

Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom

Matthieu Muffato, Alexandra Louis, Nga Thi Thuy Nguyen, Joseph Lucas, Camille Berthelot, Hugues Roest Crollius

▶ **To cite this version:**

Matthieu Muffato, Alexandra Louis, Nga Thi Thuy Nguyen, Joseph Lucas, Camille Berthelot, et al.. Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom. Nature Ecology & Evolution, In press, 10.1038/s41559-022-01956-z . hal-03943655

HAL Id: hal-03943655
https://hal.science/hal-03943655

Submitted on 17 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

nature ecology & evolution

Article

https://doi.org/10.1038/s41559-022-01956-z

Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom

Received: 17 February 2022

Accepted: 22 November 2022

Published online: 16 January 2023

Check for updates

Matthieu Muffato[1,2,4], Alexandra Louis[1,4], Nga Thi Thuy Nguyen[1], Joseph Lucas[1], Camille Berthelot[1,3,5] ✉ & Hugues Roest Crollius[1,5] ✉

Ancestral sequence reconstruction is a fundamental aspect of molecular evolution studies and can trace small-scale sequence modifications through the evolution of genomes and species. In contrast, fine-grained reconstructions of ancestral genome organizations are still in their infancy, limiting our ability to draw comprehensive views of genome and karyotype evolution. Here we reconstruct the detailed gene contents and organizations of 624 ancestral vertebrate, plant, fungi, metazoan and protist genomes, 183 of which are near-complete chromosomal gene order reconstructions. Reconstructed ancestral genomes are similar to their descendants in terms of gene content as expected and agree precisely with reference cytogenetic and in silico reconstructions when available. By comparing successive ancestral genomes along the phylogenetic tree, we estimate the intra- and interchromosomal rearrangement history of all major vertebrate clades at high resolution. This freely available resource introduces the possibility to follow evolutionary processes at genomic scales in chronological order, across multiple clades and without relying on a single extant species as reference.

Biological sequences have long been recognized as a document of evolutionary history[1], where accumulated mutations record relationships between species and the dynamics underlying their evolution. Given sufficient genetic information across species, the temporal accumulation of these mutations can be traced back in time to reconstruct sequences and genomes in their long-lost common ancestors. These ancestral reconstructions are the backbone of much of today's methodologies in molecular evolution, including phylogenetic trees[2–4] and sequence selection tests[5,6]. The reconstruction of ancestral sequences, and especially genes, has been extensively studied since the dawn of sequencing: mature methods exist to retrace the history of sequence substitutions and leverage changes in substitution dynamics to answer specific evolutionary questions. However, DNA mutations are not limited to sequence substitutions: genomes are also affected by larger scale mutational events such as duplications, deletions, sequence inversions or chromosomal rearrangements, all of which can affect genome function, species fitness and evolution. In extant species, large-scale mutations are a major determinant of disease because they can disrupt functional sequences[7–9] and reorganize functional structures within the genome[10–12]. From an evolutionary viewpoint, large-scale mutations are a well-documented source of innovations: they can produce new genetic combinations that contribute phenotypic novelty[13,14] but can

[1]Institut de Biologie de l'École Normale Supérieure, Centre National de la Recherche Scientifique Unité Mixte de Recherche 8197, Institut National de la Santé et de la Recherche Médicale U1024, Université PSL, Paris, France. [2]Present address: Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [3]Present address: Institut Pasteur, Université Paris Cité, Centre National de la Recherche Scientifique Unité Mixte de Recherche 3525, Institut National de la Santé et de la Recherche Médicale UA12, Comparative Functional Genomics group, Paris, France. [4]These authors contributed equally: Matthieu Muffato, Alexandra Louis. [5]These authors jointly supervised this work: Camille Berthelot, Hugues Roest Crollius. ✉e-mail: camille.berthelot@pasteur.fr; hrc@bio.ens.psl.eu

also have more indirect effects such as locally suppressing recombination[15,16], favouring allele hitchhiking and rapid selection[17,18]. For example, genomic rearrangements have been shown to associate with changes in brain gene expression between humans and chimpanzees[19], to underlie the evolution of intersexual development in moles[20] and variations in reproductive morphs in ruffs[21]. Despite their tremendous functional and evolutionary importance, large-scale mutational events are less extensively studied and not as well understood than sequence substitutions. In particular, the reconstruction of ancestral genomes and karyotypes lags behind that of ancestral sequences, making it difficult to study the evolutionary dynamics and impact of rearrangements, duplications and deletions over many species and within rigorous theoretical frameworks.

With the advent of massive sequencing projects ambitioning to obtain high-quality reference genomes for thousands of species across all kingdoms of life[22], evolutionary genomics faces both fresh opportunities and serious challenges to integrate this flow of data into usable comparative frameworks. Along with whole-genome alignments[23], ancestral genome and karyotype reconstructions across large clades is one of the most promising outcomes of these projects. The goal of these reconstructions is to provide a plausible organization of genomic sequences in one or many extinct common ancestors of a group of species of interest. Several palaeogenomic strategies have been explored to reconstruct the sequence content and ordering of ancestral genomes. Methods based on double-cut-and-join algorithms endeavour to reconstruct rearrangement scenarios resulting in observed extant genome structures[24,25]. These methodologies are increasingly computationally expensive and in many cases intractable for sets of large, complex genomes, which at this time have only been overcome by substantially reducing reconstruction resolution[26–28]. Other methods attempt to reconstruct a parsimonious sequence ordering in the ancestor based on orthologous sequence adjacencies in extant genomes, under the assumption that genomic rearrangements are unlikely to result in the same sequence organization several times independently. These methods can be applied to different types of markers, typically either alignable sequence blocks or individual genes, and are appropriate for small[29] and large genomes such as vertebrates or plants[30,31]. However, it is unclear whether current methods can provide high-resolution reconstructions and scale to the large genomic resources available in comparative genomics databases. At this time, only two ancestral genomic reconstruction resources are widely available to the community: AncestralGenomes[32], which provides 111 ancestral gene content reconstructions but not their order ('bags of genes'), and DESCHRAMBLER[33], which offers chromosome-complete reconstructions for 7 mammal and 14 bird ancestors but with limited subchromosomal resolution (100–300 kb sequence blocks) and dependent on a reference genome. In this study, we introduce a new resource containing 624 ancestral genomes reconstructed over the vertebrate, plant, fungi, metazoan and protist clades, at gene-scale resolution, where a third of the ancestral genomes reaches chromosomal-complete assemblies. This drastic change in magnitude is powered by an iterative, parsimony-based ancestral genome reconstruction algorithm, named AGORA (Algorithm for Gene Order Reconstruction in Ancestors), which we describe in this article. We show that AGORA is efficient, flexible and scales to integrate hundreds of large genomes, to reconstruct their common ancestors at every node in the species phylogeny with relatively modest computational costs. Along with the open-source algorithm, all precomputed ancestral genome reconstructions are publicly available in the Genomicus[34,35] database (https://www.genomicus.bio.ens.psl.eu/genomicus) and benefit from the full browsing and comparative genomics tool infrastructure of the database. The database is regularly updated since 2010 to reflect reference genome improvements and represents a perennial resource for high-quality, high-resolution ancestral genomes for the molecular evolution community across disciplines and model phylogenetic clades.

Results

A resource of ancestral genomes for evolutionary genomics

To facilitate the investigation of chromosomal and local genome dynamics across evolution, we developed an extensive resource of ancestral genome reconstructions that spans large portions of the eukaryotic tree of life. This resource is based on an algorithm named AGORA, which computes highly contiguous, near-exhaustive reconstructions of the ancestral gene order at every bifurcation in the species tree, based on gene order information in the extant species of the clade (Fig. 1a). While AGORA can be installed as a standalone package for tailored research applications, we routinely precompute and release the complete set of ancestral vertebrate genomes for every update of the Ensembl database and for a broad selection of plant and fungi clades as part of the Genomicus synteny database[36]. At the time of submission, Genomicus contains a total of 624 ancestral genomes readily available for download across the vertebrates, plants, metazoa, protists and fungi databases (Supplementary Data 1). These ancestral genomes can be explored and manipulated using the different utilities of the Genomicus web server[36] to perform karyotype comparisons, extraction and evolutionary tracing of conserved synteny blocks (Fig. 1b), and local gene–gene synteny visualization across ancestral and extant species (Fig. 1c). A partial draft version of AGORA, combined with extensive manual curation, has previously been used to reconstruct the Brassicacea[37] and Amniota[38] ancestors, illustrating several of these applications.

