1

2

3

4

5

6

7

8

9

LS^X: Automated reduction of gene-specific lineage evolutionary rate heterogeneity for multi-gene phylogeny inference

Carlos J. Rivera-Rivera[1,2] and Juan I. Montoya-Burgos[1]*

16

17

18

19

20

21

22

[1] Department of Genetics and Evolution, University of Geneva, Switzerland

[2] Institute of genetics and genomics in Geneva (iGE3), University of Geneva, Switzerland

* Corresponding author: juan.montoya@unige.ch

26 **ABSTRACT**

27 **Motivation:** LS³ is a recently published algorithm to reduce lineage evolutionary rate heterogeneity,

28 a condition that can produce inference artifacts in molecular phylogenetics. The LS³ scripts are

29 Linux-specific and the criterion to reduce lineage rate heterogeneity can be too stringent in datasets

30 with both very long and very short branches.

31

32 **Results:** LS^X is a multi-platform user-friendly R script that performs the LS³ algorithm, and has

33 added features in order to make better lineage rate calculations. In addition, we developed and

34 implemented an alternative version of the algorithm, LS⁴, which reduces lineage rate heterogeneity

35 not only by detecting branches that are too long but also branches that are too short, resulting in less

36 stringent data filtering.

37

38 **Availability:** The LS^X script LSx_v.1.1.R and the user manual are available for download at:

39 https://genev.unige.ch/research/laboratory/Juan-Montoya

40 **INTRODUCTION**

41   We recently showed that biases emerging from evolutionary rate heterogeneity among

42 lineages in multi-gene phylogenies can be reduced with a sequence data subselection algorithm to

43 the point of uncovering the true phylogenetic signal (Rivera-Rivera and Montoya-Burgos 2016).

44   In that study, we presented an algorithm called Locus Specific Sequence Subsampling (LS³),

45 which reduces lineage evolutionary rate heterogeneity gene-by-gene in multi-gene datasets. For

46 each gene alignment, LS³ implements a likelihood ratio test (LRT) (Felsenstein 1981) between two

47 models. One model assumes equal rates of evolution among all ingroup lineages (single rate model)

48 and the other model assumes that three (or more), user-defined ingroup lineages have their own

49 independent rate of evolution (multiple rates model). If the multiple rates model fits the data

50 significantly better than the single rate model, the branch lengths of the phylogeny are estimated,

51 the fastest-evolving sequence is removed, and the new reduced gene sequence dataset is tested

52 again with the LRT. This process is iterated until a set of ingroup taxa is found whose lineage

53 evolutionary rates can be explained equally well by the single rate model or the multiple rates

54 model. The fast-evolving sequences that had to be removed from each gene alignment to reach this

55 point are flagged as potentially problematic due to their higher evolutionary rate, and gene datasets

56 which never reached this point are flagged as a whole as potentially problematic (Rivera-Rivera and

57 Montoya-Burgos 2016). LS³ proved to be effective in reducing long branch attraction (LBA)

58 artifacts in simulated nucleotide data and in two biological multi-gene datasets (one in nucleotides,

59 and one in amino acids), and its potential to reduce phylogenetic biases has been recognized by

60 several other authors (Cruaud and Rasplus 2016; Suh 2016; Bleidorn 2017).

61   Concomitantly to the publication of the LS³ algorithm, a set of bash scripts were made

62 available to perform these tasks automatically (http://genev.unige.ch/en/users/Juan-Montoya/unit).

63 While these bash scripts are effective, they only run in Linux systems, and the layout and

64 programming is not entirely user-friendly. This prompted us to produce a new, re-programmed

65 version of the LS³ algorithm which is user-friendly , contains important new features, and can be

66   used across platforms. For this new implementation we also developed a second data subselection

67   algorithm based on LS³, called "LS³ *supplement*" (LS⁴) which subselects sequences for lineage

68   evolutionary rate homogeneity by not only removing very long branches, but also very short ones.

