



# *Branch length heterogeneity*

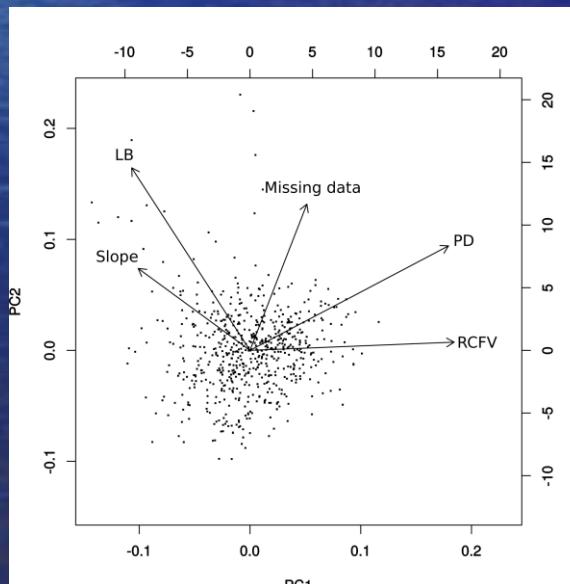


**Torsten Struck – [t.h.struck@nhm.uio.no](mailto:t.h.struck@nhm.uio.no)**  
**NHM UiO**

© nhm.uio.no

# *Long branch attraction*

Long branch attraction is caused by differences in branch lengths and NOT by an overall increased substitution rates. Often used as a proxy for long branch attraction.



Kocot et al. (2016)

# *Visualization*

METHODOLOGY ARTICLE

Open Access

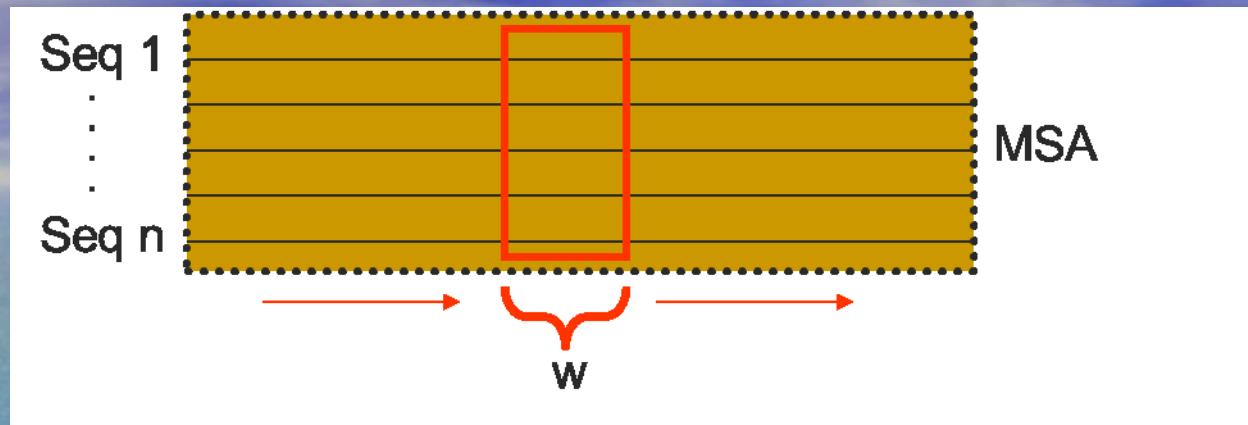
## AliGROOVE – visualization of heterogeneous sequence divergence within multiple sequence alignments and detection of inflated branch support

Patrick Kück<sup>1\*</sup>, Sandra A Meid<sup>1</sup>, Christian Groß<sup>2</sup>, Johann W Wägele<sup>1</sup> and Bernhard Misof<sup>1</sup>

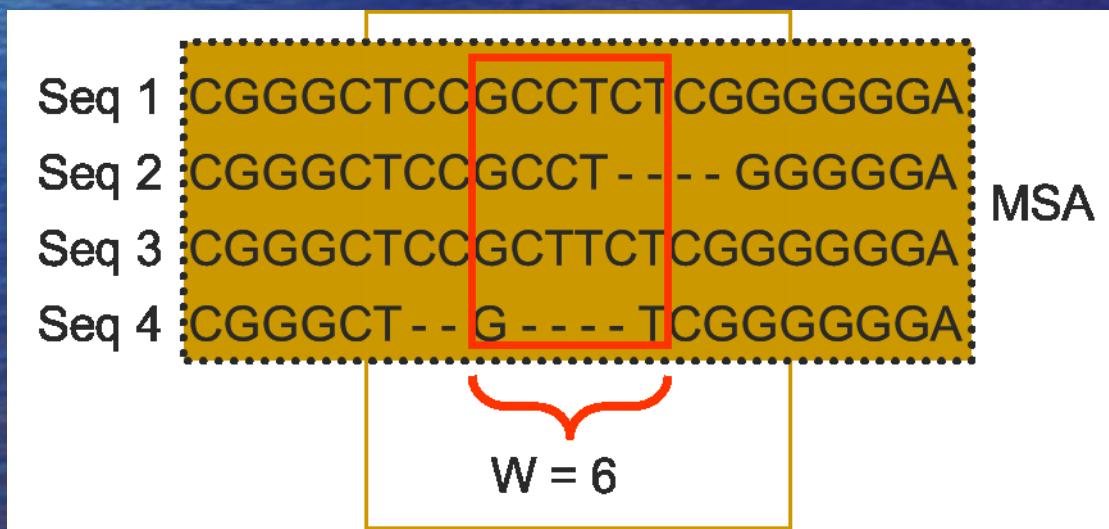
heterogeneous sequence divergence:

- increased substitution rate along a branch
- ambiguity in the alignment
- Sliding window approach
- Monte Carlo Resampling Algorithm

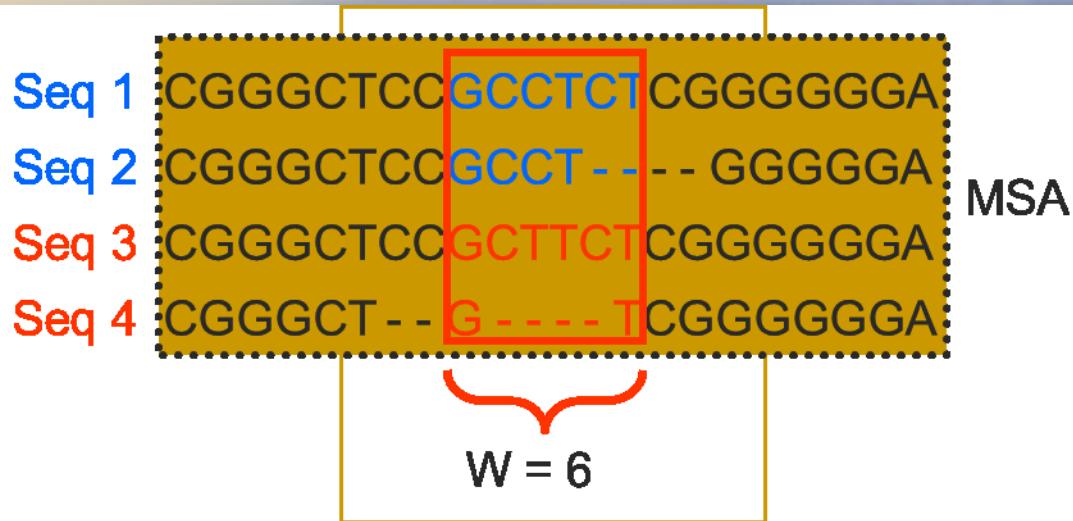
# *The procedure*



Sliding window size set to 6 (default)



# *The procedure*



Sequence pairs  $(x,y)$  are selected

Seq 1 GCCTCT  
Seq 2 GCCT - -

Seq 3 GCTTCT  
Seq 4 G - - - T

All pairwise comparisons

High sequence numbers  
are time expensive !!