AGORA is an algorithm to reconstruct ancestral gene order

AGORA is a parsimony-based algorithm that estimates the content and order of genes in the ancestor of a group of extant species for which reference genomes are available (Fig. 2 and Supplementary Fig. 1). Briefly, the method iteratively extracts commonalities between pairs of extant genomes to infer characteristics inherited from their last common ancestor and present in every ancestor along the evolutionary branches leading to each extant genome. AGORA takes as input a forest of gene phylogenetic trees, corresponding to all the gene families present in the extant genomes with their orthologous and paralogous relationships, and the gene orders in each extant genome. First, AGORA uses the phylogenies of extant genes to infer the gene content of every ancestor along the species tree (Supplementary Fig. 2). Second, AGORA compares the gene orders of every pair of extant species to identify orthologous genes adjacent and in the same orientation in both species and presumably inherited from their last common ancestor (Fig. 2a). For every ancestor in the species tree, the algorithm extracts the subset of informative pairwise extant species comparisons (Fig. 2b) and integrates the gene adjacency comparisons into a weighted graph, where nodes represent ancestral genes and edge adjacencies are supported by pairwise extant species comparisons. The weights correspond to the number of comparisons supporting that these genes were adjacent in this ancestor (Fig. 2c,d). Ideally, this process would result in a linear graph representing the ancestral gene order because genome rearrangements are unlikely to produce the same gene adjacencies independently in different lineages[39–42]. However, errors in the resolution of orthologues and paralogues in the original gene trees can result in branching in the graph. AGORA linearizes the graph by iteratively removing low-weight edges to obtain a parsimonious reconstruction of the oriented gene order in the ancestral genome (Fig. 2e). AGORA includes extensions of this algorithm to deal with larger errors in the input gene trees by identifying a set of constrained genes that are close to being single-copy in most species, and can be reliably used for gene order reconstruction. In this mode, AGORA adds the non-constrained genes in a second stage. The algorithm is presented in detail in the Supplementary Information (Supplementary Figs. 1–9). The in silico performance of AGORA has been tested on a previously used benchmark of genome evolution simulations[33], achieving 98.9% agreement with the reference (sensitivity 99.3%, precision 99.6%; Methods), similar

https://doi.org/10.1038/s41559-022-01956-z

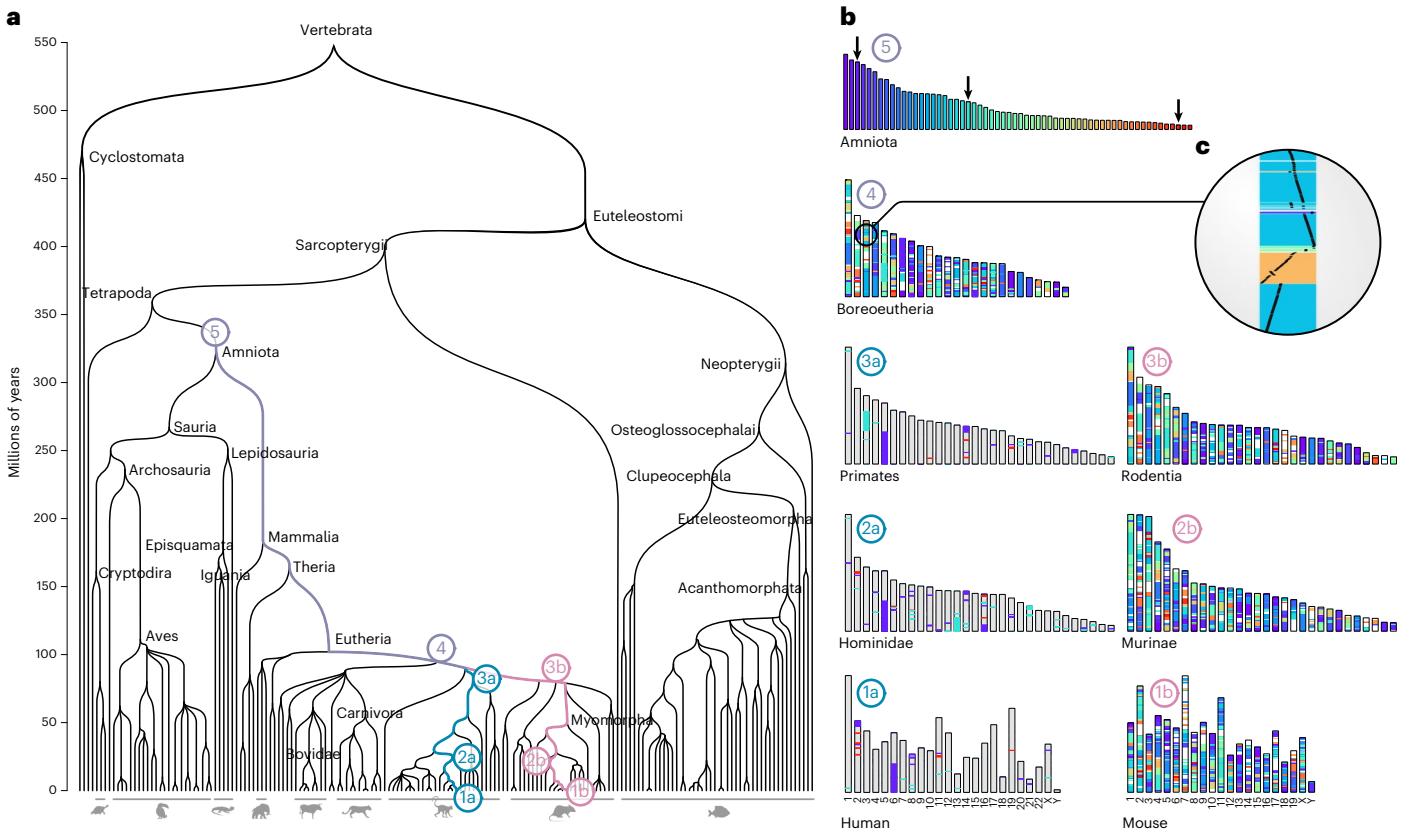


Fig. 1 | Reconstructing vertebrate ancestral genomes. a, Species phylogeny of vertebrates encompassing genomes stored in Ensembl v.92 with indications of the eight ancestral genomes detailed in **b** and the evolutionary path that they mark out. **b**,**c**, High-resolution ideograms of ancestral genome reconstructions (**b**) starting from the Amniota genome (5) and the descendant Boreoeutheria genome (4), where a region on the third chromosome is expanded to highlight the evolution of gene organization with respect to the Amniota genome (**c**). In the primate lineage (3a, 2a, 1a) only the evolution of the three Amniota chromosomes indicated by an arrow are depicted in colour, while in the Rodentia lineage (3b, 2b, 1b), the evolution of all Amniota chromosomes is shown.

to other state-of-the-art ancestral genome reconstruction methods[33]. On a different, more realistic benchmark based on simulations that are not restricted to single-copy genes, AGORA achieves 95.4% agreement, while DESCHRAMBLER's performance drops to 68.6% (Supplementary Information, 'benchmarks against simulations'), highlighting AGORA's ability to successfully deal with gene duplications and other complex evolutionary scenarios.

In practice, AGORA is highly flexible because it only requires the protein-coding gene annotations of the extant species and the set of precomputed gene trees in a standard format, which can be downloaded from a variety of genome resource initiatives for many species groups. For example, while the vertebrate ancestral genome reconstructions provided on the Genomicus server are all based on extant genomes annotated by Ensembl, plant and fungi ancestral genomes are based on genome annotations generated by a range of methods and laboratories worldwide. AGORA can be used with other markers than protein-coding genes, such as conserved non-coding elements; however, due to unreliability of phylogenetic trees for those sequences, we recommend limiting the reconstructions to the order of protein-coding genes for best performance. AGORA can also be used iteratively to assemble blocks of markers and scaffold them over several rounds of reconstruction into larger contiguous ancestral regions (CARs). We propose several workflows customized for different clades and applications as part of the AGORA package (Supplementary Fig. 1).

In this study, to demonstrate the capabilities of AGORA, we used two datasets from distant eukaryotic clades, with different numbers of species, genes and variable gene tree reliability: (1) a dataset of 93 vertebrates and 5 outgroups and their 23,528 gene trees, including a total of 1,814,614 extant protein-coding genes, and leading to the reconstruction of 81 ancestral genomes; and (2) a dataset of 58 plant genomes and 8 outgroups, corresponding to 48 ancestral genomes (Methods, Supplementary Data 4 and Supplementary File 1).

Reconstruction of key chromosome-scale ancestral genomes

For every ancestral genome, we provide two valuable results: the gene set and an assembly of their ancestral organization. To evaluate the completeness and accuracy of the ancestral gene sets, we first compared the total number of genes inferred in an ancestor to those of its descendant extant genomes. While very distant genomes can contain widely different numbers of genes, AGORA is designed to be used within clades where synteny is reasonably conserved, such as vertebrates, grasses or *Saccharomycetales* yeasts, and where genomes typically contain similar numbers of genes. We found that our methodology accurately estimated ancestral gene contents that were consistent with those of the descending clades, up to evolutionary distances of over 300 million years ago (Ma) (Fig. 3a). We also find that the vast majority of clade-relevant benchmark universal single-copy orthologue (BUSCO)[43] reference sets are present as single-copy genes in our inferred ancestral gene sets (Fig. 3b). In addition, we also confronted our inferred ancestral gene contents for seven key vertebrate ancestors to those calculated by Ancestral Genomes, another effort to estimate the ancestral gene content, but not gene order, at different evolutionary nodes[29]. Ancestral Genomes relies on the PANTHER database[44] and therefore uses an independent set of extant genomes and gene trees. AGORA and Ancestral Genomes both inferred highly similar gene contents for the same ancestors (Fig. 3c).

