Augmenting phylogenetic data

Are we approaching the brick wall?

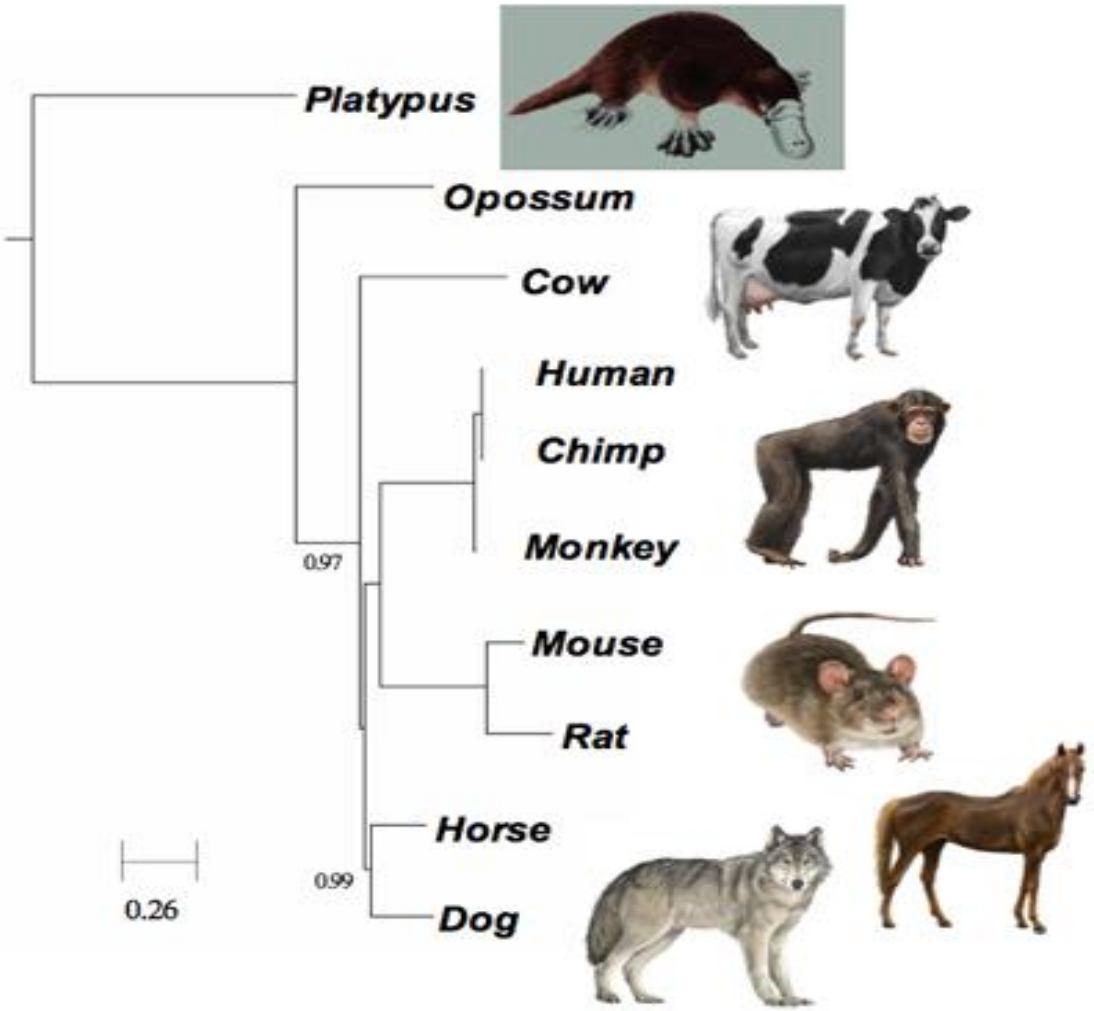
Thomas Wong | Postdoctoral Fellow

8 November 2013

CSIRO ECOSYSTEM SCIENCES
www.csiro.au

Phylogenetic tree

- A tree showing the inferred evolutionary relationships between different species



Augmenting phylogenetic data

Alignment A	Alignment A’ (more sites)
Alignment A’’ (more sequences)	Alignment A# (more sites and sequences, but usually not achievable)

Augmenting phylogenetic data

- When increasing the number of sites:
  - More available information per sequence
    - Number of sites consistent with splits in the true tree is likely to increase
    - Chance to come up the correct tree increases — ☺
- When increasing the number of sequences:
  - More available information per site
    - More precise estimates of site-specific rates of change — ☺
  - More edges in the tree
    - Number of possible trees increases
    - More difficult to come up the correct tree — ☹

Increasing the number of species

Jermiin et al. (2004) stated:

> *The length of the true internal edges will become smaller (on average) when the number of taxa in the phylogenetic data increases*
>
> *…*
>
> *consequently, more difficult to estimate the internal edges of the tree*

