosome is perhaps the most conserved mammalian chromosome. Although highly rearranged in the mouse and rat lineage, relative to the ancestral mammal X chromosome [49], the ancestral order of X chromosome markers has been remarkably conserved in the cat, dog, horse and human genomes over long periods of evolutionary time [34–50]. In the horse genome, this conservation is observed at a resolution of one marker per megabase [51]. These studies illustrate the wealth of information chromosome rearrangements hold as potential for phylogenetic inference (Box 2).

WGS-based phylogeny

Sequenced mammalian genomes provide the opportunity to accurately track historic genome rearrangements for the first time and to deduce the genomic architecture of ancestral and intermediate mammalian genomes. It is clear that genome rearrangements hold many evolutionary secrets and are extremely important for understanding genomic plasticity and fragility. Moreover, combining traditional phylogenetic and genome rearrangement studies would provide a much needed synergy for both areas. However, although there is a wealth of algorithmic and statistical tools for the study of nucleotide- and amino acid substitutions, the computational techniques for genome rearrangement studies are in their infancy.

Every genome rearrangement study involves solving a combinatorial puzzle to find a plausible series of genome rearrangements to transform one genome into another (Box 3). For unichromosomal genomes (e.g. bacterial, mitochondrial or chloroplast genomes), reconstruction usually amounts to an analysis of inversions (also known for computational purposes as reversals), which are the most frequent rearrangement event. The challenge to infer the minimum number of reversals to transform one unichromosomal genome into another is known as the reversal distance problem. For multichromosomal genomes, the most common rearrangements are reversals, translocations, fusions and fissions; the number of such rearrangements that occurs between the genomes of two species in a most parsimonious scenario is known as the genomic distance.

Using rearrangements as evolutionary characters is a relatively new approach in molecular evolution studies [52,53]. Although traditional molecular evolution studies based on point mutations have resolved many phylogenetic controversies in the past 20 years, resolving short branches remains a notoriously difficult problem because point mutations remain weak characters (subject to homoplasy). Chromosome rearrangements provide a powerful set of new evolutionary characters that can help resolve short branches. The problem, however, is how to generate and how to use these characters to construct accurate evolutionary trees.

Building blocks of genomic architectures

Before studying rearrangements, one has to identify the synteny blocks [chromosome segments of conserved order between two or more species (i.e. conserved segments)] that will be used as input to genome rearrangement algorithms. Waterston *et al.* [2] and Pevzner and Tesler [54] described two different approaches to synteny block generation that produced remarkably similar results. To construct synteny blocks, these algorithms start with a set

Box 3. Rearrangement scenarios of genomes

Finding genomic rearrangement scenarios is a difficult combinatorial problem. Early genome rearrangement studies considered breakpoints independently without revealing combinatorial dependencies between breakpoints that are created by the same rearrangement events (i.e. the breakpoints at the end of a segment formed by a single inversion). The eventual understanding of the importance of dependencies between breakpoints [69] resulted in the concept of the breakpoint graph (Figure I; [70]), which reveals correlated breakpoints in a rearrangement scenario. Based on the notion of the breakpoint graph, Hannenhalli and Pevzner [71] developed a polynomial-time algorithm for estimating the reversal distance (the most parsimonious scenario transforming one unichromosomal genome into another). This algorithm was further extended to the genomic distance problem: identification of the minimum number of rearrangements that transforms one multi-chromosomal genome, via inversions, translocations, fissions and fusions, into another [72–75]. Although these algorithms provided excellent tools to study rearrangements between two genomes, the integration of data from multiple genomes (genome phylogeny) represents a more difficult task. Initial work on multiple genomes was again based on breakpoint distances [76–79]. Recently, however, Bourque and Pevzner [80] proposed a new approach, the Multiple Genome Rearrangement (MGR) algorithm. The MGR algorithm constructs an evolutionary scenario that seeks to minimize the number of rearrangements that occur between the genomes. It is based on the Hannenhalli-Pevzner [71] theory of rearrangements and uses a fast modification of their algorithm [73,80–82]. MGR has been tested in several evolutionary studies [47,79–81] and has already produced a putative architecture of the murine ancestral genomes [49].