https://doi.org/10.1038/s41559-022-01956-z

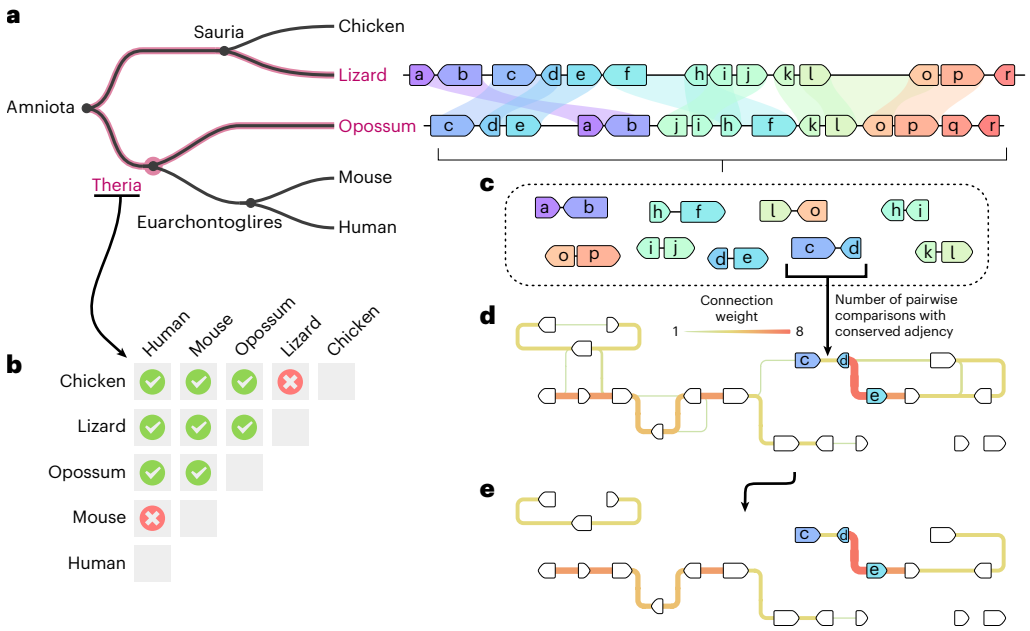


Fig. 2 | Principle of the AGORA approach. a, Conserved gene adjacencies are identified between all genome pairs that are informative for a given target ancestral genome. A portion of the lizard and opossum genomes are shown, with gene adjacencies joined by a pale coloured shape when conserved, thus supporting their prior occurrence in Theria. **b**, All comparisons between genomes that are joined in an evolutionary path intersecting the target ancestor are informative (green ticks) while comparisons between genomes that diverged after the target ancestor are uninformative (red crosses). **c**,**d**, Conserved adjacencies observed in each pairwise comparison (**c**) are collected in a graph structure (**d**) where nodes are genes and links are conserved adjacencies weighted by the number of times they have been observed in pairwise genome comparisons. **e**, The linearization of this graph by traversing the links of maximal weight provides contiguous and parsimonious ancestral gene order reconstructions.

The other output of AGORA is the reconstruction of the putative gene order in each ancestral genome along the species tree. The quality of an ancestral genome reconstruction can be evaluated by two criteria, that is, contiguity and consistency with evolutionary and biological evidence. Contiguity represents the size of the genomic regions that can be assembled into CARs, akin to measures of assembly quality for reference genome sequences. For 37 vertebrate ancestral genomes and 13 plant ancestral genomes in our test set, we obtained chromosome-scale assemblies with a small number of long CARs containing hundreds of ordered and oriented genes, corresponding to a best approximation of the ancestral karyotype (Fig. 3a). These chromosome-level assemblies include over 70% of the ancestral genes, which is comparable to well-assembled extant reference genomes in those clades (Supplementary Fig. 10). Most other ancestral genomes are assembled into fewer than 100 subchromosomal gene blocks containing over 70% of the ancestral gene content (Supplementary Fig. 11).

As expected, the contiguity of ancestral genome reconstructions was high overall in recent ancestors and decreased sharply after 100 Ma, decaying to large numbers of short, unassembled gene blocks for very ancient ancestors such as the Tetrapoda and Vertebrata ancestors (Fig. 3a). However, perhaps counterintuitively, AGORA performs better in some key older ancestors than in comparatively younger ancestral genomes. For example, the genome of Boreoeutheria, the ancestor of most placental mammals (approximately 95 Ma), is a near-complete assembly consisting of 25 large CARs covering 18,430 genes (80% of the total ancestral genome), while the genome of Afrotheria, the ancestor of the elephant and hyrax (approximately 90 Ma), is appreciably less contiguous with 70% of genes in 83 CARs. This reflects the position of these ancestors in the species tree relative to the sampling of sequenced extant genomes. As demonstrated previously[45], ancestors that precede evolutionary radiations are ideally positioned for ancestral genome reconstruction because their many outgroup and descendant lineages offer a large number of informative pairwise comparisons (N_i). Overall, AGORA's ancestral reconstruction contiguity correlates with the N_i/age ratio (Supplementary Fig. 12). Because sequencing efforts have largely targeted organisms within species-rich phyla, such as placental mammals or monocotyledon plants, the key ancestors to these widely studied subclades are particularly well reconstructed by our methodology, which should be of high value to evolutionary and functional studies. Ultimately, however, with the advent of massive sequencing undertakings such as the Vertebrate Genome Project, genome documentation in undersampled clades will increase dramatically and we expect that most ancestral genomes in the Genomicus database will eventually become chromosome-level assemblies.

Support from cytological evidence and in silico palaeogenomes

The accuracy of ancestral genome reconstructions is appreciably more difficult to evaluate than completion because the true ancestral genome sequences are inaccessible at the evolutionary scales we study. However, several ancestral genomes have garnered longstanding interest from the evolutionary genomics community, resulting in a large body of biological evidence regarding their overall organization. In vertebrates, one of the most studied ancestral genomes is Boreoeutheria, the 95 million-year-old ancestor to most placental mammals including primates, rodents, hooved mammals and carnivores, with the exception of afrotherians (elephants) and xenarthrans (tree sloths, anteaters, armadillos), along with the Eutheria ancestor (102 million-year-old, ancestral to boreoeutherian mammals and afrotherians) and the Simian ancestor (45 million-year-old, ancestral to platyrrhine and catarrhine primates). Landmark ancestral Eutheria, Boreoeutheria and Simian karyotypes have previously been reconstructed by integrating dozens of mammalian homology comparisons using fluorescent DNA probes, a technique known as chromosome painting[46,47]. This analysis suggested that the ancestral placental genome consisted of 23 pairs of chromosomes and traced the large-scale rearrangements that resulted into the karyotypic arrangement of the human genome. The Boreoeutheria ancestral genome organization inferred by AGORA contains 25 large CARs and

https://doi.org/10.1038/s41559-022-01956-z

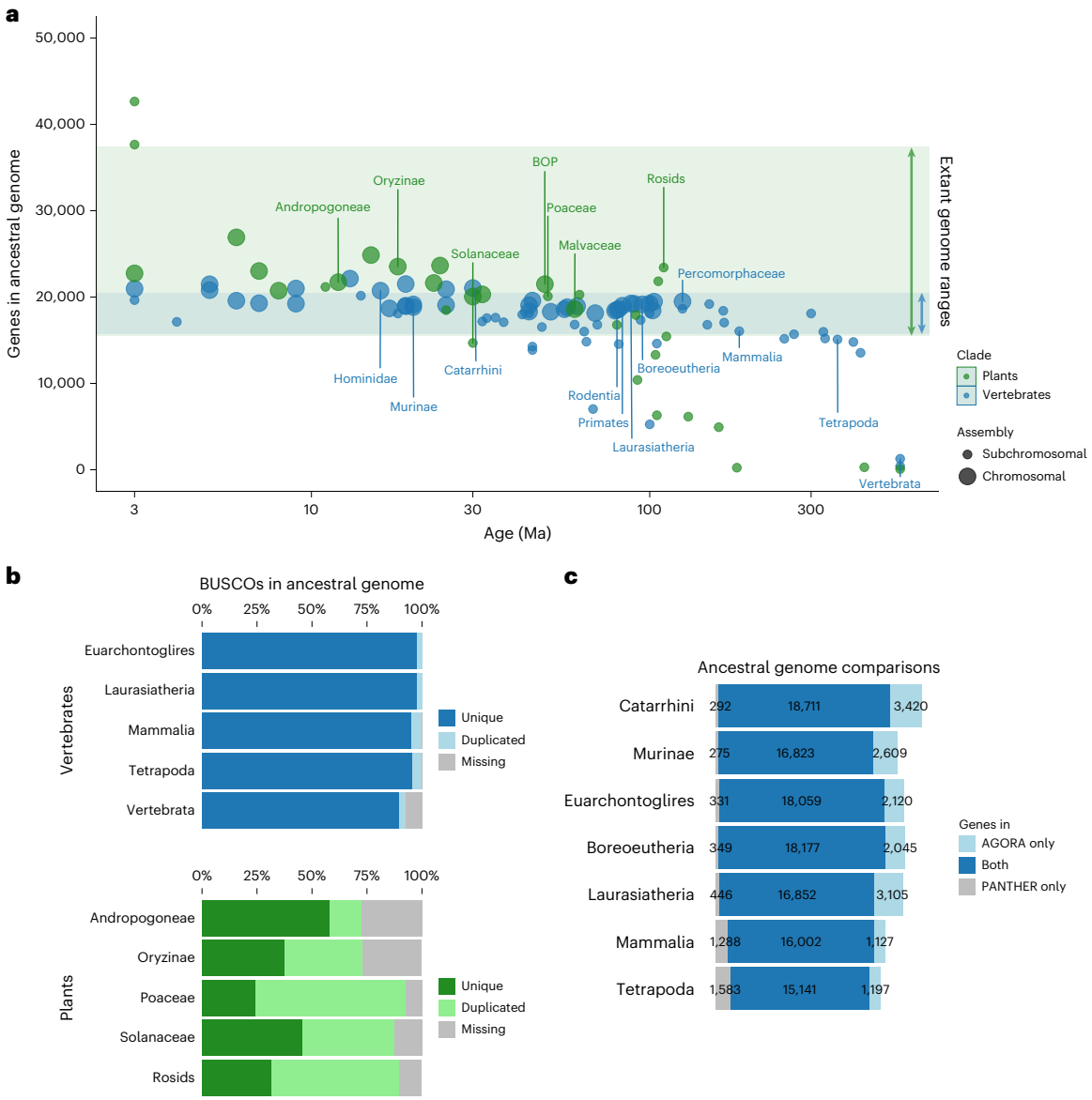
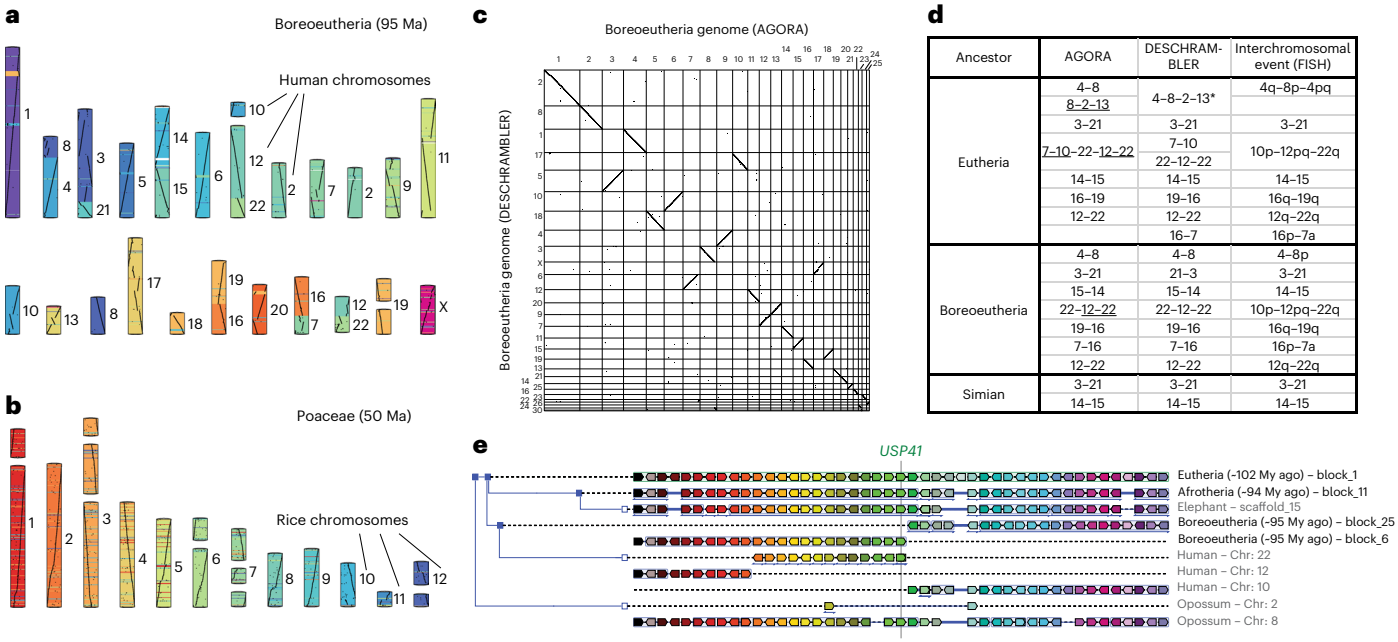


