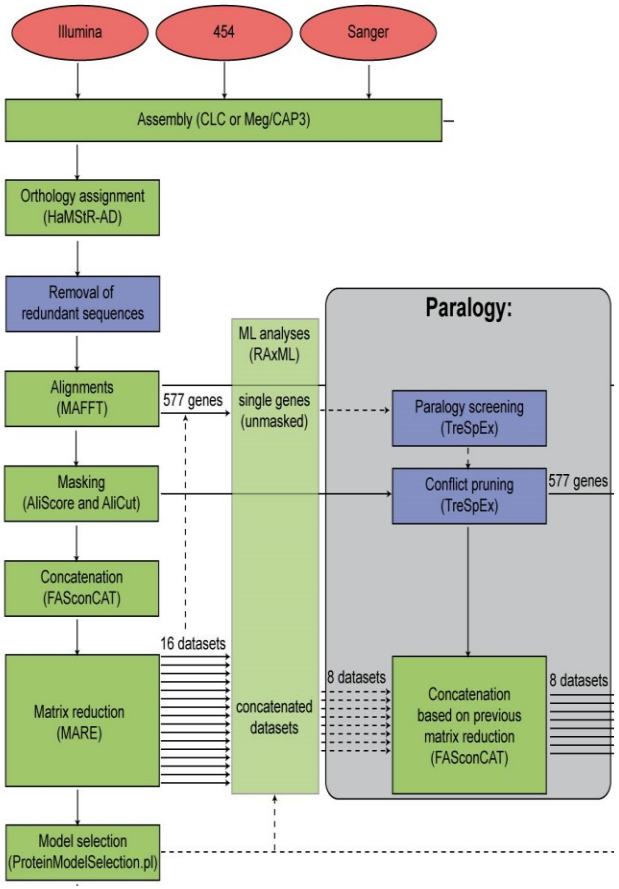
Paralogy

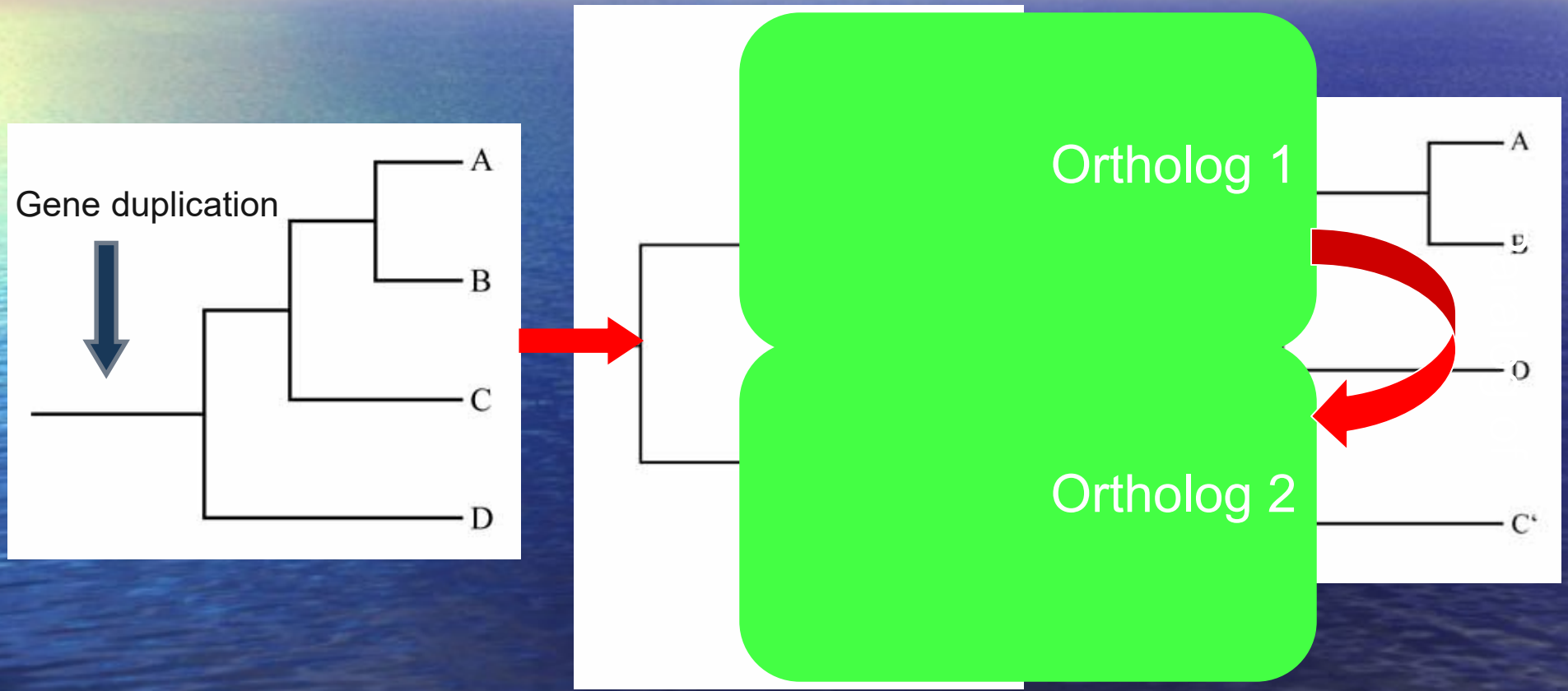
Torsten Struck – t.h.struck@nhm.uio.no
NHM UiO

Going beyond the standard



Struck (2013) PLoS one
Struck (2014) Evol. Bioinf.

Problem of paralogy

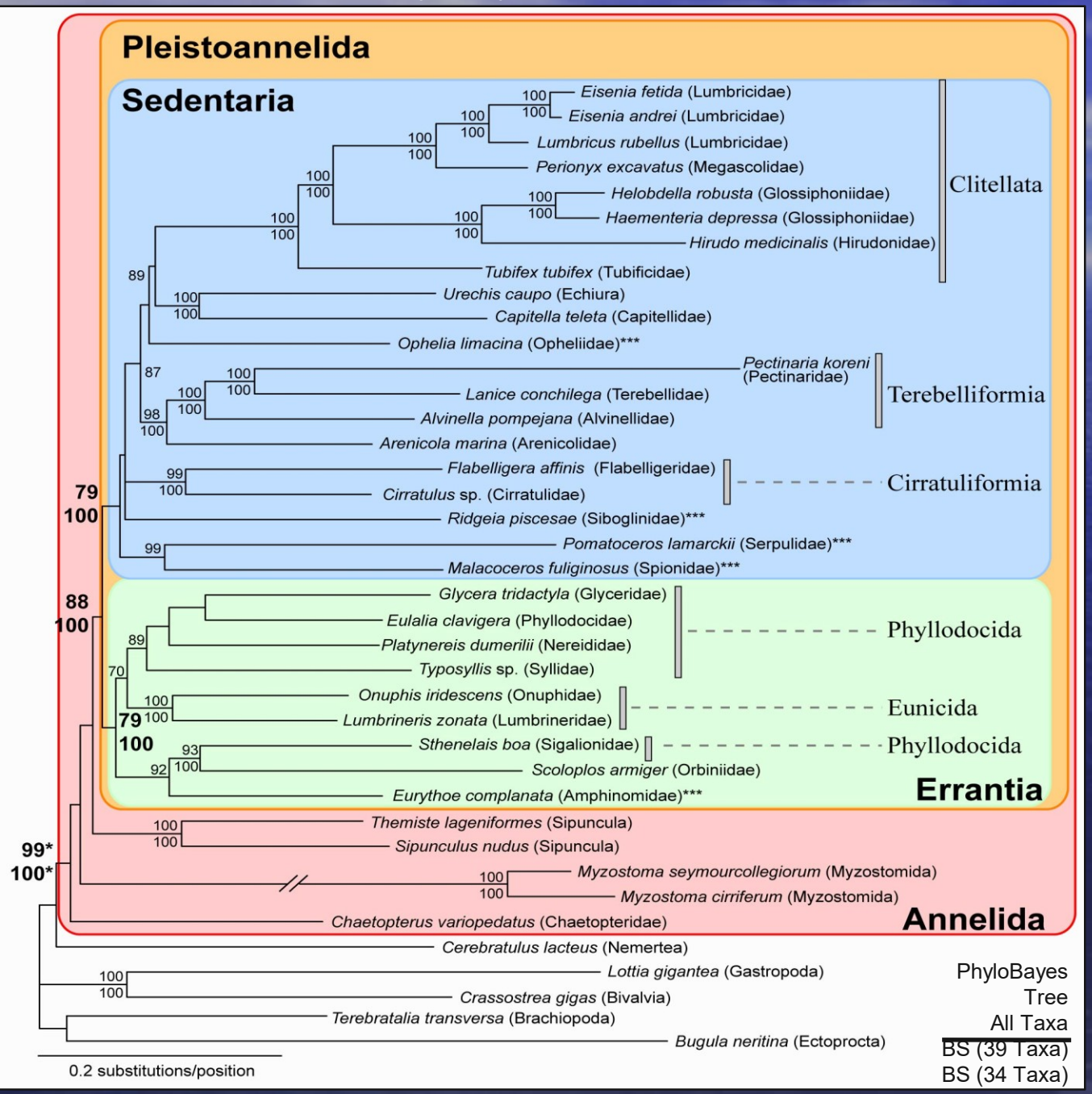


Species tree

Gene tree

Wrong species tree

Modified from Struck et al. (2011)