However, this statement has not been proved

Objectives

1. To check whether the edges will become smaller when the number of taxa increases

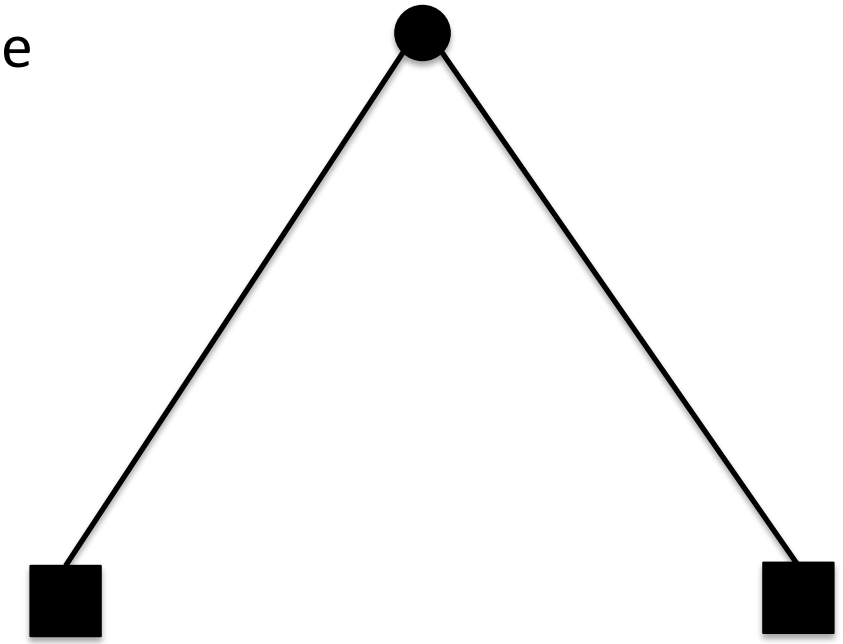
2. To see how short edges in the true tree will affect the accuracy of the phylogenetic tree inference

Simulation of trees with various number of tips

- The total number of species on the earth is estimated to be 3-100 million (May 2010)
- Hence, we simulated trees with the number of tips increasing from 3 tips to 100 million tips, in order to examine the distribution of edge lengths
- Root-to-tip distance is set to 1

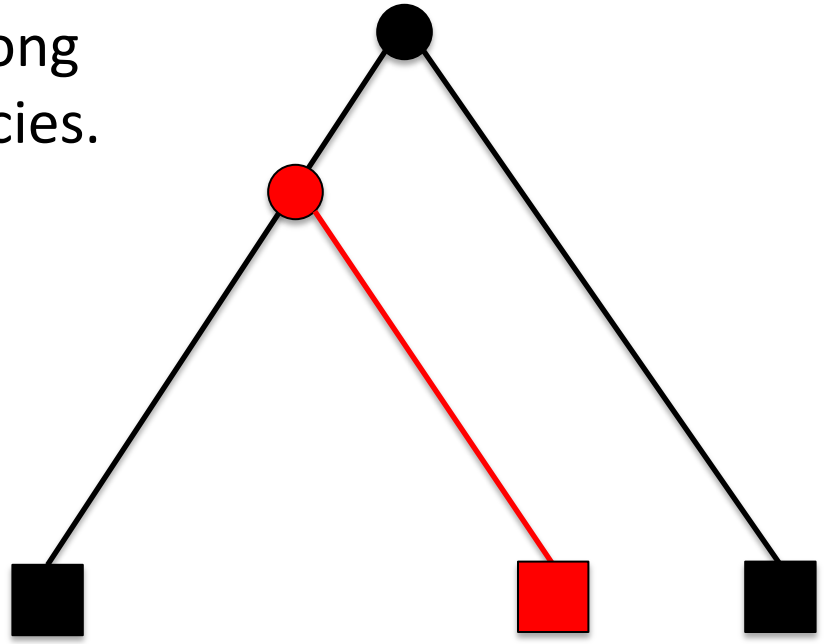
Simulation of trees with various number of tips

Start with a two-tip tree



Simulation of trees with various number of tips

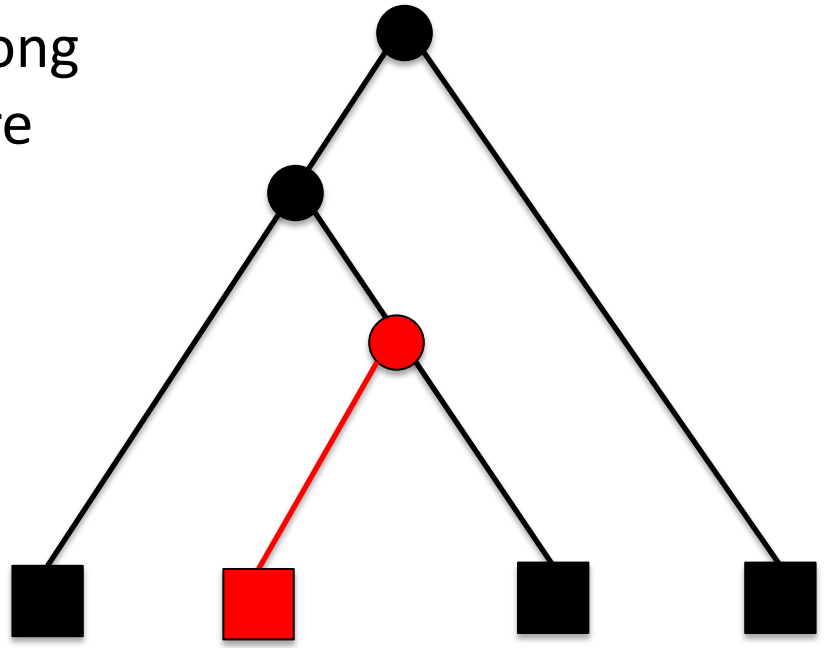
Randomly pick a position along the edges, and add one species.