Fig. 3 | Completion of ancestral genomes reconstructed by AGORA. a, Gene content and assembly continuity of 77 vertebrate and 33 plant ancestral genomes reconstructed by AGORA. The ranges of gene contents of extant vertebrate and plant genomes are highlighted in blue and green shading, respectively. Very young (<2 Ma) or very old (>550 Ma) ancestors are not represented. Chromosomal and subchromosomal assemblies are as defined in the Methods. The BOP ancestor stands for the ancestor of the Bambusoideae, Oryzoideae and Pooideae groups. **b**, Representation of BUSCO genes in AGORA's ancestral genomes. Plant genomes, which have undergone rounds of WGDs, frequently contain a large fraction of duplicated genes. **c**, Comparison of ancestral gene contents inferred by reconstructions from AGORA and PANTHER.

is highly congruent with the cytogenetically based reference karyotype (Fig. 4a). AGORA recovers all ancestral chromosomal arrangements supported by cytogenetic evidence without requiring manual assembly or curation. The only exception is the ancestral linkage of human chromosomes 10 and 12 alleged by cytogenetic data (Fig. 4c–e), which is supported neither by AGORA nor by the state-of-the-art reconstruction by DESCHRAMBLER or other in silico ancestral genome reconstruction methods[33]. Detailed manual investigation of inconsistencies between the ancestral reconstructions by AGORA and the cytogenetic references revealed that most differences are the result of the lower resolution of the chromosomal painting methodology and confirmed our proposed assembly (Supplementary Figs. 14 and 15). At the infrachromosomal scale, we found that the genomic organization of the Boreoeutheria genome inferred by AGORA is in near-perfect agreement with that of DESCHRAMBLER (Fig. 4c, Supplementary Fig. 16 and Methods). However, our reconstructed Boreoeutheria genome is more complete and includes the ancestral locations of an additional 2,023 genes (8% of the ancestral gene set) due to operating at a higher resolution. AGORA also fared better by including more species and more recent assemblies than DESCHRAMBLER. Altogether, these results support that the gene-based reconstruction algorithm of AGORA is highly consistent with current ancestral reconstruction methods, while providing a notable increase in resolution for the study of local genomic events. We further tested the robustness of AGORA to varying input datasets by reconstructing an alternative Boreoeutheria ancestral genome using gene families from hierarchical orthology groups built with OMA[48], a completely different gene orthology inference pipeline from Ensembl Compara. Both reconstructions were remarkably convergent with over 96% similarity (Supplementary Information, 'Comparison between Ensembl Compara and OMA hierarchical orthology groups' and Supplementary Fig. 17), supporting that AGORA performs well regardless of gene orthology data sources.

https://doi.org/10.1038/s41559-022-01956-z



d

Ancestor	AGORA	DESCHRAM-BLER	Interchromosomal event (FISH)
Eutheria	4–8	4–8–2–13*	4q–8p–4pq
	8–2–13		
	3–21	3–21	3–21
	7–10–22–12–22	7–10	10p–12pq–22q
		22–12–22	
	14–15	14–15	14–15
	16–19	19–16	16q–19q
	12–22	12–22	12q–22q
		16–7	16p–7a
Boreoeutheria	4–8	4–8	4–8p
	3–21	21–3	3–21
	15–14	15–14	14–15
	22–12–22	22–12–22	10p–12pq–22q
	19–16	19–16	16q–19q
	7–16	7–16	16p–7a
	12–22	12–22	12q–22q
Simian	3–21	3–21	3–21
	14–15	14–15	14–15

Fig. 4 | AGORA ancestral genome reconstructions compared to extant genomes and state-of-the-art ancestral reconstructions. a, The Boreoeutheria karyotype inferred by AGORA (the 25 largest CARs), coloured according to gene locations on human chromosomes, as indicated to the right of each CAR. **b**, The Poaceae karyotype inferred by AGORA (the 19 largest CARs), coloured according to gene locations on *Oryza sativa* chromosomes, as indicated to the right of each CAR. **c**, Collinearity of the Boreoeutheria ancestral genome reconstructed by AGORA with the genome reconstructed by DESCHRAMBLER[33]. **d**, Comparisons of computational reconstructions by AGORA and DESCHRAMBLER and Zoo-fluorescence in situ hybridization (FISH) linkage groups inferred for three key mammalian ancestors. Human chromosomes in ancestral linkage are indicated with hyphens. The Eutheria bolded linkage group 10–22–12 is documented in more detail in **e**. The underlined linkage groups are documented in Supplementary Fig. 14. DESCHRAMBLER reconstructed a linkage group between parts of human chromosomes 4, 8, 12 and 3 (asterisk) in disagreement with FISH evidence and AGORA when used on Ensembl v.92 data; however, this linkage group is also reconstructed by AGORA on Ensembl data v.102, suggesting an ambiguous ancestral linkage state (Supplementary Fig. 15). **e**, Gene adjacencies around the *USP41* gene in extant species support the linkage of fragments of human chromosomes 10, 22 and 12 in the Eutheria ancestor. Orthologous genes are shown as arrows in matching colours, pointing in the direction of transcription. Opossum and elephant have both retained the ancestral organization at this locus, which has been rearranged in the human genome.

Finally, we also examined the genome of Poaceae, the 50 million-year-old ancestor of grasses, reconstructed by AGORA to an earlier reference ancestral karyotype[49] obtained by another parsimony-based method to reconstruct ancestral adjacencies[30]. Again, the ancestral genome reconstructed by AGORA closely recapitulates the state-of-the-art knowledge regarding the organization of the ancestral grass karyotype (Fig. 4b), while providing access to a fine-scale reconstruction of the ancestral gene order.

A scalable framework to integrate genomes across phylogenies

A major strength of AGORA resides in its ability to compute the gene order of every ancestor in a phylogeny using different subsets of the same extant genome comparisons. In a context where new species genomes are being sequenced with increasing speed and accuracy, comparative genomics need methods that can integrate evolutionary information along the species tree and across lineages without relying on a single extant genome as reference. Using the legacy architecture of the Genomicus synteny database[34,35], which is updated with every new release of the Ensembl database, we tested how our methodology scales with the number of extant reference genomes available as well as their quality (Supplementary Fig. 13). Ensembl Compara v.101 included the reference sequences of 264 vertebrate species and five outgroups, for a total of 5,539,325 extant protein-coding genes organized into 62,478 gene trees. Using this information, AGORA reconstructed a total of 265 ancestral genomes along the species tree in 6 h and 50 min on a Linux machine with four central processing units and approximately 80 GB of random access memory (Supplementary Data 2). Therefore, AGORA is computationally inexpensive and can be run on a desktop machine for small-to-medium datasets. However, AGORA can also be parallelized and is optimized for usage on a computing cluster for large applications and database updates.

Overall, the quality of key ancestral genomes increases as new extant genomes are included in the database (Supplementary Fig. 13). The introduction of high-quality reference genomes in under-represented clades over time has contributed to the reconstruction of previously inaccessible ancestors, such as Strepsirrhini, the ancestor of lemurs, bushbabies and lorises, and more recently Chiroptera, the ancestor of bats. Interestingly, we observed that even the inclusion of low-contiguity, fragmented genomes markedly improves ancestral genome reconstructions. For instance, including low-contiguity genomes more than doubles the median value (G50; Methods) for the reconstructed Amniota genome (Supplementary Information, 'Impact of low-contiguity assemblies'). This is likely because different reference genomes are generally assembled independently and assembly errors rarely produce the same erroneous gene arrangements from one genome to the next. Because AGORA only considers conserved gene adjacencies as potentially ancestral, the additional information from correctly assembled scaffolds offsets the noise introduced by assembly errors, which are discarded as not conserved. Therefore, we argue that the inclusion of low-cost, fragmented reference genomes in comparative genomics databases serves a purpose beyond gene-based analyses.

