Genomicus in 2022: comparative tools for thousands of genomes and reconstructed ancestors

Nga thi thuy Nguyen, Pierre Vincens, Jean François Dufayard, Hugues Roest crollius, Alexandra Louis

Published online 18 November 2021 *Nucleic Acids Research, 2022, Vol. 50, Database issue* ***D1025–D1031***
https://doi.org/10.1093/nar/gkab1091

Genomicus in 2022: comparative tools for thousands of genomes and reconstructed ancestors

Nga Thi Thuy Nguyen[1], Pierre Vincens[1], Jean François Dufayard[2,3], Hugues Roest Crollius[1,*] and Alexandra Louis[1,*]

[1]Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France, [2]CIRAD, UMR AGAP Institut, F-34398 Montpellier, France - UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France and [3]French Institute of Bioinformatics (IFB) - South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, F-34398 Montpellier, France

Received September 15, 2021; Revised October 18, 2021; Editorial Decision October 19, 2021; Accepted October 20, 2021

ABSTRACT

Genomicus is a database and web-server dedicated to comparative genomics in eukaryotes. Its main functionality is to graphically represent the conservation of genomic blocks between multiple genomes, locally around a specific gene of interest or genome-wide through karyotype comparisons. Since 2010 and its first release, Genomicus has synchronized with 60 Ensembl releases and seen the addition of functions that have expanded the type of analyses that users can perform. Today, five public instances of Genomicus are supporting a total number of 1029 extant genomes and 621 ancestral reconstructions from all eukaryotes kingdoms available in Ensembl and Ensembl Genomes databases complemented with four additional instances specific to taxonomic groups of interest. New visualization and query tools are described in this manuscript. Genomicus is freely available at http://www.genomicus.bio.ens.psl.eu/genomicus.

INTRODUCTION

Initiated in 2010, the Genomicus database and visualization tools aim to navigate genomes in several dimensions: linearly along chromosomal axes, transversely through the comparison of genomes or genomic segments and chronologically along evolution. Over the last 12 years, we have maintained uninterrupted service, updated data regularly, and developed new tools following user suggestions. Initially developed around vertebrate genome data, it has been extended to respond to the needs of a wider community focusing on additional groups now covering all eukaryotic lineages. The present article gives a 12-year perspective on this tool, describes its historical functionalities, the novelties since the previous NAR Database issues (1–4) and plans for the future on new functionalities and data.

GENOMICUS OVER THE LAST 12 YEARS

From a vertebrate genome evolution browser to multiple dedicated servers

The development of Genomicus was initially motivated by the need to visualize vertebrate ancestral genome reconstructions inferred by our laboratory. We first developed a simple web server that stored gene positions, gene trees and enabled navigation through vertebrate genomes and their ancestors (1), in order to pinpoint possible errors in reconstructions by AGORA (Algorithm for Gene Order Reconstruction in Ancestors, https://github.com/DyogenIBENS/Agora). In 2011, Genomicus was the first online database server that allowed comparisons of hundreds of genomic segments in one single interface, and as far as we know, this is still the case. Effectively, users navigate through genomes according to three dimensions: linearly along gene neighbourhoods, transversally between genomes, and evolutionary through ancestral gene content and order reconstruction. The Ensembl database (5) provided extant genome data represented in the first Genomicus server, and further extensions to Genomicus followed the growth of Ensembl, in particular in 2013, when we published 4 additional genome browsers (2) dedicated to groups represented in Ensembl Genomes (6) (Plants, Fungi, Metazoa, Protists). Since 2011, the vertebrate-specific Genomicus browser is updated 4 times a year, following the releases of the Ensembl database, for a total of 50 complete server releases. In addition of these five instances based on Ensembl or Ensembl Genomes data, other servers are available for the study of specific genomes annotation. A tunicate instance is linked to the ANISEED database (7), as well as specific servers dedicated to amphioxus or trout comparative genomics (8,9) (Table 1). Ancestor reconstructions are

*To whom correspondence should be addressed. Tel: +33 1 44 32 23 71; Fax: +33 1 44 32 39 41; Email: alexandra.louis@bio.ens.psl.eu
Correspondence may also be addressed to Hugues Roest Crollius. Email: hrc@bio.ens.psl.eu

Table 1. All Genomicus available instances and numbers of extant and reconstructed genomes available