$$P_{xy} = N \times (N - 1) / 2$$

# *The similarity score*

$$S_{obs} = \sum_{k=i}^{i+(w-1)} \left\{ \begin{array}{ll} 1 & \text{if } seq1_i \cap seq2_i \neq 0 \text{ and non-degenerate,} \\ \frac{1}{2} & \text{if } seq1_i \cap seq2_i \neq 0 \text{ and 2-degenerate,} \\ \frac{1}{3} & \text{if } seq1_i \cap seq2_i \neq 0 \text{ and 3-degenerate,} \\ 0 & \text{if } seq1_i \cap seq2_i \neq 0 \text{ and } \geq 4\text{-degenerate,} \\ -1 & \text{if } seq1_i \cap seq2_i = 0 \end{array} \right.$$

Seq 1 GCCTCT  
Seq 2 GCCT --

1111-1-1

$$S_{obs} = 2/w (6) = 1/3$$

Seq 3 GCTTCT  
Seq 4 G - - - T

1-1-1-1-11

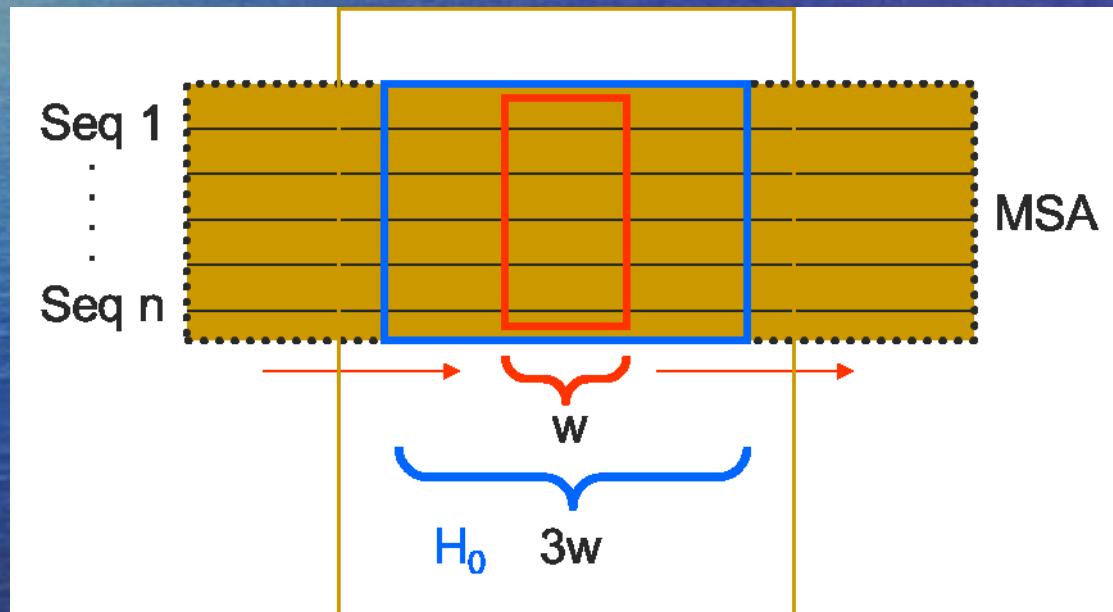
$$S_{obs} = -2/w (6) = -1/3$$

# Monte Carlo resampling

Null hypothesis:

$S_{\text{obs}}$  not different to scores of random similarity  $S_{\text{MC}}$

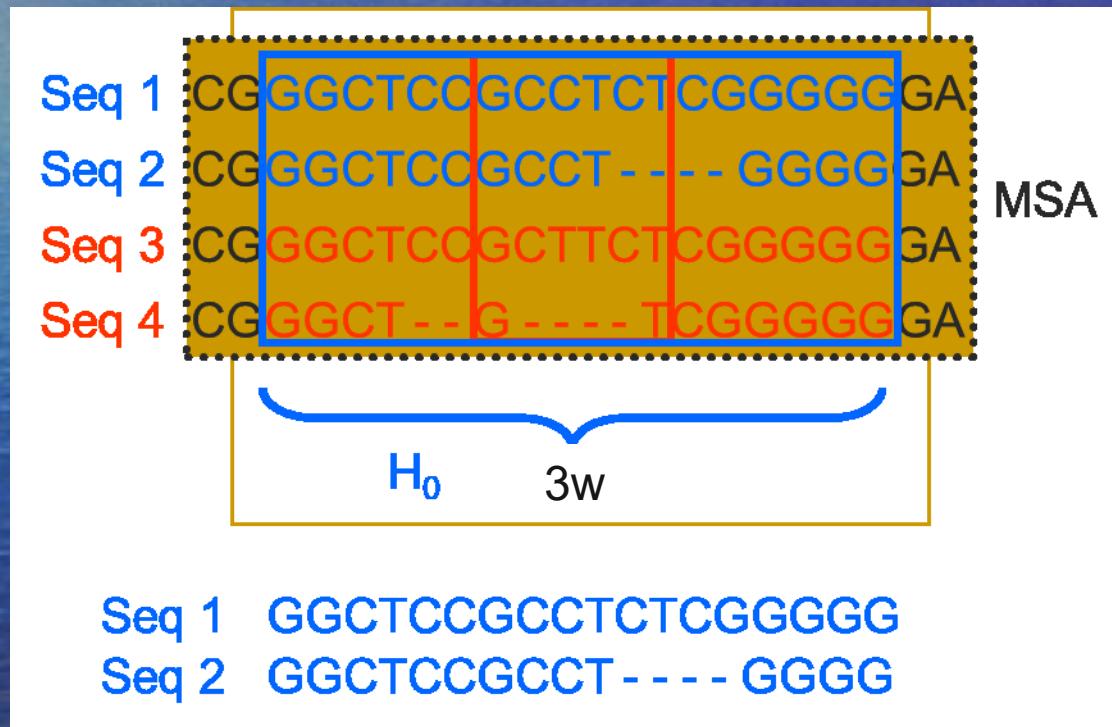
Double sized window by  $w$  in both directions



# Monte Carlo resampling

Null hypothesis:

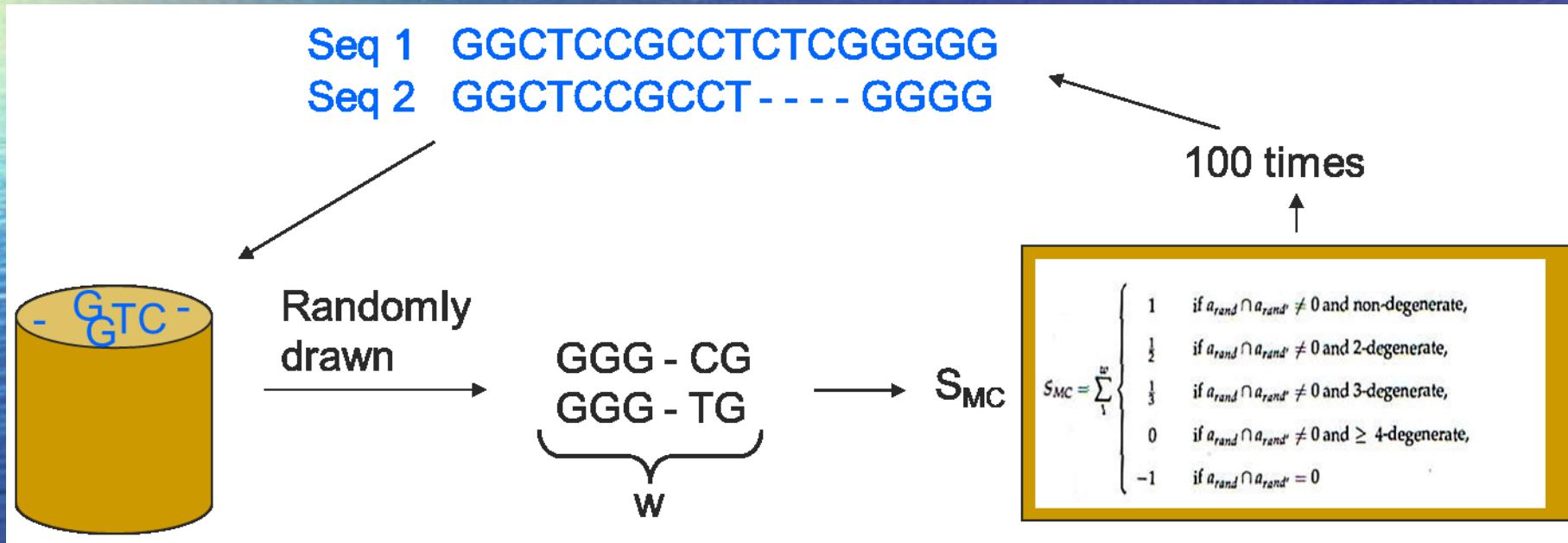
$S_{\text{obs}}$  not different to scores of random similarity  $S_{\text{MC}}$