Simulation of trees with various number of tips

Randomly pick a position along the edges, and add one more species.

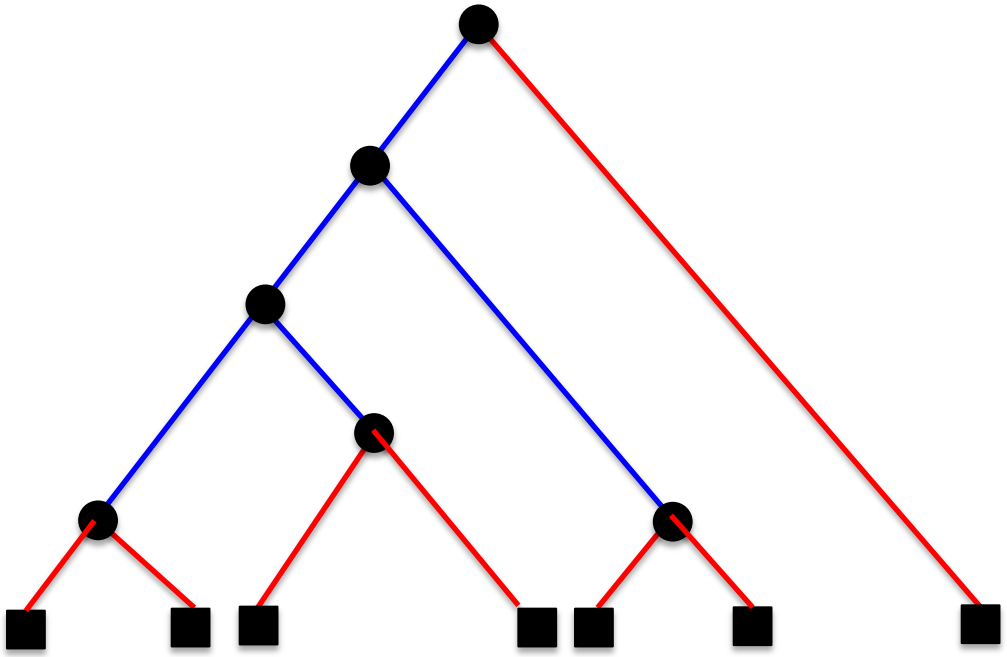
The procedure repeats until 100 million of tips are added.



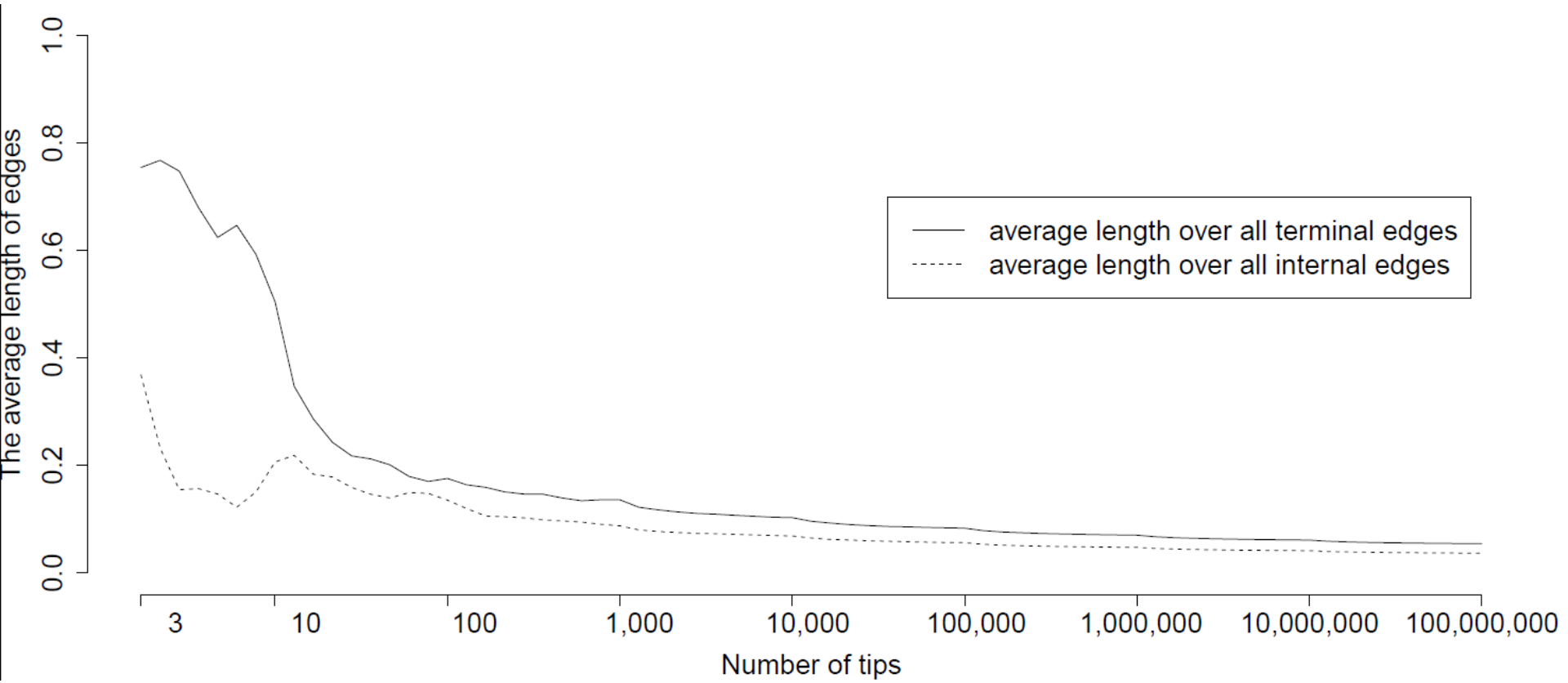
Terminologies

Terminal edges: the edges connecting to the tips (or terminals)

