BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 8 2010, pages 1119–1121
doi:10.1093/bioinformatics/btq079

Genome analysis Advance Access publication February 24, 2010

Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes

Matthieu Muffato, Alexandra Louis, Charles-Edouard Poisnel and Hugues Roest Crollius*

Dyogen Group, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Centre National de la Recherche Scientifique UMR8197, Institut National de la Santé et de la Recherche Médicale U1024, 75005 Paris, France

Associate Editor: John Quackenbush

ABSTRACT

Summary: Comparative genomics remains a pivotal strategy to study the evolution of gene organization, and this primacy is reinforced by the growing number of full genome sequences available in public repositories. Despite this growth, bioinformatic tools available to visualize and compare genomes and to infer evolutionary events remain restricted to two or three genomes at a time, thus limiting the breadth and the nature of the question that can be investigated. Here we present Genomicus, a new synteny browser that can represent and compare unlimited numbers of genomes in a broad phylogenetic view. In addition, Genomicus includes reconstructed ancestral gene organization, thus greatly facilitating the interpretation of the data.

Availability: Genomicus is freely available for online use at http://www.dyogen.ens.fr/genomicus while data can be downloaded at ftp://ftp.biologie.ens.fr/pub/dyogen/genomicus

Contact: hrc@biologie.ens.fr

Received on October 9, 2009; revised on February 15, 2010; accepted on February 19, 2010

1 INTRODUCTION

From less than 20 fully sequenced metazoan genomes 4 years ago, nearly 80 species are now represented in a variety of centralized databases. This abundance of sequence data has reinforced the role of comparative genomics as the primary approach to gain insight in the organization of a genome. Comparing sequences from different species serves several purposes: (i) to outline conserved regions, a powerful guide to rapidly focus on functional regions; (ii) to document differences among these functional sequences as a first step to understand broader biological differences (metabolic, developmental, etc.) between organisms; and (iii) to identify evolutionary events that have interrupted the gene colinearity between the genomes of two species since their last common ancestor.

To document and study the latter, the inference of ancestral gene orders starting from extant species provides important reference points; yet no visualization tool currently allows comparisons between an ancestral genome to one or more of its modern descendant. Existing software still limit the comparison to two or three extant genomes at a time, and are restricted to a limited range of species (Byrne and Wolfe, 2005; Courcelle *et al.*, 2008; Derrien *et al.*, 2007; Dong *et al.*, 2009; Jensen *et al.*, 2009; Lyons *et al.*, 2008; Pan *et al.*, 2005; Sinha and Meller, 2007).

To address these issues, we have developed Genomicus, a browser dedicated to the study of synteny and the conservation of gene order among multiple genomes (currently 52 metazoan genomes and the yeast *Saccharomyces cerevisiae*). Importantly, Genomicus also integrates reconstructed ancestral synteny blocks at 44 ancestral nodes.

2 METHODS

2.1 Data integration

Most of the genome data displayed in Genomicus is already stored, integrated and publicly available from the Ensembl database (Hubbard *et al.*, 2009) but without extensive synteny visualization tools. The two main types of information that are required by Genomicus are gene positional information in their respective genomes and phylogenetic relationships (orthology, paralogy) between genes. Genomicus then edits Ensembl phylogenetic trees (Vilella *et al.*, 2009) in three ways. First, duplication nodes with a Duplication Consistency Score (Vilella *et al.*, 2009) below a threshold, that is optimized to increase the synteny between extant genomes, are selected. In such cases, duplication nodes are shifted towards terminal branches unless stopped by an intermediate, strong, duplication node. Second, we have added *Boreoeutheria*, *Euarchontoglires* and *Atlantogenata* ancestral nodes in existing trees of placental mammals (Prasad *et al.*, 2008). Third, we have added some extant species that are not currently referenced in Ensembl (*Branchiostoma floridae*, *Nematostella vectensis* and *Oikopleura dioica*), together with their respective ancestral nodes. For each of these new species, best reciprocal blast comparisons [best reciprocal hit (BRH)] are performed between predicted proteins and the proteins from a set of key species already referenced in Genomicus. Comparisons that are internally consistent (mutual orthology relationships are respected) allow a given protein to be added in the same phylogenetic tree as that of its BRH. In rare cases, a new protein may act as outgroup to two existing trees and fuse them through a new duplication node.

2.2 Reconstruction method

Ancestral syntenic blocks are reconstructed by a complex procedure that will be described in details elsewhere (M.Muffato *et al.*, manuscript in preparation). Briefly, parsimonious scenarios are estimated based on pairwise comparisons of gene order between all available sequenced genomes (1378 comparisons in Genomicus v56.01). For a given ancestor, all ancestral genes that are identified as conserved neighbours in at least one such comparison become linked nodes in a graph. A weight (with values comprised between 1 and 1378) reflecting the number of times this situation was observed in all the comparisons is then applied to each link.

At this stage, inconsistencies may appear in the form of ancestral genes connected to more than two neighbours. To resolve these, the weighted graph

*To whom correspondence should be addressed.