# Monte Carlo resampling

Null hypothesis:

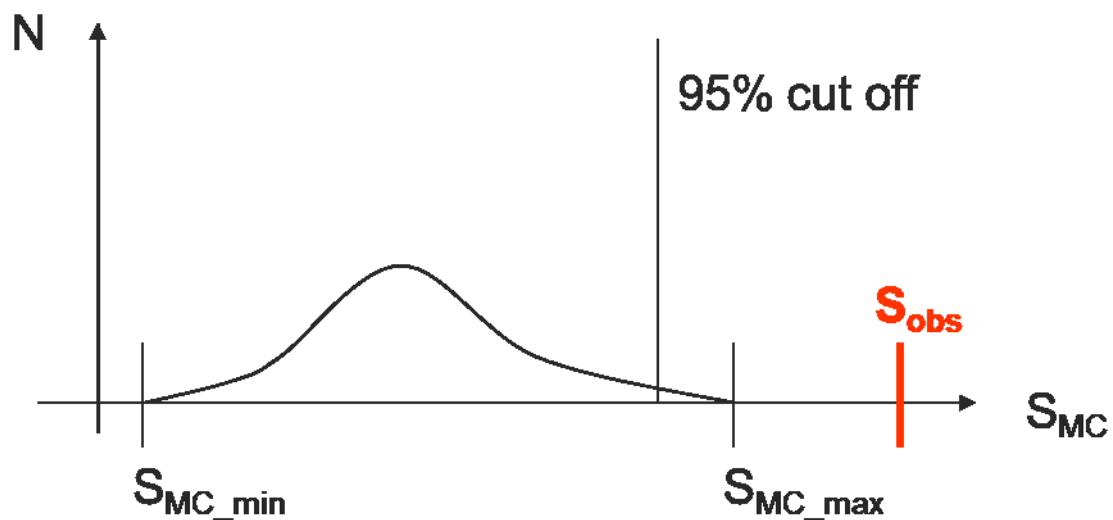
$S_{\text{obs}}$  not different to scores of random similarity  $S_{\text{MC}}$



# *Monte Carlo resampling*

Null hypothesis:

$S_{\text{obs}}$  not different to scores of random similarity  $S_{\text{MC}}$

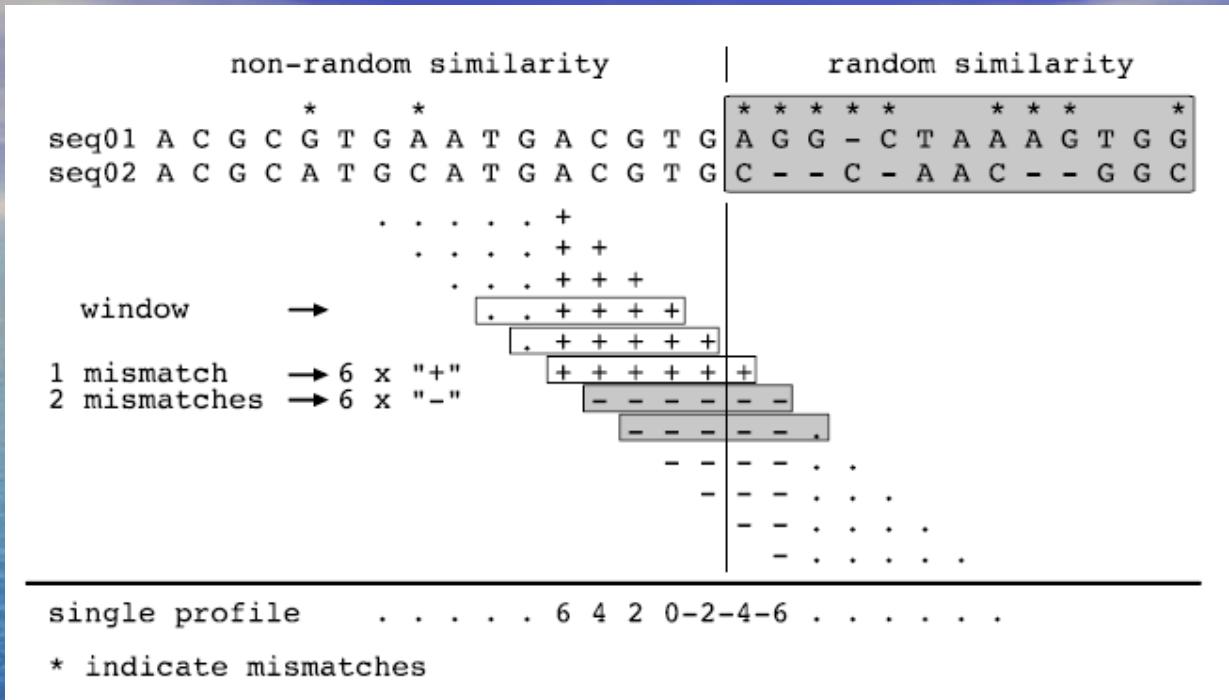


Within the window:

If non-random,  
every position gets +1.

If random,  
every position gets -1.

# *Similarity score for position*



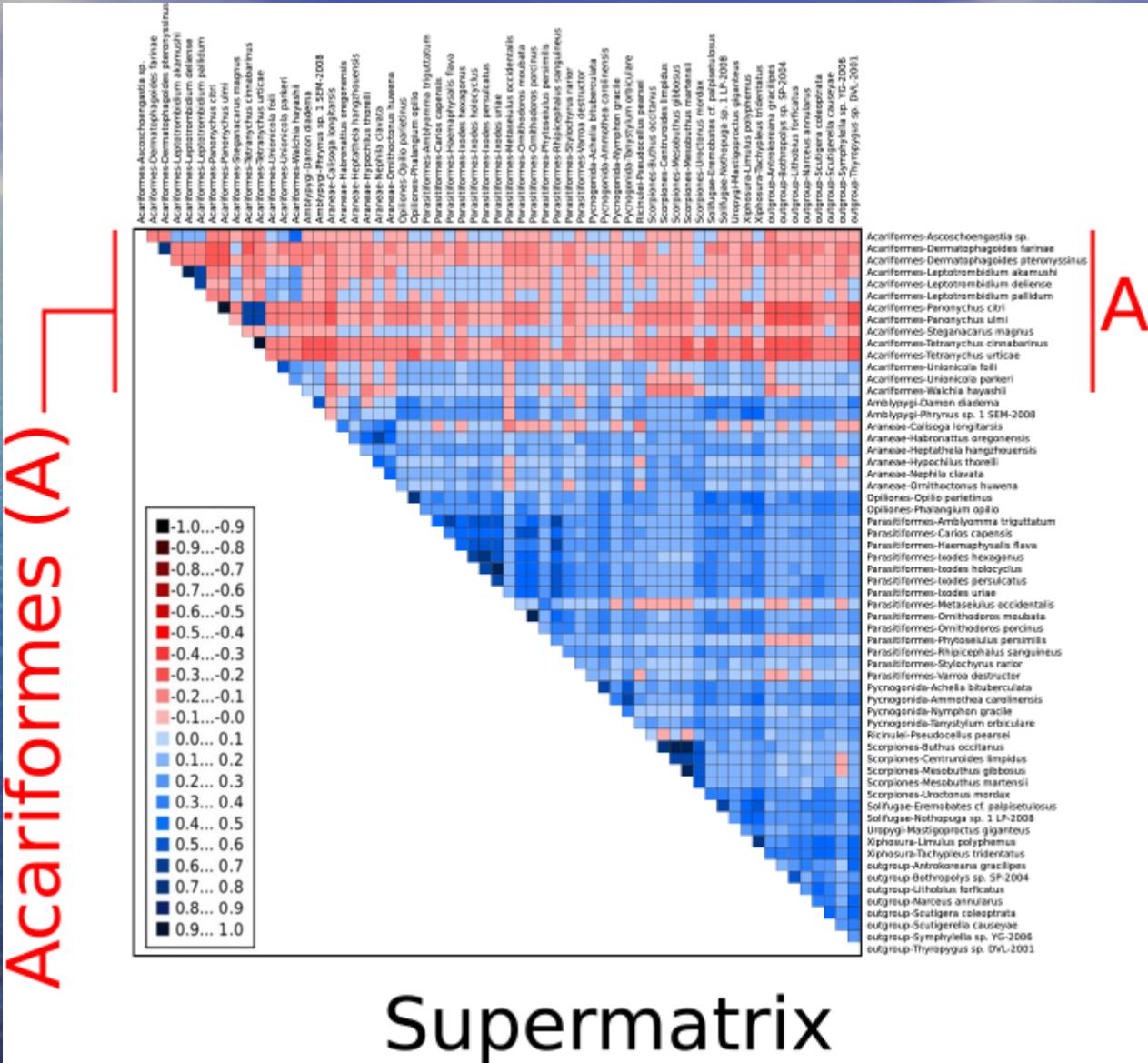
For each position signs are summed up and normalized by the sliding window size.