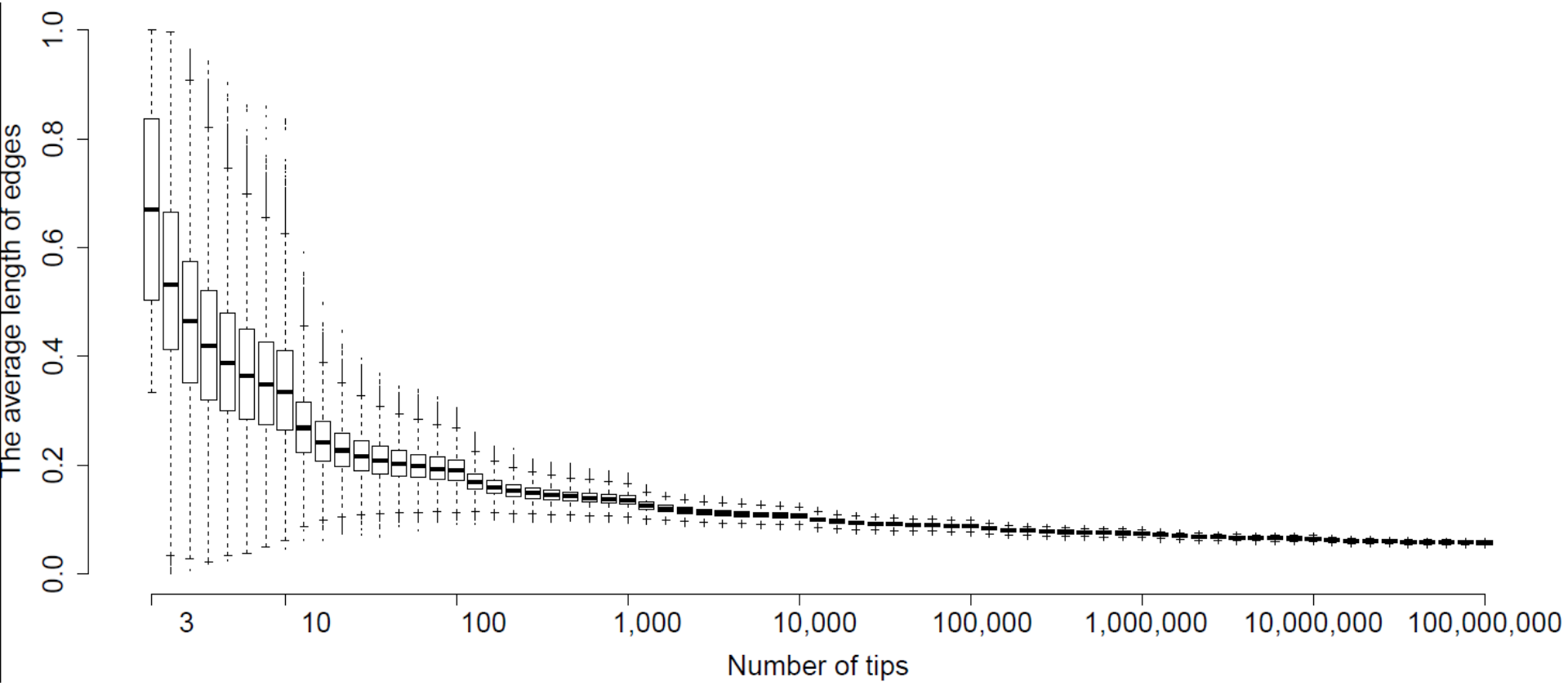
Internal edges: the edges between the internal nodes.