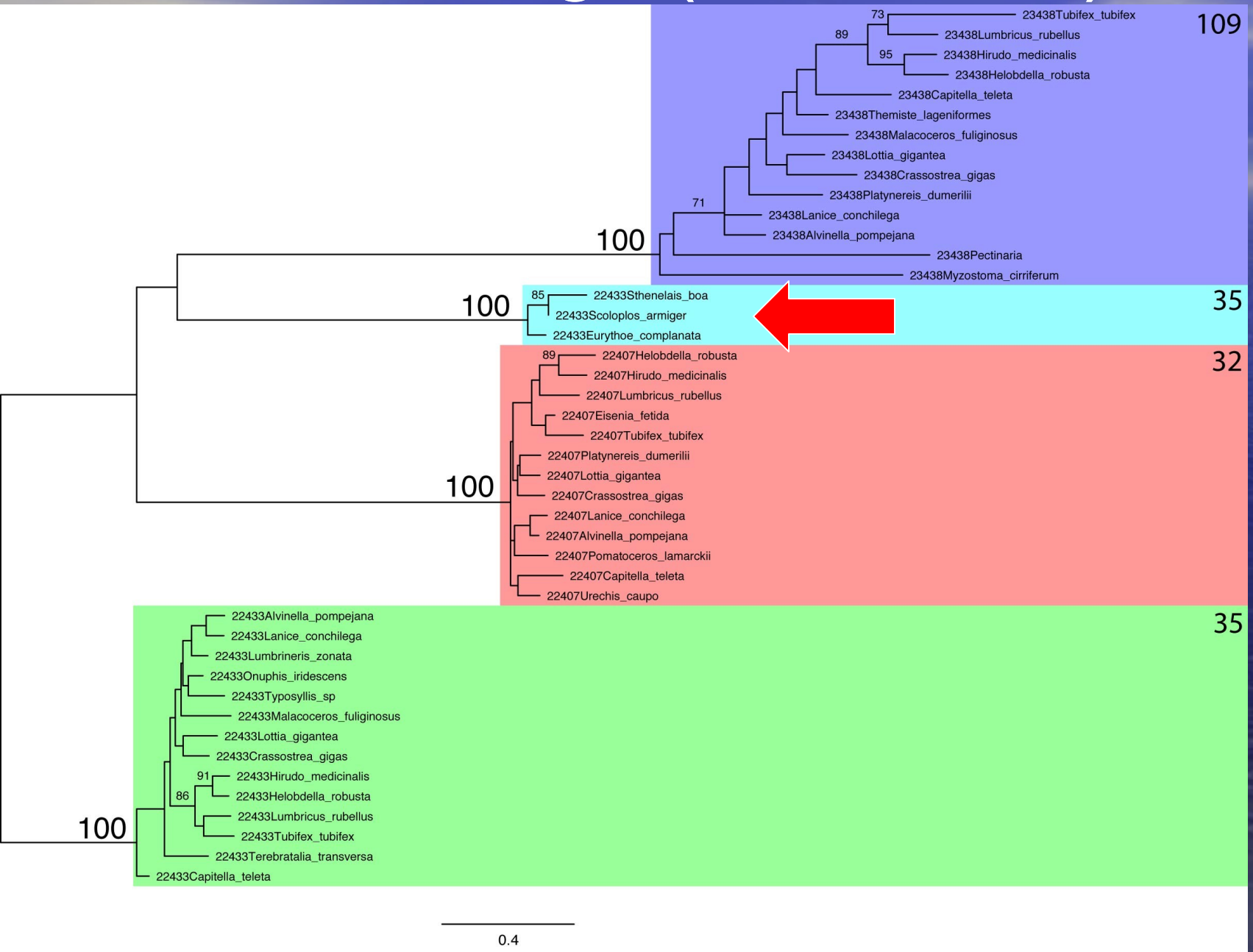
Phylogenomics

17 new EST-Datasets
Approx. 1,300 clones
231 ortholog genes
47,952 aa-positions
41.7% coverage per taxon
20 polychaete families [inclusive of Siboglinidae (Pogonophora)]
Clitellata, Sipuncula, Echiura, Myzostomida

PhyloBayes/RAxML
Only bootstrap values ≥ 70%
Sensitivity analyses
*** unstable taxa

•BS-values without Myzostomida due to long-branch issues in RAxML

Paralogs (32/109/35)



Paralogs (32/109/35)

Blast against *Bos taurus* transkriptome
= Proteasome subunit alpha

reference taxa	32 (23428)	109 (22407)	35 (22433)
Helobdella	PSMA5	PSMA3	PSMA7/8
Capitella	PSMA5	PSMA3	PSMA7/8
Lottia	PSMA5	PSMA3	PSMA7/8

Critcial taxa	35 (22433)
Eurythoe	PSMA2
Scoloplos	PSMA2
Sthenelais	PSMA2

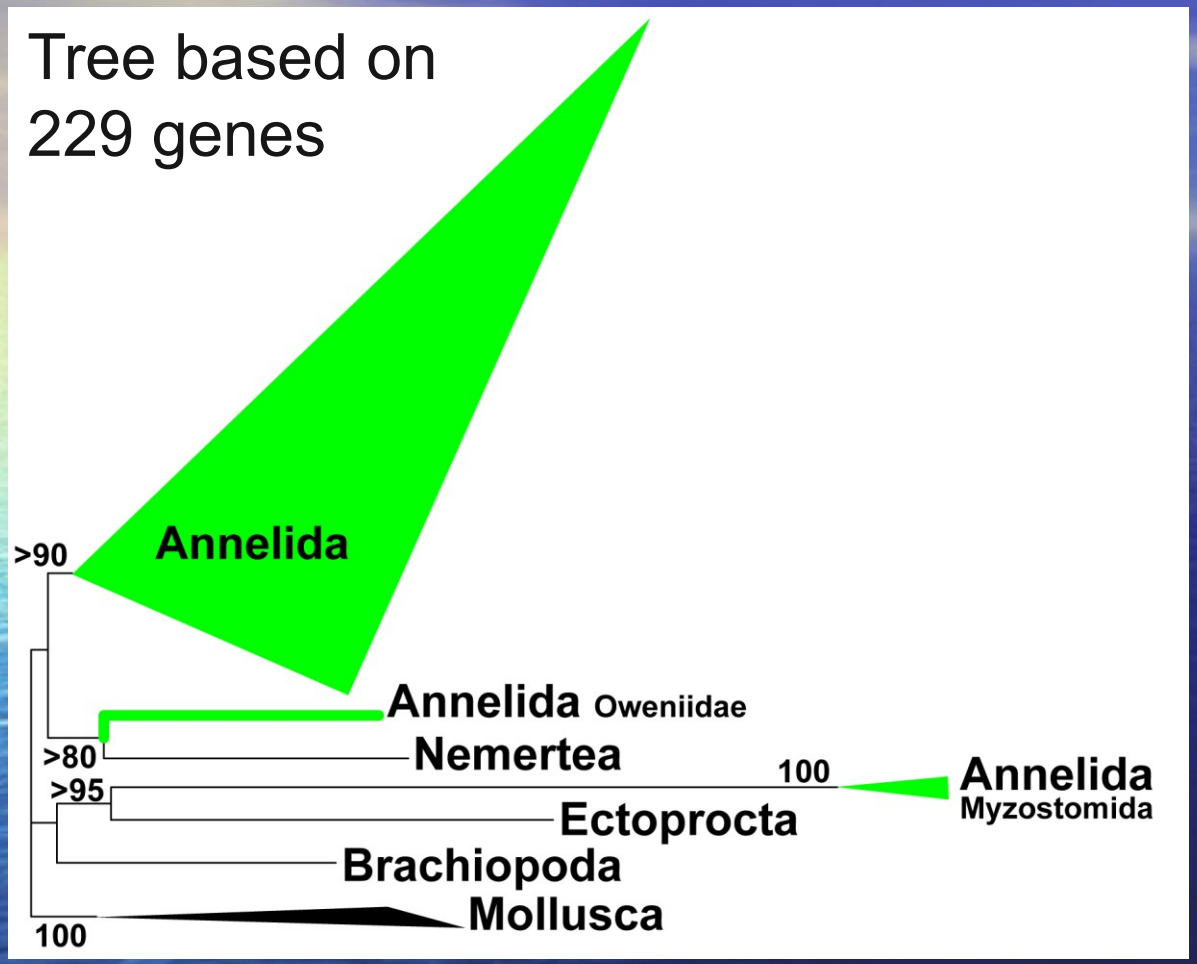
Blast of *Bos taurus* sequence against *Helobdella* transcriptome:
PSMA2 (NM_001034662): *Helobdella* sequence 185484 e^-39
That is the sequence of orthology group 22433

What has happened?