Arithmetic mean of the position value for each pair of sequences.

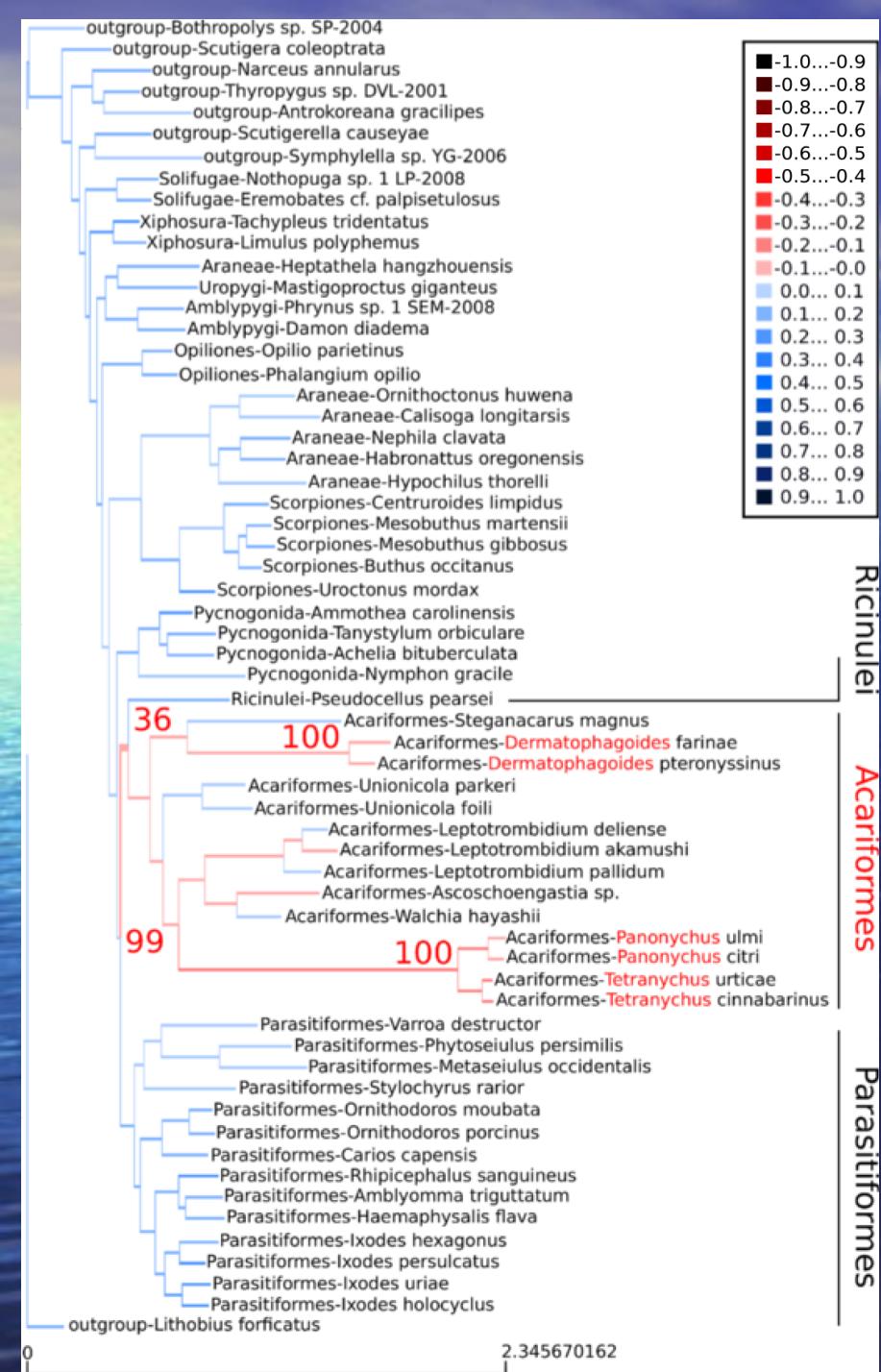
Compile values in a similarity matrix.

# Similarity matrix

Acariformes (A)

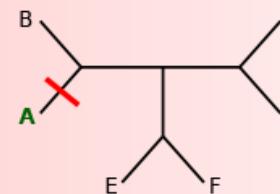


# Visualization on a given tree



## Terminal Branches

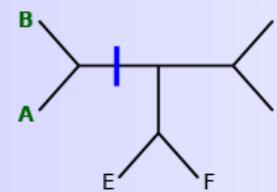
$$S1 (A|BCDEF) = \sum A:B \ A:C \ A:D \ A:E \ A:F / 5$$