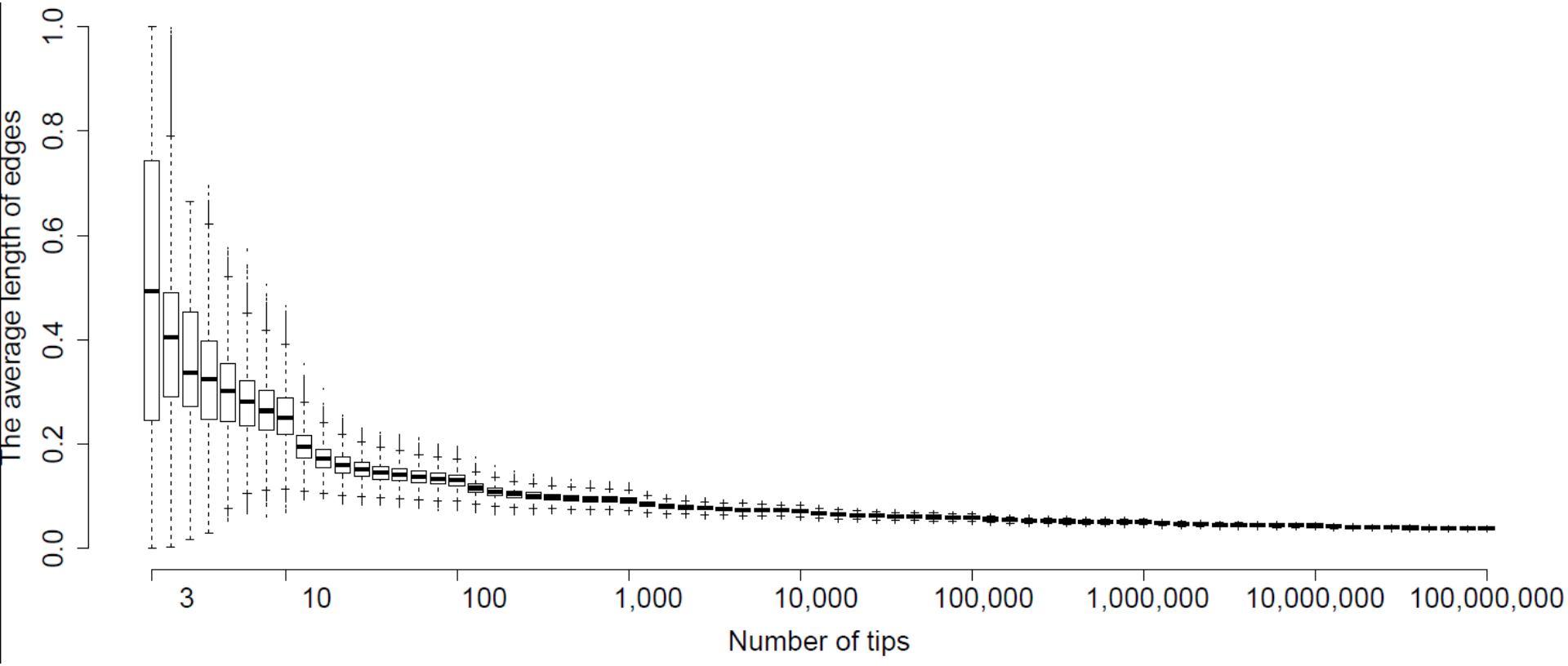
The average length over all terminal edges and that over all internal edges (in one simulation)



The average length over all terminal edges in 20,000 simulations



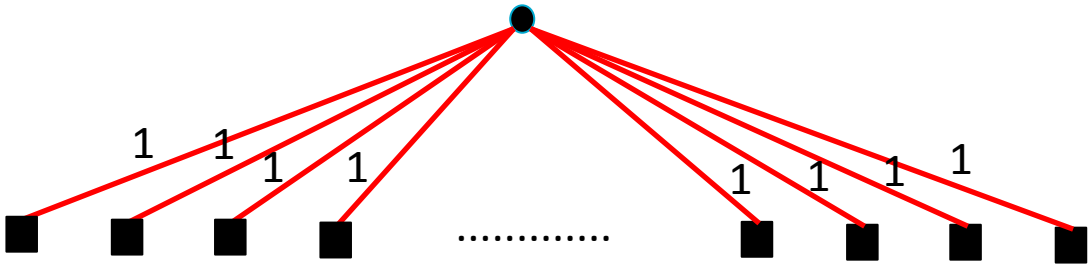
The average length over all internal edges in 20,000 simulations



Mathematical analysis

- Can we prove that the edges will tend to 0 when the number of tips tends to infinity in ALL cases?
- Unfortunately, the answer is no.

Counter example 1



- The lengths of all terminal edges are 1.
- The average length of terminal edges is 1, even when the number of species tends to infinity.

Counter example 2

Number of species: k

Total length of all internal edges: (1-c)(k/2)

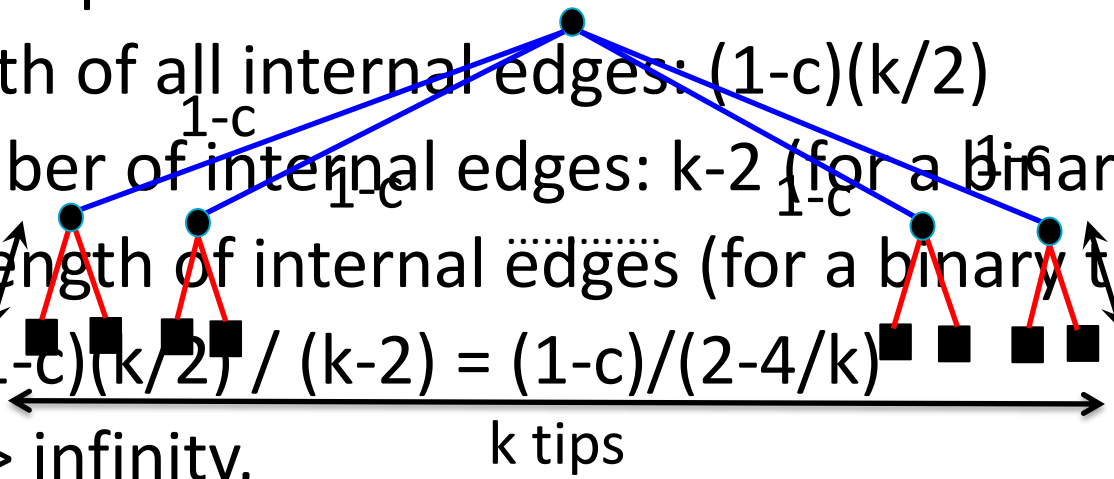
Total number of internal edges: k-2 (for a binary tree)

