

Manuscript ID: BIOINF-2007-1530

Associate Editor: Martin Bishop.

MANTIS, a phylogenetic framework for multi-species genome comparisons  
**Supplementary Material.**

Raphaël Helaers<sup>1#</sup>, Athanasia C. Tzika<sup>1#</sup>, Yves Van de Peer<sup>2</sup> & Michel C. Milinkovitch<sup>1\*</sup>

1. *Laboratory of Evolutionary Genetics, Institute for Molecular Biology & Medicine, Université Libre de Bruxelles, Gosselies, Belgium.*
2. *Bioinformatics & Evolutionary Genomics, Department of Plant Systems Biology, Ghent University, VIB, Gent, Belgium*

# These two authors contributed equally to this work

\* **Corresponding author:** Michel C. Milinkovitch

Laboratory of Evolutionary Genetics (CP 300), Institute for Molecular Biology & Medicine,  
Université Libre de Bruxelles,

B6041, Gosselies, Belgium

fax: +32 2 650 9950; phone: +32 2 650 9956

email: [mcmilink@ulb.ac.be](mailto:mcmilink@ulb.ac.be)

## Supplementary Methods

### *The MANTIS Software Pipeline*

The MANTIS software system integrates the following MySQL databases: (i) ENSEMBL protein trees, as given by the *ENSEMBL-Compara* database; (ii) MANTIS tree structure, *i.e.*, the best-supported species tree used for all MANTIS functionalities; (iii) character mapping (gains/losses) and genome content computed both with and without considering duplication events (see below); (iv) ENSEMBL gene descriptions; (v) *eGenetics* and GNF *Homo sapiens* expression data from the *ENSEMBL-Mart* database; (vi) HMDEG expression data available for ENSEMBL genes, converted from HMDEG categories to eVOC anatomical systems; (vii) PANTHER molecular function and biological process categories for *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, and *Drosophila melanogaster*; and (viii) conversion table from ENSEMBL gene IDs to ENSEMBL protein or transcript IDs, as well as to Entrez and Unigene IDs. Integration, manipulation, and interrogation of these databases are performed through a user-friendly interface whose functionalities are described below together with the MANTIS pipeline.

### *Biological processes and molecular functions*

Statistical significance of category over- or under-representation is computed on the basis of the category distribution of the reference-species genes: *e.g.*, a category  $C$  is over-represented in gains (or losses) when  $k(C)$ , the observed number of gained (or lost) genes of category  $C$ , is greater than  $p(C)K$ , the expected number of corresponding events (where  $p(C)$  is the proportion of genes of category  $C$  in the reference species, and  $K$  is the total number of gains (or losses) on the branch considered). Statistical significance is determined by the calculation of a  $p$ -value following the binomial statistics (under the null hypothesis, the number of genes mapped to  $C$  is distributed binomially with probability parameter  $p(C)$ ):

$$p\text{-value} = \sum \binom{K}{k} p(C)^k (1 - p(C))^{K-k} \quad (1)$$

where the sum runs from  $k(C)$  to  $K$  in the case of over-representation (*i.e.*, when the number of observed gains/losses is greater than expected under the null hypothesis), and from 0 to  $k(C)$  in the case of under-representation. All categories with a  $p$ -value  $< 0.05$  are considered ‘*significantly*’ over- or under- represented.

Given the cost of computing the p-value using the binomial distribution, MANTIS performs classical approximations as follows: (i) the Normal approximation is used when  $K$  is sufficiently large ( $>20$ ) and  $p(C)$  is not too close to 0 or 1 (i.e.,  $0.05 < p(C) < 0.95$ , and both  $Kp(C)$  and  $K(1-p(C))$  are  $> 5$ ), otherwise, (ii) the Poisson approximation is used when the mean and variance are similar (i.e., the  $Kp(C)$  value is between 90 and 110% of the  $Kp(C)(1-p(C))$  value) and  $K$  is sufficiently large (here,  $> 150$ ), otherwise (iii) the exact binomial computation (see above) is used if ( $K < 1000$ ), otherwise (iv) the Beta approximation is used as last resort.

### *Queries*

If a query includes several statements, each one is executed separately, and the results are grouped according to the selected logical operators: the intersection for the 'and' operator, the asymmetric difference for the 'and not' operator, the union for the 'or' operator, and the symmetric difference for the 'xor' operator (Figure 4). Operators priorities are also defined by the user (given three statements A, B, and C, the query '(A and B) or C' is not equivalent to the query 'A and (B or C)').