Genomicus instance	url	Number of extant genomes in last release	Number of reconstructed ancestors in last release
Vertebrate	http://genomicus.biologie.ens.fr/genomicus	199	195
Plants	http://genomicus.biologie.ens.fr/genomicus-plants	99	59
Fungi	http://genomicus.biologie.ens.fr/genomicus-fungi	478	222
Metazoa	http://genomicus.biologie.ens.fr/genomicus-metazoa	117	81
Protist	http://genomicus.biologie.ens.fr/genomicus-protists	136	64
Tunicates	http://genomicus.biologie.ens.fr/genomicus-tunicates	21	20
Trout	http://genomicus.biologie.ens.fr/genomicus-trout	15	13
Amphioxus	http://genomicus.biologie.ens.fr/genomicus-amphioxus	12	11
Pre-2R Vertebrate	http://genomicus.biologie.ens.fr/genomicus-69.10	61	52

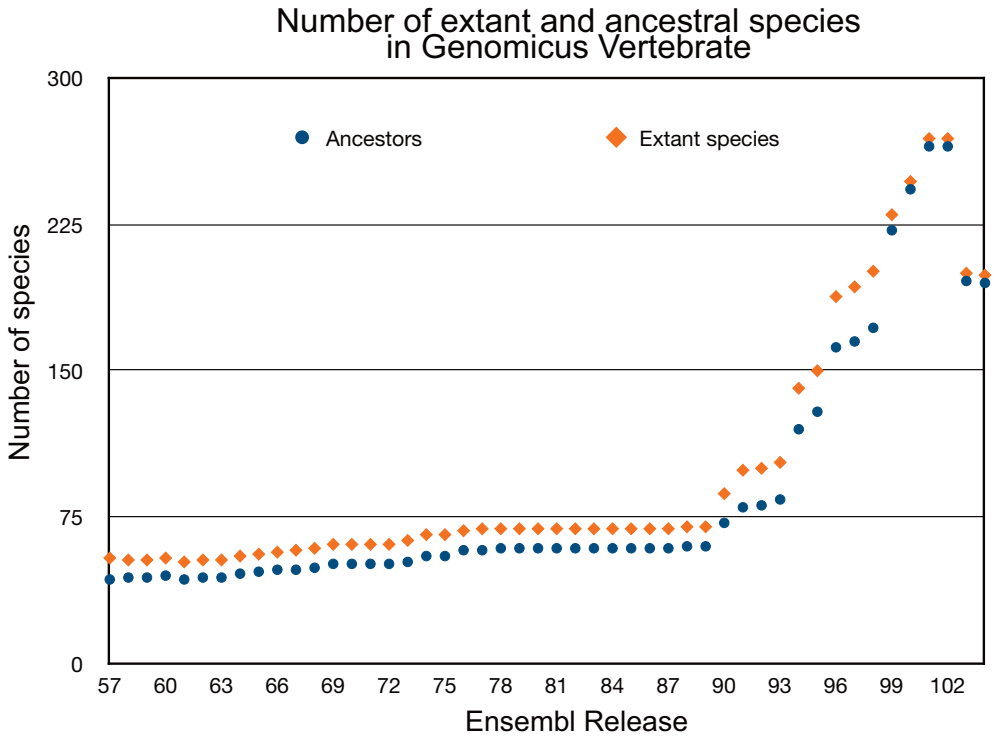

Figure 1. Evolution of the number of extant and reconstructed genomes represented in Genomicus Vertebrate. The number of genomes has more than quadrupled in 12 years. Genomicus follows the Ensembl releases four times a year since 2010.

available for plants since 2014 (10), and only since the current release for protists, metazoan and fungi.

Evolution of the number of genomes

The number of extant genomes represented in Genomicus vertebrates increases with each quarterly update, and has more than quadrupled in ten years, in par with the number of reconstructed ancestors (Figure 1). Starting from 54 extant genomes (43 ancestral genomes) in release 53, it reached 247 extant genomes in release 100 (243 ancestral genomes). The use and representation of data as well as the fluidity of queries on the server remains efficient, despite the amount of information requested and transferred. Other Genomicus servers have also seen an increase in the amount of data, with a focus on less frequent but more major releases, including ancestral genome reconstructions in the current one.

Main search and visualization tools available on Genomicus servers

Over the last 12 years, the code behind Genomicus servers has evolved in response to user feedbacks and the needs of the comparative genomics community. Interfaces to query the database and visualization modules have also been developed and integrated into the different servers in response to the availability of new data (Conserved Non Coding Elements, ancestral reconstructions). If our first purpose was to visualize multiple genome comparisons, it was quite obvious that pairwise genome studies would be of interest. Therefore, in addition to PhyloView and AlignView described in previous Genomicus article (1), matrix dotplot of homologous genes (MatrixView), karyotype comparisons (KaryoView) and PhylDiagView, a tool to compute conserved gene order between two genomes (modern or ancestral) have been deployed (3,4) (Figure 2).