Average length of internal edges (for a binary tree):

= (1-c)(k/2) / (k-2) = (1-c)/(2-4/k)

When k -> infinity,

= (1-c) / 2



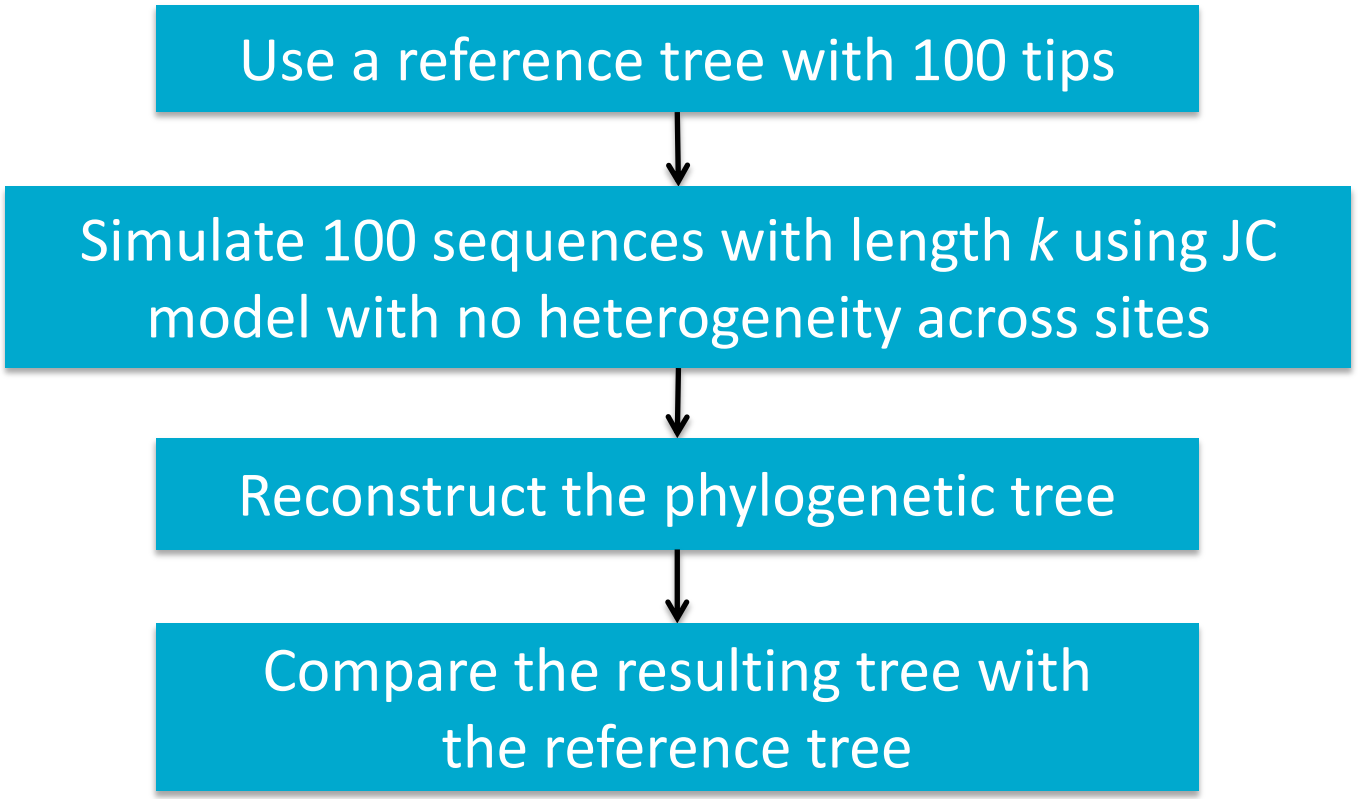
The effect of the short edges

Questions:

How do the short edges affect the accuracy of inferring phylogenetic tree?