Ancestral genomes as backbones for evolutionary studies

In this section, we experimented the paradigm shift. This consisted of studying genome evolution from the perspective of multiple reconstructed ancestral genomes. We first revisited known observations

https://doi.org/10.1038/s41559-022-01956-z

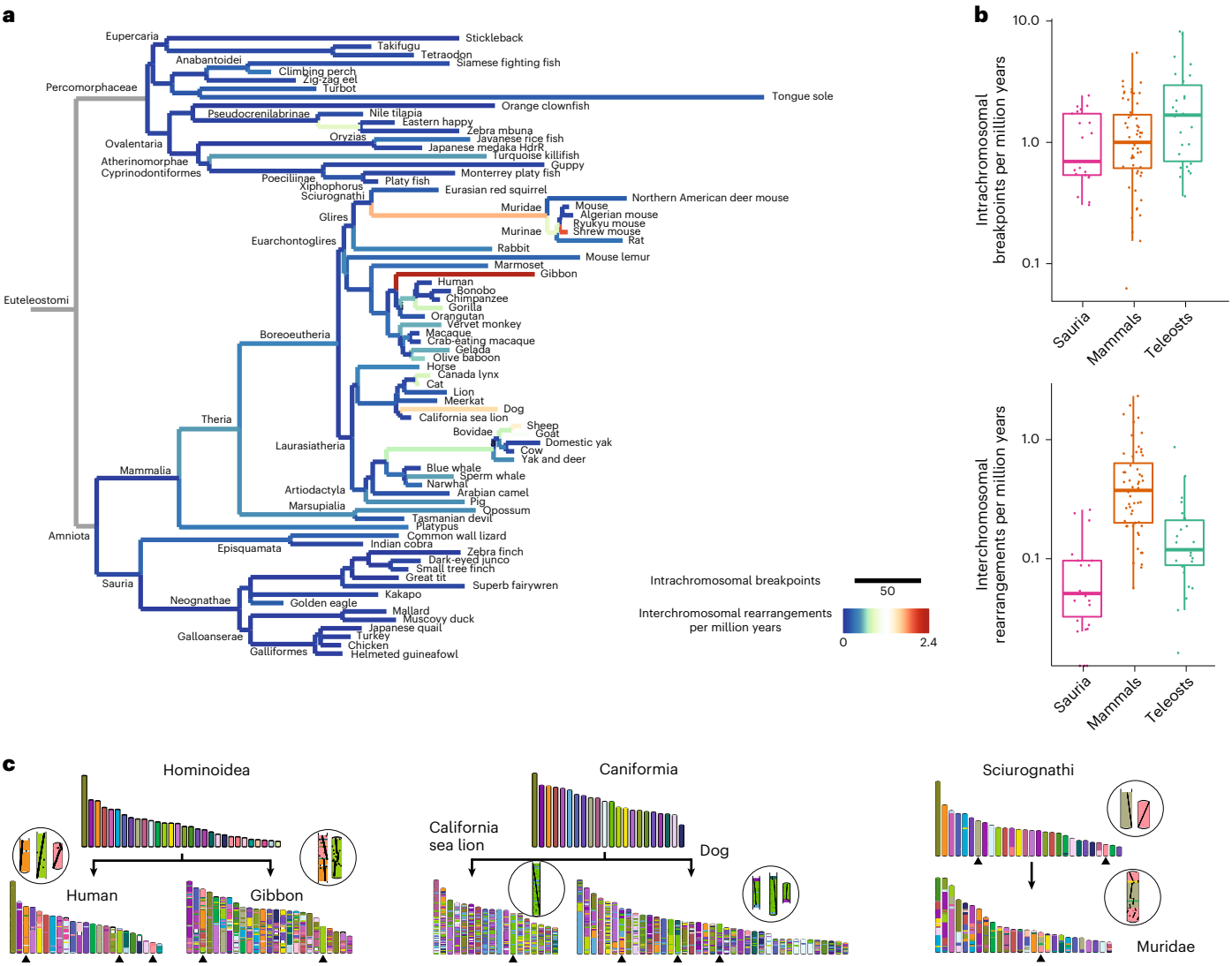


Fig. 5 | Vertebrate genome evolutionary dynamics. a, Phylogenetic tree of 74 extant and 73 ancestral genomes, where branch lengths represent the number of breakpoints computed between successive nodes. The colours represent the rate of interchromosomal rearrangements. The branches in grey connect ancestral genomes to their Euteleostomi root, which is too fragmented as a genome reconstruction to serve as reference for computing breakpoints and rearrangements. **b**, Distributions of breakpoints and rearrangement rates represented in **a**, broken down into three taxonomic groups: saurians (birds and reptiles); mammals; and teleosts (fish). The centre line of each plot corresponds to the median, the box edges correspond to the first and third quartiles, the whiskers correspond to 1.5× the interquartile range. **c**, Examples of rearrangements in lineages with notable rates of evolution. The black upward arrowheads point to chromosomes that are shown enlarged in the circles, with individual orthologous genes drawn as black dots.

generated by traditional comparative genomics based on extant genomes. As a case study, we used ancestral reconstructions to investigate the patterns of karyotypic rearrangements that occurred during the evolution of mammals, birds and ray-finned fish (Fig. 5). These three groups represent the main jawed vertebrate (Euteleostomi) lineages, whose respective chromosomal dynamics have been documented using comparative genetics, cytogenetics and genomics approaches across different taxonomic groups. We selected 73 well-reconstructed ancestors and their 74 extant descendants (15 birds and reptiles, 41 mammals, 18 fish; Methods) from the Genomicus Vertebrates database v.102, which contains a total of 269 extant and 265 ancestral genomes. We then compared consecutive genomes on all 131 branches of the phylogenetic tree, representing a combined time of about 5 billion years of independent evolution, and traced gene adjacencies that were rearranged on each branch (Methods). In total, we identified 5,749 rearrangement breakpoints that occurred along the 131 branches (average rate 1.17 breakpoint per million years), most of which are intrachromosomal. We also identified 1,370 interchromosomal rearrangements (translocations, fusions or fissions) with an average rate of 0.28 rearrangement per million year (Fig. 5a and Supplementary Data 3). These rearrangement rates are lower bound values because rearrangements occurring between genes without disrupting gene order or orientation cannot be observed (Discussion). Comparing rates per million years, and restricting the analysis to the 105 branches longer than 5 Ma to avoid small sample distortions, we confirm that birds and reptiles have more stable chromosomal structures than mammals, as reported previously[50,51], with lower rates of interchromosomal rearrangements ($P = 3.8 \times 10^{-6}$, Wilcoxon rank-sum test; Fig. 5b). Fish in turn display higher intrachromosomal breakpoint rates than mammals, birds and reptiles (teleosts versus saurians, $P = 0.0181$; teleosts versus mammals, $P = 0.0532$, Wilcoxon rank-sum test), which is consistent with the rediploidization process following the whole-genome duplication (WGD) that occurred in this phylum[52], yet they display a uniformly high karyotypic stability.