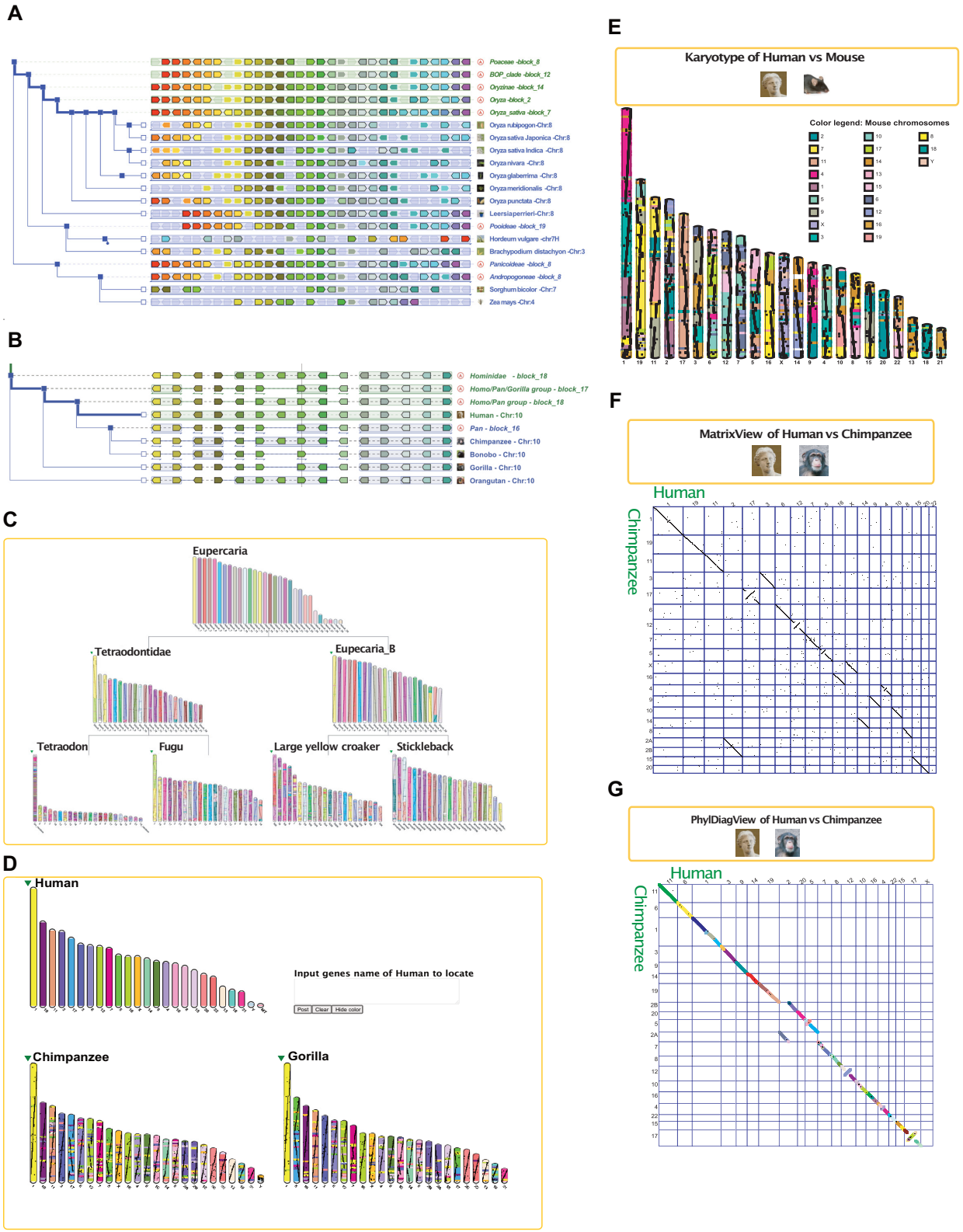

Figure 2. Overview of the tools available in Genomicus. (**A**) PhyloView shows the order and orientation of homologous genes around a reference gene in different genomes that share the same ancestral root. When these neighbouring genes are orthologs or paralogs of genes in the reference genome, they are shown with matching colours. (**B**) AlignView shows an alignment between (i) the genes contained within the genomic region of the reference gene and (ii) all their respective orthologs in other genomes (unlike PhyloView where genomes that do not have an ortholog of the reference gene are not displayed). (**C**) MultikaryoView shows the evolution of karyotypes from an ancestor to its descendant genomes. (**D**) MultikaryoView without ancestors represents extant genomes compared against a chosen reference karyotype. (**E**) Pairwise karyotype comparison, between extant or ancestral genomes. (**F**) MatrixView is a dotplot matrix where each dot represents an ortholog between two genomes. (**G**) PhylDiagView computes exact boundaries of syntenic blocks between two genomes.

