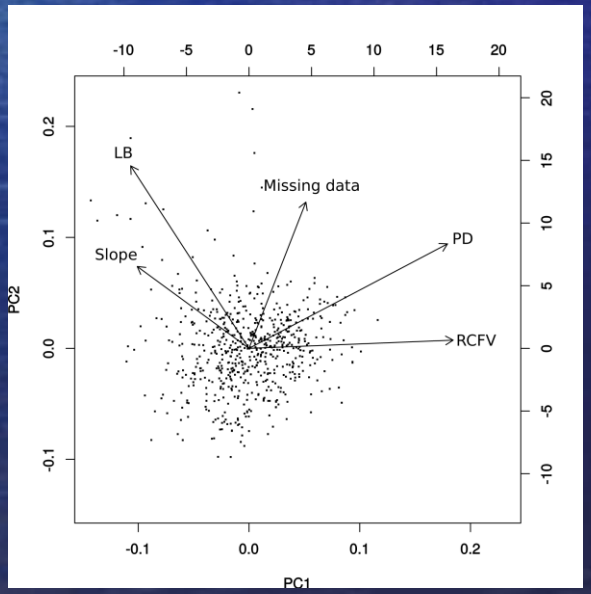
Branch length heterogeneity

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

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

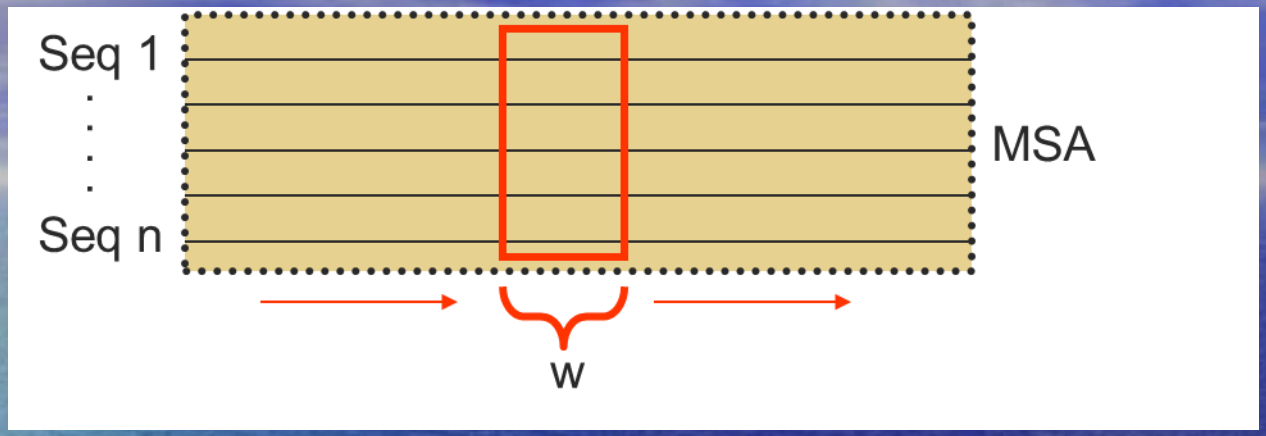
AliGROOVE – visualization of heterogeneous sequence divergence within multiple sequence alignments and detection of inflated branch support

Patrick Kück[1*], Sandra A Meid[1], Christian Groß[2], Johann W Wägele[1] and Bernhard Misof[1]

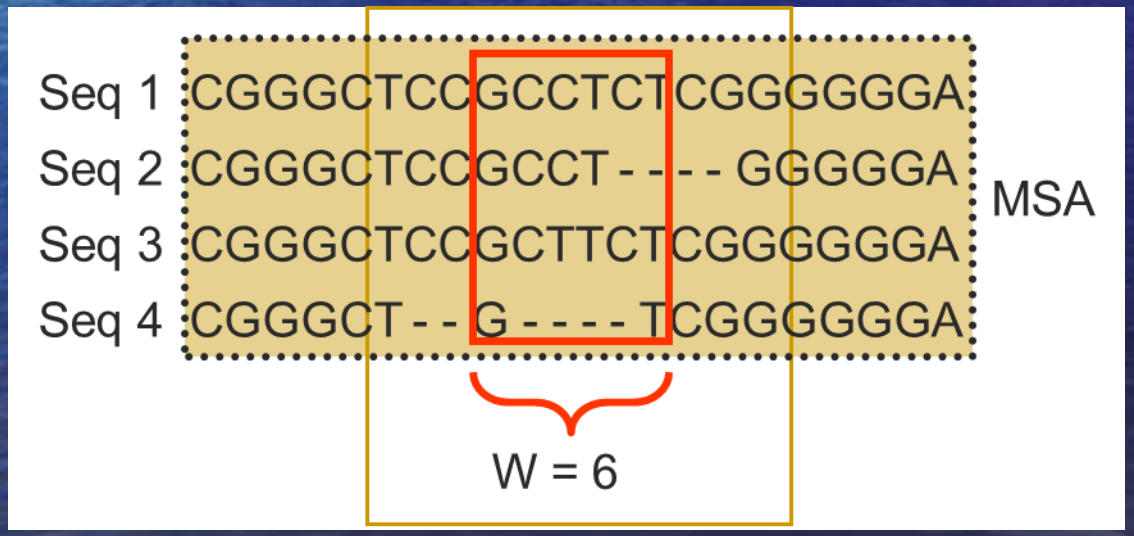
heterogeneous sequence divergence:

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

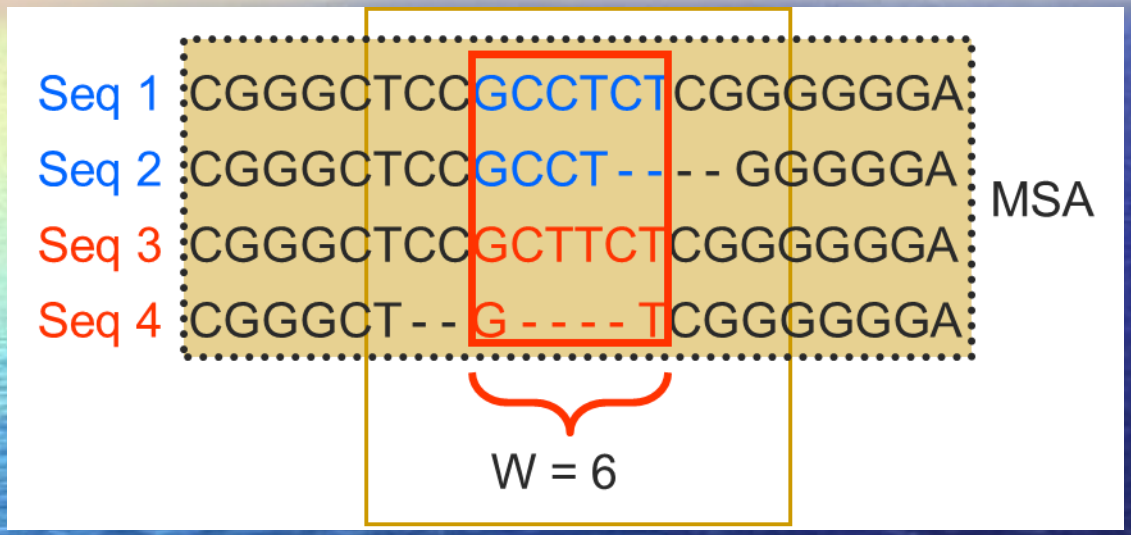
The procedure



Sliding window size set to 6 (default)



The procedure



All pairwise comparisons

High sequence numbers are time expensive !!