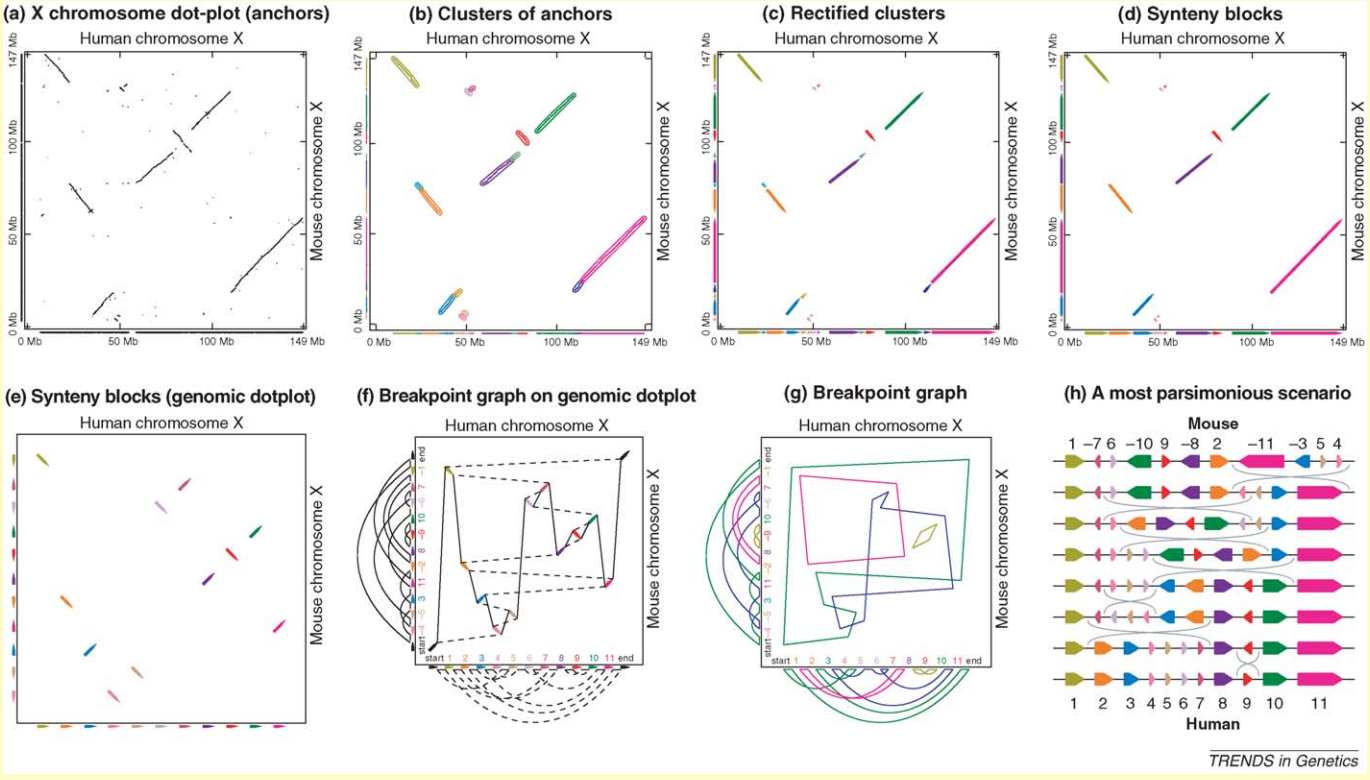

Figure I, Box 1. Two programs, GRIMM-Synteny [54] and Multiple Genome Rearrangement (MGR) [63], applied to human and mouse X chromosome genomic sequences, from local similarities, synteny blocks, breakpoint graph to rearrangement scenario (Figure I). **(a)** Genomic dot-plot showing two-way regions of best similarity (anchors) between human (*x*-axis) and mouse (*y*-axis) X chromosome sequences. The anchors are enlarged for visibility. **(b)** Clusters of anchors are identified after filtering out spurious regions of similarity (e.g. probable paralogs) shown as dots in (a). **(c)** Minor rearrangements within clusters are rectified until the clusters form a single diagonal line, thus forming **(d)** synteny blocks. **(e)** Each synteny block is then assigned an equal length as genome rearrangement units. **(f)** A two-dimensional breakpoint graph reveals which breakpoints (the endpoints of the synteny blocks) are related through connector lines (the path). Superimposition of the human and mouse paths form the breakpoint graph. The solid lines connect human synteny blocks, whereas broken lines connect mouse synteny blocks. **(g)** The synteny blocks are removed from the two-dimensional breakpoint graph to reveal the cycles. The four cycles (boxes) in the breakpoint graph, shown by different colors, are used to create the most parsimonious rearrangement scenario **(h)** for human and mouse X-chromosomes, computed by MGR. For more details, see Ref. [54].