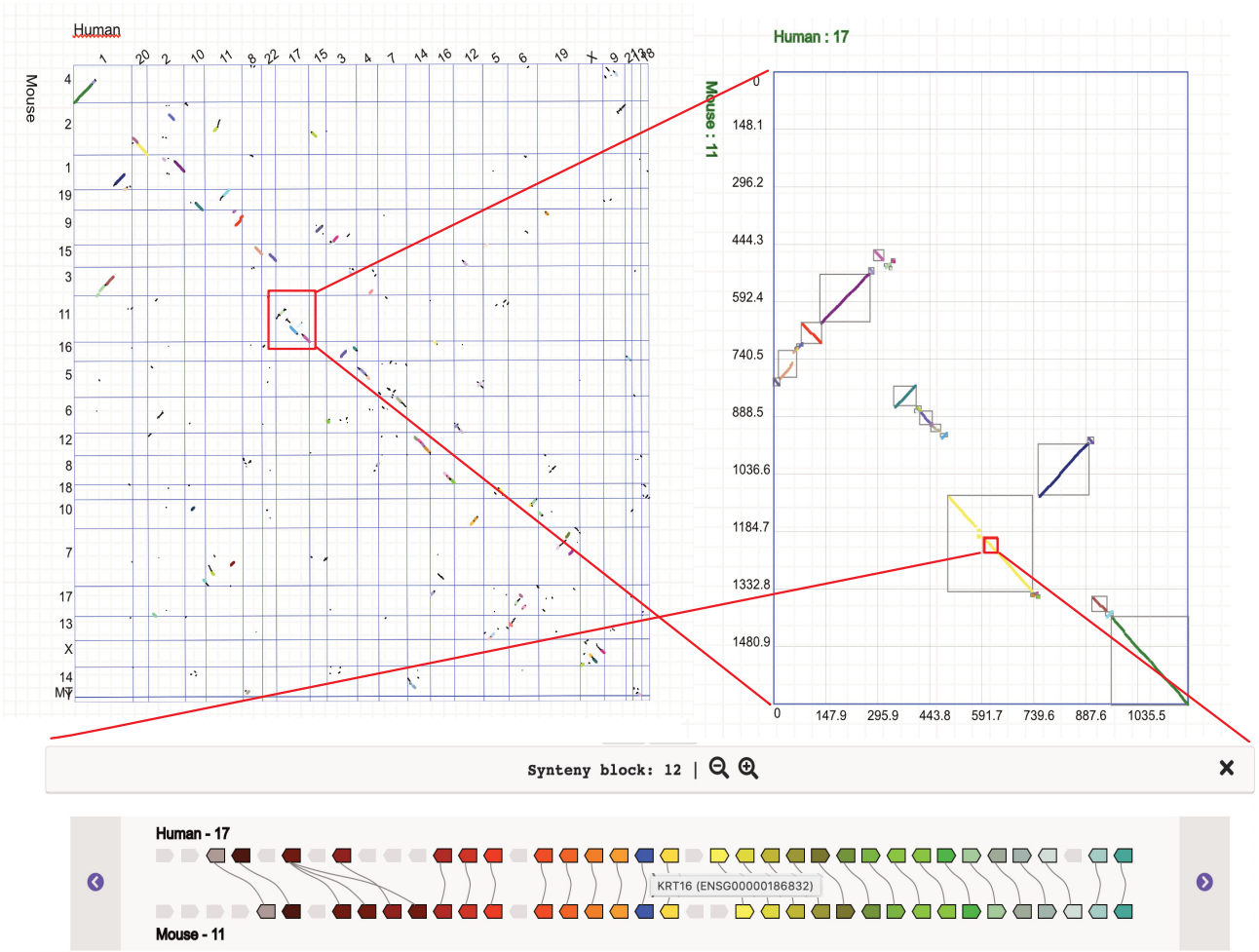

Figure 3. PhylDiagView with the new zoom functionality. Starting from a whole genome comparison matrix representing syntenic blocks, users can first zoom on a chromosome-to-chromosome comparison between the two genomes and then on a specific syntenic block of interest. An alignment of genes in this region is then represented, where each gene and its homologs are coloured the same. Mouse-over will show the gene name and a mouse click will open the complete PhyloView interface for more information on this gene context.