Furthermore, the 'Count mapping' / 'Count Functions' actions add a count condition (i.e., a comparison operator and a threshold; e.g., '> 2') to the mapping/functions criteria: occurrences having the same mapping/function type are grouped and counted, and only the genes that correspond to this criterion are displayed in the result sheet. This functionality allows to easily identify, e.g., the genes that have been gained in branch x (say, the root of the mammalian lineage) and lost less than twice within mammals. The request can be fine-tuned using the 'Display' commands (affecting the 'Gene' and/or 'Mapping' and/or 'Branch' and/or 'Function' fields): e.g., if a gene is lost at more than one branch, choosing to display the branch field will return one result for each branch at which the gene was lost instead of only one result with just the information that the gene was lost. Conversely, the selection of a 'Group & Count' field causes the merging of results with the same information in the remaining fields (and the count is displayed, instead): e.g. if genes, mapping, branches, and functions are selected to be displayed and genes to be 'Grouped & Counted', the result sheet will include a list of unique combinations of branches, gains or losses, and functions with the count of genes associated with each combination (e.g., 83 losses of genes with function x in branch y).

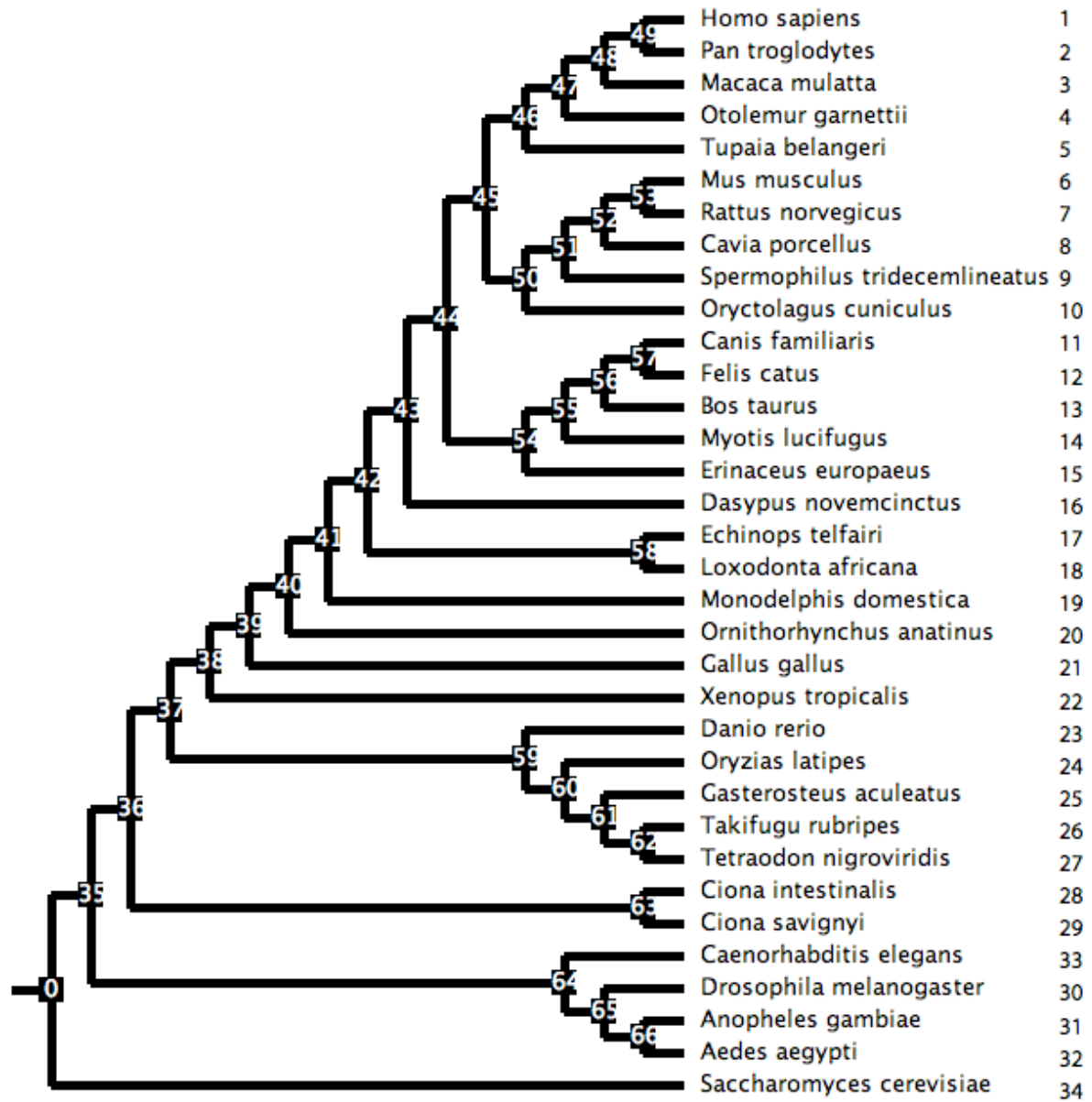
### **Supplementary Figure 1 legend.**

The “true” species tree, *i.e.*, the best-supported tree on the basis of the available literature (Bashir, et al., 2005; Halanych, 2004; Springer, et al., 2004), that is used in MANTIS. Numbers at the right of the tree indicate the priority order of the species for assignment of functional and expression data to MANTIS characters. Species sorting is performed partly according to phylogenetic proximity with human then mouse, *i.e.*, the two species with the largest amount of functional and expression data available.

### **Supplementary Figure 2 legend.**

Example of a complex MANTIS query: “*List the genes, among the 10,000 Entrez IDs that I provide, that are specifically expressed in the Nervous system (human EST data), are assigned to any Developmental process, were gained between the origin of vertebrates and the origin of eutherians, and are present in Human as well as in my two laboratory model species (i.e., mouse and dog)*”. The statement is built as follows: List the genes, among the 10,000 Entrez IDs that I provide (left part of Statement 1), that are assigned as ‘*Developmental*’ in the list of human ‘*Biological Processes*’ ontology terms (right part of Statement 1) and that have been gained in the most recent common ancestor (MRCA) of Vertebrates (branch 37) or Tetrapods (branch 