https://doi.org/10.1038/s41559-022-01956-z

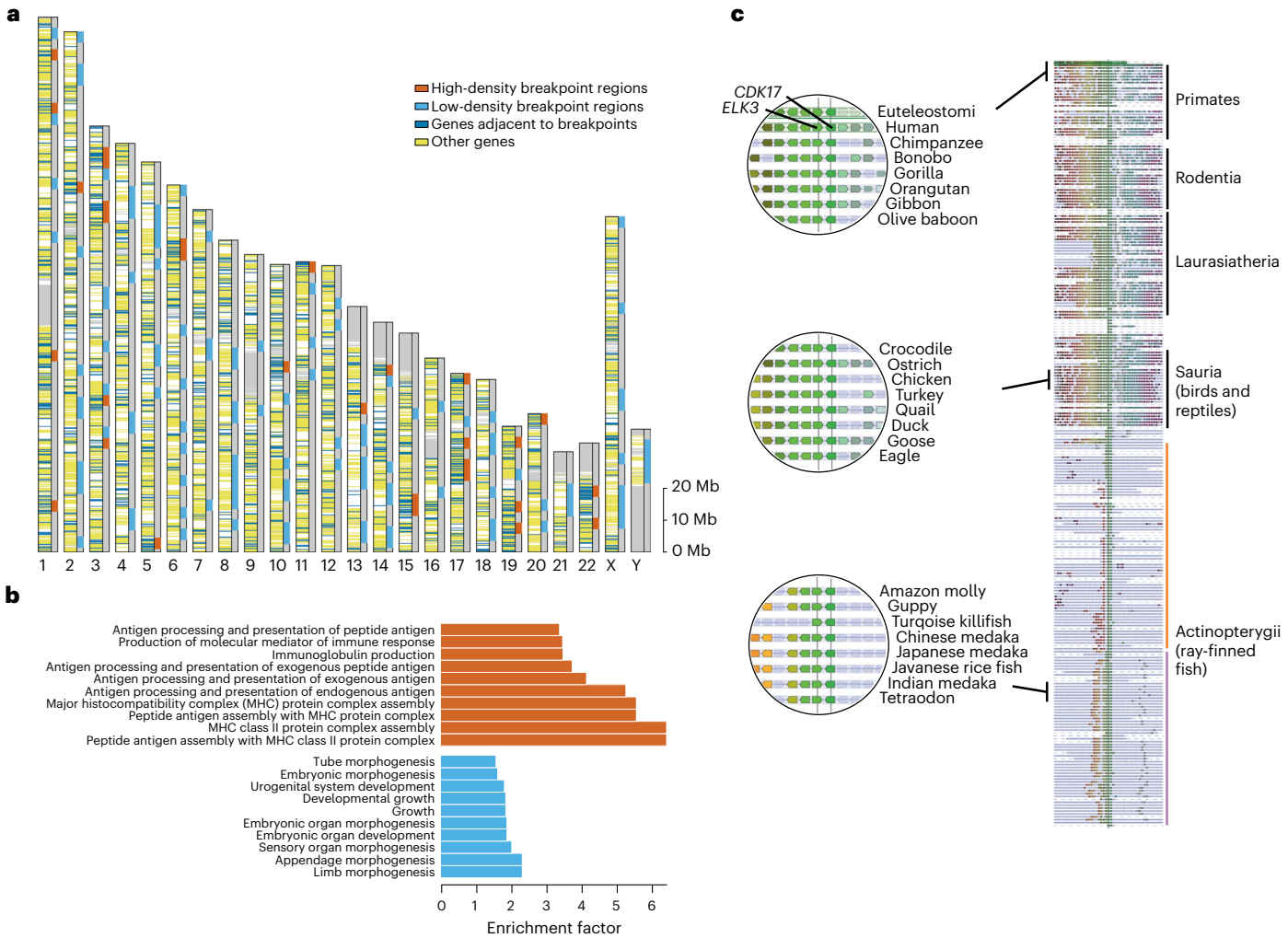


Fig. 6 | Breakpoint map and functional associations. a, Human chromosome map showing the 1,985 genes that flanked a breakpoint at least once during the evolution of Boreoeutheria (dark blue) and the genes that were never adjacent to a breakpoint (yellow). The genome was divided in 5 Mb non-overlapping windows and the red boxes show the top 28 windows (5%) richest in breakpoints, while the light blue boxes indicate the 123 boxes without any breakpoints. **b**, A GO enrichment test showed that breakpoint-rich and breakpoint-poor windows are enriched in genes with very contrasted GO biological functions. The 10 most enriched terms with an FDR < 5.10^{-2} and associated with fewer than 1,500 human genes in the complete genome for each gene category are shown. **c**, The *ELK3-CDK17* genes represent the most conserved gene adjacency in bony vertebrates (Euteleostomi). Right, The *EDK3-CDK17* locus is shown in a Genomicus v.106 phyloview, at the centre of a 45-gene window in the 187 extant Euteleostomi genomes where it is conserved. Genes and loci are not drawn to scale and each gene is represented by an arrow of fixed size pointing in the direction of transcription. The orange and purple vertical lines outline the two groups of teleost fish loci separated by the 3R WGD. Left, The three circles show a zoomed-in region with the *EDK3-CDK17* gene pair in head-to-head orientation, aligned over the vertical black line in each genome.

Interestingly, a few branches in placental mammals stand out as having strikingly high rearrangement rates. For instance, the gibbon lineage is the outlier of our analysis, having experienced 60 interchromosomal rearrangements in 25 My, confirming previous observations that this is a fast-evolving lineage compared, for example, to the human lineage[53] (Fig. 5c). The dog genome was also subject to high rates of rearrangement, especially compared to its sister branch leading to the slowly evolving sea lion genome, which only changed through three chromosome fusions compared to their Caniformia ancestor[54]. The lineage leading to the Muridae is notable for a high rate of intrachromosomal breakpoints combined with multiple interchromosomal rearrangements but associated to a stable chromosome number, which is consistent with cytogenetic studies of different murid clades[55,56]. These examples revisit lineages that are known to be subject to fast evolutionary rates, underlining how AGORA reconstructions agree with current knowledge and represent a sound basis to explore and understand genome evolution.

A key feature of AGORA reconstructions is that they are independently derived for each ancestor, enabling the investigation of evolutionary events in internal branches, between successive ancestral genomes (Fig. 5c, Sciurognathi to Muridae). We exploited this feature to investigate whether rearrangement breakpoints accumulate in specific genomic regions in mammals, where they may present an evolutionary advantage by providing new gene combinations. We collected 2,466 rearrangement breakpoints that occurred across all the boreoeutherian mammal lineages shown in Fig. 5a. This 'breakpoint map' recapitulates almost 1.4 cumulated billion years of genome reorganization, projected on the human genome as a reference. In total, 1,985 human genes are flanked by at least 1 breakpoint (Fig. 6a) and high and low breakpoint density regions are evident. To characterize these further, we identified the 5 Mb windows in the human genome with the highest density of breakpoints (top 5%) and those without breakpoints. A Gene Ontology (GO) analysis showed that high breakpoint intensity occurs near genes involved in the acquired immune system,

while breakpoints are depleted in regions flanking genes involved in embryonic development (Fig. 6b), confirming on a broad scale previous observations[57,58]. Genomic regions involved in immunity are fast-evolving at the sequence level, typically interpreted as evidence of positive selection: in this study, we show that these regions are also fast-rearranging; further investigation may reveal whether genomic reorganization acts in concert with sequence evolution to produce functional novelty in these regions.

Finally, we took advantage of the unique standpoint provided by ancestral genomes to investigate which gene-to-gene interval is the most conserved in all bony vertebrates. Scanning the Euteleostomi ancestral genome, we selected the gene adjacency with the strongest support from 8,173 pairwise genome comparisons used to reconstruct this genome. The adjacency between the *ELK3* and *CDK17* genes is ancestral to bony vertebrates and remains conserved in 187 out of 192 descendant genomes available in the Ensembl 106 database (Fig. 6c). Interestingly, *ELK3* and *CDK17* coexpress sense–antisense messenger RNA transcripts in mouse neuronal cells[59]. Additionally, *ELK3* introns contain enhancer sequences that putatively target the *CDK17* promoter[60]. Potential complex regulatory functions may be associated with this locus because a sense–antisense transcript produced in the same cells can lead to double-stranded RNA, and in this case, also overlap the *ELK3* coding exon. Further investigation should reveal whether the same coexpression occurs in all Euteleostomi, which would suggest an ancestral function established early in vertebrate evolution and a possible explanation for the extensive linkage conservation at this locus.

Discussion

Biology is a historical science but this historical dimension is often ignored because the records required to document ancestral states are missing. Without this chronological perspective, the reasons why contemporary biological systems are organized as they are will continue to elude us. In practice, this information gap hinders our ability to integrate conclusions across different living models and to draw the full benefits of comparative genomics. Ancestral genomes are fundamental blocks of the conceptual framework aiming to address this problem. They complement fossils as biological time points because they are a theoretical representation of the precise divergence between two lineages, while fossils represent true extinct species but whose exact phylogenetic position is often unclear. Because ancestral genomes encapsulate all the genes present in the ancestral organism and their structural organization, they will enable detailed investigations of developmental and metabolic pathways evolution, such as the expansion and contraction of specific gene families over time; the contribution of genome structure changes to evolutionary transitions and speciations; and the tracing of evolutionary innovations through reorganization of functional gene arrangements. Additionally, ancestral genomes can act as unique reference points to compare multiple descendant genomes, removing the bias of relying on an extant genome as central reference. This property makes them powerful tools to identify, measure and study lineage-specific genomic events and clade-wide trends.

Reconstructions of ancestral genomes by AGORA have a number of limitations. First, the method relies on the assumption of parsimony, which is widely used for both cytogenetic and marker-based bioinformatics reconstructions. This premise is reasonable because intergenic breakpoints are rare (fewer than ten per million years in eukaryotes) and conservative scenarios involving the fewest steps are likely to be correct in the vast majority of cases. However, breakpoint reuse can occur[61] and will violate this assumption, which may result in non-reconstructed gene adjacencies (false negatives) in the AGORA reconstructions but will not create erroneous adjacencies (false positives). Thus, breakpoint reuse may cause reconstructions to be more fragmented but should not induce incorrect links between markers. It will, however, cause an underestimation of breakpoint rates as presented in Fig. 5, although there is no evidence that it should distort the relative rates between taxonomic groups. Conversely, a false positive adjacency present in a given ancestor but absent in the previous one and in the next one in a chronology, will give a false signal of breakpoint reuse. Other limits are less due to the method but are inherent to the underlying data. For example, in this study we used gene trees to define the set of ancestral genes to be ordered into chromosomes and to locate the set of descendant genes in extant genomes (orthogroups). Although we showed that two different sources of orthogroups (Ensembl and OMA) generate essentially the same Boreoeutheria genome, this may be different for more ancient genomes or poorer-quality gene trees. In particular, incorrect placement of duplication events will affect the number of ancestral genes and incorrect partitioning of extant copies under their ancestral duplicate copy will affect the adjacencies that can be deduced. This issue is amplified after WGD events, where all genes are duplicated at once, but can be mitigated by tree edition steps as implemented in SCORPiOs[62]. WGD are not obstacles per se for genome reconstruction. Several instances occurred in vertebrate, plant and fungi genome evolution and AGORA can reconstruct ancestral genomes at speciation nodes immediately flanking the event. This is the case, for example, between the Protacanthopterygii and Salmoninae in fish that flank a single WGD, and between the Malvids and Brassicaceae ancestors in plants that flank two successive WGDs. In each case, the classical 'double-conserved synteny' pattern[52,63,64] is clearly visible across ancestral chromosome segments hundreds of genes long (Supplementary Fig. 18). The density of markers (that is, protein-coding genes) also limits the resolution of the reconstructions because intermarker space consists of blind spots where inversions contained within cannot be observed. As algorithms mature, ancestral genomes such as those presented in this study could become enriched with many more features, including non-coding sequences such as ancestral repeat elements, non-coding RNA genes or regulatory elements, and serve as organizational maps for reconstructed[65] or fossil nucleotide sequences. Reaching this goal could alleviate some limitations of gene-based ancestral reconstructions by providing a much-increased resolution.