of local similarities (also called anchors) between multiple genomes (Box 3). Several software tools have recently become available to generate such anchors for entire mammalian genomes [55–58]. After the set of anchors is constructed, the goal is to determine large-scale synteny blocks by combining anchors that are close to each other even if their ordering in the different genomes is inconsistent as a result of microrearrangements. These are defined as rearrangements of anchors or markers within a synteny block, whereas macrorearrangements are defined as rearrangements of the order and orientations of the synteny blocks. Mouse and rat genomic projects revealed the previously unknown phenomenon of frequent microrearrangements (i.e. inversions with a short span) in mammalian evolution. Microrearrangements present many previously unexplored evolutionary characters that provide new insights into mammalian evolution. The GRIMM-synteny algorithm [54] has an important feature: it preserves information about microrearrangements within synteny blocks and enables the analysis of the microrearrangement history of every synteny block.

Rearrangement analysis and difficult phylogeny problems

Two decades ago, Nadeau and Taylor [59] introduced the random breakage model of genome rearrangements and estimated the number of human-mouse conserved segments to be ~180, based on a few genes mapped in common between the two species. This prediction [59] has withstood the test of time, providing close similarity to recent estimates from gene-dense genetic and radiation hybrid maps, and the annotated number of conserved segments between the human and mouse genome sequences [2]. However, the increased resolution has now revealed, with a new level of precision, that the random breakage model is unable to explain the numerous breakpoint clumps in the genome. The analysis of human and mouse complete genome sequences [60,61] implies that, in addition to the obvious 'visible' synteny blocks, many 'hidden' synteny blocks exist, whose length was too short to be revealed by previous genetic or radiation hybrid maps. The existence of a surprisingly large number of hidden synteny blocks (typically <1 Mb) provides an argument in favor of a different fragile breakage model of chromosome evolution [60]; one that postulates that the breakpoints often occur within relatively short fragile regions (hotspots of rearrangements). Although GRIMM-Synteny algorithm generates synteny blocks of all sizes, only the blocks with a length above a certain threshold can be used in genome rearrangement analysis. Ideally, one should use all blocks for a comprehensive rearrangement analysis. Moreover, short 'hidden' synteny blocks are extremely important for studies of genomic fragility. However, many of the short synteny blocks are likely to be artifacts caused by spurious similarities as a result of segmental duplications or errors in genome assembly. Distinguishing between such spurious similarities and 'real' or hidden synteny blocks (i.e. the intervals between two consecutive rearrangement end-points) remains a problem.

Constructing a complete set of conserved synteny blocks for many mammalian genomes has far-reaching consequences. First, this set would immediately reveal the set of fragile regions (as regions with high concentrations of short synteny blocks or breakpoints). Second, the rearrangement events derived from the complete set of conserved synteny blocks can form a set of evolutionary characters. Therefore, if the complete set of conserved synteny blocks for several species is known, then one could use these characters to reconstruct a synteny-based evolutionary tree for these species and to reconstruct the genomic architecture of their ancestors. This approach opens a new avenue to resolve the controversial branches of the evolutionary tree of mammals. Such short branches (e.g. branches spanning less than a few million years) can be difficult to resolve by traditional phylogenetic analysis of nucleotide sequences. However, they might be easier to resolve if supported by chromosome rearrangement

Box 4. Mammalian BAC libraries currently available or in production

MONOTREMES
Echidna, *Tachyglossus aculeatus*
Platypus, *Ornithorhynchus anatinus*

MARSUPIALS
American opossum, *Didelphis virginianus*
Laboratory opossum, *Monodelphis domestica*
Tammar wallaby, *Macropus eugenii*