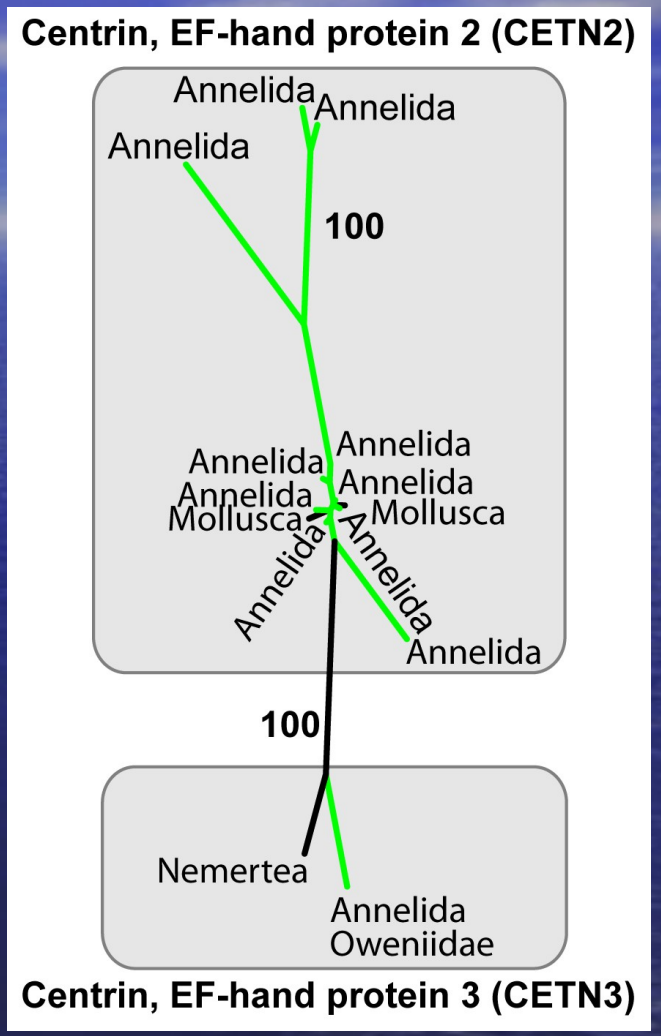
	PSMA7/8	PSMA2
Helobdella	present	absent
Eurythoe	absent	present


Eiseniafetida
Eiseniaandrei
Lumbricusrubellus
Perionyxexcavatus
Haementeriadepressa
Helobdellarobusta
Hirudomedicinalis
Tubifextubifex
Laniceconchilega
Pectinaria
Alvinellapompejana
Arenicolamarina
Capitellateleta
Urechiscaupo
Ophelialimacina
Flabelligeraaffinis
Cirratulidae
Ridgeiapiscesae
Malacocerosfuliginosus
Pomatocerosla marckii
Scoloplosarmiger
Glyceratridactyla
Eulaliaclavigera
Platynereisdumerilii
Typosyllissp
Onuphisiridescens
Lumbrineriszonata
Sthenelaisboa
Sipunculusnudus
Themistelageniformes
Eurythoecomplanata
Chaetopterusvariegatus
Crassostreagigas
Lottiagigantea
Cerebratuluslacteus
Terebrataliatransversa
Myzostomaseymourcollegiorum
Myzostomacirriferum
Bugulaneritina
100
100
100
100
100
100
100
100
27
100
100
95
41
100
18
33
100
29
29
100
18
86
68
51
62
93
100
100
69
100
86
78
71
99
100
49
100
0.08

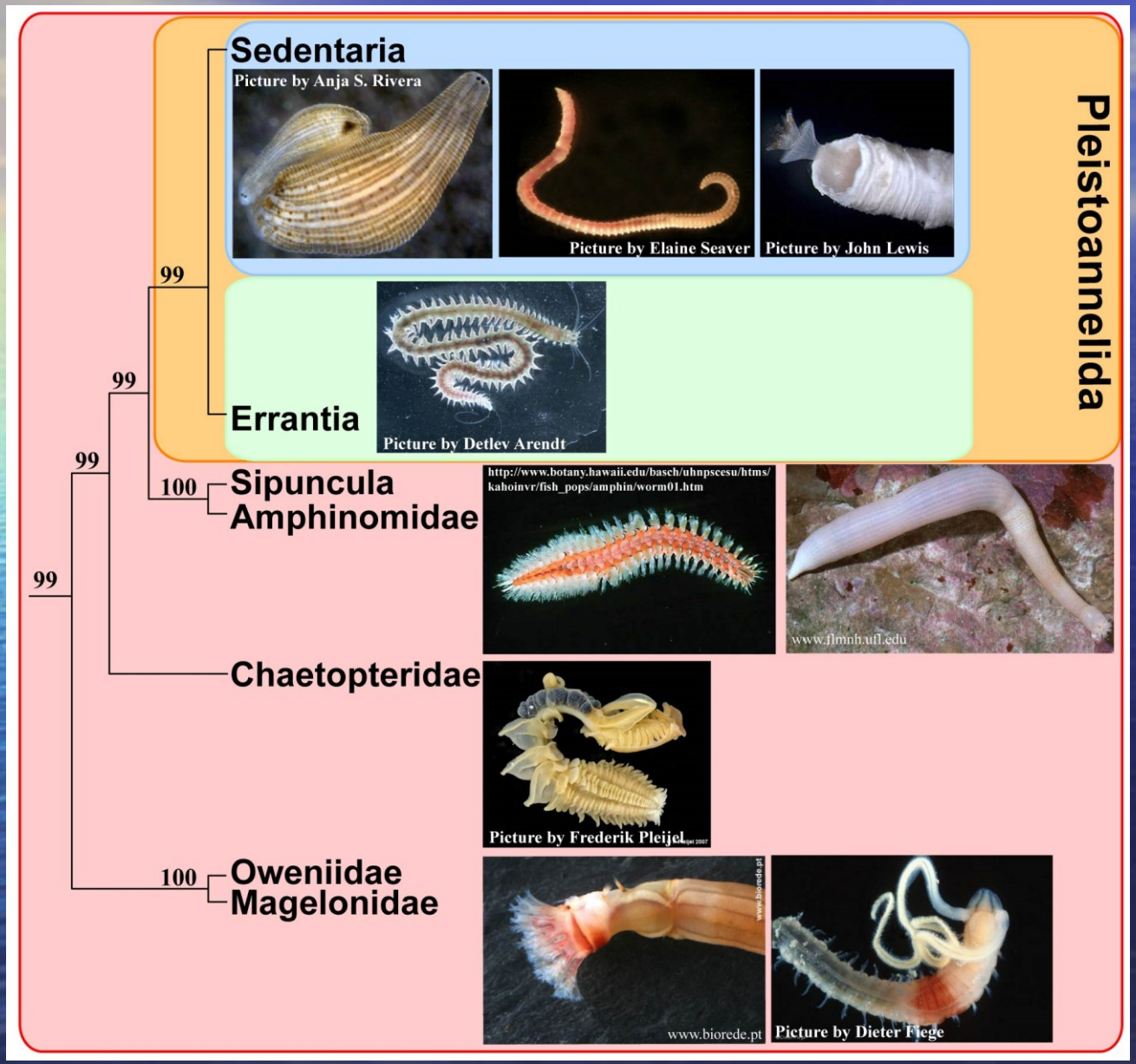
Problem of paralogy



Struck (2013) PLoS one



More data and taxa of Annelida



28 new NGS datasets
620 orthologous genes
155422 aa

RAxML

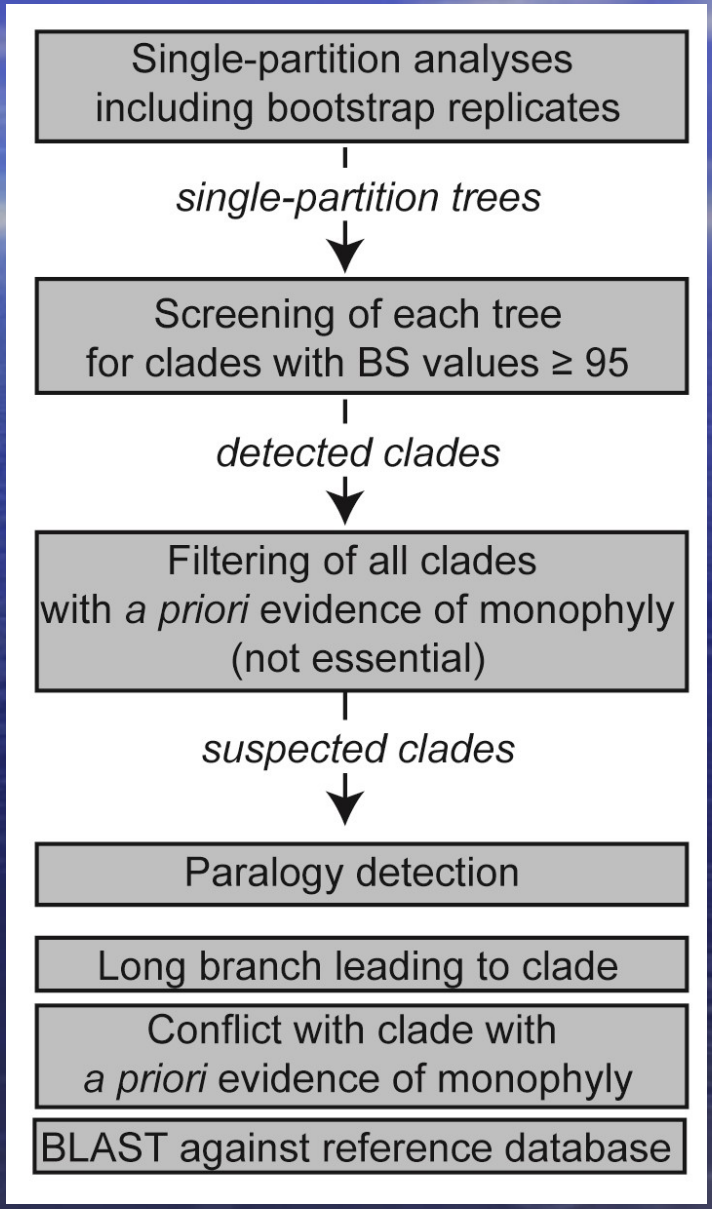
Struck et al. (2011) Nature; Weigert et al. (2014) MBE