69

70   **LS^X DESCRIPTION**

71   The new script is called LS^X, is entirely written in R (R Core Team 2016), and uses PAML

72   (Yang 2007) and the R packages *ape* (Paradis et al. 2004; Popescu et al. 2012) and *adephylo*

73   (Jombart et al. 2010). If PAML, R, the R packages and Rscript are installed and functional, the

74   script runs regardless of the platform, with all parameters given in a single raw text control file. As

75   with the bash implementation of LS³, LS^X reads alignments in PHYLIP format, but unlike the

76   former, it produces for each gene a version of the alignment including only the sequences that

77   satisfied the lineage rate homogeneity criterion. In LS^X, the best model of sequence evolution can be

78   given for each gene (in the original implementation only a single model could be selected for all

79   genes), thus improving the accuracy of the branch length estimations and likelihood calculations

80   under PAML. In addition, the users can select more than three lineages of interest for the lineage

81   evolutionary rate heterogeneity test (Fig. 1*a,b*), while the bash implementation was hardcoded to

82   work with only three lineages of interest.

83

84   **LS³: AN LS⁴ SUPPLEMENT**

85   Within LS^X we also implemented LS⁴, a second data subselection algorithm that is

86   optimized for datasets in which not only extremely long branches but also extremely short branches

87   are present in the starting phylogenetic tree. While both LS³ and LS⁴ follow the same general

88   algorithm, they differ in the criterion for choosing the sequence to be removed in the sequence

89   subselection steps. Under LS³, the fastest-evolving of the ingroup sequences is removed in each

90   iteration, as determined by its calculated sum of branch length (SBL), starting from the stem of the

91   ingroup (Fig. 2*a*). This can lead to the flagging of too much data in a dataset containing extremely

92 slow-evolving sequences (Fig. 2*b*). In such cases, under LS³, not only the fast-evolving sequences

93 will be removed, but also the sequences with intermediate evolutionary rates which are still

94 evolving "too fast" relative to the extremely slow-evolving ones (Fig. 2*b*). While this approach is

95 not incorrect, simply very stringent, it can lead to the flagging of too many sequences in certain

96 datasets and to poorly-resolved phylogenies when concatenating and analyzing the remaining data.
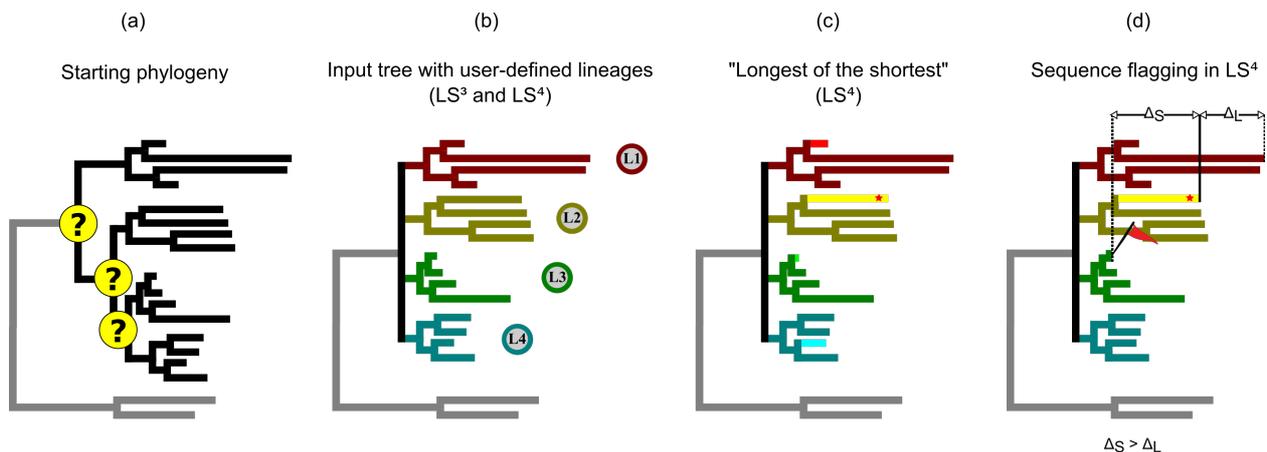