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

Sequence pairs (x,y) are selected

Seq 1 GCCTCT
Seq 2 GCCT - -

Seq 3 GCTTCT
Seq 4 G - - - - T

The similarity score

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

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

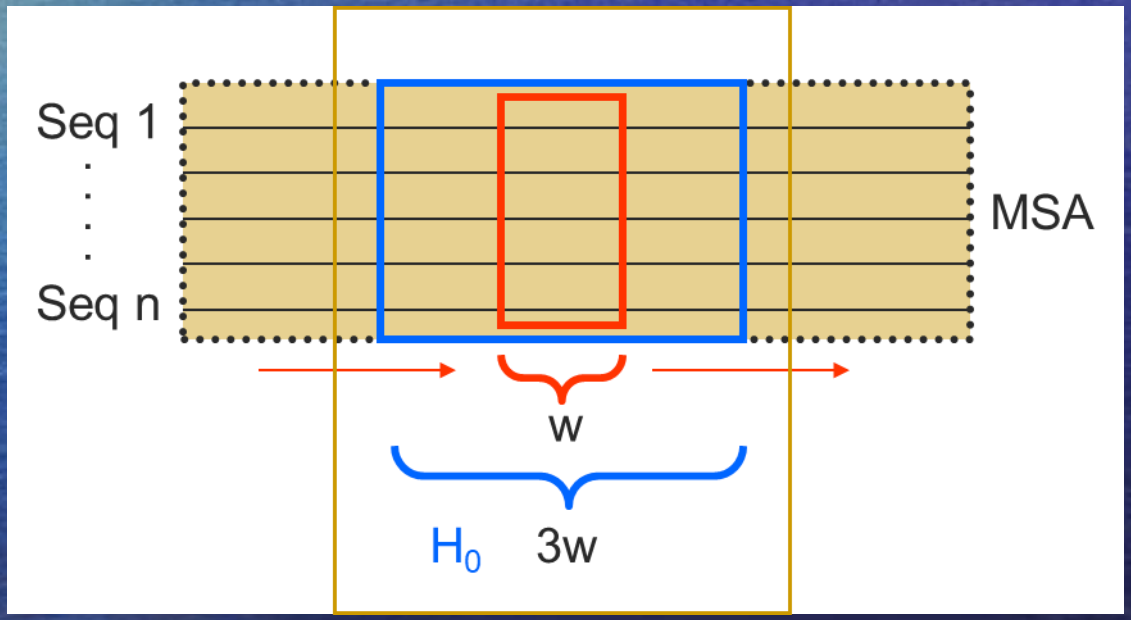
S_{obs} = -2/w (6) = -1/3

Monte Carlo resampling

Null hypothesis:

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

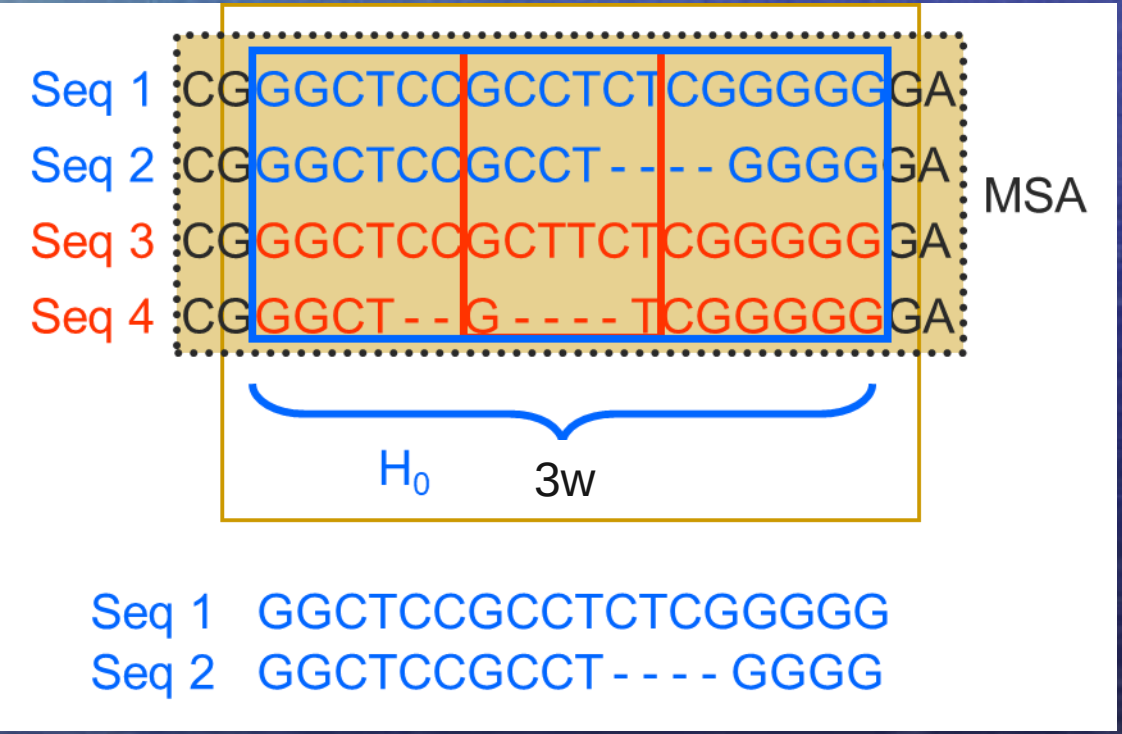
Double sized window by w in both directions



Monte Carlo resampling

Null hypothesis:

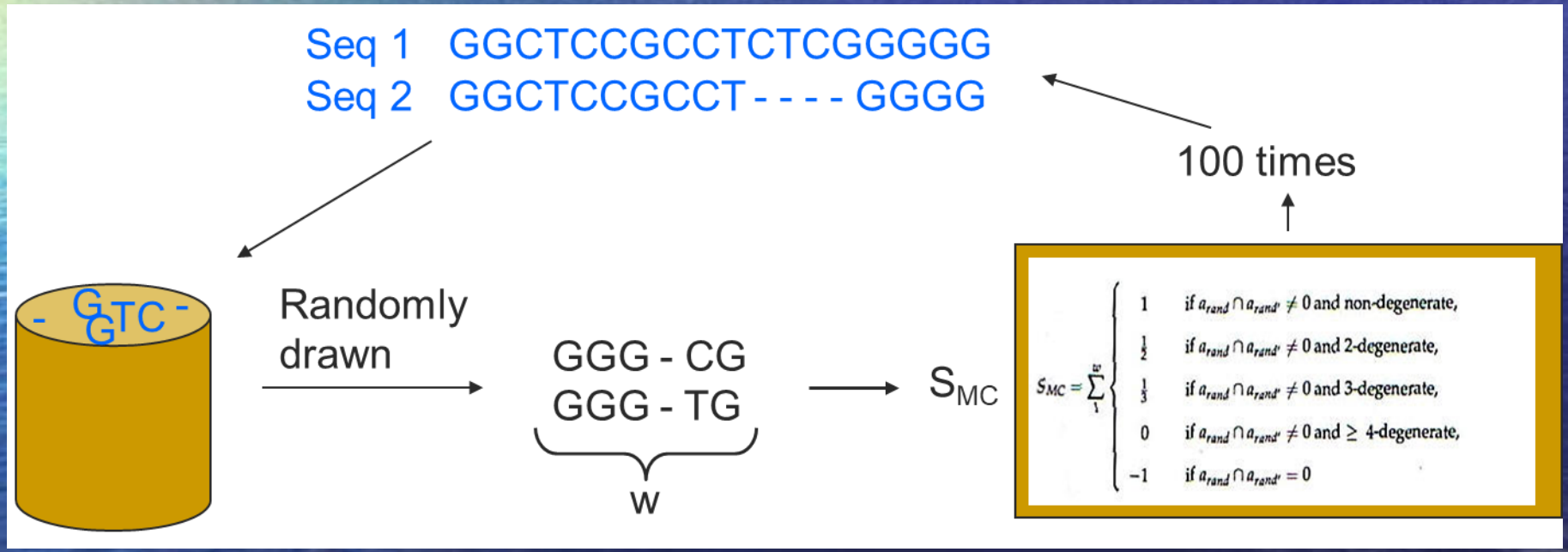
S_{obs} not different to scores of random similarity S_{MC}



Monte Carlo resampling

Null hypothesis:

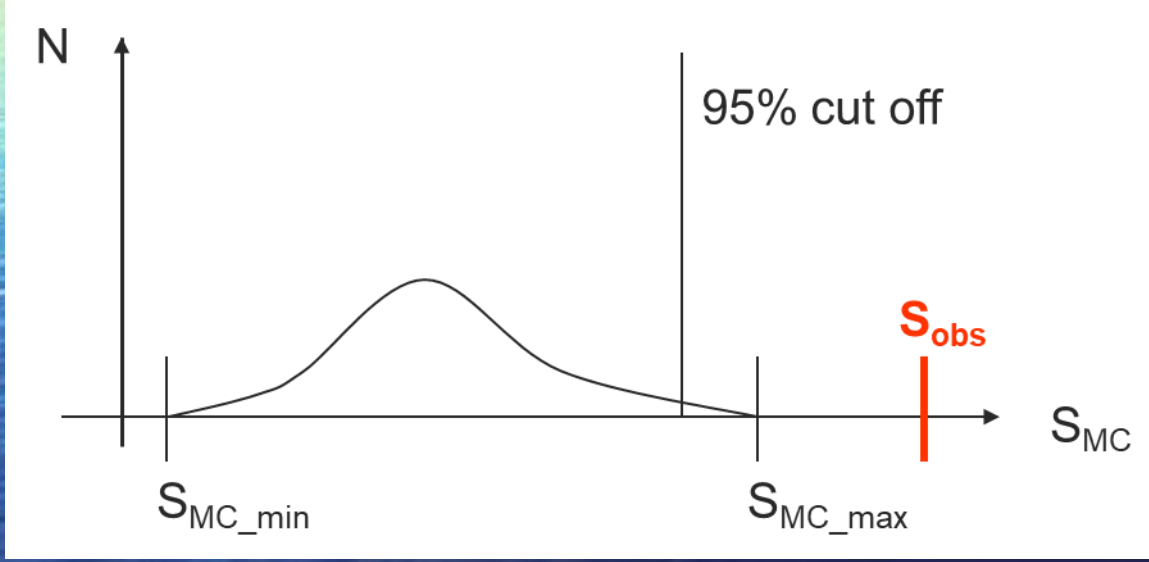
S_{obs} not different to scores of random similarity S_{MC}



Monte Carlo resampling

Null hypothesis:

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

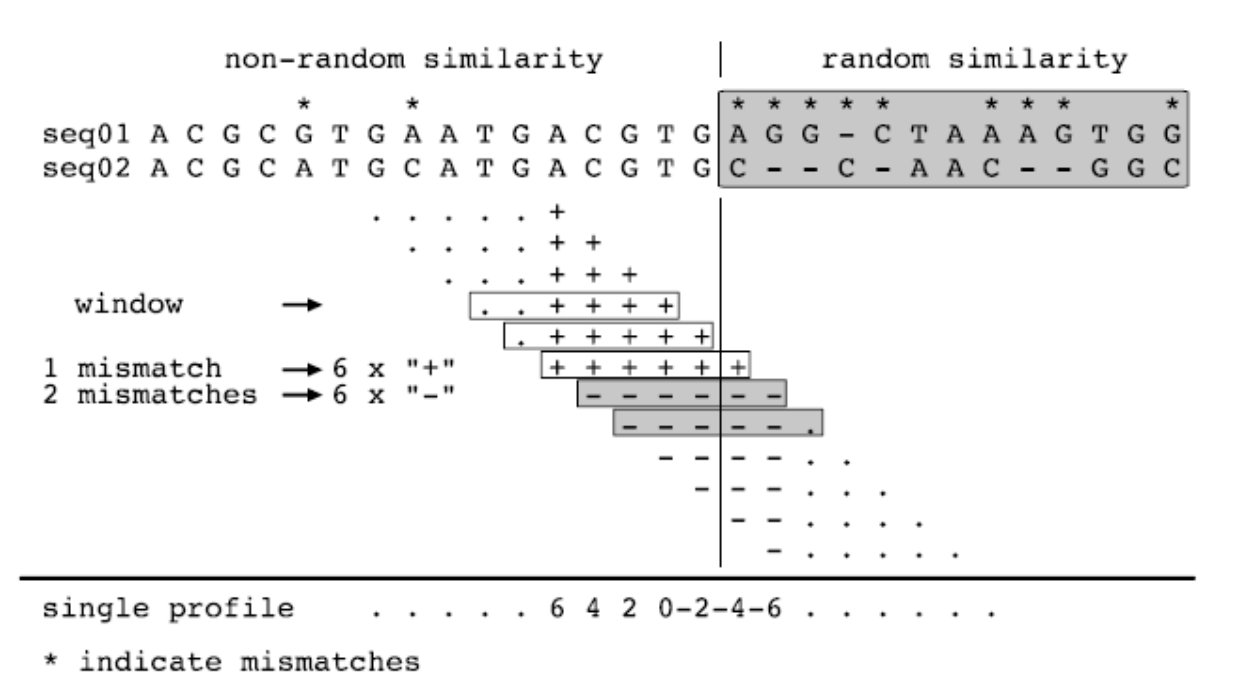


Within the window:

If non-random,
every position gets +1.

If random,
every position gets -1.