38) or Amniotes (branch 39) or Mammals (branch 40) or Therians (branch 41) or Eutherians (branch 42) (central part of Statement 1). Among the resulting genes, I want to see only those that are also assigned (following human EST data) as ‘*Nervous*’ in the list of ‘*Gene Expression*’ ontology terms (Statement 2, linked with an ‘AND’ to Statement 1). Among the resulting genes, I want to see only those that are present in human, mouse, and dog (Statement 3). The statement 3 requires identifying the branches in which the genes have not been lost, hence, the ‘AND NOT’ operator linking the result of statements 1&2 with the statement 3. An alternative Statement 3 would have been to use the ‘Presence’ field (instead of the ‘Mapping -> Losses’ field) and listing “*Homo sapiens*”, “*Mus musculus*”, and “*Canis familiaris*” in the branch field; an ‘AND’ operator between statements 2 and 3 should have then been used. Selections of processes and branches are facilitated by the ‘*Category selection*’ (upper left) and ‘*Branch selection*’ (lower left) graphical tools, respectively. The 9 genes generated by the query meet all criteria and can be exported or used for a new query.

Supplementary Figure 1.



## Supplementary Figure 2.

The figure illustrates a multi-step bioinformatics query process. It starts with a 'Category selection' window where 'Developmental processes' is chosen. This leads to a 'Query' window where three statements are defined: Statement 1 (Gene: 1417, 9421, 58158, 22326; Branch: 37,38,39,40,41,42; Function: Developmental processes), Statement 2 (Not considered), and Statement 3 (Not considered). Logical operators 'AND' and 'AND NOT' are used between statements. A 'Branch selection' window shows a phylogenetic tree with 'Homo sapiens' selected. The 'Query result 1' window displays details for gene ENSG00000171532, including its description, orthologs, and biological processes. A table of orthologous genes is provided below.

Main gene	Gene
ENSG00000100053	ENSG00000100053
ENSG00000113196	ENSG00000113196
ENSG00000123307	ENSG00000123307
ENSG00000148704	ENSMUSG00000006270
ENSG00000148704	ENSG00000148704
ENSG00000171532	ENSMUSG00000038255
ENSG00000171532	ENSG00000171532
ENSG00000183423	ENSG00000183423
ENSG00000185610	ENSG00000185610

## Supplementary References

- Bashir, A., Ye, C., Price, A.L. and Bafna, V. (2005) Orthologous repeats and mammalian phylogenetic inference, *Genome Res.* %R 10.1101/gr.3493405, **15**, 998-1006.
- Halanych, K.M. (2004) The New View of Animal Phylogeny, *Annual Review of Ecology, Evolution, and Systematics*, **35**, 229-256.
- Springer, M.S., Stanhope, M.J., Madsen, O. and de Jong, W.W. (2004) Molecules consolidate the placental mammal tree, *Trends in ecology & evolution (Personal edition)*, **19**, 430-438.