A:B	A:C	A:D	A:E	A:F
-----	-----	-----	-----	-----

## Internal Branches

$$S3 (AB|CDEF) = \sum A:C \ A:D \ A:E \ A:F \ B:C \ B:D \ B:E \ B:F / 8$$



A:C	A:D	A:E	A:F
B:C	B:D	B:E	B:F

# *Branch length heterogeneity*

Traditionally: Tip-to-root distances

New measurement: LB (long branch) score

- root independent
- can already be used for uncorrected distances

$$LB_i = \frac{\overline{PD}_i}{\overline{PD}_a} - 1 * 100$$

PD = patristic distance

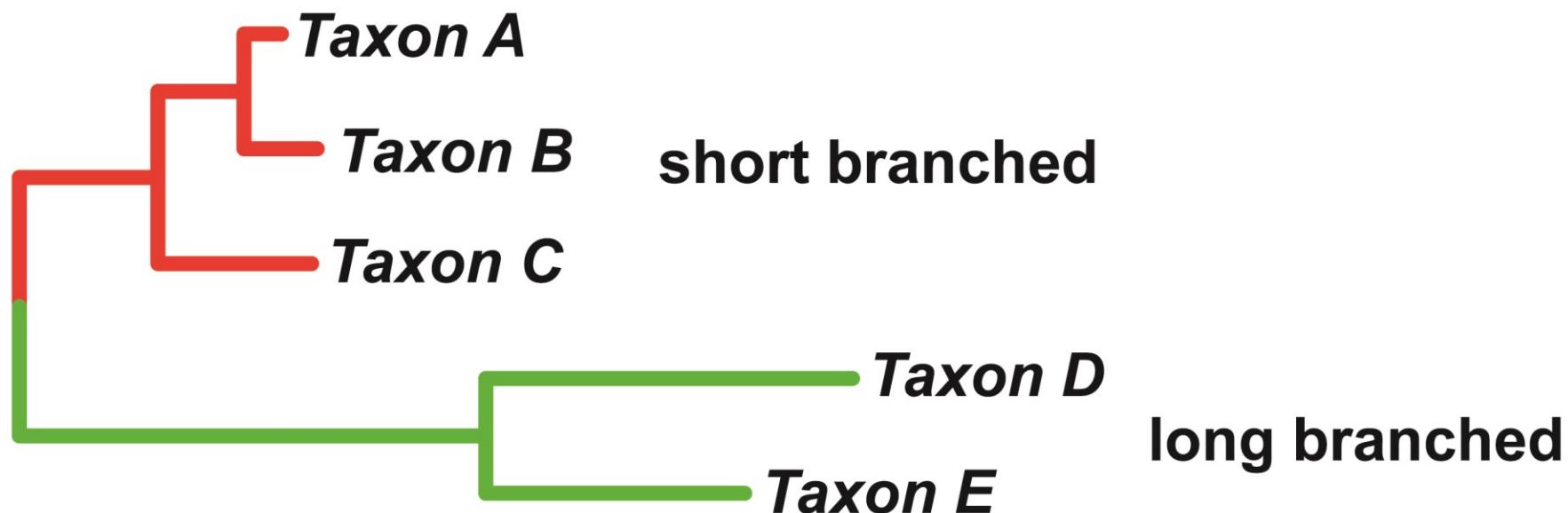
Struck (2014); Struck et al. (2014)

# *Branch length heterogeneity*

Traditionally: Tip-to-root distances

New measurement: LB (long branch) score

- root independent
- can already be used for uncorrected distances

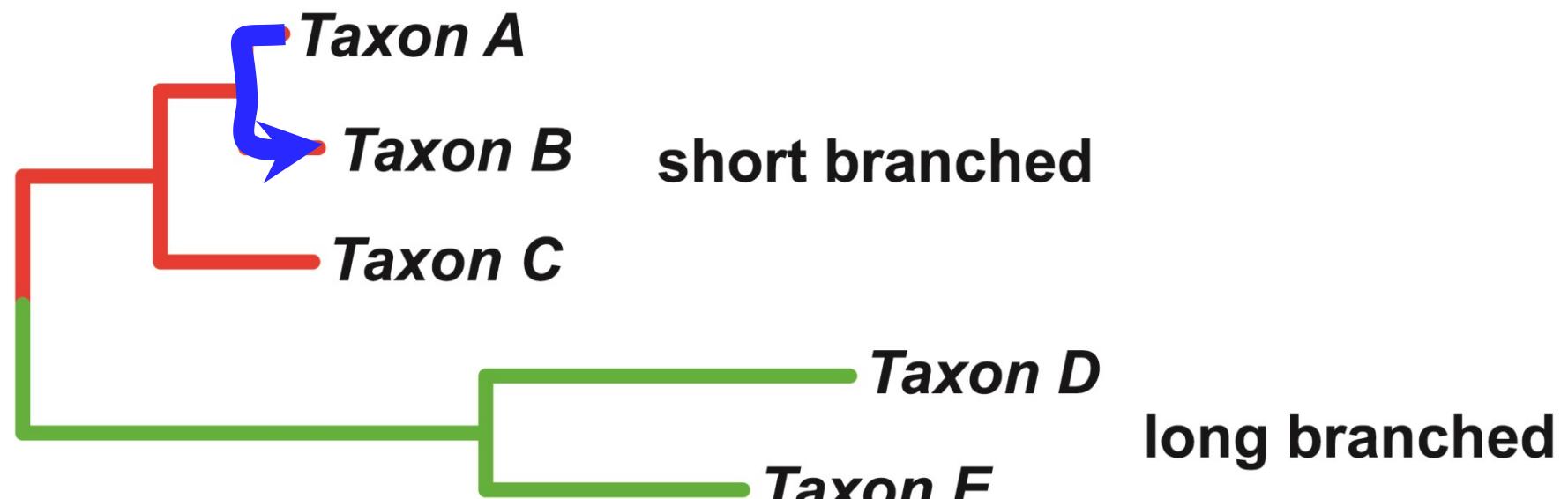


# *Branch length heterogeneity*

Traditionally: Tip-to-root distances

New measurement: LB (long branch) score

- root independent
- can already be used for uncorrected distances

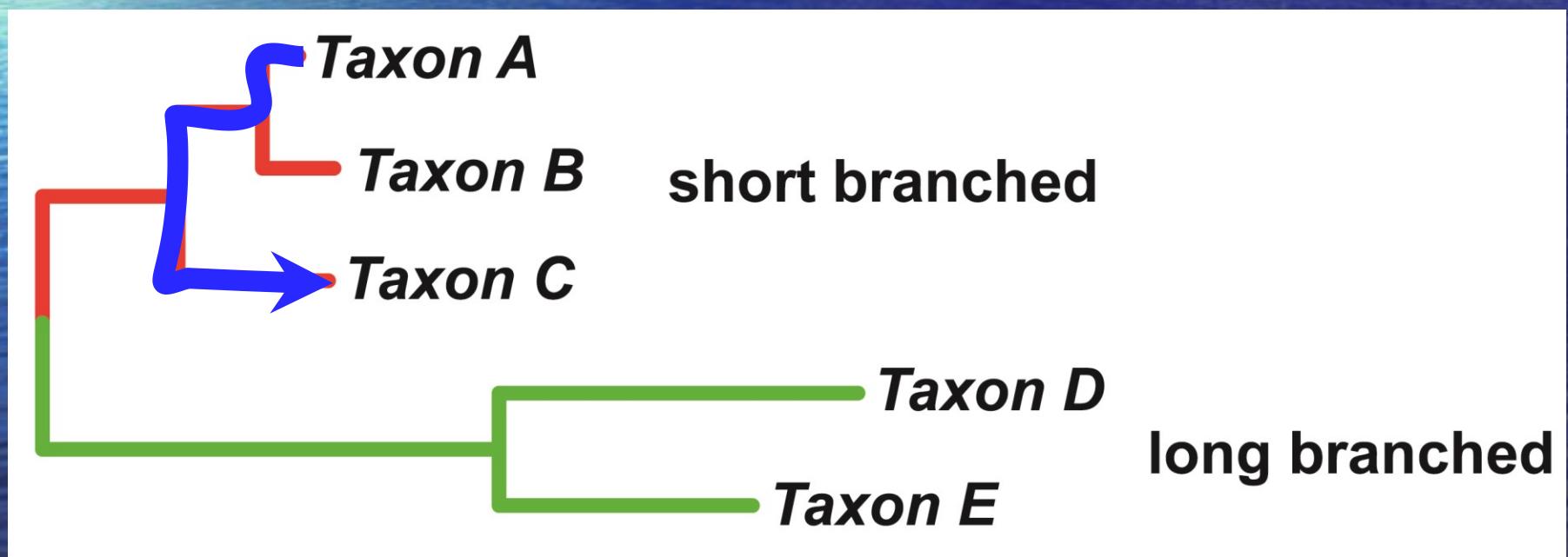


# *Branch length heterogeneity*

Traditionally: Tip-to-root distances

New measurement: LB (long branch) score

- root independent
- can already be used for uncorrected distances

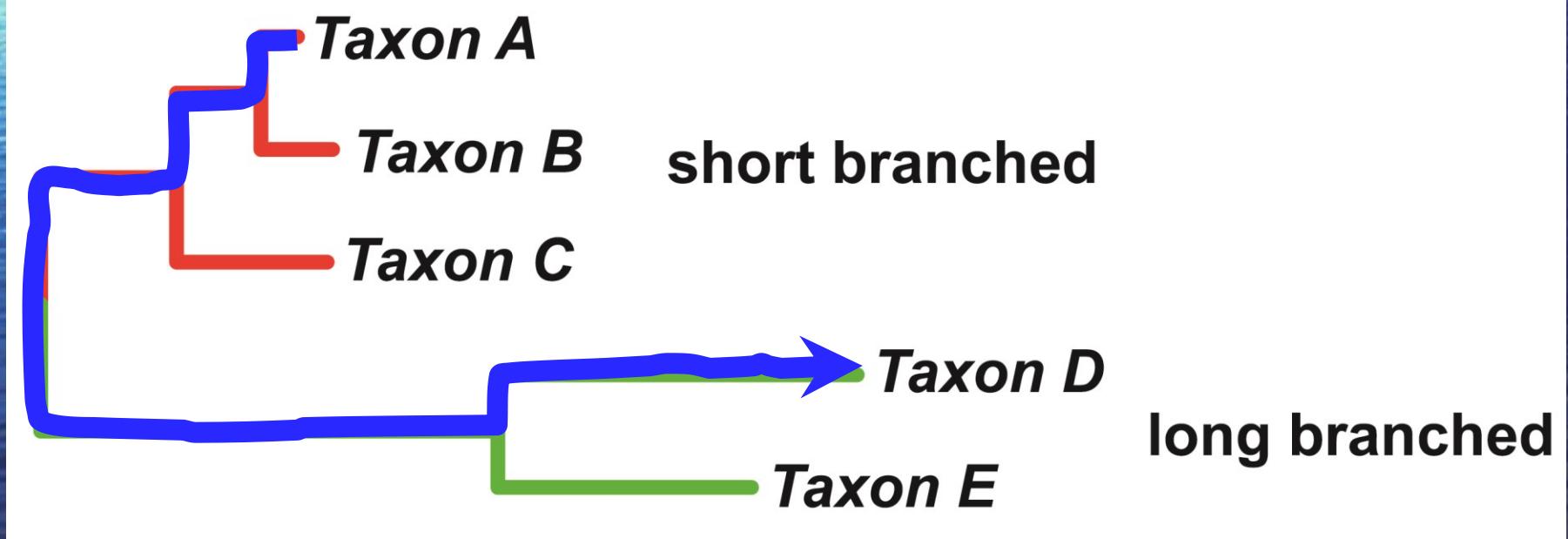


# *Branch length heterogeneity*

Traditionally: Tip-to-root distances

New measurement: LB (long branch) score

- root independent
- can already be used for uncorrected distances

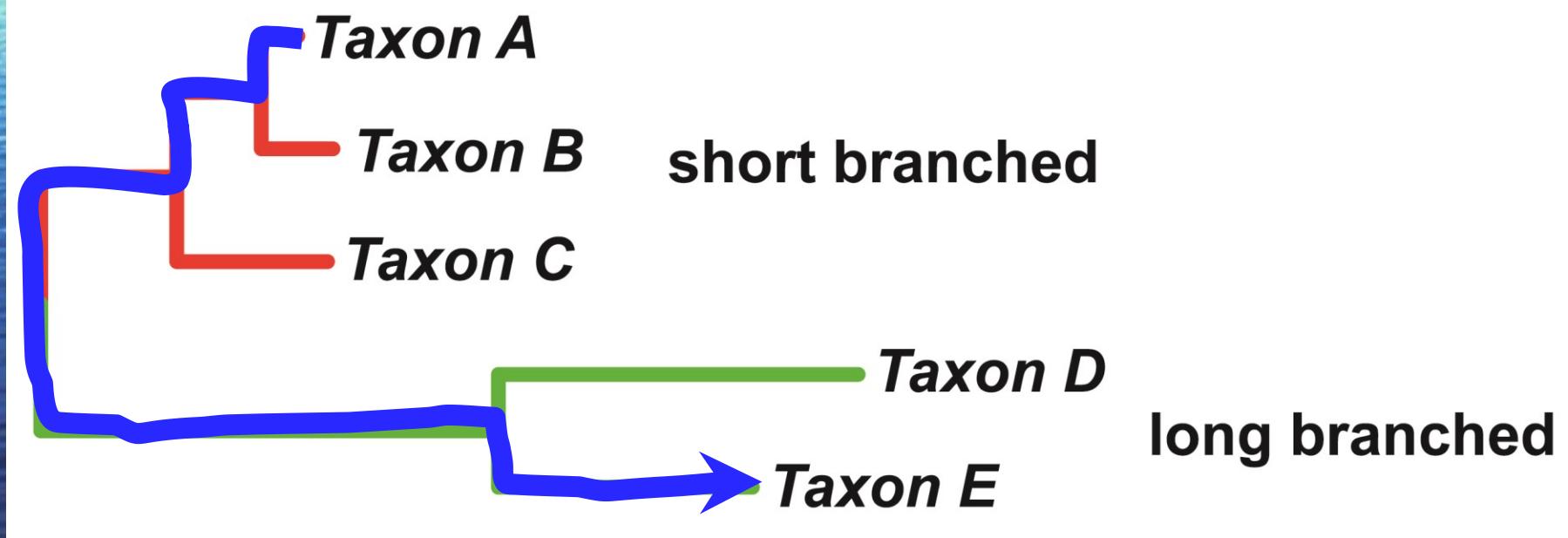


# *Branch length heterogeneity*

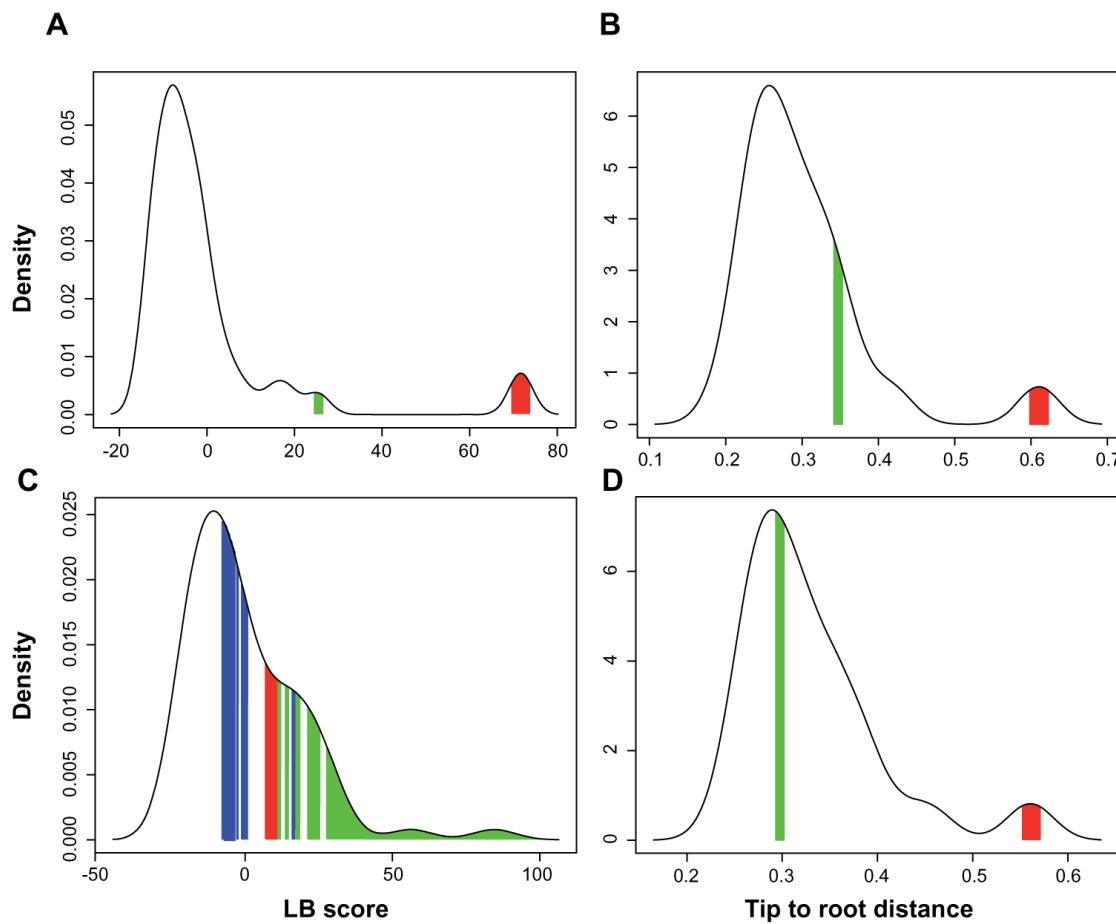
Traditionally: Tip-to-root distances

New measurement: LB (long branch) score

- root independent
- can already be used for uncorrected distances



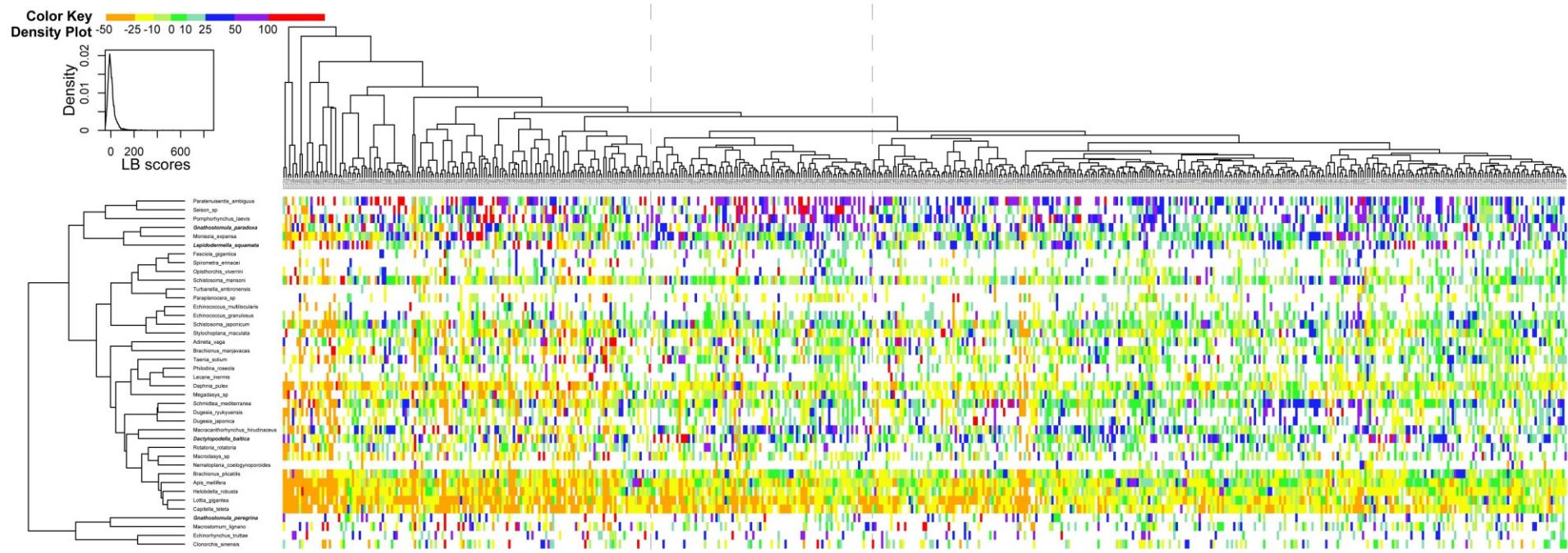
# *Importance of root*



**Figure 6.** Density plots generated with *R* of taxon-specific LB scores (A and C) or tip-to-root distances (B and D) of the trees shown in Figure 9 of Struck<sup>11</sup> (A and B) or Figure 3 of Ryan et al.<sup>81</sup> (C). Tip-to-root distances in (D) are based on a tree rerooted with the ectoproct *Bugula* instead of the brachiopod *Terebratalia* and the nemertean *Cerebratulus* as in Struck<sup>11</sup> and (B).

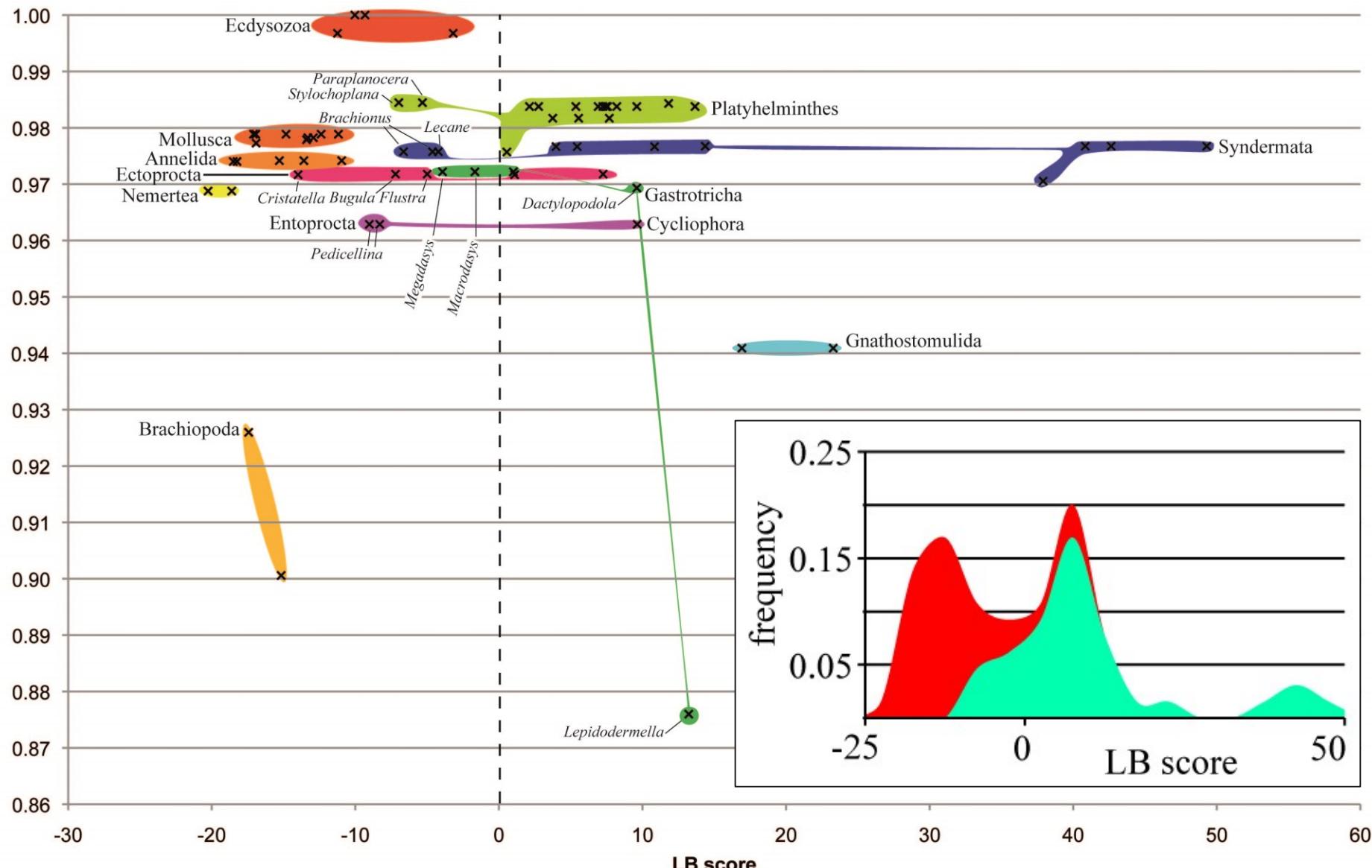
**Notes:** Red = values of either Myzostomidae (A, B, and D) or Ctenophora (C). Green = values of either the outgroup taxon *Bugula* (Ectoprocta) (A, B, and D) or all outgroup taxa (C). Blue = values of Porifera (C).

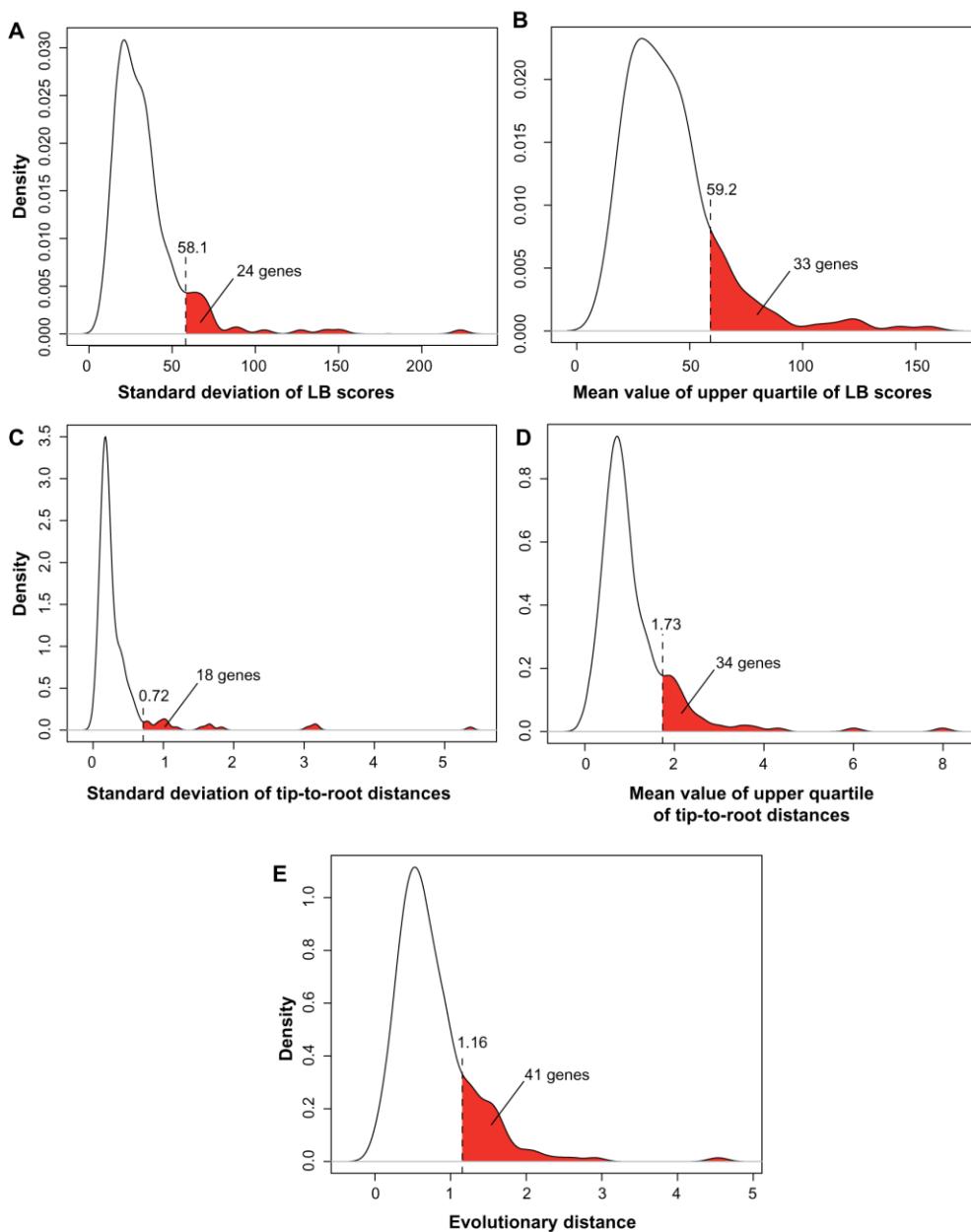
# *Heatmap of heterogeneity*



Struck et al. (2014)

# Other displays of heterogeneity





**Figure 7.** Density plots generated with *R* of different gene-specific long-branch indices for the 229 genes present in the data set of Struck.<sup>11</sup> (A) Standard deviation of LB scores measuring heterogeneity; (B) average of the upper quartile of LB scores representing the taxa with the longest branches; (C) standard deviation of tip-to-root distances measuring heterogeneity; (D) average of the upper quartile of tip-to-root distances representing the taxa with the longest branches; and (E) average PD as a proxy for genes affected by long-branch attraction. Red areas indicate deviations from the normal distribution.

# TreSpEx



Libertas Academica  
FREEDOM TO RESEARCH

Open Access: Full open access to  
this and thousands of other papers at  
<http://www.la-press.com>.

## Evolutionary Bioinformatics

### TreSpEx—Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information

Torsten H. Struck

Zoological Research Museum Alexander Koenig, Bonn, Germany.

#### 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

# *TreSpEx – LB scores and such*

eg, RAxML

Phylogenetic reconstruction  
of each dataset of interest



TreSpEx

**Calculation of long-branch indices**  
for each tree and taxon:  
a) LB score  
b) tip-to-root distance  
c) evolutionary distance  
(not for taxa)



eg, R

Further statistical analyses

## Command-line options

- fun <e> | calls the desired function like RAxML  
(e = calculation of long branch indices)
- ipt <file name> | file containing the list of names of tree files in Newick format with branch length
- tf <file name> | file containing the list of taxon names, which occur across all trees analysed
- path <path> | specifies the path to the files (optional)

```
perl TreSpEx.v1.pl -fun e –ipt Name-of-file –tf Name-of-file –path Path-from-root
```

# *TreSpEx – LB scores and such*

eg, RAxML

Phylogenetic reconstruction  
of each dataset of interest



**TreSpEx**

**Calculation of long-branch indices**  
for each tree and taxon:  
a) LB score  
b) tip-to-root distance  
c) evolutionary distance  
(not for taxa)



eg, R

Further statistical analyses

## Example files

ipt:

RAxML\_bipartitions.21904Ali.phy\_fastBoot\_Out.txt  
RAxML\_bipartitions.21912Ali.phy\_fastBoot\_Out.txt  
RAxML\_bipartitions.21942Ali.phy\_fastBoot\_Out.txt  
RAxML\_bipartitions.21952Ali.phy\_fastBoot\_Out.txt  
RAxML\_bipartitions.21981Ali.phy\_fastBoot\_Out.txt  
RAxML\_bipartitions.22055Ali.phy\_fastBoot\_Out.txt  
RAxML\_bipartitions.22068Ali.phy\_fastBoot\_Out.txt  
RAxML\_bipartitions.22075Ali.phy\_fastBoot\_Out.txt  
RAxML\_bipartitions.22083Ali.phy\_fastBoot\_Out.txt  
RAxML\_bipartitions.22089Ali.phy\_fastBoot\_Out.txt  
RAxML\_bipartitions.22090Ali.phy\_fastBoot\_Out.txt

tf:

Bugula\_neritina\_EST  
Arenicola\_marina  
Terebratalia\_transversa  
Eurythoe\_complanata\_EST  
Ophelia\_limacina  
Myzostoma\_seymourcollegiorum  
Lottia\_gigantea  
Tubifex\_tubifex  
Eisenia\_fetida  
Platynereis\_dumerilii  
Haementeria\_depressa  
Cirratulus\_sp

# *TreSpEx – LB scores and such*

## Output files:

- LB\_score\_calculation.log | log file
- LBscore\_Results/LB\_scores\_perTaxon.txt | LB scores of each taxon for each gene
- LBscore\_Results/TR\_scores\_perTaxon.txt | Tip\_to\_Root scores of each taxon for each gene
- LBscore\_Results/LB\_scores\_summary\_perPartition.txt | scores (LB\_score\_upper\_quartile, LB\_score\_Heterogeneity, Tip\_to\_Root\_upper\_quartile, Tip\_to\_Root\_Heterogeneity, Average\_PD) for each gene
- LBscore\_Results/PD\_matrices/ | folder containing the pairwise distance matrix for each gene