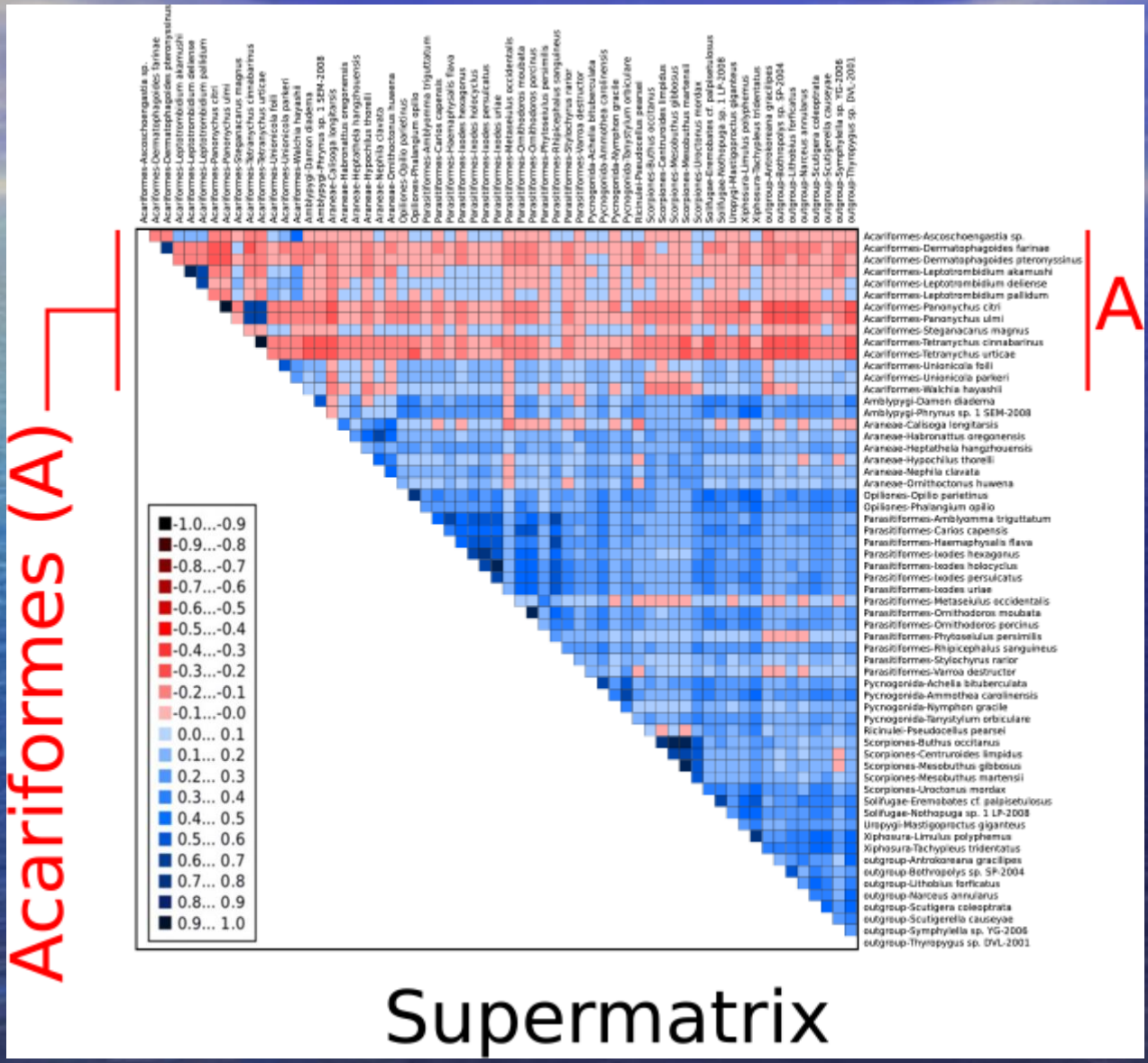
Similarity score for position



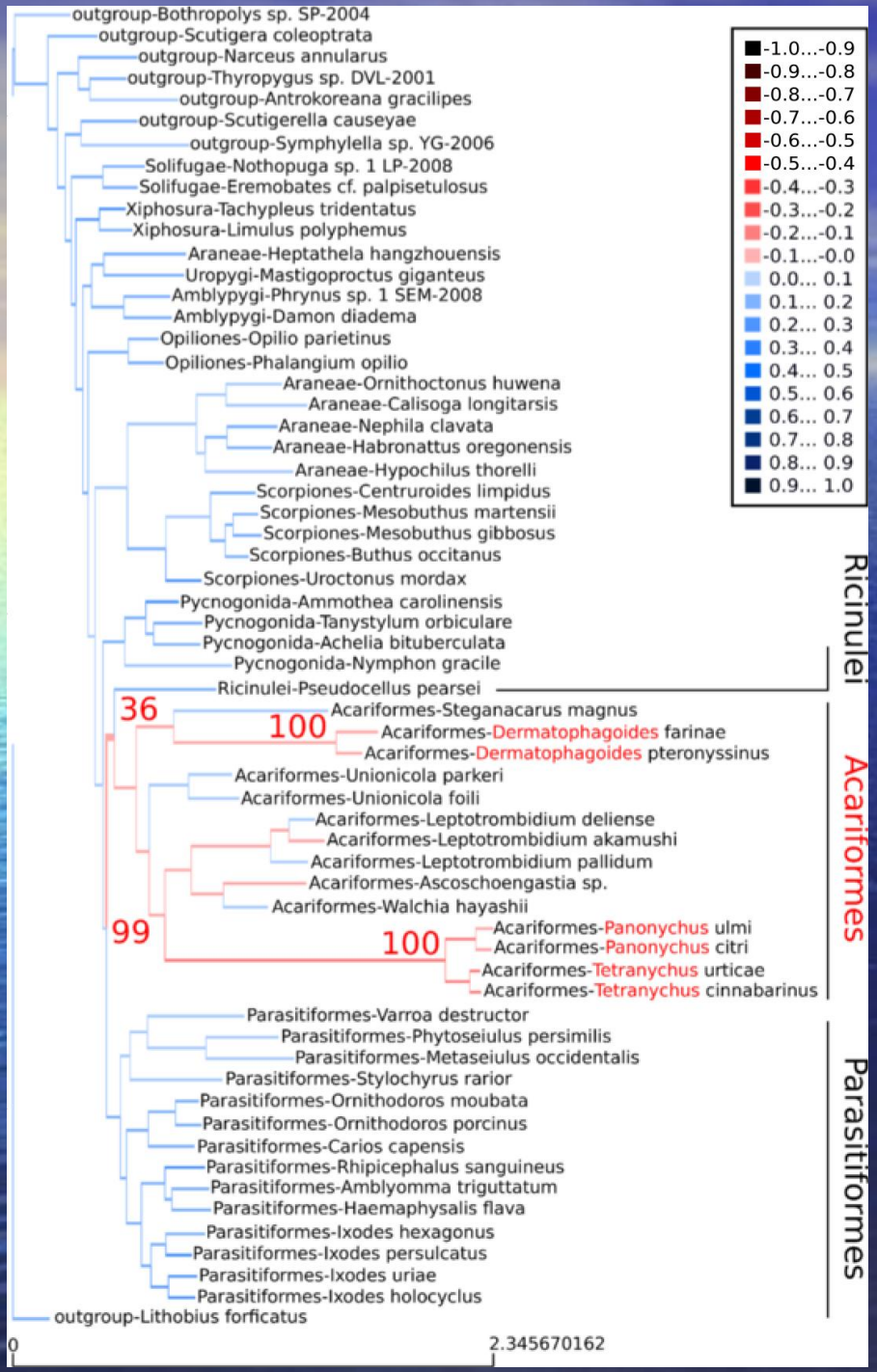
For each position signs are summed up and normalized by the sliding window size.

Arithmetic mean of the position value for each pair of sequences.
Compile values in a similarity matrix.

Similarity matrix

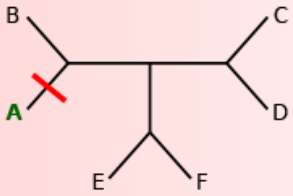


Visualization on a given tree



Terminal Branches

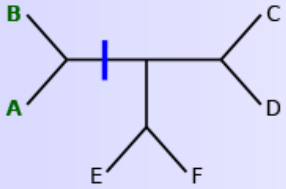
S1 (A|BCDEF) = ∑A:B A:C A:D A:E A:F / 5



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

Internal Branches

S3 (AB|CDEF) = ∑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

$$LB_i = \left(\frac{\overline{PD_i}}{PD_a} - 1 \right) * 100$$

PD = patristic distance

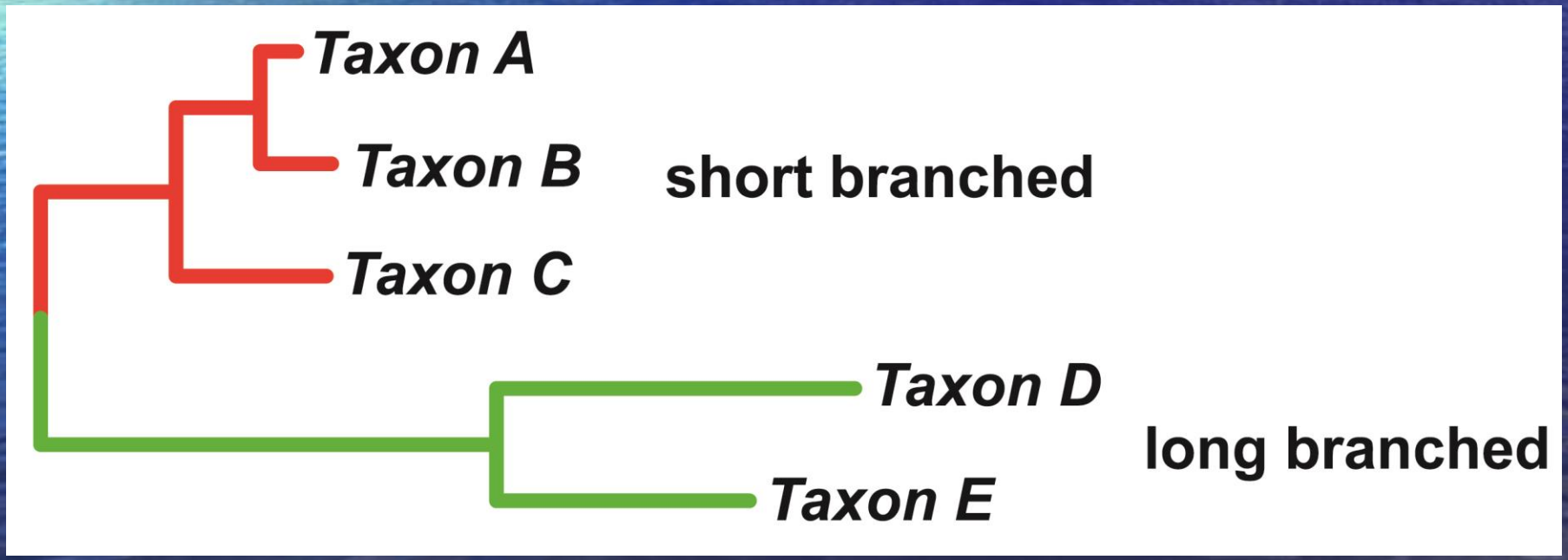
Struck (2014); Struck et al. (2014)