GENOMICUS 2022 NOVELTY

Ancestors for fungi, metazoan and protists

Year 2021 saw new releases for Genomicus-Plants (based on Ensembl Genomes 49), Genomicus-Fungi (based on Ensembl Genomes 43), Genomicus-Protists and Genomicus-Metazoa (both based on Ensembl Genomes 51). For the first time for fungi, protists and metazoan, we computed ancestral genome reconstructions with AGORA (https://github.com/DyogenIBENS/Agora), and made them available in the respective servers (Table 1).

Zooming on PhylDiag synteny blocks

In 2018, we developed a new interface for on-the-fly computation of synteny between two genomes (PhylDiag View) (4), which generates exact gene boundaries of syntenic blocks according to extensive parameterization possibilities. Users may now select a region anywhere inside a chromosome-to-chromosome comparison matrix, and zoom to see the local gene organization and identify the exact gene that flanks the breakpoint (Figure 3). In this new graphical interface, gene names are visible by mouse-over, and orthologous genes in both genomes bear the same colour and are linked by a line to facilitate the interpretation of the conservation of gene order and orientation in the locus. Links to PhyloView are available with a mouse-click on the gene.

MultiKaryoview without ancestors

We developed multiKaryoview in 2017 to represent the evolution of karyotypes, from a reconstructed given ancestor to descendant extant genomes. However, it became clear that users often wish to only compare extant genomes and explore the homology of one genome against multiple others. To this end, MultiKaryoView 'without ancestor' (Figure 4) is now available to compare the karyotypes of extant genomes, with one them used as reference and without the complete history that includes ancestral genomes. Genes are represented by small black dots, leading to diagonal lines inside chromosomes when their order is