Hughes et al. (2018)

- paralogy originating from inferred WGDs in ancestral vertebrates or teleosts
- two sets of topological constraints (monophyly of all teleosts; monophyly of Ostariophysi)
- topology tests (AU tests) of the constrained tree against the unconstrained ML topology
- rejection of constraint topology ($p < 0.05$) was expected in the presence of paralogy

Detection of paralogy

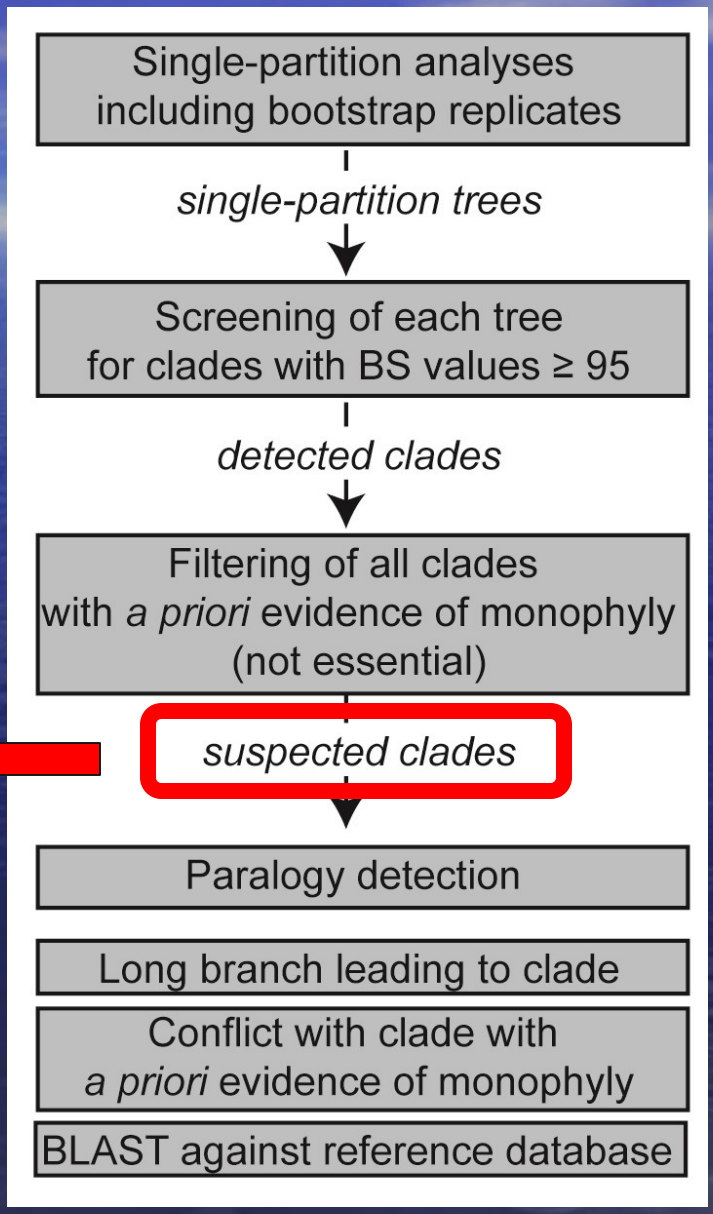
Paralogous sequences can be problematic in large-scale analyses of hundreds of genes



Struck (2013) PLoS one; Struck et al. (2014) MBE; Struck (2014) Evol. Bioinf.

Detection of deviating single genes

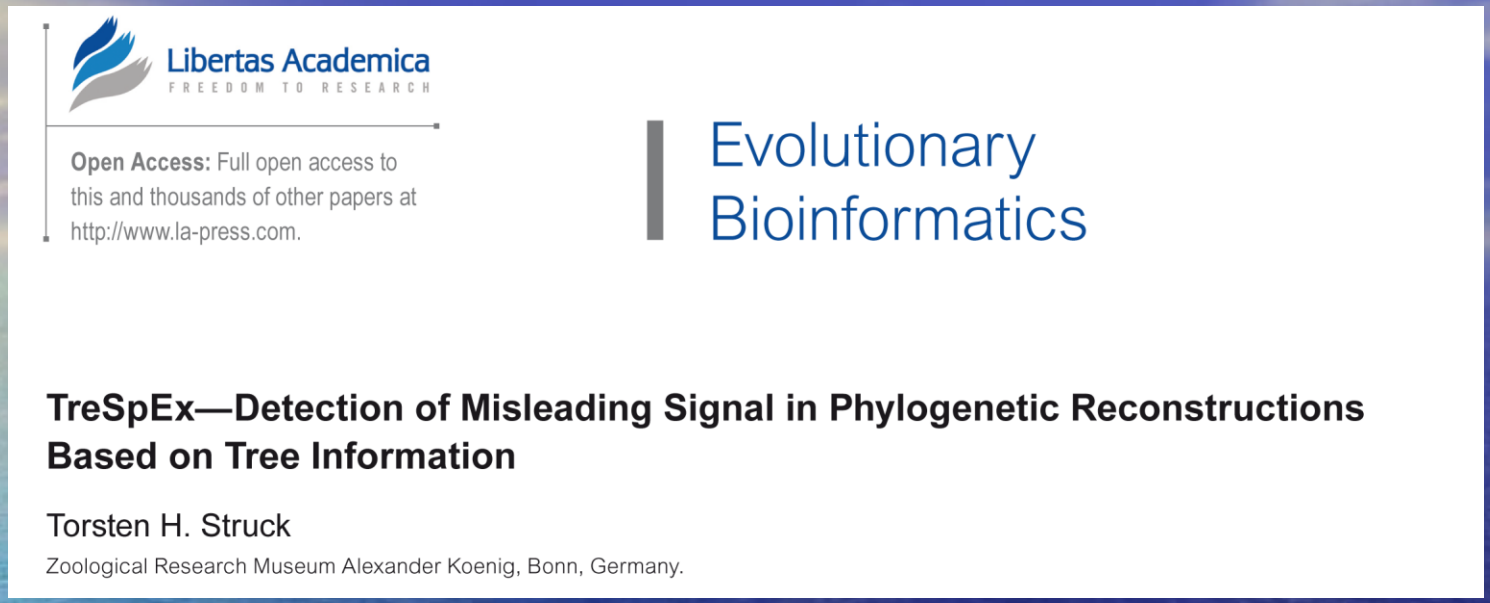
Can be used to detect clades in single genes deviating from a given relationship. Based on bootstrap support



Paralogy is only one possible cause of deviation.
Others are incomplete lineage sorting, horizontal gene transfer or ancestral hybridization.

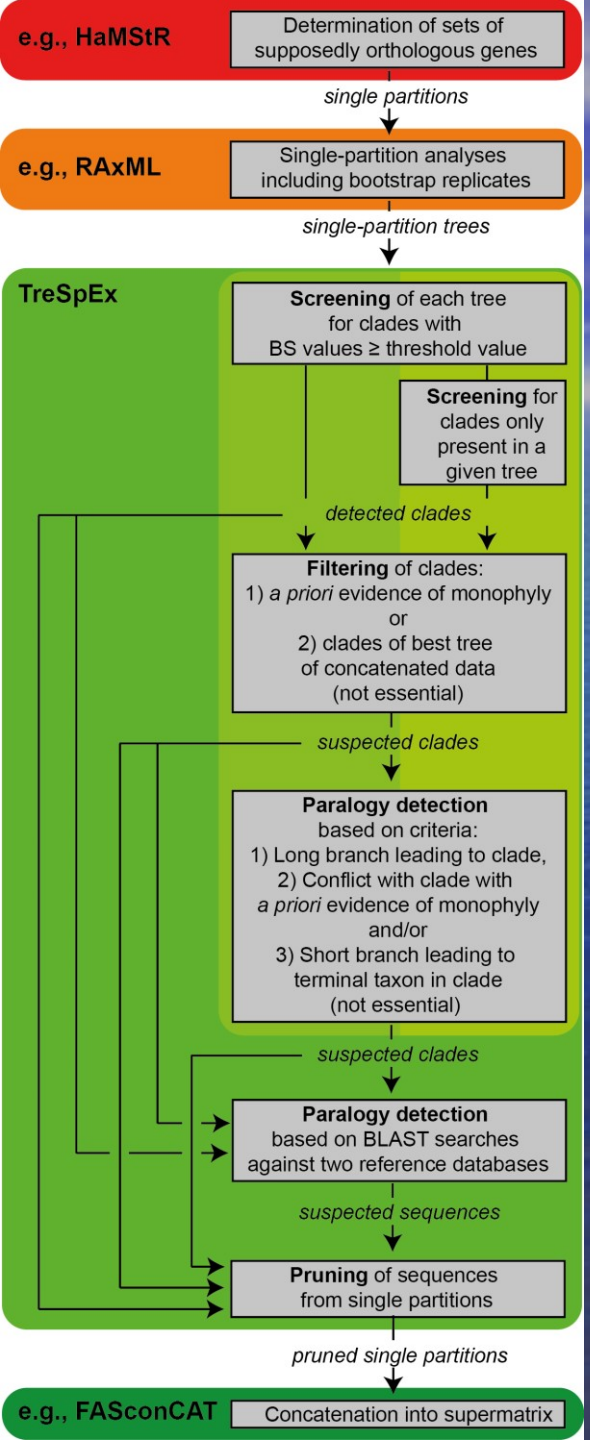
Struck (2013) PLoS one; Struck et al. (2014) MBE; Struck (2014) Evol. Bioinf.