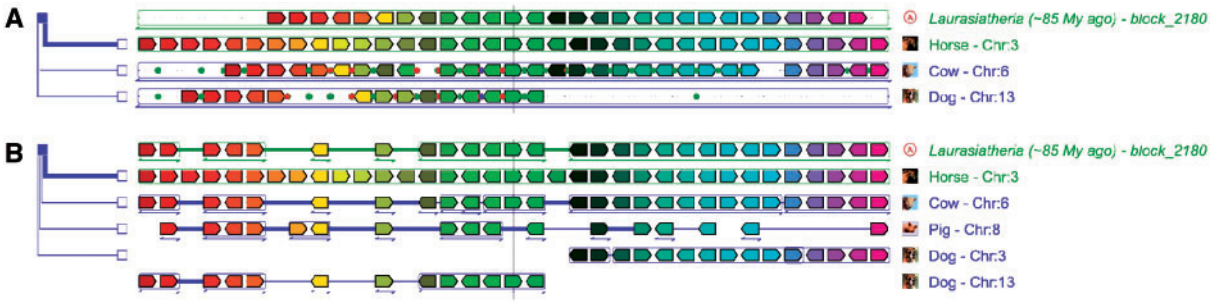

Fig. 1. PhyloView (**A**) and AlignView (**B**) of the horse PHOX2B gene as reference. In both views, the horse PHOX2B gene and its orthologs is shown in light green over a thin vertical line. In (A), the right part of dog chromosome 13 is not syntenic with the horse and cow chromosomes (and therefore neither with their ancestral one in *Laurasiatheria*). In (B), AlignView shows that this region underwent a dog-specific translocation onto chromosome 3. Furthermore, the pig locus can be analysed with (B) but not in (A), because PhyloView is based on the phylogenetic tree of PHOX2B, which is not annotated in the pig genome, whereas AlignView shows genes that are orthologous to genes across the locus of reference species, not just the reference gene. Coloured circles between genes represent conserved CNEs.

is processed using a top-down greedy algorithm where the links of highest weight are selected first and are used to select the most likely gene-to-gene connection in case of multiple choice.

This produces a set of linear paths in the graph connecting ancestral genes based on the number of times their respective descendants are observed as extant neighbours. We performed extensive simulations and benchmarked our methods against several alternative methods: MGR (Bourque and Pevzner, 2002), MGRA (Alekseyev and Pevzner, 2009) and InferCars (Ma *et al.*, 2006). Our method is the only approach able to satisfactorily analyse data with the volume (53 species and 888 217 extant genes) and complexity (duplications, deletions) found in the complete set of sequenced vertebrate genomes. The reconstructed gene order is correct in >95% of the cases (specificity), and includes between 70% and 95% of the expected ancestral gene pairs (sensitivity).

2.3 Systems and technical aspects

Genomicus is composed of Perl scripts and modules, executed with mod_perl on an Apache2 server and querying an MySQL database. The pages embed inline-SVG drawings in XHTML while the JavaScript usage is limited to an information panel retrieved with AJAX calls. Users with browsers that are not yet compliant with open web technologies require the Google Chrome Frame extension (http://www.google.com/chromeframe).

3 USAGE AND 'VIEWS'

The home page invites the user to enter its gene of interest and will by default show a graphical representation in PhyloView.

- *PhyloView* shows the chosen reference gene in the centre of the display with 15 neighbouring genes on both side, as well as orthologs and paralogs of the query gene in their own respective genomic regions, also with 15 neighbouring genes. When these neighbouring genes are orthologs or paralogs of genes in the reference species, they are shown with matching colours. Some species may appear twice if a copy of the reference gene underwent a duplication (shown as a red square) within the evolutionary range presented on the display.
- *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 species. Here also, the 'query' gene is centred and the colour code is used to indicate orthologs between different genomes. A species spanning multiple lines means that the reference gene content is distributed over multiple chromosomes (or scaffolds; see the case of dog in Fig. 1B).

In both views, the tree can be edited (by expanding, collapsing, hiding, showing chosen nodes) to clarify the view. Genomicus also displays orthologous conserved non-coding elements (CNEs) at three levels of conservation. Finally, gene and loci information can be reached with links to other browsers such as Ensembl, UCSC and NCBI.

4 FUTURE DEVELOPMENTS

The main perspectives are to extend the functionalities and the breadth of species displayed in Genomicus. In particular, a 'chromosome painting' view showing extant and ancestral karyotypes that are colour coded according to a species of interest is currently in development. Genomicus will also follow the 'Ensembl Genomes' project and will therefore extend its scope to include plant and fungal genomes.

ACKNOWLEDGEMENTS

We thank Pierre Vincens and his group at the Ecole Normale Supérieure for providing hosting and security services for this project.

Funding: Agence Nationale pour la Recherche (ANR-07-GANI-008-01).

Conflict of Interest: none declared.

REFERENCES

Alekseyev,M.A. and Pevzner,P.A. (2009) Breakpoint graphs and ancestral genome reconstructions. *Genome Res.*, **19**, 943–957.

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

Byrne,K.P. and Wolfe,K.H. (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.*, **15**, 1456–1461.

Courcelle,E. *et al.* (2008) Narcisse: a mirror view of conserved syntenies. *Nucleic Acids Res.*, **36**, D485–D490.

Derrien,T. *et al.* (2007) AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. *Bioinformatics*, **23**, 498–499.

Dong,X. *et al.* (2009) Synorth: exploring the evolution of synteny and long-range regulatory interactions in vertebrate genomes. *Genome Biol.*, **10**, R86.

Hubbard,T.J. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.

Jensen,L.J. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.

Lyons,E. *et al.* (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.*, **148**, 1772–1781.

Ma,J. *et al.* (2006) Reconstructing contiguous regions of an ancestral genome. *Genome Res.*, **16**, 1557–1565.

Pan,X. *et al.* (2005) SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics*, **21**, 3461–3468.

Prasad,A.B. *et al.* (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol. Biol. Evol.*, **25**, 1795–1808.

Sinha,A.U. and Meller,J. (2007) Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, **8**, 82.

Vilella,A.J. *et al.* (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.