PLACENTALS
Afrotheria
Proboscidea
African savanna elephant, *Loxodonta africana*
Xenarthra
Nine-banded armadillo, *Dasypus novemcinctus*
Laurasiatheria
Carnivora
Domestic dog, *Canis familiaris*
Domestic cat, *Felis catus*
Clouded leopard, *Neofelis nebulosa*
Perissodactyla
Domestic horse, *Equus caballus*
Cetartiodactyla
Formosan muntjac, *Muntiacus reevesi*
Indian muntjac, *Muntiacus muntjac*
Domestic cattle, *Bos taurus*
Domestic sheep, *Ovis aries*
Domestic pig, *Sus scrofa*
Domestic goat, *Capra hircus*
Chiroptera
Horseshoe bat, *Rhinolophus ferrumequinum*
Brown bat, *Myotis lucifugus*
Flying fox, *Pteropus livingstoni*
Eulipotyphla
Hedgehog, *Atilerex albiventris*
Shrew, *Sorex araneus*
Euarchontoglires
Lagomorpha
Rabbit, *Oryctolagus cuniculus*
Rodentia
Deer mouse, *Peromyscus maniculatus*
Hamster, *Cricetulus griseus*
Mouse, *Mus musculus*
Rat, *Rattus norvegicus*
Ground squirrel, *Spermophilus tridecemlineatus*
Primates
Baboon, *Papio hamadryas*
Black lemur, *Eulemur macaco*
Chimpanzee, *Pan troglodytes*
Colobus monkey, *Colobus guereza*
Dusky titi, *Callicebus moloch*
Galago, *Otolemur garnetti*
Gibbon, *Hylobates concolor*
Gorilla, *Gorilla gorilla*
Ring-tailed lemur, *Lemur catta*
Macaque, *Macaca mulatta*
Marmoset, *Callithrix jacchus*
Mouse lemur, *Microcebus murinus*
Sumatran orangutan, *Pongo pygmaeus*
Owl Monkey, *Aotus trivirgatus/nancymai*
Squirrel monkey, *Saimiri boliviensis*
Tarsier, *Tarsius bancanus*
Vervet monkey, *Cercopithecus aethiops*
Scandentia
Tree shrew, *Tupaia minor*
For more information, see http://www.genome.gov/10001852 and http://bacpac.chori.org/.

(or even microrearrangement) events occurring on short branches of the mammalian tree. Given that thousands of microrearrangements were identified in the human–mouse–rat genome comparisons [3,45], it is likely that every short branch in the mammalian tree has at least one such microrearrangement.

Future expectations of genome scale phylogenetics

Currently five mammalian species (human, mouse, rat, chimp and dog) have a 'completed' genome, in either finished or high-quality draft form, whereas four additional species (cow, rhesus macaque, opossum and platypus) are in various stages of production. Plans are now underway for low-coverage low-coverage draft sequencing of additional placental mammals including species from unsampled clades in the placental tree, for example, Afrotheria and Xenarthra, plus accelerated lineages inside these clades to aid the annotation of the human and mouse genomes [31,62–65]. In addition, BAC libraries have been constructed for species of nearly every placental mammalian lineage and should enable targeted sequence acquisition across a diverse phylogenetic spectrum [32] (Box 4).

With the acquisition of targeted or WGS from a diverse assemblage of mammals, researchers can look forward to the challenges of whole genome phylogenetics. Assembly of multiple ordered genomes with chromosomes largely aligned at the nucleotide level will also enable glimpses into the mechanisms driving chromosome breakage and the genomic properties conferring long-term maintenance of synteny, which are only now gleaned from gene maps [45,47,48,51] and pairwise sequence comparisons [2,54]. Future Zoo-FISH analysis from many species of the remaining orders of mammals, combined with potentially thousands of phylogenetically informative chromosomal rearrangements and rare genomic changes in each additional mammalian genome that is mapped and sequenced, will soon provide us with a full resolution of the major branches of the mammalian family tree.

Acknowledgements

We thank Guillaume Bourque, Eduardo Eizirik, Mark Springer, Roscoe Stanyon, Emma Teeling and Glenn Tesler for ongoing discussions and collaboration on these topics. We also thank the editor, Terry Robinson and two anonymous reviewers for constructive comments on the manuscript. We thank Joan Menninger and Bill Nash for the image in Figure I (Box 2). This publication has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health under contract N01-CO-12400.