TreSpEx



Different methods related to tree-based measurements:

- **Detection and pruning of paralogous sequences**
- Detection of conflict
- Detection of long-branched taxa and partitions
- Detection of saturation and phylogenetic signal



Paralogy detection

Command-line options

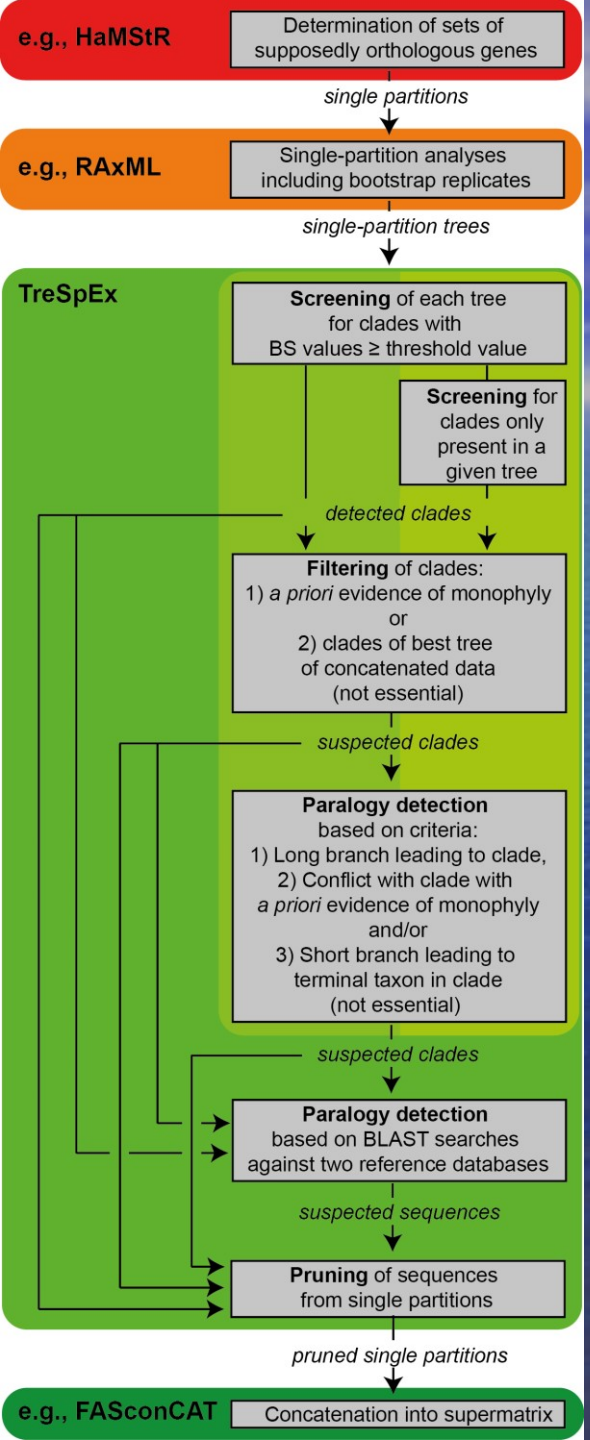
-fun <a, b, c or d> | calls the desired function like RAxML

(a = *A priori* screening)

(b = *A posteriori* screening)

(c = Blast search against reference database)

(d = Pruning of suspicious sequences from the individual datasets)



Paralogy detection

Command-line options for fun a

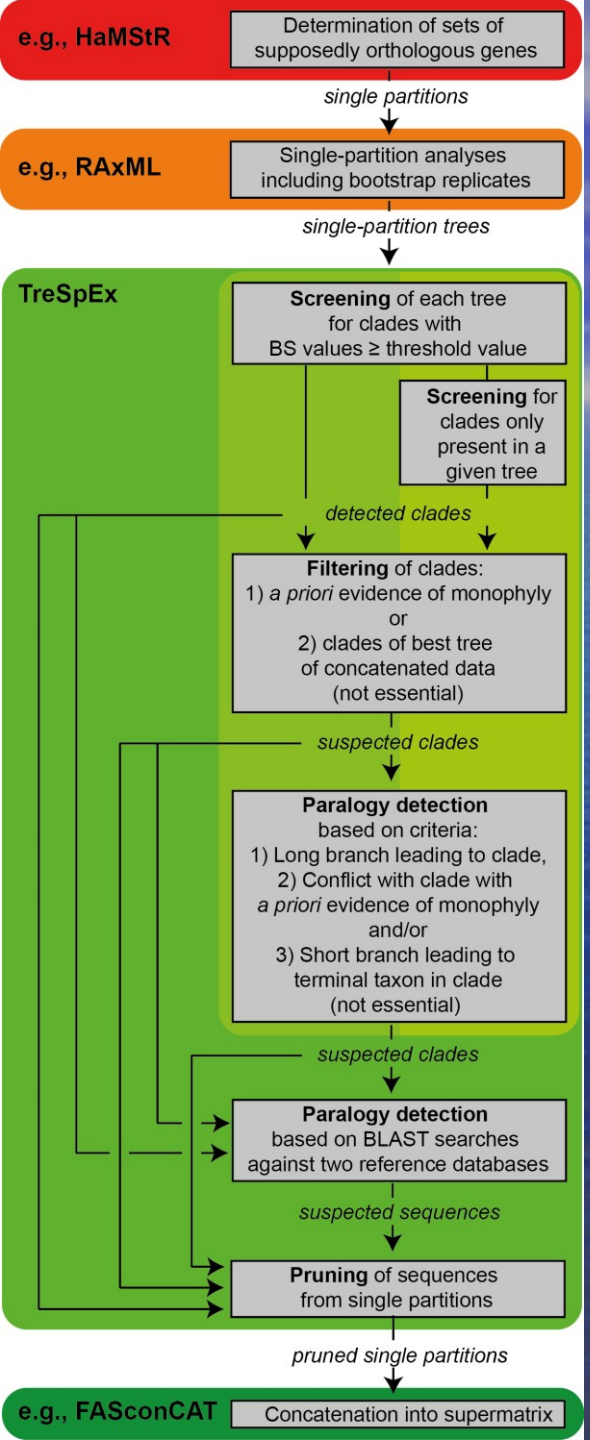
-ipt <file name> | file containing the list of names of tree files in Newick format with branch length and bootstrap values

-lowbs <value> and –upbs <value> | provide the lower and upper threshold of bootstrap values for the screening (e.g. from 95 to 100 or from 95 to 95)

-path <path> | specifies the path to the files (optional)

For filtering clades:

-gts <Y or N> | to mask clades and additionally in the same folder a file or files beginning with *GroupedTaxa* and ending with *.txt* (e.g., GroupedTaxaFamily.txt) has to be provided. A nested approach is NOT possible.



Paralogy detection

Command-line options for fun a

Criteria-based sorting:

-possc <0,1,2 or 3> | strong conflict to a priori clade;
0 = turned off;
1, 2 or 3 = turned on & position of priority provided

-poslb <0,1,2 or 3> | long branch leading to clade

-possb <0,1,2 or 3> | short branch leading to terminal taxon

Additional options for –possc:

-gts Y is required plus the accompanying file or files

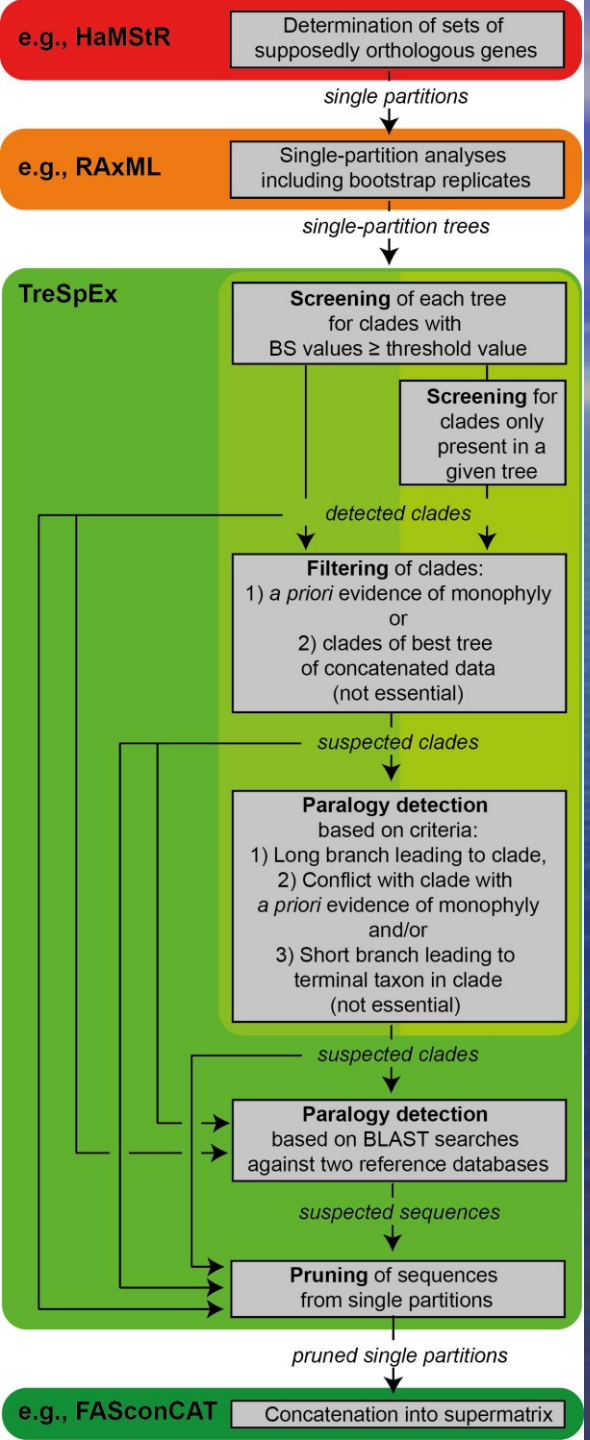
Additonal options for –poslb:

–lowbl <value> and –upbl <value> | lower and upper limit of the ratio the leading branch can be longer than the average of all internal branches

Additional options for –possb:

-blt <value> | upper limit of the terminal branch length

–maxtaxa <value> | maximum number of taxa present in a detected clade with short branches



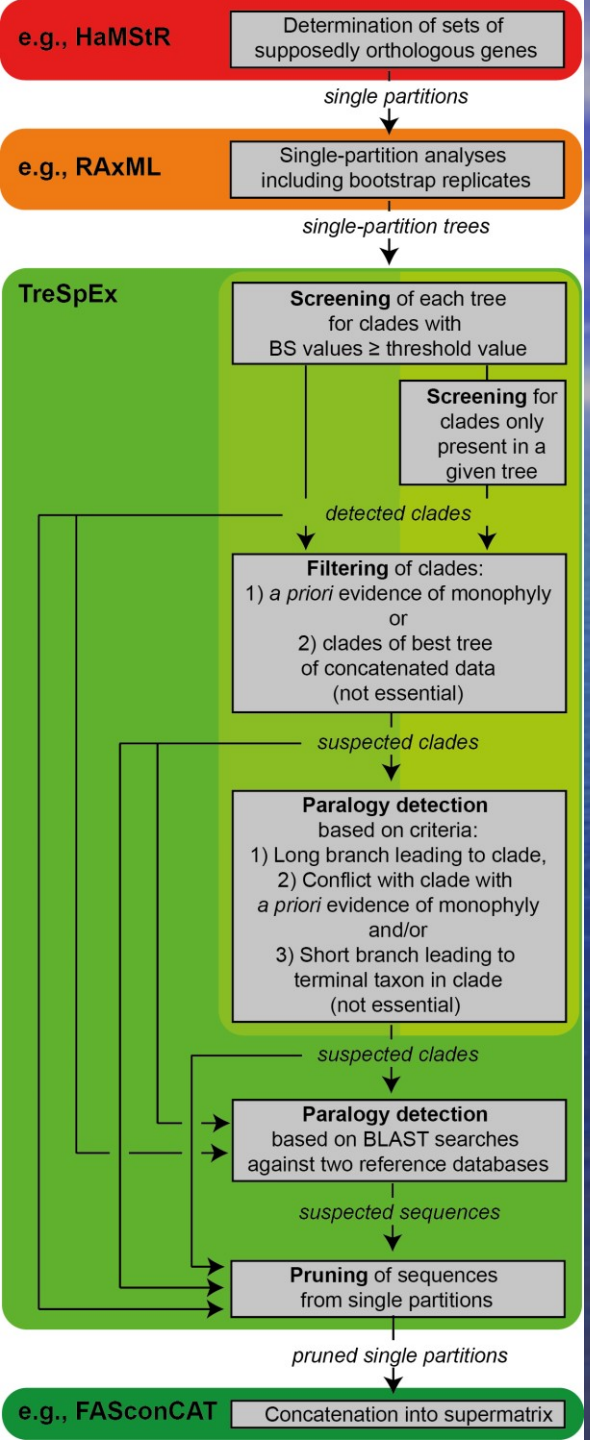
Paralogy detection

Command-line options for fun b

All the same as for -fun a, but in contrast herein the screening is done only for clades provided via a tree file. All detected clades not congruent with these clades will be masked. The tree file has not be completely resolved.

Additional option:

–trf <Name-of-file> name of tree file, which contains a single tree in Newick format



Paralogy detection

Command-line options for fun c

-ppf <file name> | a file with a list of the suspicious clades, as generated by TreSpEx using -fun a or b (e.g., Pruned_PotentialParalogsBootstrap.txt).

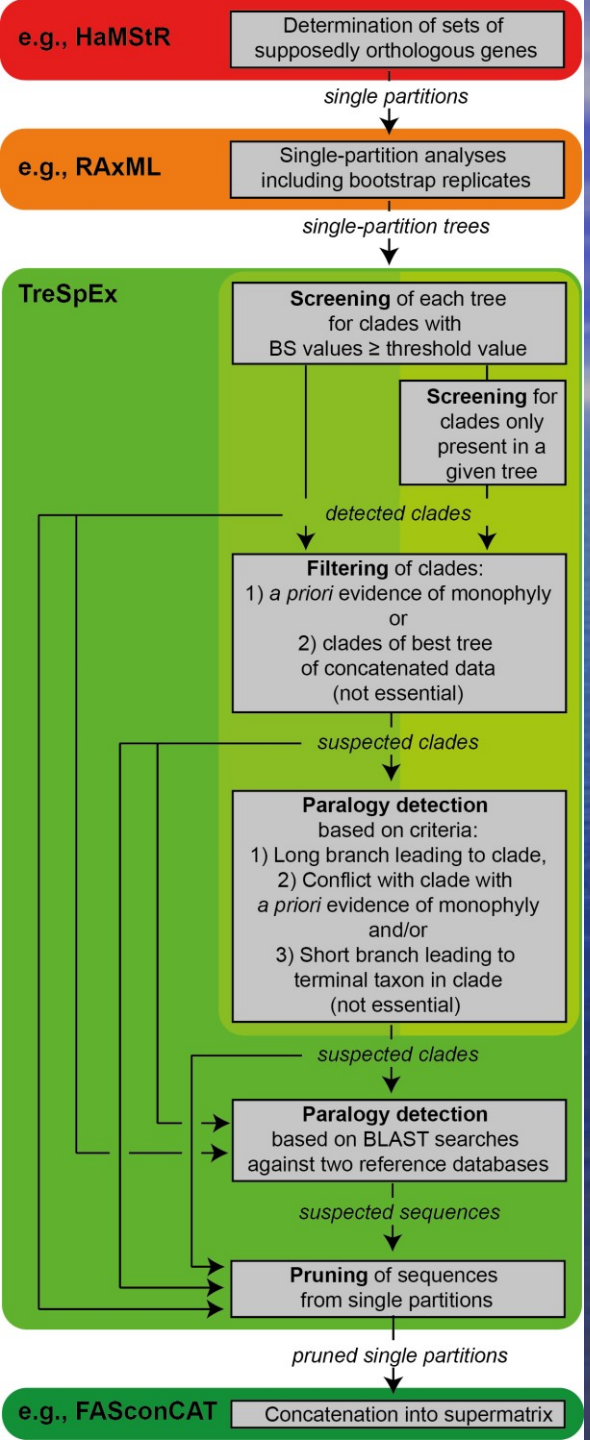
-ipt <file name> | file containing the list of names of tree files in Newick format with branch length and bootstrap values

-ipa <file name> | file containing the list of names of alignment files in phylip format in same order as in the ipt file

-path <path> | specifies the path to the files (optional)

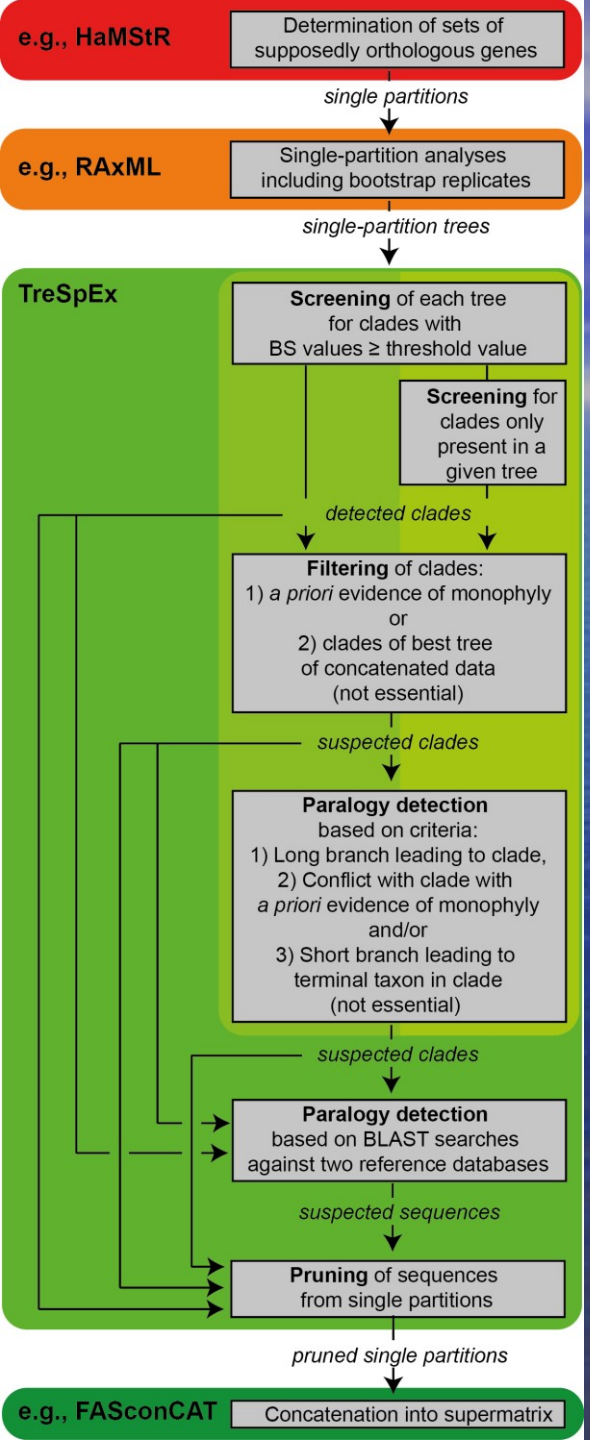
-db1 <Name-of-database> and -db2 <Name-of-database> | two reference databases, against which the blast search is conducted, within the folder blast/bin/Databases or in the same folder as the other files. Ideally, these databases are non-redundant, that is each orthologous sequence is represented in the database only by one sequence. Precompiled databases are available for