Branch length heterogeneity

Traditionally: Tip-to-root distances

New measurement: LB (long branch) score

- root independent
- can already be used for uncorrected distances

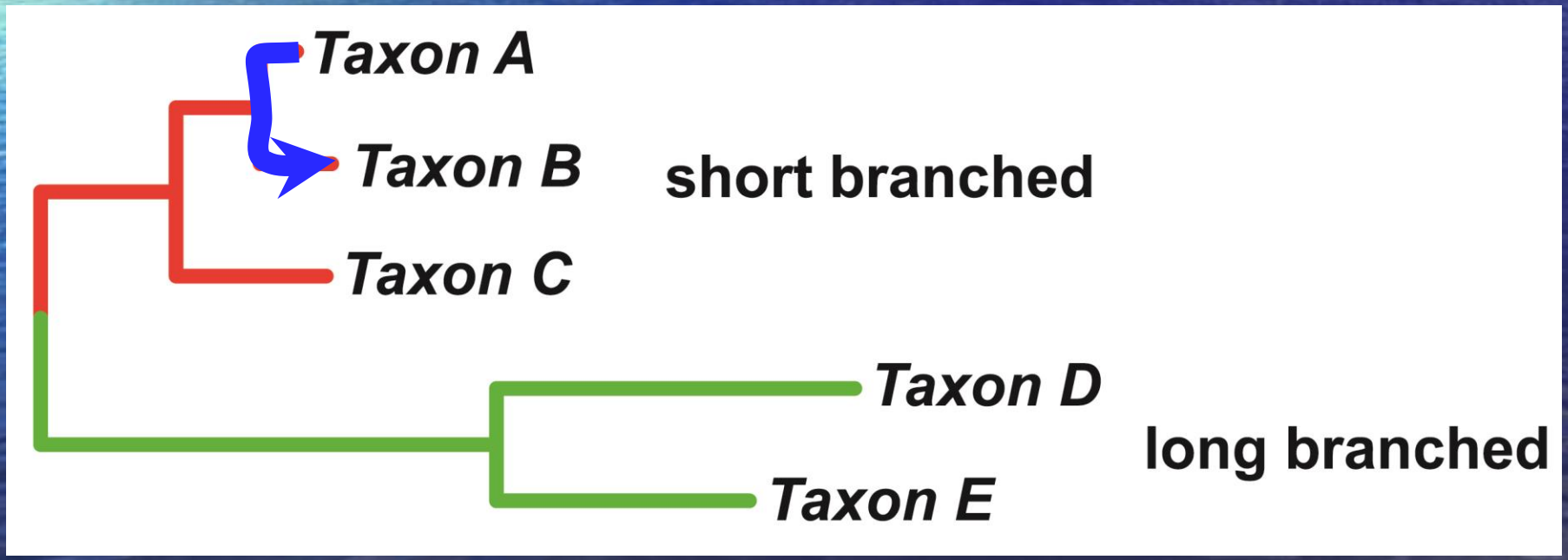


Branch length heterogeneity

Traditionally: Tip-to-root distances

New measurement: LB (long branch) score

- root independent
- can already be used for uncorrected distances

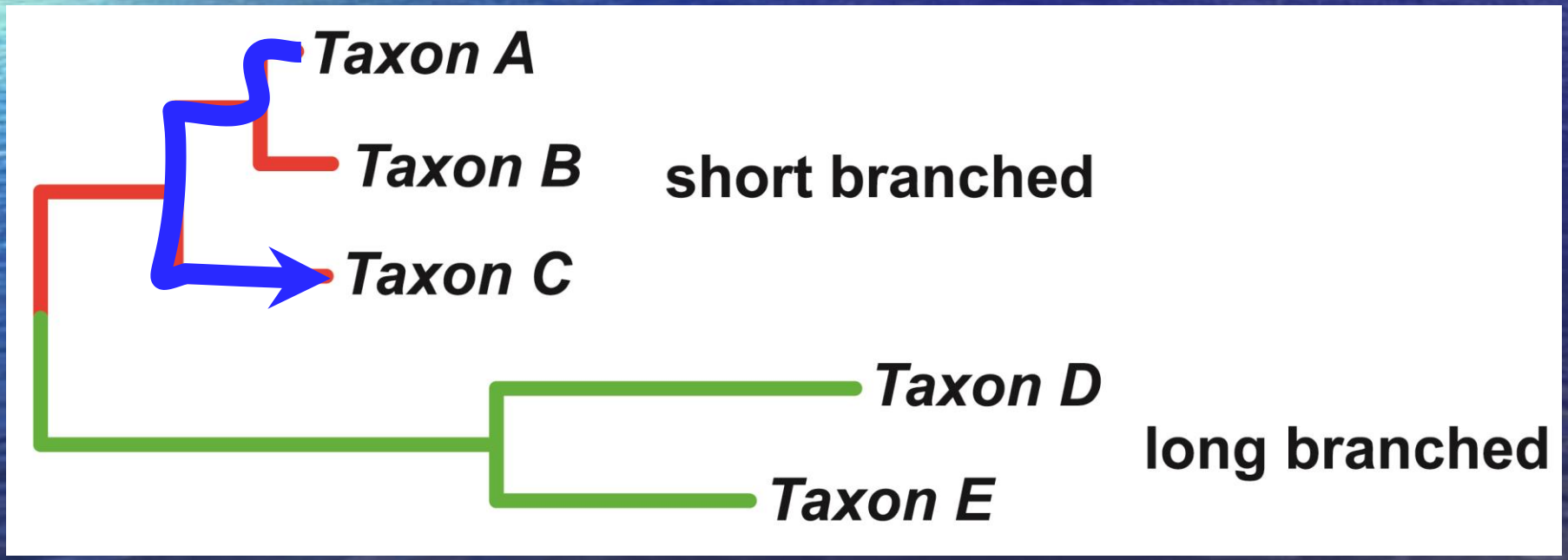


Branch length heterogeneity

Traditionally: Tip-to-root distances

New measurement: LB (long branch) score

- root independent
- can already be used for uncorrected distances

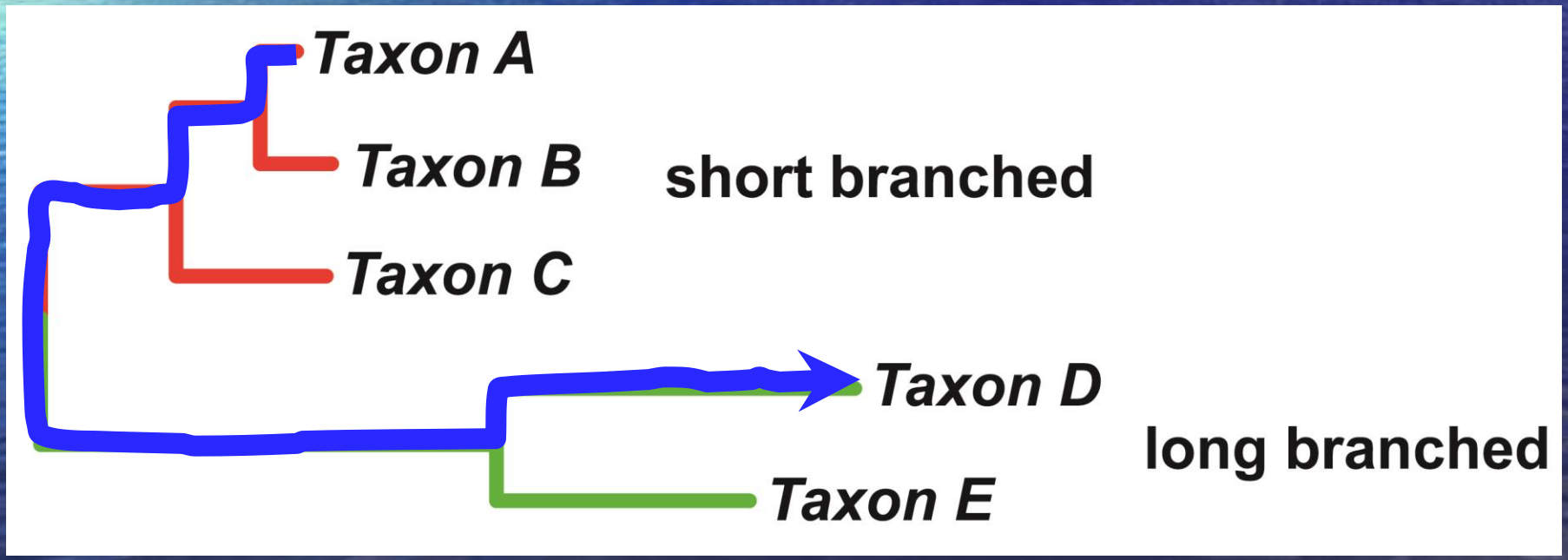


Branch length heterogeneity

Traditionally: Tip-to-root distances

New measurement: LB (long branch) score

- root independent
- can already be used for uncorrected distances

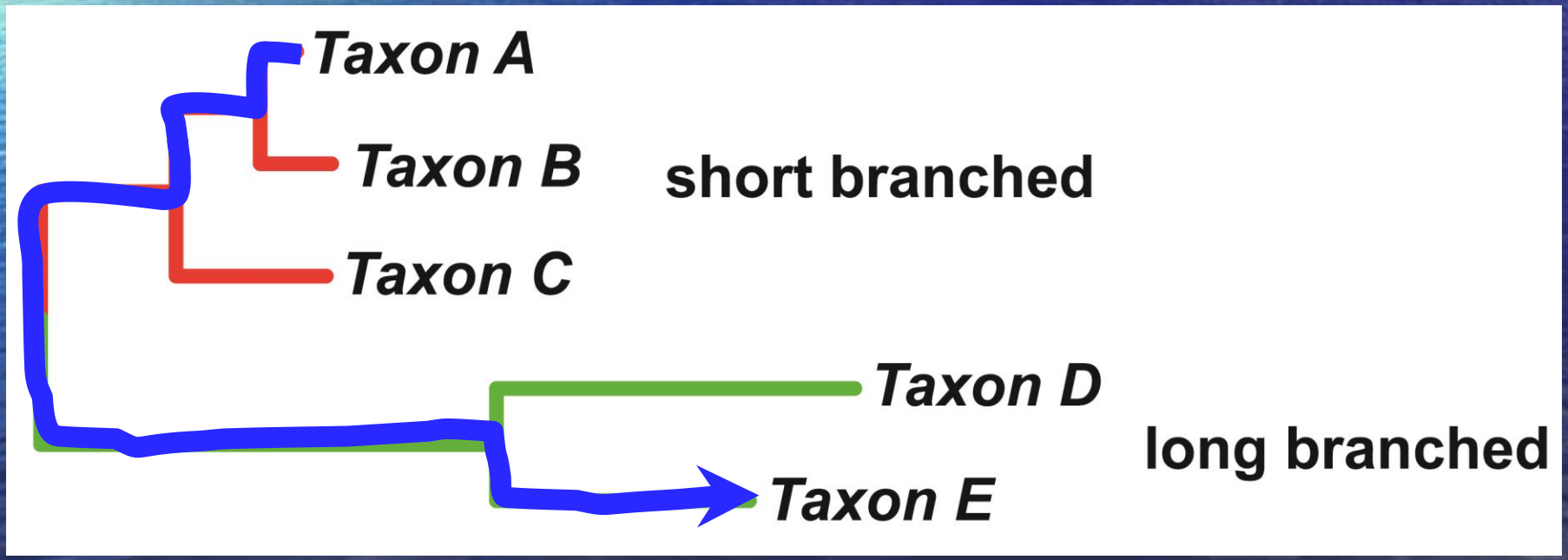


Branch length heterogeneity

Traditionally: Tip-to-root distances

New measurement: LB (long branch) score

- root independent
- can already be used for uncorrected distances



Importance of root

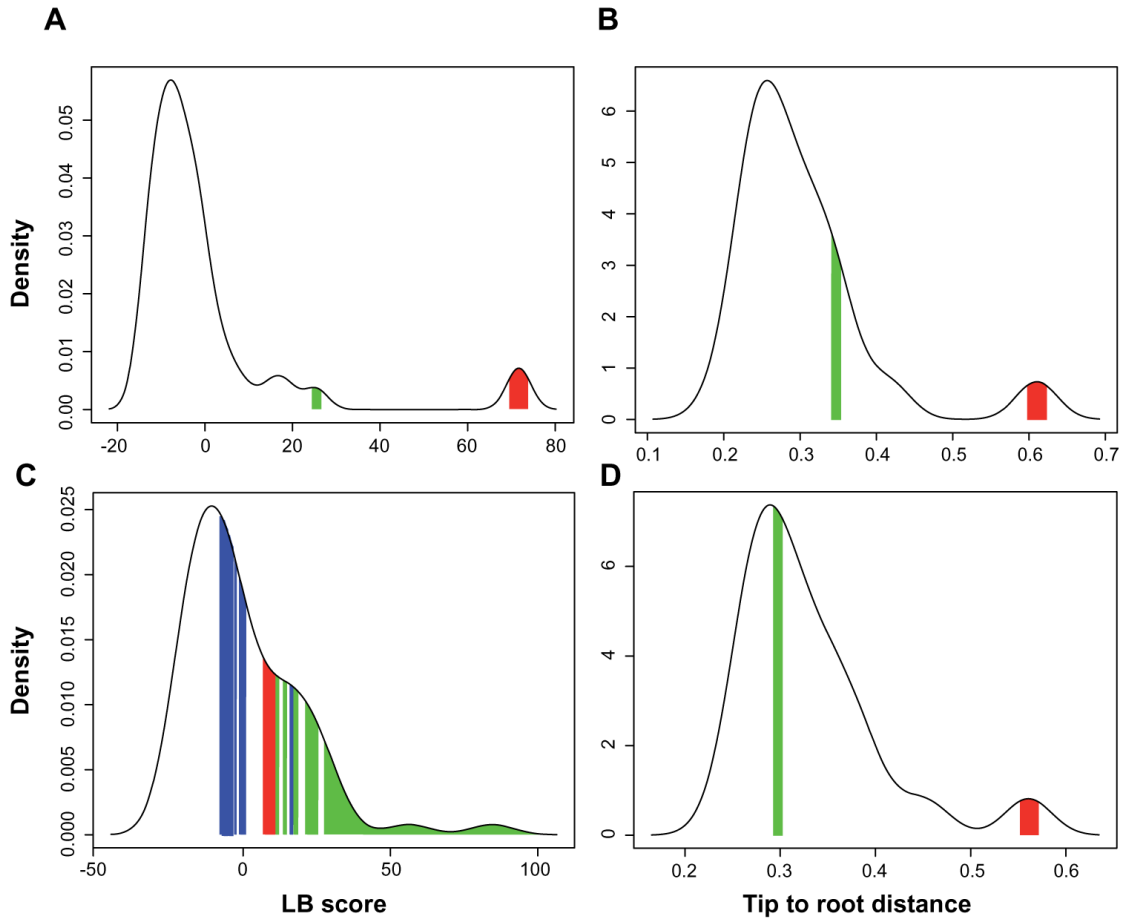
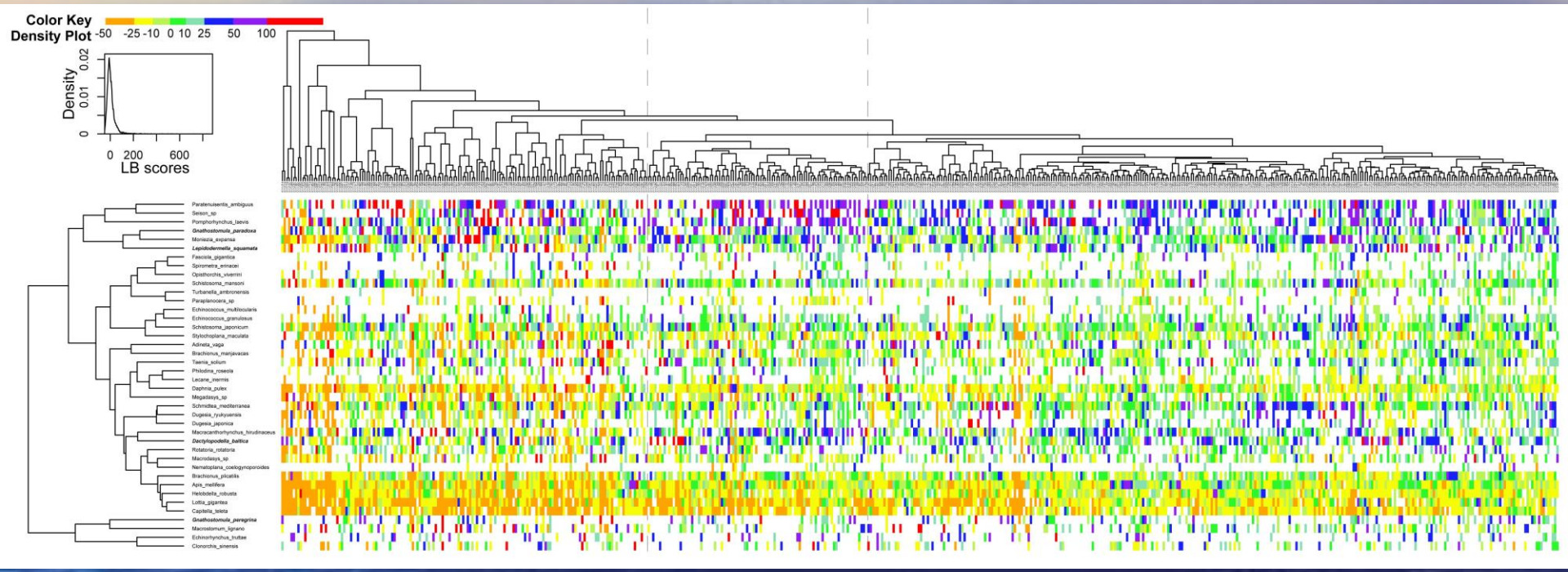