References

1 O'Brien, S.J. *et al*. (1999) The promise of comparative genomics in mammals. *Science* 286, 458–481
2 Waterston *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
3 Gibbs, R.A. *et al*. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428, 493–521
4 Novacek, M.J. (2001) Mammalian phylogeny: genes and supertrees. *Curr. Biol.* 11, R573–R575
5 Helgen, K.M. (2003) Major mammalian clades: a review under consideration of molecular and palaeontological evidence. *Mamm. Biol.* 68, 1–15
6 Kocher, T.D. *et al*. (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proc. Natl. Acad. Sci. U. S. A.* 86, 6196–6200
7 Springer, M.S. *et al*. (1997) Endemic African mammals shake the phylogenetic tree. *Nature* 388, 61–64
8 Springer, M.S. *et al*. (1997) The interphotoreceptor retinoid binding protein gene in therian mammals: implications for higher level relationships and evidence for loss of function in the marsupial mole. *Proc. Natl. Acad. Sci. U. S. A.* 94, 13754–13759
9 Stanhope, M.J. *et al*. (1998) Molecular support for multiple origins of insectivora and for a new order of endemic African insectivore mammals. *Proc. Natl. Acad. Sci. U. S. A.* 95, 9967–9972
10 Stanhope, M.J. *et al*. (1998) Highly congruent molecular support for a diverse superordinal clade of endemic African mammals. *Mol. Phylogenet. Evol.* 9, 501–508
11 Hedges, S.B. *et al*. (1996) Continental breakup and the ordinal diversification of mammals. *Nature* 381, 226–229
12 Kumar, S. and Hedges, S.B. (1998) A molecular timescale for vertebrate evolution. *Nature* 392, 917–920
13 Madsen, O. *et al*. (2001) Parallel adaptive radiations in two major clades of placental mammals. *Nature* 409, 610–614
14 Murphy, W.J. *et al*. (2001) Molecular phylogenetics and the origins of placental mammals. *Nature* 409, 614–618
15 Murphy, W.J. *et al*. (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* 294, 2348–2351
16 Eizirik, E. *et al*. (2001) Molecular dating and biogeography of the early placental mammal radiation. *J. Hered.* 92, 212–219
17 Amrine-Madsen, H. *et al*. (2003) A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. *Mol. Phylogenet. Evol.* 28, 225–240
18 Springer, M.S. *et al*. (2003) Placental mammal diversification and the Cretaceous–Tertiary boundary. *Proc. Natl. Acad. Sci. U. S. A.* 100, 1056–1061
19 Amrine-Madsen, H. *et al*. (2003) Nuclear gene sequences provide evidence for the monophyly of australidelphian marsupials. *Mol. Phylogenet. Evol.* 28, 186–196
20 Nilsson, M.A. *et al*. (2003) Radiation of extant marsupials after the K/T boundary: evidence from complete mitochondrial genomes. *J. Mol. Evol.* 57 (Suppl. 1), S3–12
21 Szalay, X. (1982) A new appraisal of marsupial phylogeny and classification. In *Carnivorous marsupials* (Archer, M. ed.), pp. 621–640, Royal Zoological Society of New South Wales, Mosman, Australia
22 Woodburne, M.O. (1984) Families of marsupials: relationships, evolution and biogeography. In *Mammals: notes for a short course, studies in geology* (Vol. 8) (Broadhead, T.W. ed.), pp. 48–71, University of Tennessee, Knoxsville
23 Graves, J.A. and Westerman, M. (2002) Marsupial genetics and genomics. *Trends Genet.* 18, 517–521
24 Woodburne, M.O. *et al*. (2003) The evolution of tribospheny and the antiquity of mammalian clades. *Mol. Phylogenet. Evol.* 28, 360–385
25 Janke, A. *et al*. (1997) The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. *Proc. Natl. Acad. Sci. U. S. A.* 94, 1276–1281
26 Janke, A. *et al*. (2002) Phylogenetic analysis of 18S rRNA and the mitochondrial genomes of the wombat, *Vombatus ursinus*, and the spiny anteater, *Tachyglossus aculeatus*: increased support for the Marsupionta hypothesis. *J. Mol. Evol.* 54, 71–80
27 Killian, J.K. *et al*. (2001) Marsupials and eutherians reunited: genetic evidence for the Theria hypothesis of mammalian evolution. *Mamm. Genome* 12, 513–517
28 Belov, K. *et al*. (2002) Echidna IgA supports mammalian unity and traditional Therian relationship. *Mamm. Genome* 13, 656–663
29 Misawa, K. and Nei, M. (2003) Reanalysis of Murphy *et al*.'s data gives various phylogenies and suggests overcredibility of Bayesian trees. *J. Mol. Evol.* 57 (Suppl. 1), S290–296
30 Misawa, K. and Janke, A. (2003) Revisiting the Glires concept – phylogenetic analysis of nuclear sequences. *Mol. Phylogenet. Evol.* 28, 320–327
31 Pennisi, E. (2004) Mammalian biology: more genomes but shallower coverage. *Science* 304, 1227
32 Thomas, J.W. and Touchman, J.W. (2002) Vertebrate genome sequencing: building a backbone for comparative genomics. *Trends Genet.* 18, 104–108
33 Thomas, J.W. *et al*. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424, 788–793