Paralogy detection



Available databases (have to be placed in the blast/bin/Databases folder):

Anolis_carolinensis
Apis_mellifera
Bombyx_mori
Bos_taurus
Branchiostoma_floridae
Caenorhabditis_elegans
Capitella_teleta
Ciona_intestinalis
Danio_rerio
Daphnia_pulex
Drosophila_melanogaster
Gallus_gallus
Helobdella_robusta
Homo_sapiens
Lottia_gigantia
Monodelphis_domestica
Mus_musculus
Ornithorhynchus_anatinus
Pediculus_humanus
Schmidtea_mediterranea
Taeniopygia guttata
Tribolium_castaneum
Xenopus_tropicalis



Paralogy detection

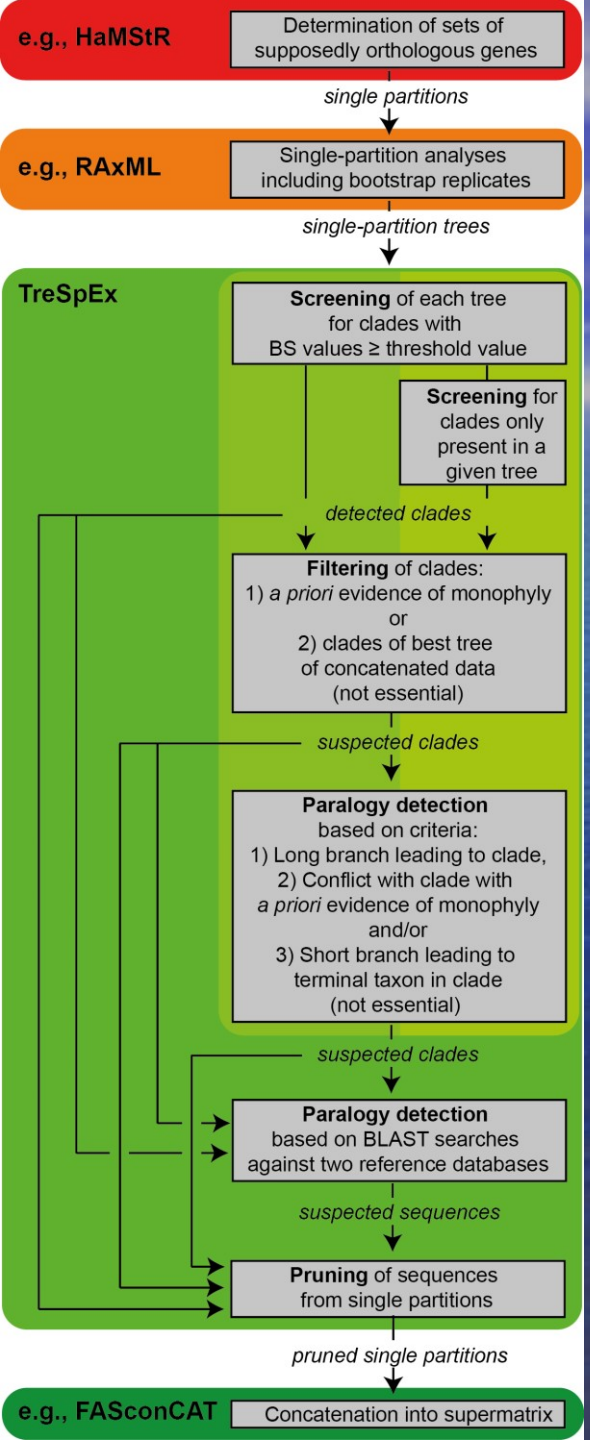
Command-line options for fun c

–evalue <value> | The threshold of the e value for the blast searches (e.g., -evalue 1e-20)

TreSpEx will automatically determine, which blast search has to be conducted given the database and the alignment file and also if blast-searchable database has to be generated.

Sorting options (no hits, certain, no paralogy and uncertain):

-ltp <value> and -utp <value >| lower and upper sorting threshold values in proportion of identical blast results (e.g., -ltp 0.0 -utp 1.0)



Paralogy detection

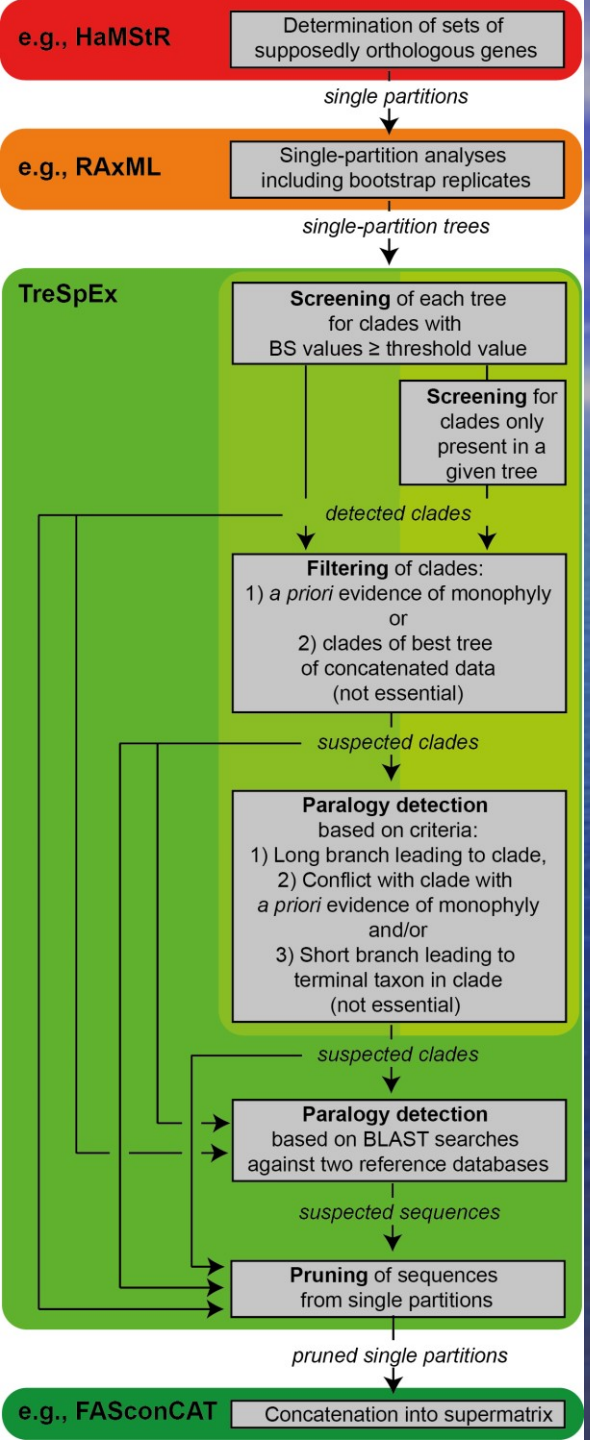
Sorting procedure

For each suspicious sequence the proportion of best hits in the searches of the other non-suspicious sequences is determined, which areidentical with the best hit of that suspicious sequence.

If for ALL suspicious sequences in the searches against ONE of the two databases this proportion is equal to or lower than this lower threshold (ltp) this detected clade is regarded as a certain case of PARALOGY.

If for ALL suspicious sequences in the searches against BOTH databases this proportion is equal to or higher than this upper threshold (utp) this detected clade is regarded as a certain case of NONE PARALOGY.

All remaining cases are considered as UNCERTAIN.