Because genome sequencing costs continue to decrease, reference genomes are becoming widely available for model and non-model species alike. At the time of writing, the NCBI database accounts for a total of 8,505 eukaryote, 32,172 bacterial and 1,909 archaeal whole-genome sequencing projects and dedicated efforts such as the Vertebrate Genome Project[66] promise to deliver extensive phylogenetic coverage across many clades. Integrating sequence and genome organization evolution over such massive phylogenetic samplings remains a challenge. Many phylogenomics projects still rely on sequence alignments as a means to study how genome organization evolves[33,51]. Aligning whole genomes is computationally expensive, and while new methodologies are emerging to step up to the challenge[23,67], the requirements to handle hundreds of genomes remain out of reasonable reach for many. Additionally, identifying conserved and rearranged regions from whole-genome alignments becomes technically difficult as phylogenetic distance increases, especially in large genomes where an important fraction of the sequence is non-coding and repetitive. Due to these limitations, the evolution of genome organization is typically studied at large scale, but low resolution, and/or in a limited sampling of species, often those included in publicly available, reference multispecies alignments. Marker-based ancestral genome reconstructions provide an alternative to methods based on whole-genome alignments by relying on gene phylogenies instead, which require much more modest computational infrastructures and scale up to hundreds of genomes with relative ease. In the future, as polymorphism information becomes available for more extant species, we may expect to see ancestral genomes move on from unique references to compendiums, representing structural genomic variation present at any given point in

time and opening the door from increasingly sophisticated population genomics models of molecular evolution.

Methods

Data collection

Genes and gene trees were downloaded from Ensembl v.92 (ref. 68) and Ensembl Plants v.41 (ref. 69). Ensembl v.92 gene trees were edited for poorly supported duplication nodes as described previously[70], as part of the standard build procedure for the Genomicus synteny database. Of note, this step only marginally improves ancestral genome reconstructions and is not a prerequisite to use AGORA. The species trees for the extant and ancestral genomes from Ensembl v.92 and Ensembl Plants v.41 are described in Supplementary File 1.

Ancestral genome reconstructions

Ancestral gene sets and gene orders were reconstructed for 82 ancestors on Ensembl v.92 data using AGORA with 2 passes and a tree parameter of 0.35, and for 41 plant ancestors in 2 multi-integration passes without tree selection (Supplementary Data 4). The details of the AGORA algorithm, validations by evolutionary simulations, suggested procedure to select an optimal tree parameter and advances compared to earlier publications are detailed in the Supplementary Information ('AGORA method').

Statistics on ancestral genomes

Ancestral genome contiguity was measured using the L70 and G50 metrics. L70 is the smallest number of CARs adding up to 70% of the total genome length, measured in gene units. G50 is the length of the ancestral CAR such that 50% of the total genome length, measured in gene units, is contained in larger CARs. Vertebrate chromosomal assemblies have an L70 < 100 and G50 > 450 and plant chromosomal assemblies have an L70 < 20 and a G50 > 450. These values correspond to well-assembled extant genomes (Supplementary Fig. 10) from these respective clades. Other assemblies were considered subchromosomal.

Comparisons to reference ancestral gene sets

We downloaded the BUSCO sets v.3 (ref. 43) based on OrthoDB v.9 (ref. 71). BUSCO gene identifiers were converted to Ensembl gene IDs using the conversion tables provided by the OrthoDB. A BUSCO orthogroup is a set of near 1-to-1 orthologous genes across sequenced genomes of a relevant phylum. An ancestral gene inferred by AGORA was identified as a BUSCO if two or more of its extant descendant genes were contained in the same orthogroup. When a single ancestral gene had descendants in more than one BUSCO orthogroup, we chose the orthogroup with the highest overlap. We then computed the number of BUSCOs matched to a single ancestral gene, to two or more ancestral genes (dubious duplication) and absent from the ancestral genome reconstructed by AGORA (missing gene). Independent ancestral gene sets were downloaded from Ancestral Genomes[32], based on PANTHER v.13.1 (ref. 44). Because Ancestral Genomes and AGORA use different sets of extant species, we only considered ancestral genes with descendants in one of their common species for comparison (human for all ancestors except Murinae and Laurasiatheria where mouse and dog were used, respectively). Ancestral Genomes ancestral genes were converted from UniProt knowledge base IDs to Ensembl gene IDs using the correspondence tables provided by Ensembl BioMart and compared with the gene sets in the ancestral genomes reconstructed by AGORA.

Comparison between AGORA and DESCHRAMBLER eutherian ancestor

We compared AGORA's v.92 eutherian reconstructions to DESCHRAMBLER's[33] (300 kb resolution: APCF_hg19_merged.map from http://bioinfo.konkuk.ac.kr/DESCHRAMBLER/). Because DESCHRAMBLER uses segments of the human genome as units of the reconstruction and was based on the hg19 genome assembly, we converted those regions to their protein-coding gene content and selected the genes still found in Ensembl v.92 and descendants of ancestral boreoeutherian genes. The Oxford grid plot was generated with the AGORA src/misc.compareGenomes.py script in 'matrix' mode.

Vertebrate evolutionary dynamics

Ancestral genomes reconstructed by AGORA from Ensembl v.102 were filtered to retain the most contiguous reconstructions, resulting in 73 ancestral genomes with G50 > 230 and L70 < 40. Conserved syntenic blocks between successive ancestral genomes in internal branches, and between ancestral genomes and their extant descendant in terminal branches, were computed with PhylDiag[72]. Ends of blocks corresponding to likely evolutionary breakpoints were identified using ad hoc scripts. Orthologous genes between successive genomes were also compared in terms of their assignation to scaffolds or chromosomes larger than 200 genes using AGORA's src/misc.compareGenomes.py script in 'printOrthologousChrom' mode. Groups of at least 20 genes relocating to more than 1 chromosome in a descendant genome, and inversely groups of at least 20 genes from 2 or more ancestral chromosomes relocating on the same descendant chromosome, were considered interchromosomal rearrangements. Breakpoint and rearrangement rates per million years were computed using branch length estimates from TimeTree[73]. A full description of the parameters and selection thresholds are provided in the Supplementary Information ('Vertebrate genome evolutionary dynamics').

GO analysis

Human genes from Ensembl 106 contained in 5 Mb windows with the 5% highest number of breakpoints or with no breakpoints were tested for GO term enrichments (biological function) against the rest of the human genes, using the PANTHER web server[44] (version 17.0). Enrichment was tested with Fisher's exact test; terms with a false discovery rate (FDR) < 0.05 were retained. Control experiments with random selections of windows with the same gene densities as found in the 0-breakpoint windows and 5% richest windows did not show significant enrichment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Ancestral genomes have been precomputed for approximately 200 vertebrate (depending on the release), 41 plant and 222 fungi genomes and are available on the Genomicus database FTP server (ftp://ftp.bio.ens.psl.eu/pub/dyogen/genomicus/). These ancestral genomes can also be explored visually within the Genomicus[35] synteny browser (http://www.genomicus.bio.ens.psl.eu/genomicus). Ancestral genomes and the data used in this article for analysis are archived on a *Zenodo* repository (https://doi.org/10.5281/zenodo.7479507)[74].

Code availability

The source code of AGORA, user instructions and a test dataset are available for download from https://github.com/DyogenIBENS/Agora.

References

1. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**, 357–366 (1965).
2. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
3. Yang, Z. Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* **42**, 294–307 (1996).
4. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, 2000).