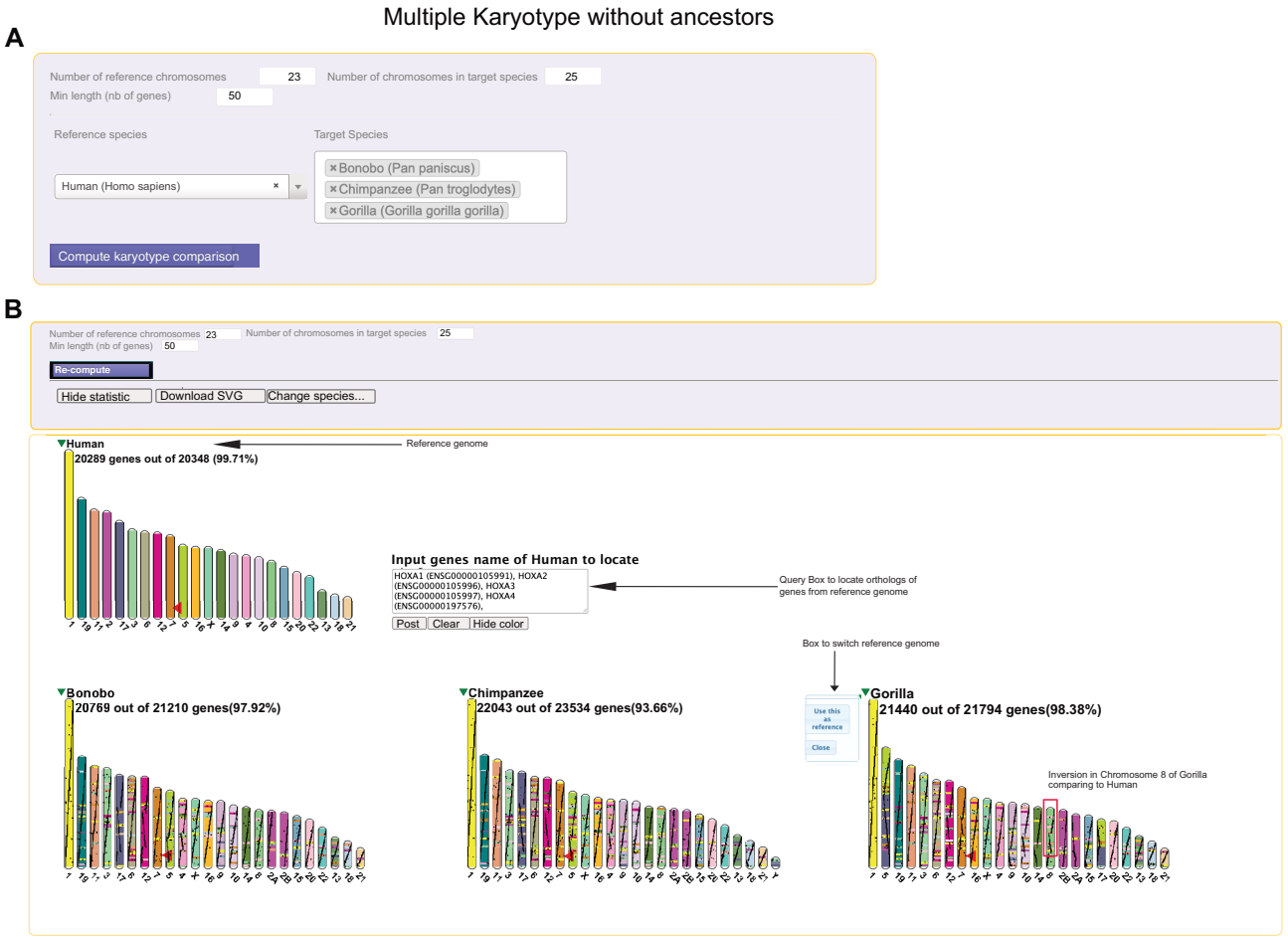

Figure 4. MultiKaryoview 'without ancestors' with Human as reference, Bonobo, Chimpanzee and Gorilla as target genome. (**A**) Users can select a reference extant genome, different target genomes, and the parameters of represented karyotypes (the number of chromosomes in the reference genome, the number of chromosomes in descendant genomes, the minimum number of genes in chromosomes). (**B**) By default, the karyotype of the reference genome will be drawn on top of the display window. All the target karyotypes are coloured according to the reference human genome. By clicking on the little green arrow near the genome name, users can choose to change the 'reference genome'. Users can search for specific location of genes in the reference genome thanks to the input box, and locate their orthologs if they exist with the red arrow. In the example, a cluster of HOXA genes is located on chromosome 7 of each genome. A specific inversion is located on chromosome 8 of Gorilla and does not appear in Chimpanzee nor Bonobo.

similar to the reference and line breaks representing rearrangements. They are overlaid with the colour of the chromosome of the reference genome where the homologous gene is located, thus pointing to large scale homologous chromosome segments. A mouse-over on a selected chromosome segment highlights all other segments homologous to the reference chromosome of the same colour. Users can also dynamically highlight the positions of genes from the reference genome and visualize the positions of their orthologs, if they exist, on the karyotype of other genomes in the display.

Search and visualization of specific evolutionary scenario in gene families

In Genomicus, all visualization tools, inference of ancestral gene organization and computation of synteny are based on the availability of a forest of gene trees reconciled with a species tree (downloaded from Ensembl or Ensembl Genomes databases, or computed with the same procedure as in the EnsemblCompara pipeline (11)). While each gene family has its own evolutionary history made of gene duplication and gene deletions or absence thereof, some evolutionary patterns exist where multiple gene families will share an identical history in parts of their tree. Being able to query the entire forest of trees to extract all the families that fit a specific pattern is a powerful strategy to isolate at once the extant genes that result from this pattern. To this end, users can now search Genomicus Vertebrate using the RapGreen software (12) (Figure 5). The tree pattern-matching part of the software instantly parses all the gene trees available in the Genomicus server and retrieves specific evolutionary scenario defined by the user. These include a given tree topology with optional speciation and duplication nodes, explicit taxa at the leaves and/or on ancestral nodes, constraints on the branches limiting additional speciations or duplications. For example, users can search for all families that retained paralogues in the river trout (*Salmo truta*) and in the rainbow trout (*Oncorhynchus mykiss*) after the Salmonidae (4R) whole genome duplication (see Fig-