Figure 1. Schematic representation of the procedure for flagging sequences in LS⁴. In (*a*), the general phylogeny for this group of taxa, with the nodes in question highlighted. For for LS³ and LS⁴, an input tree is given in which the nodes into question are collapsed, and the lineages involved are identified (*b*). In (*c*), LS⁴ identifies the shortest branch for each clade (highlighted), and then identifies the longest among them (red star). The sequence to be removed in each iteration of LS⁴ is the one with the tip furthest from the tip of the "longest of the shortest" branch (*d*), resulting in the flagging of both extremely long and extremely short branches.

98   In order to allow for the inclusion of more data while still reaching lineage rate

99 homogeneity, LS⁴ employs a different criterion which considers both too fast- and too slow-

100 evolving sequences for removal. Under LS⁴, when the SBLs for all ingroup sequences of a given

101 gene are calculated, they are grouped by the user-defined lineage of interest to which they belong.

102 The shortest branch of each lineage of interest is identified, and then the longest among them across

103 all ingroup lineages ("the longest of the shortest", see Fig. 1*c*) is picked as a benchmark. Because in

104 both LS³ and LS⁴ each lineage of interest has to be represented by a minimum of species (defined

105 by the user), this "longest of the shortest" branch represents the slowest evolutionary rate at which

106 all lineages could converge. Then, in the sequence subselection steps, the sequence to be removed is

107 the ingroup sequence for which the absolute value of the difference between its SBL and the SBL of

108 the benchmark sequence is the largest. In other words, for a given gene, the sequence removed in

109 each iteration of LS⁴ is that which produces the tip furthest from the benchmark, be it faster- or

110   slower-evolving (Fig. 1*d*). We are confident that the criterion used in LS^4 will result in less genes

111   flagged completely (as compared to LS^3) when extremely slow-evolving sequences are present (Fig.
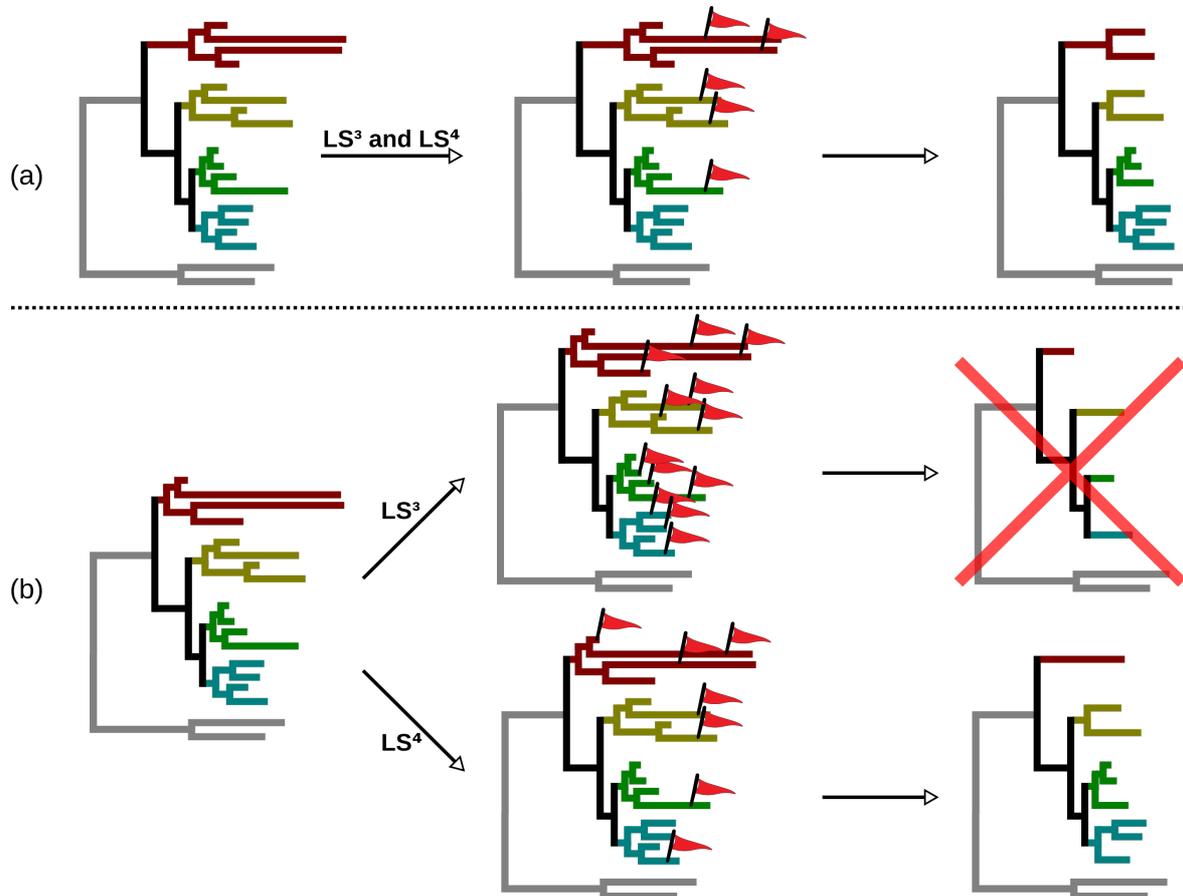
112   2*b*).