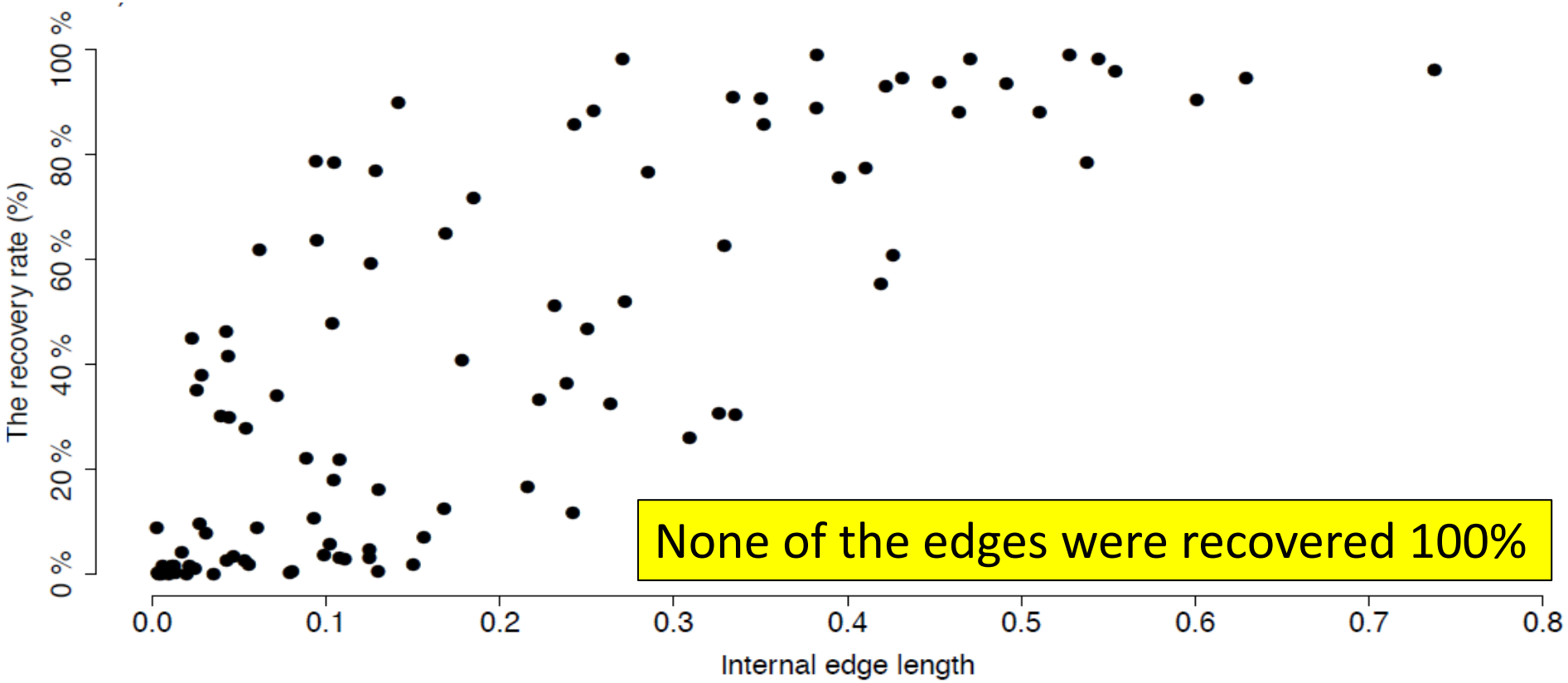
Are the short edges difficult to be recovered correctly?

Experiment

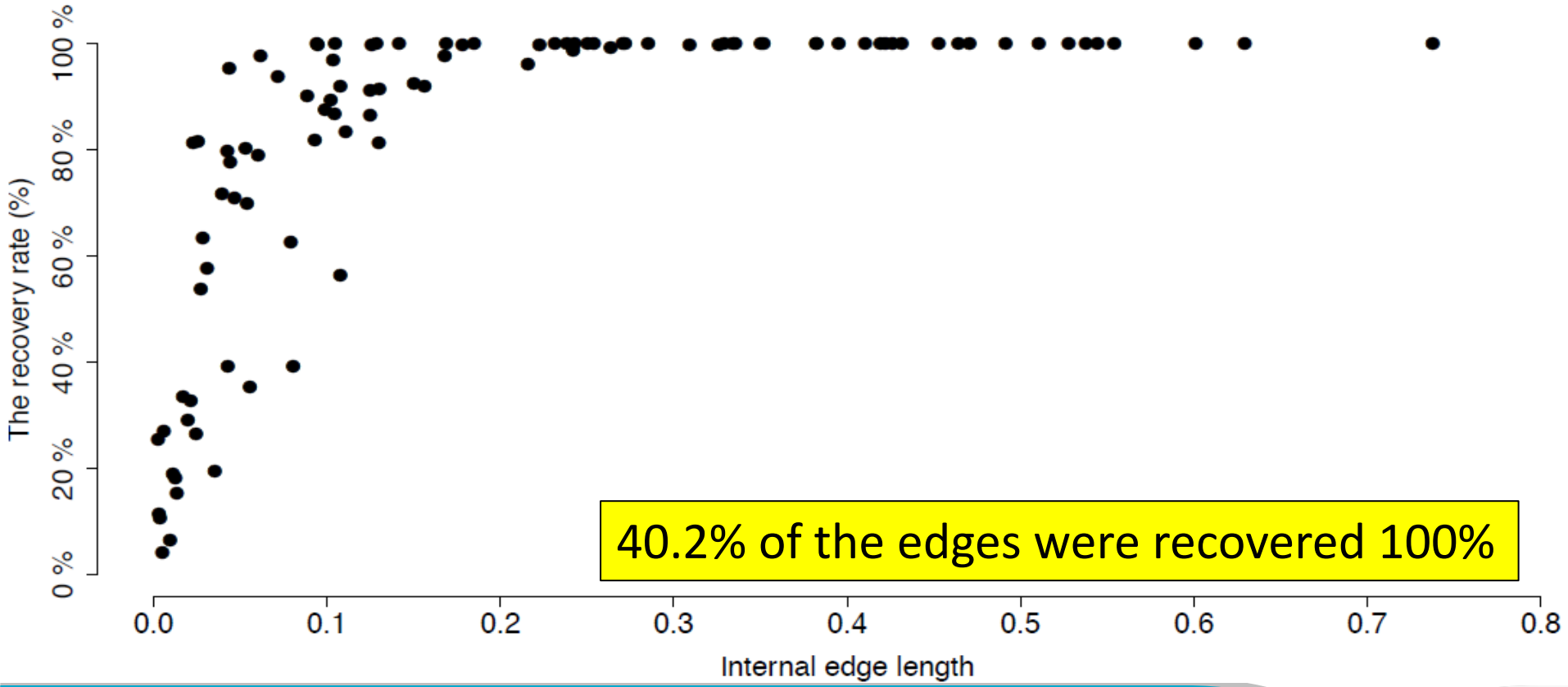


Repeat the procedure 1000 times for k = 100, 1000 and 10000.