Paralogy detection

Command-line options for fun c

–evalue <value> | The threshold of the e value for the blast searches (e.g., -evalue 1e-20)

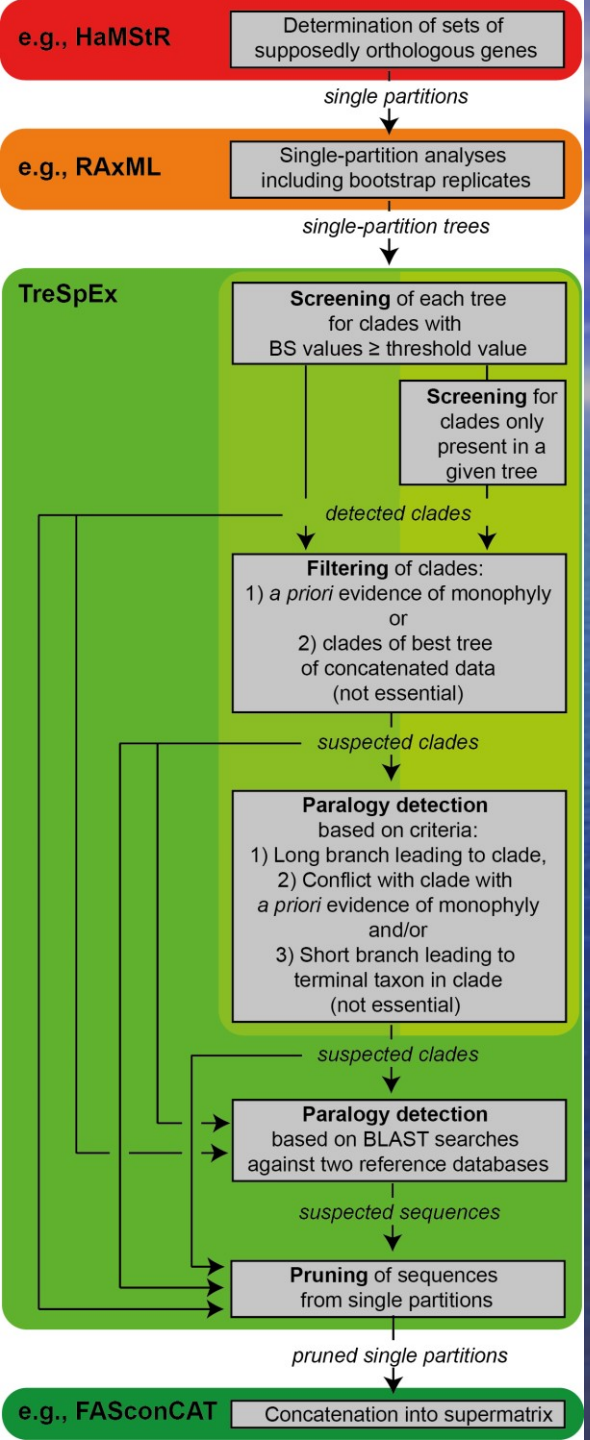
TreSpEx will automatically determine, which blast search has to be conducted given the database and the alignment file and also if blast-searchable database has to be generated.

Sorting options (no hits, certain, no paralogy and uncertain):

-ltp <value> and -utp <value >| lower and upper sorting threshold values in proportion of identical blast results (e.g., -ltp 0.0 -utp 1.0)

For explorative reasons:

–ediff <value> | This value provides the difference in the order of magnitude (e.g., -ediff 5) by which the e value can be higher to be still considered part of the confidence set of best hits.



Paralogy detection

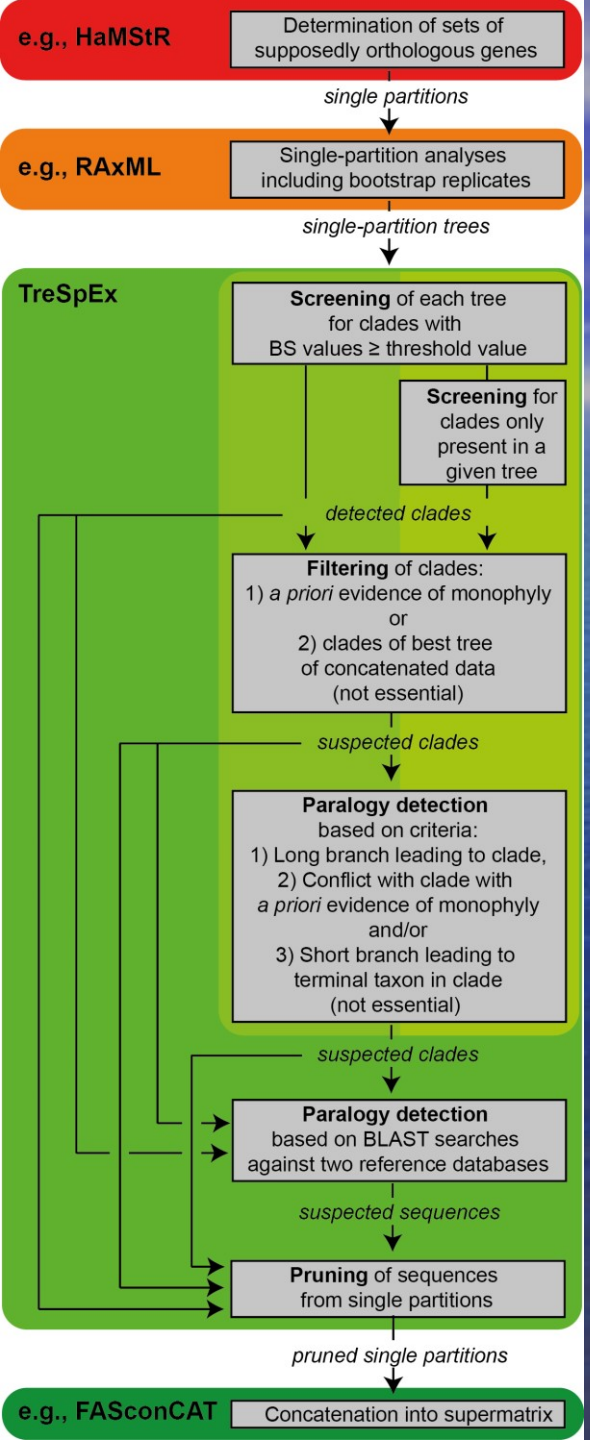
Command-line options for fun d

-ppf <file name> | a file with a list of the suspicious clades, as generated by TreSpEx using -fun a, b or c (Certain_Pruned_PotentialParalogsBootstrap.txt).

-ipt <file name> | file containing the list of names of tree files in Newick format with branch length and ootstrap values

-ipa <file name> | file containing the list of names of alignment files in phylip format in same order as in the ipt file

-path <path> | specifies the path to the files (optional)



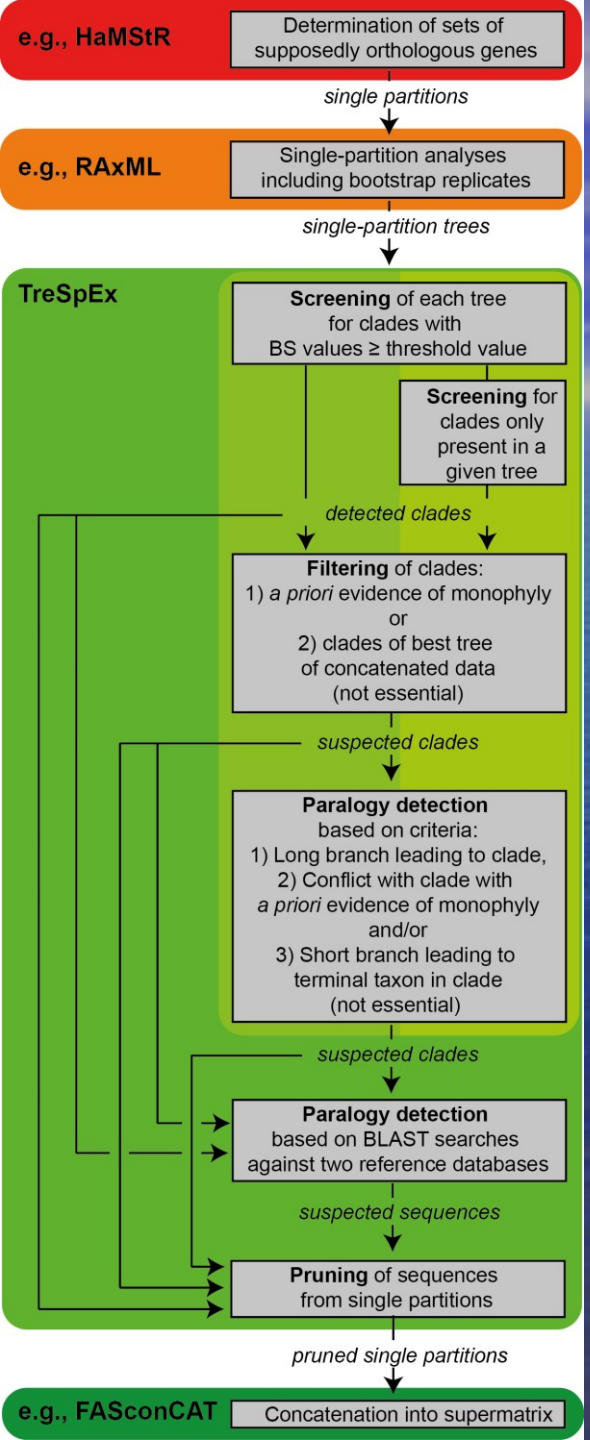
Example commands

A priori screening

perl TreSpEx.v1.pl -fun a -gts Y or N -lowbs Value -upbs Value -possc 0,1,2,3 -poslb 0,1,2,3 -lowbl Value -upbl Value -possb 0,1,2,3 -maxtaxa Value -blt Value -ipt Name-of-file -path Path-from-root

A posteriori screening

perl TreSpEx.v1.pl -fun a -gts Y or N -lowbs Value -upbs Value -possc 0,1,2,3 -poslb 0,1,2,3 -lowbl Value -upbl Value -possb 0,1,2,3 -maxtaxa Value -blt Value -ipt Name-of-file -path Path-from-root



Example commands

Blast search against reference database

perl TreSpEx.v1.pl -fun c -ppf Name-of-file -ipt Name-of-file -ipa Name-of-file –db1 Name-of-database –db2 Name-of-database -ediff value -ltp value -utp value -evalue value -path Path-from-root

Pruning of suspicious sequences from the individual datasets

perl TreSpEx.v1.pl -fun d -ppf Name-of-file -ipt Name-of-file -ipa Name-of-file -path Path-from-root