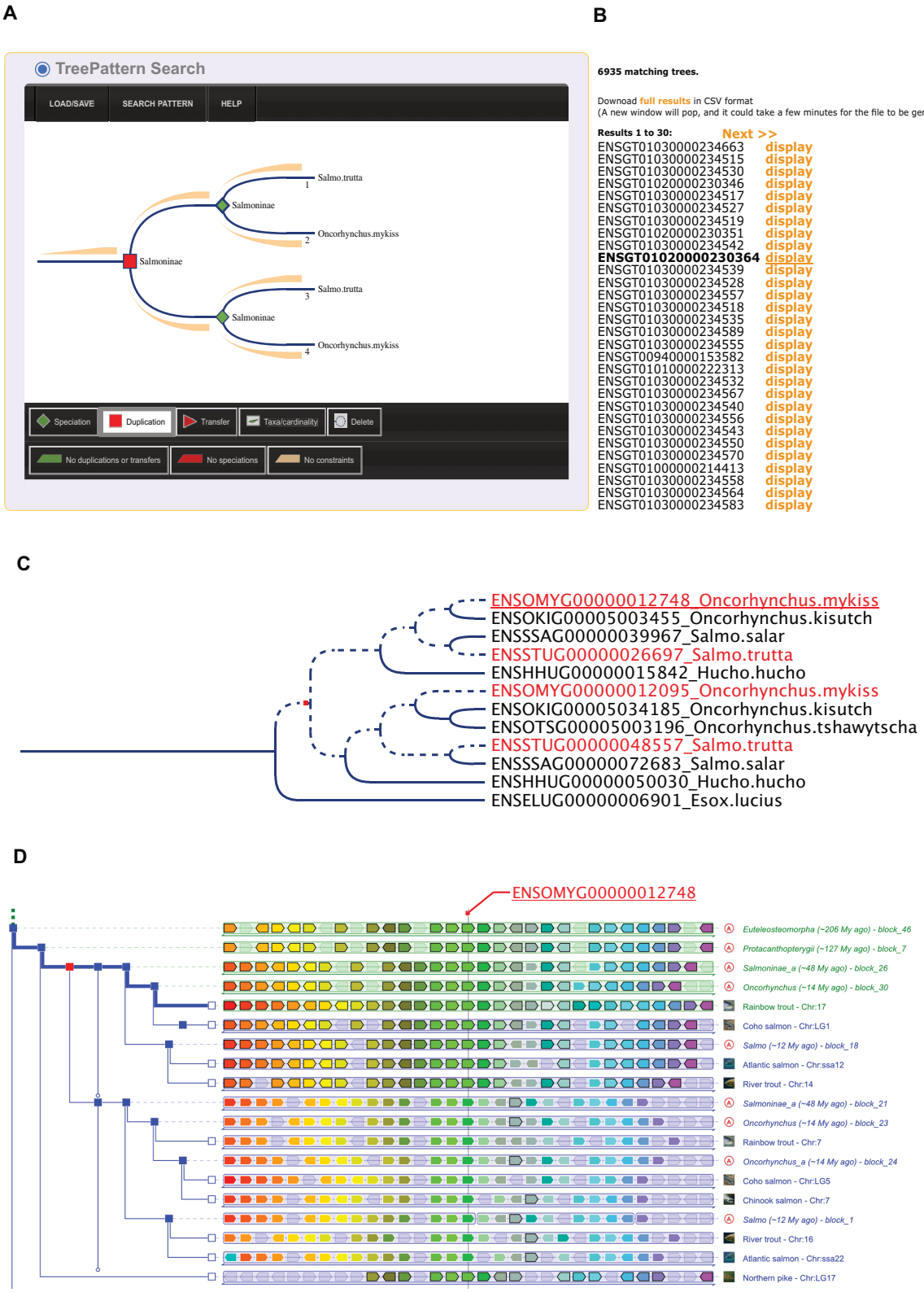

Figure 5. Schematic presentation of a tree pattern search use-case applied on Genomicus vertebrate (**A**) the pattern editor is accessible by choosing the TreePattern search option. Then a pattern can be created and edited using the toolbox, or a demonstration pattern can be chosen in the file menu. Here the red square represents a duplication node in Salmoninae and green lozenges represent speciation nodes leading to *Salmo truta* and *Oncorhynchus mykiss*. (**B**) Results are displayed by clicking on the Search Pattern button. Each tree family can be visualized, or full results can be downloaded in CSV format. (**C**) InTreeGreat viewer displaying one of the resulting pattern (the dotted lines represent the requested pattern, matching genes are highlighted in red. The small red dot represents the duplication node Salmoninae). A mouse click on a leaf (here underlined) will link to PhyloView. (**D**) Representation in Genomicus PhyloView to visualize the comparative genomic context of the query reference gene and the conservation on either side of the duplication node (red quare) in each Salmoninae descendant genome, marking the 4R whole genome duplication.

ure 5A for pattern). Families corresponding to the query pattern are listed, can be downloaded (Figure 5B) and can be visualized in InTreeGreat (Figure 5C) (part of the RapGreen package) that highlights in red the pattern in the specific gene tree and links genes to PhyloView in Genomicus (Figure 5D).

CONCLUSION AND FUTURE DEVELOPMENTS

Genomicus is online since 2010 and is increasingly used by the community. It is still unique for its ability to compare hundreds of extant genomes and is the first tool to provide access to inferred ancestral gene orders for five different eukaryote kingdoms. It is now part of the catalogue of bioinformatic tools labelled by ELIXIR. We now plan to install a mirror server in the French node of ELIXIR (Institut Français de Bioinformatique IFB) to optimize access in case of temporary down-time on the main server. The Treepattern search tool will be soon available for all Genomicus instances.