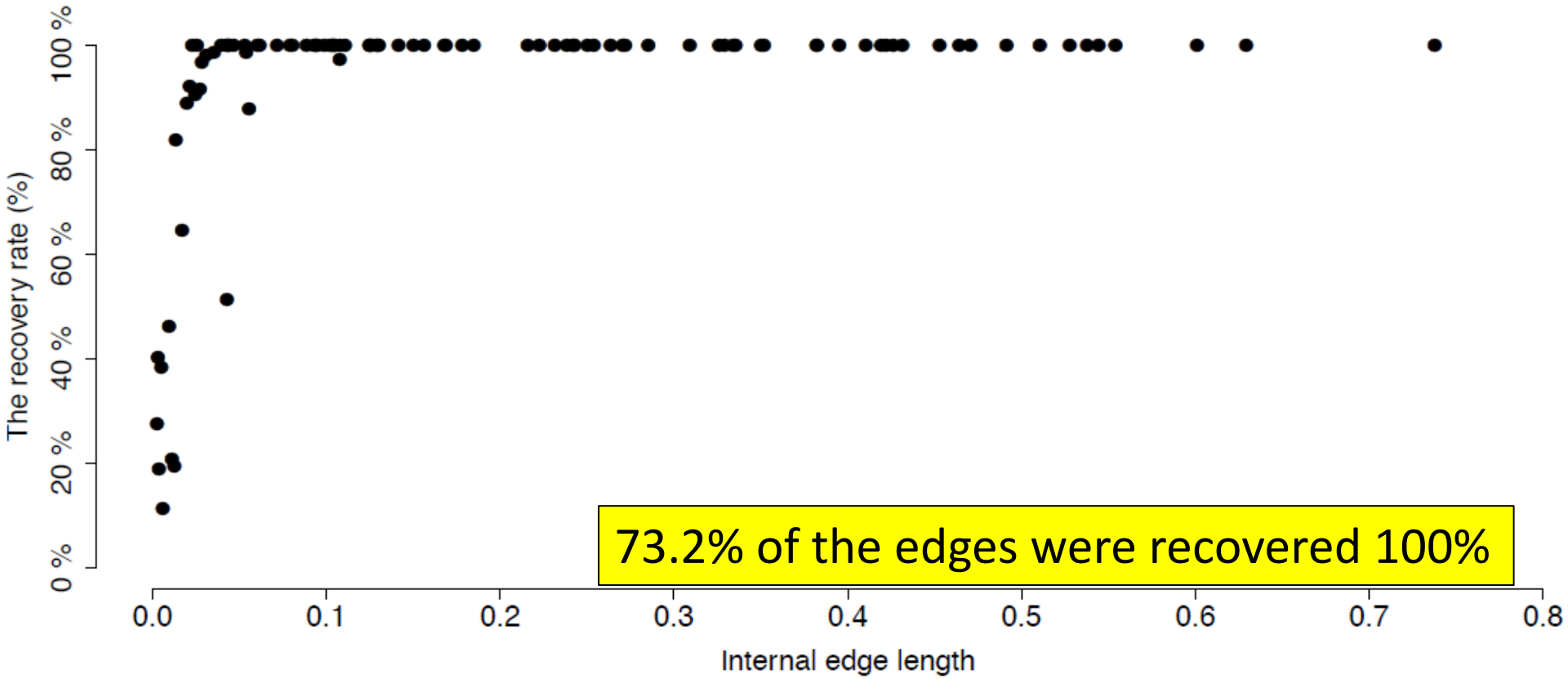
The recovery rate of different length of internal edges in 1,000 simulations (Sequence length = 100)



The recovery rate of different length of internal edges in 1,000 simulations (Sequence length = 1,000)



The recovery rate of different length of internal edges in 1,000 simulations (Sequence length = 10,000)



Conclusions

- Practically, the average length of edges can reach some relatively small values when the number of tips increases.

- The edge length is inversely correlated to the recovery rate.
  - ➢Short edges can deteriorate the accuracy of phylogenetic tree construction.

- Increasing the sequence length has significant positive effect on tree construction for short edges.
  - ➢When there are large number of species, it is necessary to increase the sequence length (i.e. number of sites in the alignment).

Acknowledgement

- Lars S Jermiin, CSIRO Ecosystem Sciences
- Leon Poladian, School of Mathematics and Statistics, University of Sydney

Thank you

Ecosystem Sciences/ Bioinformatics & Phylogenomics Team
Thomas Wong
Postdoctoral Fellow

t +61 2 6246 4057
e Thomas.Wong@csiro.au