Figure 2. A schematic representation of the different ways that LS^3 and LS^4 reach lineage rate homogeneity for a given gene sequence dataset. In (*a*), a dataset with a general rate homogeneity with the exception of several faster evolving branches. In this case, both methods will flag the faster evolving sequences, and reach the same taxon subset with homogeneous rates of evolution. In (*b*), a dataset with general lineage rate heterogeneity, and with a very short branch (the top branch of that tree). In such a case, LS^3 will remove all of the faster evolving sequences, until only the slowest sequence of each clade of interest remains. At this point, that gene sequence dataset is flagged completely as problematic because lineage rate heterogeneity is still too strong. In contrast, LS^4 will remove the faster evolving sequences and also the slowest one, thus reaching lineage rate homogeneity for this dataset.

ACKNOWLEDGEMENTS

118

REFERENCES

119   **REFERENCES**

120   Bleidorn C. 2017. *Phylogenomics: An Introduction*. page 9. Springer International Publishing AG

121   Cruaud A, Rasplus JY. 2016. Testing cospeciation through large-scale cophylogenetic studies. *Curr.*
122   *Opin. Insect Sci.* 18:53–59.

123   Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J.*
124   *Mol. Evol.* 17:368–376.

125   Jombart T, Balloux F, Dray S. 2010. Adephylo: New Tools for Investigating the Phylogenetic Signal
126   in Biological Traits. *Bioinformatics* 26:1907–1909.

127   Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R
128   language. *Bioinformatics* 20:289–290.

129   Popescu AA, Huber KT, Paradis E. 2012. Ape 3.0: New tools for distance-based phylogenetics and
130   evolutionary analysis in R. *Bioinformatics* 28:1536–1537.

131   Rivera-Rivera CJ, Montoya-Burgos JI. 2016. LS[3]: A Method for Improving Phylogenomic
132   Inferences When Evolutionary Rates Are Heterogeneous among Taxa. *Mol. Biol. Evol.*
133   33:1625–1634.

134   Suh A. 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of
135   *Neoaves. Zool. Scr.* 45:50–62.

136   R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for
137   Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

138   Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–
139   1591.