5. Suzuki, Y. & Gojobori, T. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **16**, 1315–1328 (1999).
6. Yang, Z. & Nielsen, R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.* **25**, 568–579 (2008).
7. Lupski, J. R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet.* **1**, e49 (2005).
8. Rowley, J. D. Letter A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243**, 290–293 (1973).
9. Lupiáñez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
10. Tawn, E. J. & Earl, R. The frequencies of constitutional chromosome abnormalities in an apparently normal adult population. *Mutat. Res.* **283**, 69–73 (1992).
11. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
12. Despang, A. et al. Functional dissection of the *Sox9–Kcnj2* locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).
13. Schultz, J. & Dobzhansky, T. The relation of a dominant eye color in *Drosophila melanogaster* to the associated chromosome rearrangement. *Genetics* **19**, 344–364 (1934).
14. Wilson, A. C., Sarich, V. M. & Maxson, L. R. The importance of gene rearrangement in evolution: evidence from studies on rates of chromosomal, protein, and anatomical evolution. *Proc. Natl Acad. Sci. USA* **71**, 3028–3030 (1974).
15. Sturtevant, A. H. Genetic factors affecting the strength of linkage in *Drosophila*. *Proc. Natl Acad. Sci. USA* **3**, 555–558 (1917).
16. Dobzhansky, T. & Sturtevant, A. H. Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics* **23**, 28–64 (1938).
17. Joron, M. et al. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**, 203–206 (2011).
18. Lowry, D. B. & Willis, J. H. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8**, e1000500 (2010).
19. Muñoz, A. & Sankoff, D. Detection of gene expression changes at chromosomal rearrangement breakpoints in evolution. *BMC Bioinformatics* **13**, S6 (2012).
20. M. Real, F. et al. The mole genome reveals regulatory rearrangements associated with adaptive intersexuality. *Science* **370**, 208–214 (2020).
21. Loveland, J. L. et al. Functional differences in the hypothalamic-pituitary-gonadal axis are associated with alternative reproductive tactics based on an inversion polymorphism. *Horm. Behav.* **127**, 104877 (2021).
22. Lewin, H. A. et al. Earth Biogenome Project: sequencing life for the future of life. *Proc. Natl Acad. Sci. USA* **115**, 4325–4333 (2018).
23. Armstrong, J. et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
24. Yancopoulos, S., Attie, O. & Friedberg, R. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* **21**, 3340–3346 (2005).
25. Avdeyev, P., Jiang, S., Aganezov, S., Hu, F. & Alekseyev, M. A. Reconstruction of ancestral genomes in presence of gene gain and loss. *J. Comput. Biol.* **23**, 150–164 (2016).
26. Chauve, C., Gavranovic, H., Ouangraoua, A. & Tannier, E. Yeast ancestral genome reconstructions: the possibilities of computational methods II. *J. Comput. Biol.* **17**, 1097–1112 (2010).
27. Tannier, E., Zheng, C. & Sankoff, D. Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics* **10**, 120 (2009).
28. Avdeyev, P., Jiang, S. & Alekseyev, M. A. Linearization of median genomes under the double-cut-and-join-indel model. *Evol. Bioinform. Online* **15**, 1176934318820534 (2019).
29. Vakirlis, N. et al. Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome Res.* **26**, 918–932 (2016).
30. Chauve, C. & Tannier, E. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput. Biol.* **4**, e1000234 (2008).
31. Ma, J. et al. Reconstructing contiguous regions of an ancestral genome. *Genome Res.* **16**, 1557–1565 (2006).
32. Huang, X. et al. Ancestral genomes: a resource for reconstructed ancestral genes and genomes across the tree of life. *Nucleic Acids Res.* **47**, D271–D279 (2019).
33. Kim, J. et al. Reconstruction and evolutionary history of eutherian chromosomes. *Proc. Natl Acad. Sci. USA* **114**, E5379–E5388 (2017).
34. Muffato, M., Louis, A., Poisnel, C.-E. & Roest Crollius, H. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* **26**, 1119–1121 (2010).
35. Nguyen, N. T. T., Vincens, P., Dufayard, J. F., Roest Crollius, H. & Louis, A. Genomicus in 2022: comparative tools for thousands of genomes and reconstructed ancestors. *Nucleic Acids Res.* **50**, D1025–D1031 (2022).
36. Nguyen, N. T. T., Vincens, P., Roest Crollius, H. & Louis, A. Genomicus 2018: karyotype evolutionary trees and on-the-fly synteny computing. *Nucleic Acids Res.* **46**, D816–D822 (2018).
37. Murat, F. et al. Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol.* **16**, 262 (2015).
38. Sacerdot, C., Louis, A., Bon, C., Berthelot, C. & Roest Crollius, H. Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol.* **19**, 166 (2018).
39. Boore, J. L. The use of genome-level characters for phylogenetic reconstruction. *Trends Ecol. Evol.* **21**, 439–446 (2006).
40. Rokas, A. & Holland, P. W. Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* **15**, 454–459 (2000).
41. Drillon, G., Champeimont, R., Oteri, F., Fischer, G. & Carbone, A. Phylogenetic reconstruction based on synteny block and gene adjacencies. *Mol. Biol. Evol.* **37**, 2747–2762 (2020).
42. Zhao, T. et al. Whole-genome microsynteny-based phylogeny of angiosperms. *Nat. Commun.* **12**, 3498 (2021).
43. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
44. Mi, H. et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **49**, D394–D403 (2021).
45. Blanchette, M., Green, E. D., Miller, W. & Haussler, D. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.* **14**, 2412–2423 (2004).
46. Ferguson-Smith, M. A. & Trifonov, V. Mammalian karyotype evolution. *Nat. Rev. Genet.* **8**, 950–962 (2007).
47. Stanyon, R., Stone, G., Garcia, M. & Froenicke, L. Reciprocal chromosome painting shows that squirrels, unlike murid rodents, have a highly conserved genome organization. *Genomics* **82**, 245–249 (2003).
48. Altenhoff, A. M. et al. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res.* **49**, D373–D379 (2021).

49. Murat, F. et al. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* **20**, 1545–1557 (2010).
50. Romanov, M. N. et al. Reconstruction of gross avian genome structure, organization and evolution suggests that the chicken lineage most closely resembles the dinosaur avian ancestor. *BMC Genomics* **15**, 1060 (2014).
51. O'Connor, R. E. et al. Reconstruction of the diapsid ancestral genome permits chromosome evolution tracing in avian and non-avian dinosaurs. *Nat. Commun.* **9**, 1883 (2018).
52. Jaillon, O. et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**, 946–957 (2004).
53. Carbone, L. et al. Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513**, 195–201 (2014).
54. Beklemisheva, V. R. et al. The ancestral carnivore karyotype as substantiated by comparative chromosome painting of three pinnipeds, the walrus, the Steller sea lion and the Baikal seal (Pinnipedia, Carnivora). *PLoS ONE* **11**, e0147647 (2016).
55. Pereira, A. L. et al. Extensive chromosomal reorganization in the evolution of new world muroid rodents (Cricetidae, Sigmodontinae): searching for ancestral phylogenetic traits. *PLoS ONE* **11**, e0146179 (2016).
56. Romanenko, S. A. et al. Multiple intrasyntenic rearrangements and rapid speciation in voles. *Sci. Rep.* **8**, 14980 (2018).
57. Ullastres, A., Farré, M., Capilla, L. & Ruiz-Herrera, A. Unraveling the effect of genomic structural changes in the rhesus macaque—implications for the adaptive role of inversions. *BMC Genomics* **15**, 530 (2014).
58. Becker, T. S. & Lenhard, B. The random versus fragile breakage models of chromosome evolution: a matter of resolution. *Mol. Genet. Genomics* **278**, 487–491 (2007).
59. Kerr, N. et al. The expression of ELK transcription factors in adult DRG: Novel isoforms, antisense transcripts and upregulation by nerve damage. *Mol. Cell. Neurosci.* **44**, 165–177 (2010).
60. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, bax028 (2017).
61. Larkin, D. M. et al. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res.* **19**, 770–777 (2009).
62. Parey, E. et al. Synteny-guided resolution of gene trees clarifies the functional impact of whole genome duplications. *Mol. Biol. Evol.* **37**, 3324–3337 (2020).
63. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**, 617–624 (2004).
64. Berthelot, C. et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat. Commun.* **5**, 3657 (2014).
65. Paten, B. et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843 (2008).
66. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
67. Armstrong, J., Fiddes, I. T., Diekhans, M. & Paten, B. Whole-genome alignment and comparative annotation. *Annu. Rev. Anim. Biosci.* **7**, 41–64 (2019).
68. Cunningham, F. et al. Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
69. Yates, A. D. et al. Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Res.* **50**, D996–D1003 (2022).
70. Peres, A. & Roest Crollius, H. Improving duplicated nodes position in vertebrate gene trees. *BMC Bioinformatics* **16**, A9 (2015).
71. Zdobnov, E. M. et al. OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **49**, D389–D393 (2021).
72. Lucas, J. M., Muffato, M. & Roest Crollius, H. PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. *BMC Bioinformatics* **15**, 268 (2014).
73. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
74. Muffato, M. et al. Data related 'Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom'. *Zenodo* https://sandbox.zenodo.org/record/1089175 (2022).

Acknowledgements

We thank P. Vincens for the coordination of computing resources and A. Peres for computer code engineering. This work was supported by grants from the French Government and implemented by the Agence Nationale de la Recherche (ANR) (ANR-07-GANI-008-01 GENOVERT, ANR-10-BINF-01-03 ANCESTROME, ANR-10-LABX-54 MEMOLIFE and ANR-10-IDEX-0001-02 PSL* Research University) to H.R.C.

Author contributions

M.M., A.L., C.B. and H.R.C. conceived and designed the experiments. M.M., A.L., C.B. and H.R.C. performed the experiments. M.M., A.L., C.B. and H.R.C. analysed the data. M.M., A.L., N.T.T.N., J.L., C.B. and H.R.C. contributed materials and analysis tools. M.M., A.L., C.B. and H.R.C. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-022-01956-z.

Correspondence and requests for materials should be addressed to Camille Berthelot or Hugues Roest Crollius.

Peer review information *Nature Ecology & Evolution* thanks Kai Ye, Pavel Avdeyev and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

nature portfolio

Corresponding author(s): Roest Crollius, Hugues

Last updated by author(s): Nov 9, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	no software was used
Data analysis	The main analyses are based on the AGORA software, which is the topic of this publication and was developed by the authors. AGORA is deposited on a GitHub server and full details are provided in the the manuscript. Additional analyses are performed using custom python and R scripts but are not central to the research. They can be provided on request. All the data generated during the analyses are provided on a Zenodo server, full details on how to access it are in the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The source code of AGORA, user instructions and a test dataset are available for download from https://github.com/DyogenIBENS/Agora. Ancestral genomes have

been precomputed for ~200 vertebrate (depending on the release), 41 plant, and 222 fungi genomes and are available on the Genomicus database FTP server (ftp://ftp.bio.ens.psl.eu/pub/dyogen/genomicus/). These ancestral genomes can also be explored visually within the Genomicus35 synteny browser (http://www.genomicus.bio.ens.psl.eu/genomicus). Ancestral genomes and data used in this article for analysis are archived on a Zenodo depository (https://sandbox.zenodo.org/record/1089175).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data where this information has been collected, and consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study concerns the evolution of genome structures. The analyses describe breakpoint data grouped by taxon and shown as distributions in boxplots. The data is descriptive. A series of Fishers's exact test was perfoemd with adjustments for multiple testing.
Research sample	The analyses were performed on publicly available genome data stored in the Ensembl database covering several hundred eukaryote species. They were downloaded from http://www.ensembl.org/index.html, http://fungi.ensembl.org/index.html, http://plants.ensembl.org/index.html, http://protists.ensembl.org/index.html, http://metazoa.ensembl.org/index.html
Sampling strategy	The sampling was based on available data, and the sample size is not relevant to the study's conclusions
Data collection	No data was collected
Timing and spatial scale	No data was collected
Data exclusions	No data was excluded
Reproducibility	The data was analysed exclusively by bioinformatics methods. Such results are, by definition, 100% reproducible.
Randomization	There was no randomization because such a procedure was not relevant to the study
Blinding	There was no blinding because such a procedure was not relevant to the study

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging