AUTOMATIC RECONSTRUCTION OF 42 ANCESTRAL METAZOAN GENOMES BY COMPARATIVE GENOMICS.

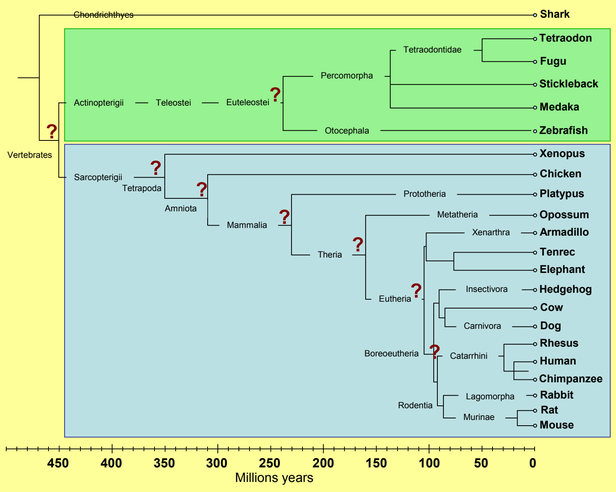
Matthieu Muffato[1], Alexandra Louis[1], Hugues Roest Crollius[1]

[1]École Normale Supérieure, CNRS UMR 8541, 46 rue d'Ulm, 75005 Paris, France.

PURPOSE

Our current understanding of biological processes is limited to contemporary processes occurring in living organisms. Yet Biology is a historical science: all current biological processes are the result of complex evolutionary events.

Large scale genome sequencing makes it possible, through comparative genomics, to gain knowledge on ancestral biological genome organization.



Here we present a general method to reconstruct ancestral genomes (gene organization) which uses as raw material the genome sequences and gene annotations of extant species, and interprets the data using parsimony and a phylogenetic classification of species to infer the most-plausible ancestral state at any given branch point.

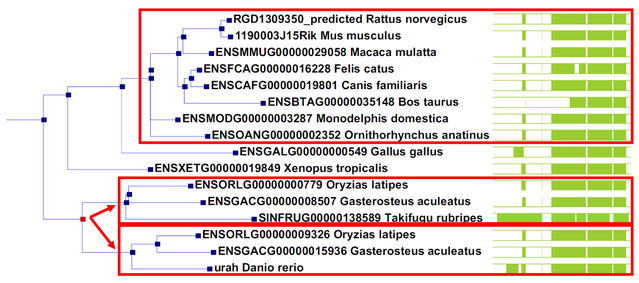
Applied to 44 vertebrate genomes, the method produces a new framework to study genome evolution in a broad chronological perspective.

MATERIAL

The Ensembl database provides the genome sequence and the gene annotation for 45 vertebrates species, and also phylogenetic trees for all the genes. At each node of a tree, we can define (i) that the ancestor at that node possessed a copy of the gene (ii) the orthology and paralogy relationship of that ancestral gene and other genes in the same tree.

We enriched the dataset with three non-vertebrate species: Amphioxus (*Branchiostoma floridae*), Oikopleura (*Oikopleura dioica*) and the sea anemone (*Nematostella vectensis*) with the following procedure:

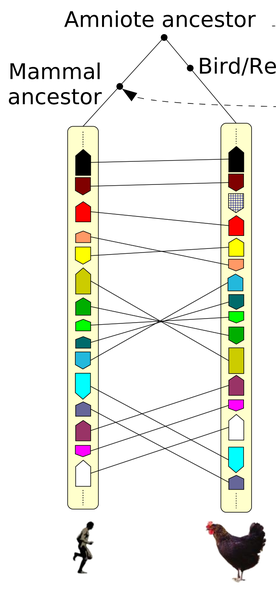
- Compute best reciprocal blast comparisons (BRH) against a set of representative species (human, zebrafish, …)
- Insert a gene from a new species only when the BRH are consistent both internally and with the Ensembl gene trees.
- Create a duplication node when the new gene acts as an outgroup to two existing trees.



In this phylogenetic tree that recapitulates the evolution of a gene since the vertebrate ancestor, the ancestral mammalian genome possessed a single copy of this gene whereas the ancestral fish genome possessed two copies.

METHODS

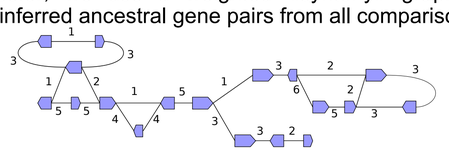
1. Pairwise comparisons



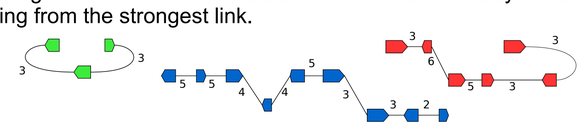
Conserved gene pairs are defined through the comparisons of every chromosomes of every pair of species. For every two modern species, we define ancestral synteny blocks (same exact gene order and orientations) that exist since their last common ancestor and remained intact through intermediate ancestors.

2. Integration

For each ancestor, we use a weighted «synteny» graph that includes all the inferred ancestral gene pairs from all comparisons.



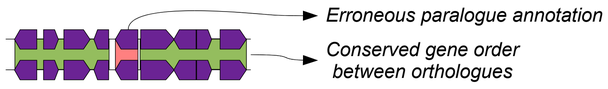
Cycles and genes with multiple links are treated by iteratively selecting links that are consistent with those already selected, starting from the strongest link.



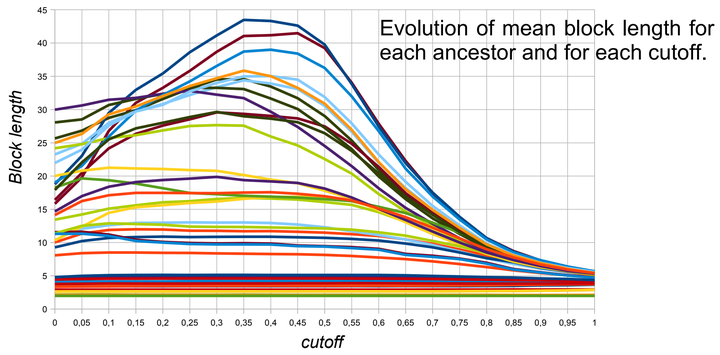
This procedure ensures that none of the deleted links would have connected two extremities of two synteny blocks. The strategy results in maximally connected, non-overlapping segments, or «integrated» synteny blocks.

3. Duplications adjustment

The correct identification of duplication nodes is a notorious problem in phylogeny reconstruction. Information on synteny can greatly assist in correctly identifying duplication nodes, and was used here in a systematic procedure.



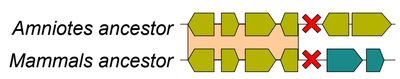
In Ensembl gene trees, a *duplication confidence score* indicates how well a duplication node is supported. We computed optimal cut-off values for retaining duplication nodes based on maximal synteny block length. Nodes below the cut-off were displaced towards terminal branches or until a duplication node with a score higher than the cut-off is encountered.



Most ancestral nodes show an optimal blocks length for a given duplication node threshold. Applying these thresholds to all Ensembl trees increases the length of synteny blocks by approximately 25%.

4. Breakpoint extraction

For every pair of genomes (modern or ancestral), any non-conserved gene pair potentially reveals a breakpoint.



In real data, to eliminate false positives, we have to:

- Eschew genes that are singletons in one genome (no neighbor has been identified)
- Eschew genes that underwent a duplication
- Select breakpoints where the initial state (resp. final) is conserved at a few older (resp. more recent) nodes

774 breakpoints were identified in the mammalian lineage and are analyzed in Berthelot et al. work.

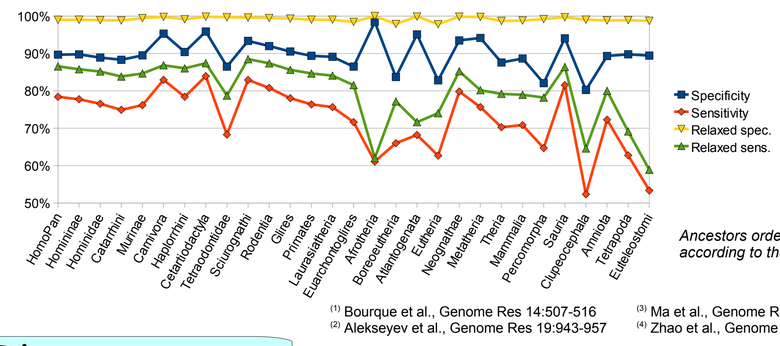
RESULTS

Performances

First, we performed simulations without duplications nor gene loss, in order to benchmark our method against others that cannot deal with such events: MGR[(1)], MGRA[(2)] and inferCars[(3)]. Each simulation deals with 21 vertebrate species, and varies in number of genes (100 to 10000), number of chromosomes (5 to 20), rearrangement rates (from 0.2x to 3x, 1x is defined in [(4)]). The table shows the results for the Boreoeutheria ancestor, according to the simulated genome size. The interpretation of this benchmark is limited because programs are either unable to produce a result due to the size of the genomes or a high rearrangement rate (N/A in the table), or they output nearly perfect reconstructions.

Number of genes	100	500	1000	5000	10000
MGR	Perfect	Perfect	Perfect *(rate ≤ 1x)* N/A *(other rates)*	N/A *(all rates)*	N/A *(all rates)*
MGRA	Perfect	Perfect	Perfect *(rate ≤ 2x)* N/A *(other rates)*	Perfect *(rate ≤ 0.5x)* N/A *(other rates)*	Perfect (at 0.2x) N/A *(other rates)*
inferCars	Perfect	Perfect	Perfect	N/A *(all rates)*	N/A *(all rates)*
Our method	Perfect	Perfect	99.99% correct *(all rates)*	99.99% correct *(all rates)*	99.99% correct *(all rates)*

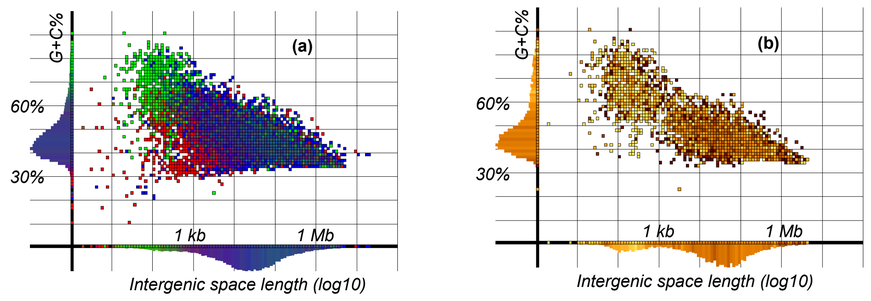
In a second set of simulations, where duplications events and gene losses were allowed, we tested only our method since others cannot handle asymmetric gene contents between nodes. In a significant number of cases, co-localization of gene losses and duplications lead to false predictions. In those cases, predicted gene pairs were not neighbors, but still relatively close to each other (generally less than 4 genes away). We thus also computed relaxed sensitivity and specificity, counting these gene pairs as correct ones.



(1) Bourque et al., Genome Res 14:507-516
(2) Alekseyev et al., Genome Res 19:943-957
(3) Ma et al., Genome Res 16:1557-1565
(4) Zhao et al., Genome Res 19:934-942

Divergent genes

In (a), human intergenic spaces are classified into bins according to their length and their G+C content. The color code shows the dominant transcriptional orientations in a given bin: green indicates divergent genes, red is for convergent genes and blue for co-oriented genes. A cluster of short (< 1 kb) and G+C rich (> 60%) intergenic spaces, composed mainly of divergent genes (green) appears clearly.



In (b), only intergenic spaces between divergent genes are plotted. Here, the color code reflects the age of the intergenic space, i.e. the most ancient ancestral genome that already possessed this interval (lighter orange means older).

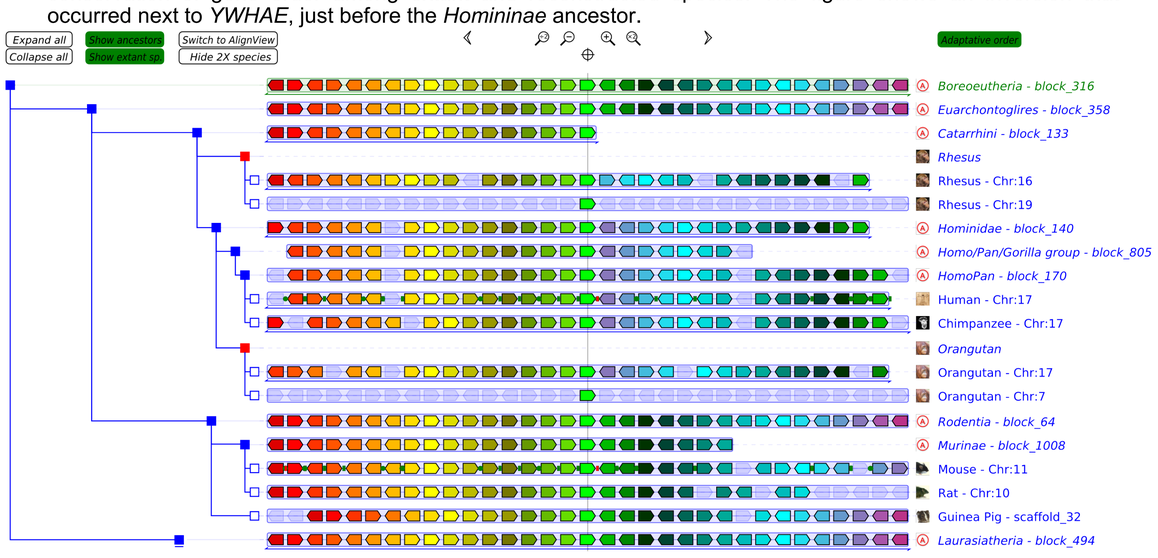
The cluster of short and GC-rich intervals flanked by divergent genes is much more ancient than the rest of the divergent intervals of the human genome.

The table shows a summary of the properties of intervals that belong to the above cluster.

	Selected pairs	Whole genome
Number	848	19316
Age	310 My	220 My
Size	250 bp	90 kb
G+C	65,5 %	46,4 %
% that contains CpG (CpG obs/exp > 0.375)	98,6 %	59,8 % in uni. pairs 84,2 % in div. pairs > 1kb
% of seq. with CNEs	27,3 %	5,5%
% of seq. with Repeats	5,8 %	46,6 %

GENOMICUS

Genomicus (http://www.dyogen.ens.fr/genomicus/) is a synteny browser that has been created to study the conservation of gene order among extant and reconstructed species. The figure shows an inversion that occurred next to *YWHAE*, just before the *Homininae* ancestor.



STATISTICS

The table shows a number of statistics and measures collected on the gene sets and synteny blocks of some ancestors. One can observe the impact of the age of the ancestor, or the quality of the sequence of the descending extant species. (*lengths are measured in number of genes*)

Ancestor	Age (My)	AncGenes	Blocks	Genes in blocks	(in %)	Nb Intervals	(in %)	Median size	N50 size	Max size	Mean size
Chordata	550	10404	149	317	3,05	168	1,62	2	2	6	2,13
Euteleostomi	420	17781	3009	10522	59,18	7513	42,30	3	4	28	3,50
Tetrapoda	359	17910	2506	12855	71,78	10349	57,85	3	7	51	5,13
Amniota	326	18689	1287	14906	79,76	13619	72,95	5	24	178	11,58
Theria	166	19338	948	16249	84,03	15301	79,21	8	37	198	17,14
Boreoeutheria	95	21322	428	17840	83,67	17412	81,74	10	124	345	41,68
Euarchontoglires	90	20989	403	17891	85,24	17488	83,40	12	130	292	44,39
Catarrhini	31	21292	544	18359	86,22	17815	83,75	13	87	387	33,75
Homo/Pan group	5	21075	628	18829	89,34	18201	86,45	12	83	406	29,98
Rodentia	80	19266	604	17585	91,27	16981	88,23	14	62	417	29,11
Laurasiatheria	88	20475	455	17939	87,61	17484	85,48	15	103	287	39,43
Afrotheria	94	18089	3480	11986	66,26	8506	47,08	3	4	28	3,44
Neognathae	105	14971	761	13041	87,11	12280	82,13	8	38	291	17,14
Tetraodontidae	65	18451	1693	16162	87,59	14469	78,50	5	18	103	9,55