34 Kirkness, E.F. *et al*. (2003) The dog genome: survey sequence and analysis. *Science* 301, 1898–1903
35 Nikaido, M. *et al*. (2003) Ancient SINEs from African endemic mammals. *Mol. Biol. Evol.* 20, 522–527
36 O'Brien, S.J. and Stanyon, R. (1999) Phylogenomics. Ancestral primate viewed. *Nature* 402, 365–366
37 Stanyon, R. *et al*. (2002) Chromosome painting reveals that galagos have highly derived karyotypes. *Am. J. Phys. Anthropol.* 117, 319–326
38 Nash, W.G. *et al*. (1998) Comparative genomics: tracking chromosome evolution in the family Ursidae using reciprocal chromosome painting. *Cytogenet. Cell Genet.* 83, 182–192
39 Nash, W.G. *et al*. (2001) The pattern of phylogenomic evolution of the Canidae. *Cytogenet. Cell Genet.* 95, 210–224
40 Graphodatsky, A.S. *et al*. (2002) Comparative molecular cytogenetic studies in the order Carnivora: mapping chromosomal rearrangements onto the phylogenetic tree. *Cytogenet. Genome Res.* 96, 137–145
41 Yang, F. *et al*. (2003) Reciprocal chromosome painting among human, aardvark and elephant (superorder Afrotheria) reveals the likely eutherian ancestral karyotype. *Proc. Natl. Acad. Sci. U. S. A.* 100, 1062–1066
42 Froenicke, L. *et al*. (2003) Towards the delineation of the ancestral eutherian genome organization: comparative genomic maps of human and the African elephant (*Loxodonta africana*) generated by chromosome painting. *Proc. R. Soc. Lond. B. Biol. Sci.* 207, 1331–1340
43 Svartman, M. *et al*. (2004) A chromosome painting test of the basal eutherian karyotype. *Chromosome Res.* 12, 45–51
44 Murphy, W.J. *et al.* (2001) Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biol*, 2, Review 0005 (http://genomebiology.com/2001/2/6/0005)
45 Murphy, W.J. *et al*. (2003) Evolution of human chromosome 1 and its homologues in placental mammals. *Genome Res* 13, 1880–1888
46 Richard, F. *et al*. (2003) Reconstruction of the ancestral karyotype of eutherian mammals. *Chromosome Res.* 11, 605–618
47 Murphy, W.J. *et al*. (2003) Analysis of mammalian genome rearrangements using multispecies comparative maps. *Human Genomics* 1, 30–39
48 Larkin, D.M. *et al*. (2003) A cattle–human comparative map built with BAC-ends and human genome sequence. *Genome Res.* 13, 1966–1972
49 Bourque, G. *et al*. (2004) Reconstruction the genomic architecture of mammals: lessons from human, mouse and rat genomes. *Genome Res.* 14, 507–516
50 Murphy, W.J. *et al*. (1999) Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping. *Genome Res.* 9, 1223–1230
51 Raudsepp, T. *et al*. (2004) Exceptional conservation of horse-human gene order on X chromosome revealed by high-resolution radiation hybrid mapping. *Proc. Natl. Acad. Sci. U. S. A.* 101, 2386–2391
52 Blanchette, M. *et al*. (1999) Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.* 49, 193–203
53 Boore, J. (1999) Animal mitochondrial genomes. *Nucleic Acids Res.* 27, 1767–1780
54 Pevzner, P. and Tesler, G. (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* 13, 37–45
55 Mayor, C. *et al*. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* 16, 1046–1047
56 Schwartz, S. *et al*. (2000) Pip-Maker: A Web server for aligning two genomic DNA sequences. *Genome Res.* 10, 577–586
57 Ma, B. *et al*. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics* 18, 440–445
58 Kent, W.J. (2002) BLAT – The BLAST-like alignment tool. *Genome Res.* 12, 656–664
59 Nadeau, J.H. and Taylor, B.A. (1984) Lengths of chromosome segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. U. S. A.* 81, 814–818
60 Pevzner, P. and Tesler, G. (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. U. S. A.* 100, 1056–1061
61 Kent, W.J. *et al*. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U. S. A.* 100, 11484–11489