Figure 6. Density plots generated with *R* of taxon-specific LB scores (**A** and **C**) or tip-to-root distances (**B** and **D**) of the trees shown in Figure 9 of Struck[11] (**A** and **B**) or Figure 3 of Ryan et al.[81] (**C**). Tip-to-root distances in (**D**) are based on a tree rerooted with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

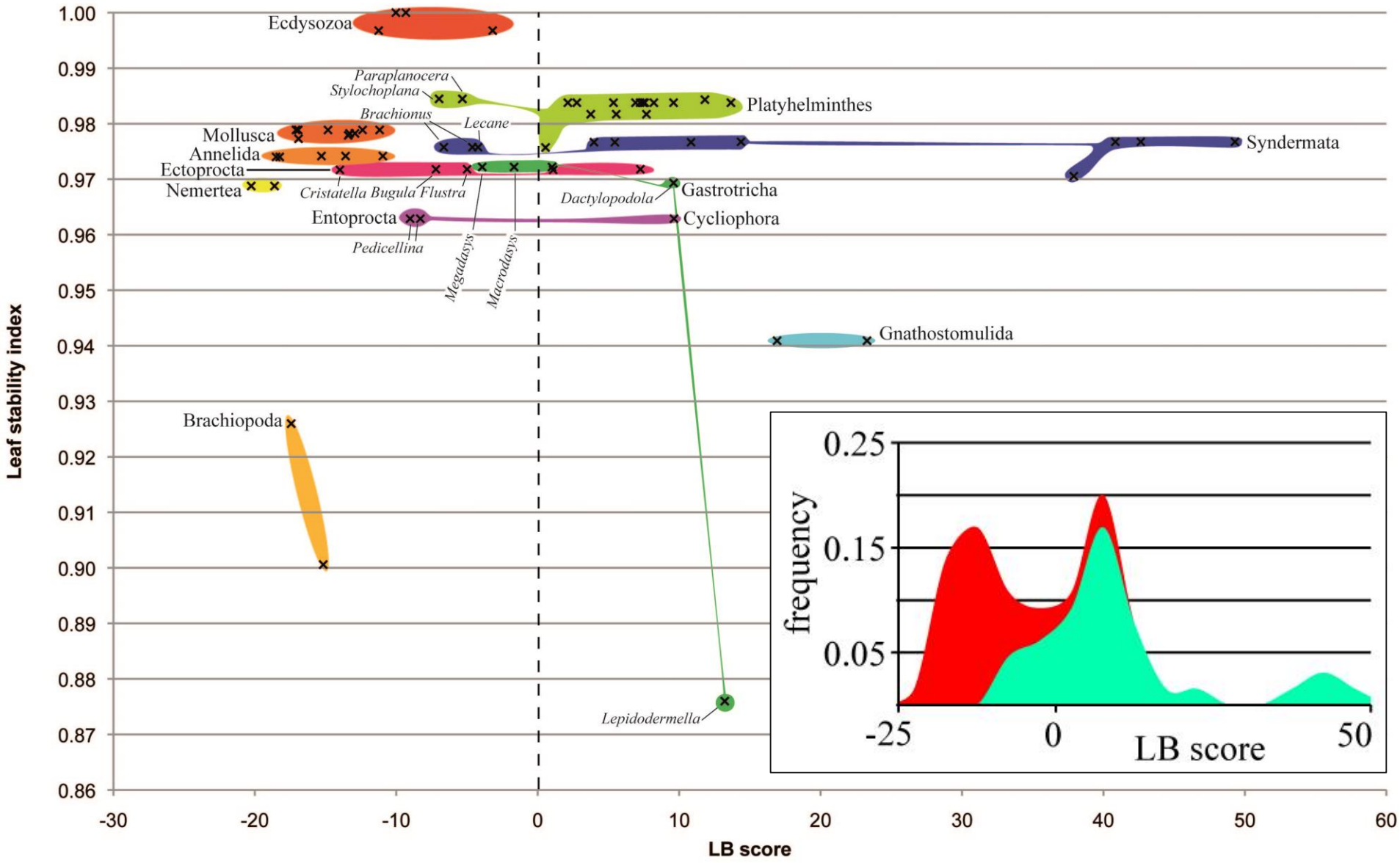
Struck (2014)

Heatmap of heterogeneity



Struck et al. (2014)

Other displays of heterogeneity



Struck et al. (2014)

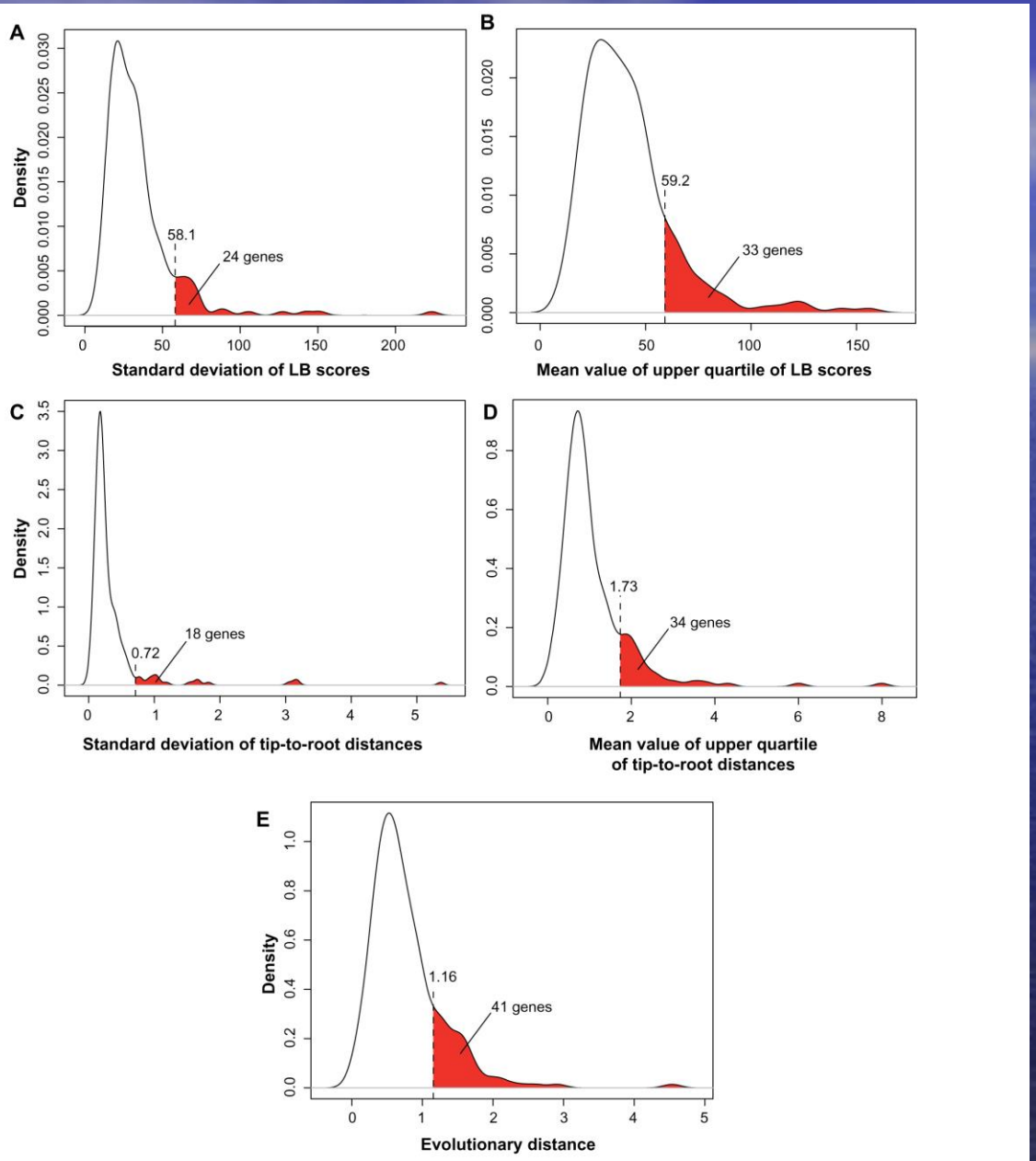
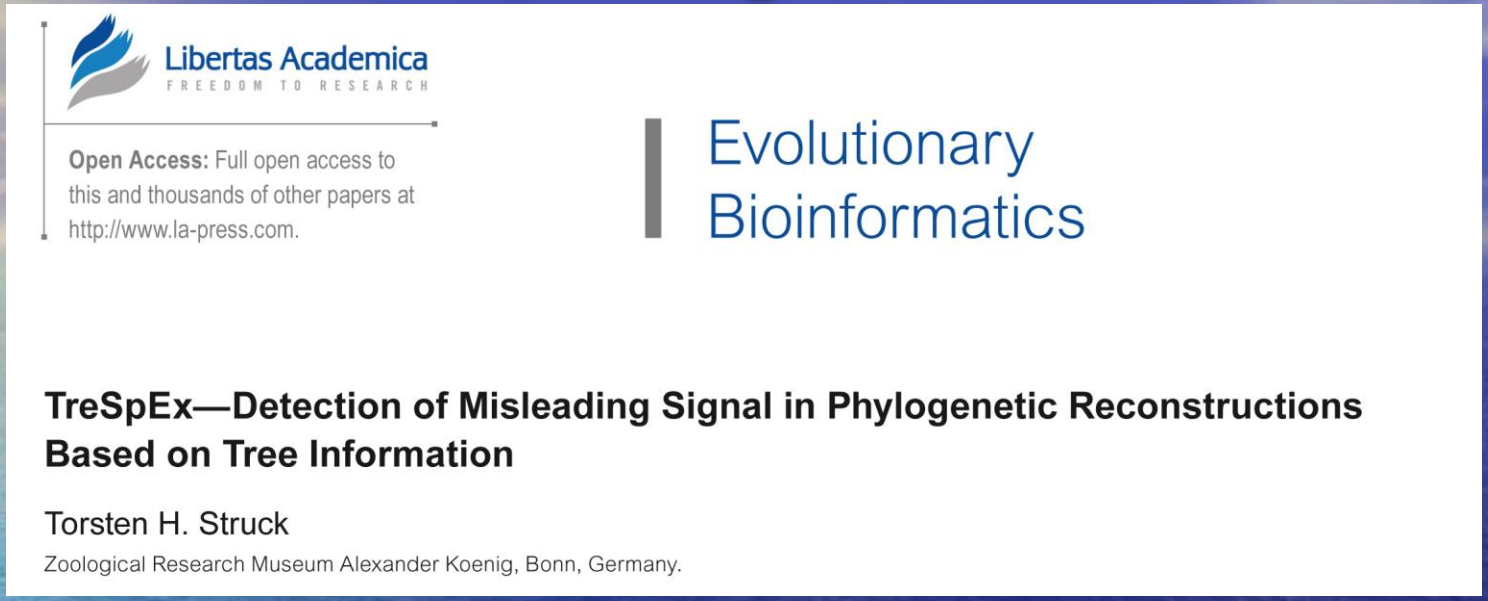