GENOMICUS SOFTWARE IMPLEMENTATION

Genomicus is composed of Perl (version 5.30) scripts and modules, alongside with Python (version 2.7) libraries from other projects in our laboratory (LibsDyogen, PhylDiag) for PhylDiag computation, executed with mod_perl on an Apache2 (version 2.4) server and querying a MariaDB (version 10.5.12) database. The web server is running on an Ubuntu server 20.04. The pages embed inline-SVG drawings in XHTML while the JavaScript (version 1.9.1) usage is limited to an information panel retrieved with AJAX calls.

The tree pattern search web interface is implemented as a Javascript/PHP user interface to edit patterns and explore results connected with a Java daemon running on OpenJDK 11.0.9.1 that implements the tree pattern matching algorithm for parsing and exploring the tree forest in a minimal amount of time.

The interface is optimized for Chrome navigators but it also runs on Safari, Opera and Firefox. The source codes for Genomicus and the MariaDB schema can be obtained upon request by email for local implementation.

DATA AVAILABILITY

All public Genomicus servers are accessible from http://genomicus.bio.ens.psl.eu/genomicus/. All data can be downloaded on ftp://ftp.biologie.ens.fr/pub/dyogen/. Web servers and FTP data can be freely accessed by all users without a login requirement.

ACKNOWLEDGEMENTS

We are particularly thankful to the IT team of IBENS for assistance with computer systems administration and network management.

FUNDING

Program ≪ Investissements d'Avenir ≫ launched by the French Government and implemented by ANR [ANR-10-LABX-54 MEMOLIFE, ANR-10-IDEX-0001-02 PSL* Université Paris]; RENABI-IFB call Funding [ANR-11-INSB-0013]. Funding for open access charge: CNRS.
Conflict of interest statement. None declared.

REFERENCES

1. Muffato,M., Louis,A., Poisnel,C.-E. and Roest Crollius,H. (2010) Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinforma. Oxf. Engl.*, **26**, 1119–1121.
2. Louis,A., Muffato,M. and Roest Crollius,H. (2013) Genomicus: five genome browsers for comparative genomics in eukaryota. *Nucleic Acids Res.*, **41**, D700–D705.
3. Louis,A., Nguyen,N.T.T., Muffato,M. and Roest Crollius,H. (2014) Genomicus update 2015: KaryoView and MatrixView provide a genome-wide perspective to multispecies comparative genomics. *Nucleic Acids Res.*, **43**, D682–D689.
4. Nguyen,N.T.T., Vincens,P., Roest Crollius,H. and Louis,A. (2018) Genomicus 2018: karyotype evolutionary trees and on-the-fly synteny computing. *Nucleic Acids Res.*, **46**, D816–D822.
5. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
6. Kersey,P.J., Allen,J.E., Armean,I., Boddu,S., Bolt,B.J., Carvalho-Silva,D., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C. *et al.* (2016) Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.*, **44**, D574–D580.
7. Dardaillon,J., Dauga,D., Simion,P., Faure,E., Onuma,T.A., DeBiasse,M.B., Louis,A., Nitta,K.R., Naville,M., Besnardeau,L. *et al.* (2019) ANISEED 2019: 4D exploration of genetic data for an extended range of tunicates. *Nucleic Acids Res.*, **439**, 965.
8. Berthelot,C., Brunet,F., Chalopin,D., Juanchich,A., Bernard,M., Noel,B., Bento,P., Da Silva,C., Labadie,K., Alberti,A. *et al.* (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat. Commun.*, **5**, 3657.
9. Marlétaz,F., Firbas,P.N., Maeso,I., Tena,J.J., Bogdanovic,O., Perry,M., Wyatt,C.D.R., de la Calle-Mustienes,E., Bertrand,S., Burguera,D. *et al.* (2018) Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature*, **564**, 64–70.
10. Louis,A., Murat,F., Salse,J. and Roest Crollius,H. (2015) GenomicusPlants: a web resource to study genome evolution in flowering plants. *Plant Cell Physiol.*, **56**, e4.
11. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
12. Dufayard,J.-F., Sidibe-Bocs,S., Guignon,V., Larivière,D., Louis,A., Oubda,N., Rouard,M., Ruiz,M. and de Lamotte,F. (2021) RapGreen, an interactive software and web package to explore and analyze phylogenetic trees. *NAR Genomics and Bioinformatics*. **3**, lqab088.