62 O'Brien, S.J. *et al*. (2001) On choosing mammalian genomes for sequencing. *Science* 292, 2264–2266
63 Cooper, G.M. and Sidow, A. (2003) Genomic regulatory regions: insights from comparative sequence analysis. *Curr. Opin. Genet. Dev.* 13, 604–610
64 Lindblad-Toh, K. (2004) Genome sequencing: three's company. *Nature* 428, 475–476
65 Cooper, G.M. *et al*. (2003) Quantitative estimates of sequence divergence for comparative analysis of mammalian genomes. *Genome Res.* 13, 813–820
66 Robinson, T.J. *et al*. (2004) Cross-species chromosome painting in the golden mole and elephant shrew: support for the mammalian clades Afrotheria and Afroinsectiphilla but not Afroinsectivora. *Proc. R. Soc. Lond. B. Biol. Sci.* 271, 1477–1484
67 Ersfeld, K. (2004) Fiber-FISH: fluorescence *in situ* hybridization on stretched DNA. *Methods Mol. Biol.* 270, 395–402
68 Tsend-Ayush, E. *et al*. (2004) Plasticity of human chromosome 3 during primate evolution. *Genomics* 83, 193–202
69 Kececiogly, J. and Sankoff, D. (1995) Exact and approximation algorithms for the inversion distance between two permutations. *Algorithmica* 13, 180–210
70 Bafna, V. and Pevzner, P.A. (1993) Genome rearrangements and sorting by reversal. *34th Symposium on Foundations of Computer Science*, pp. 148–157
71 Hannenhalli, S. and Pevzner, P.A. (1995) Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *Proceedings of the 27th Annual ACM Symposium on the Theory of Computing*, pp. 178–189
72 Hannenhalli, S. and Pevzner, P.A. (1995) Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science*, pp. 581–592
73 Tesler, G. (2002) GRIMM: Genome rearrangements web server. *Bioinformatics* 18, 492–493
74 Ozery-Flato, M. and Shamir, R. (2003) Two notes on genome rearrangment. *J Bioinform Computat Biol* 1, 71–94
75 Pevzner, P. (2000) *Computational Molecular Biology: An algorithmic approach*. MIT Press, Cambridge, MA
76 Blanchette, M. *et al*. (1997) Breakpoint phylogenies. In *Genome Informatics Workshop* (Miyano, S. and Takagi, T., eds), pp. 25–34, Japan Univ. Acad. Press, Tokyo
77 Sankoff, D. and Blanchette, M. (1997) The median problem for breakpoints in comparative genomics. In *Computing and Combinatorics, Proceedings of COCOON '97, Lecture Notes in Computer Science*, pp. 251–263. Springer Verlag, New York
78 Moret, B. *et al.* (2001) A new implementation and detailed study of breakpoint analysis. In *Pac. Symp. Biocomput.* (*PSB 2001*) pp. 583–594, AAAI Press, Menlo Park, California
79 Bourque, G. and Pevzner, P.A. (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.* 12, 26–36
80 Bader, D.A. *et al*. (2001) A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comput. Biol.* 8, 483–491
81 Andelfinger, G. *et al*. (2004) Detailed four-way comparative mapping and gene order analysis of the canine *ctvm* locus reveals evolutionary chromosome rearrangements. *Genomics* 83, 1053–1062
82 Tesler, G. (2002) Efficient algorithms for multichromosomal genome rearrangements. *J. Comp. Sys. Sci.* 65, 587–609
83 van Dijk, M. *et al*. (1999) The virtues of gaps: xenarthran (Edentate) monophyly supported by a unique deletion in alpha A-crystallin. *Syst. Biol.* 48, 94–106
84 Poux, C. *et al*. (2002) Sequence gaps join mice and men: phylogenetic evidence from deletions in two proteins. *Mol. Biol. Evol.* 19, 2035–2037
85 de Jong, W. *et al*. (2003) Indels in protein coding sequences of Euarchontoglires constrain the rooting of the eutherian tree. *Mol. Phylogenet. Evol.* 28, 328–340
86 Springer, M.S. *et al.* A molecular view on relationships among the extant orders of placental mammals. In *Origin, Timing, and Relationships Among the Major Clades of Extant Placental Mammals* (Rose, K.D. and Archibald, J.D., eds), Johns Hopkins University Press (in press)