Figure 7. Density plots generated with *R* of different gene-specific long-branch indices for the 229 genes present in the data set of Struck.[11] (**A**) Standard deviation of LB scores measuring heterogeneity; (**B**) average of the upper quartile of LB scores representing the taxa with the longest 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.

Struck (2014)

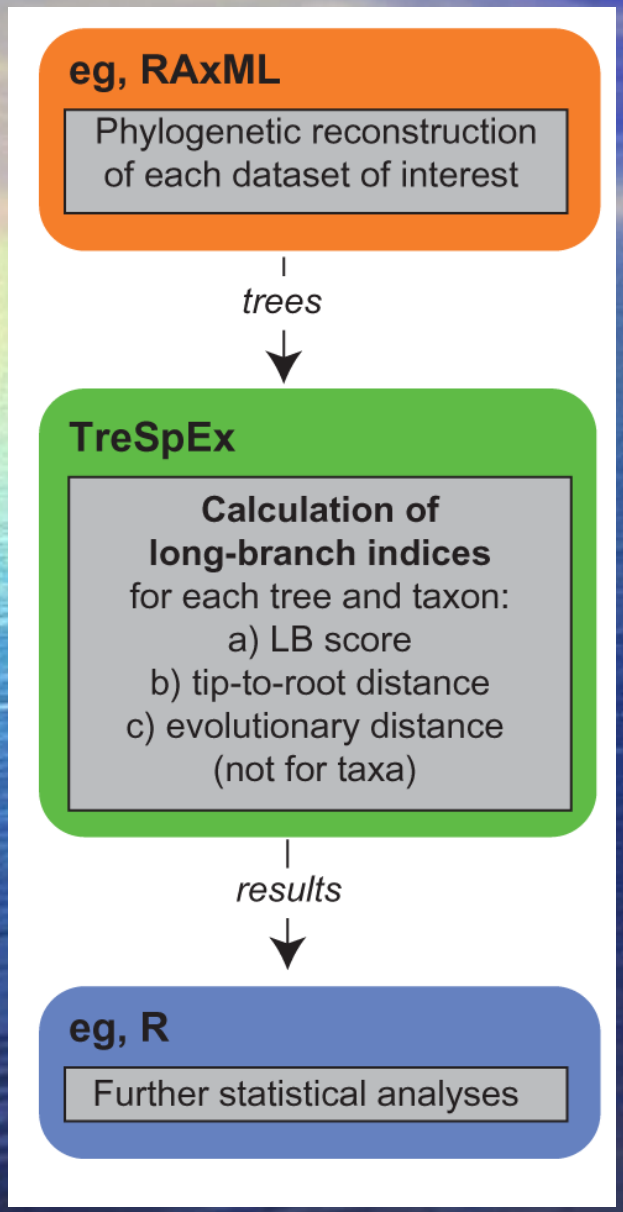
TreSpEx



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

TreSpEx – LB scores and such

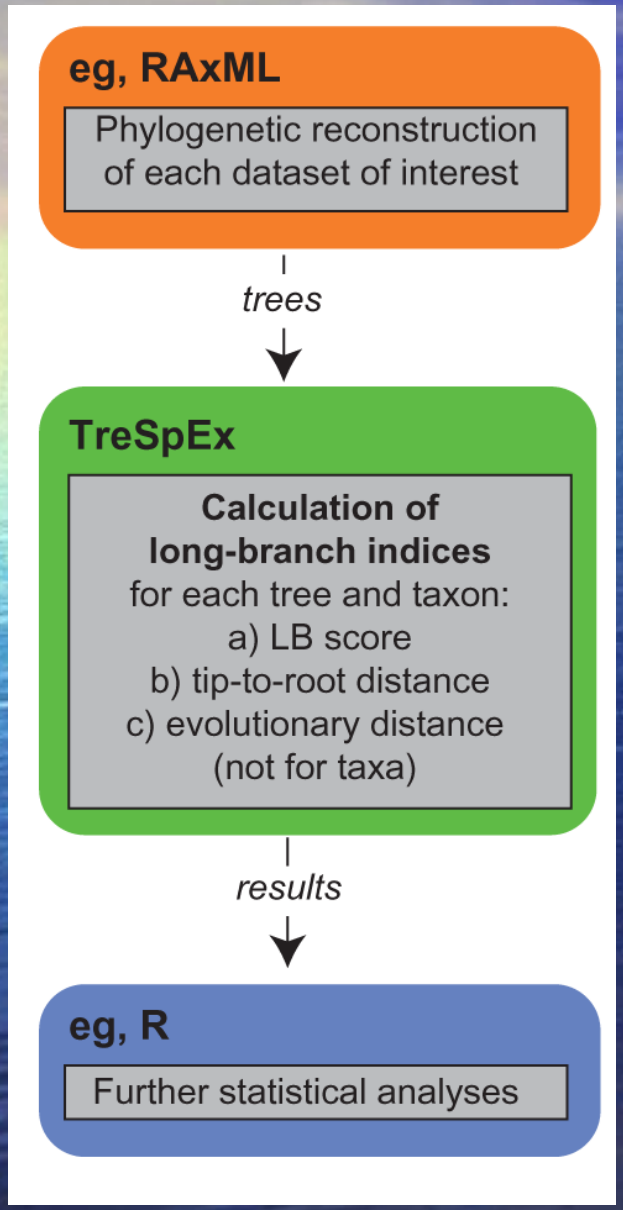


Command-line options

-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



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