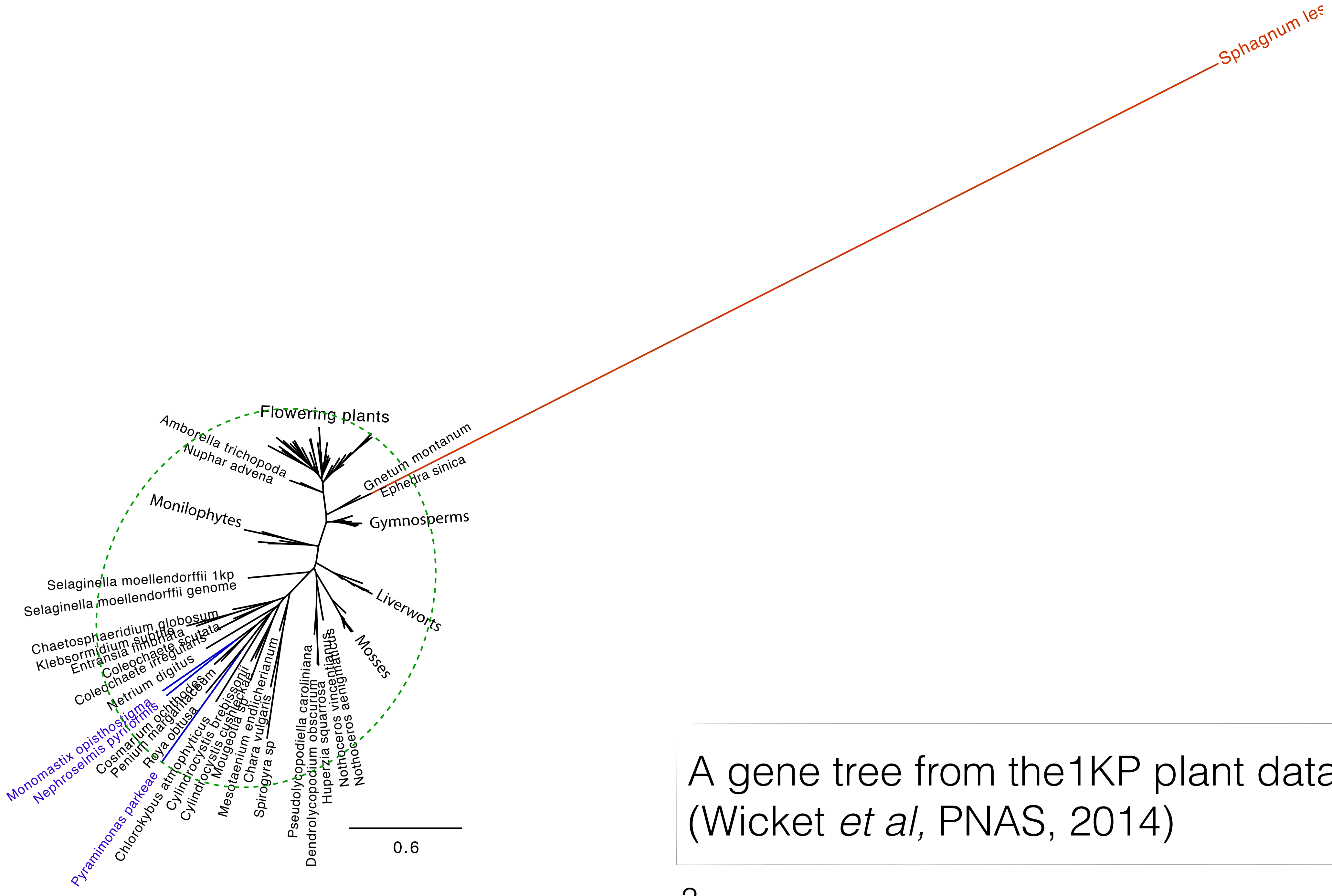
TreeShrink: efficient detection of outlier tree leaves

Uyen Mai
Siavash Mirarab

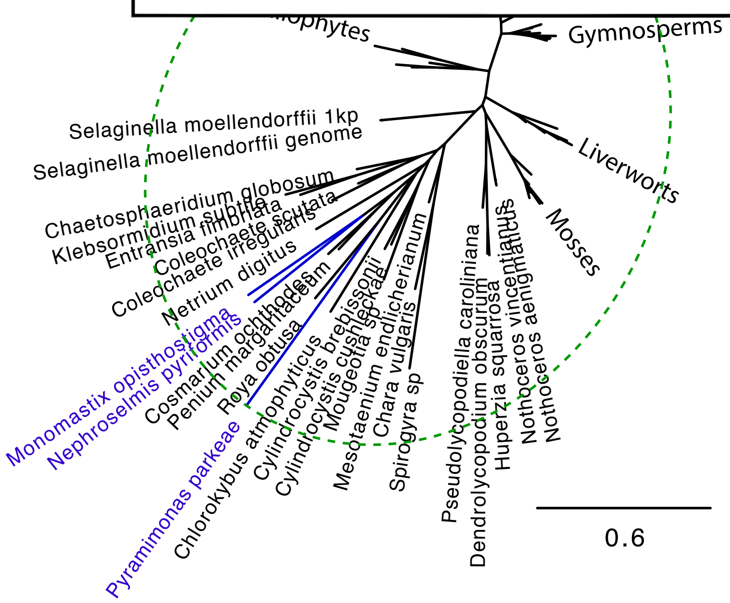
University of California at San Diego

Long branches are suspect



A gene tree from the1KP plant dataset
(Wicket *et al*, PNAS, 2014)

Long branches are suspect

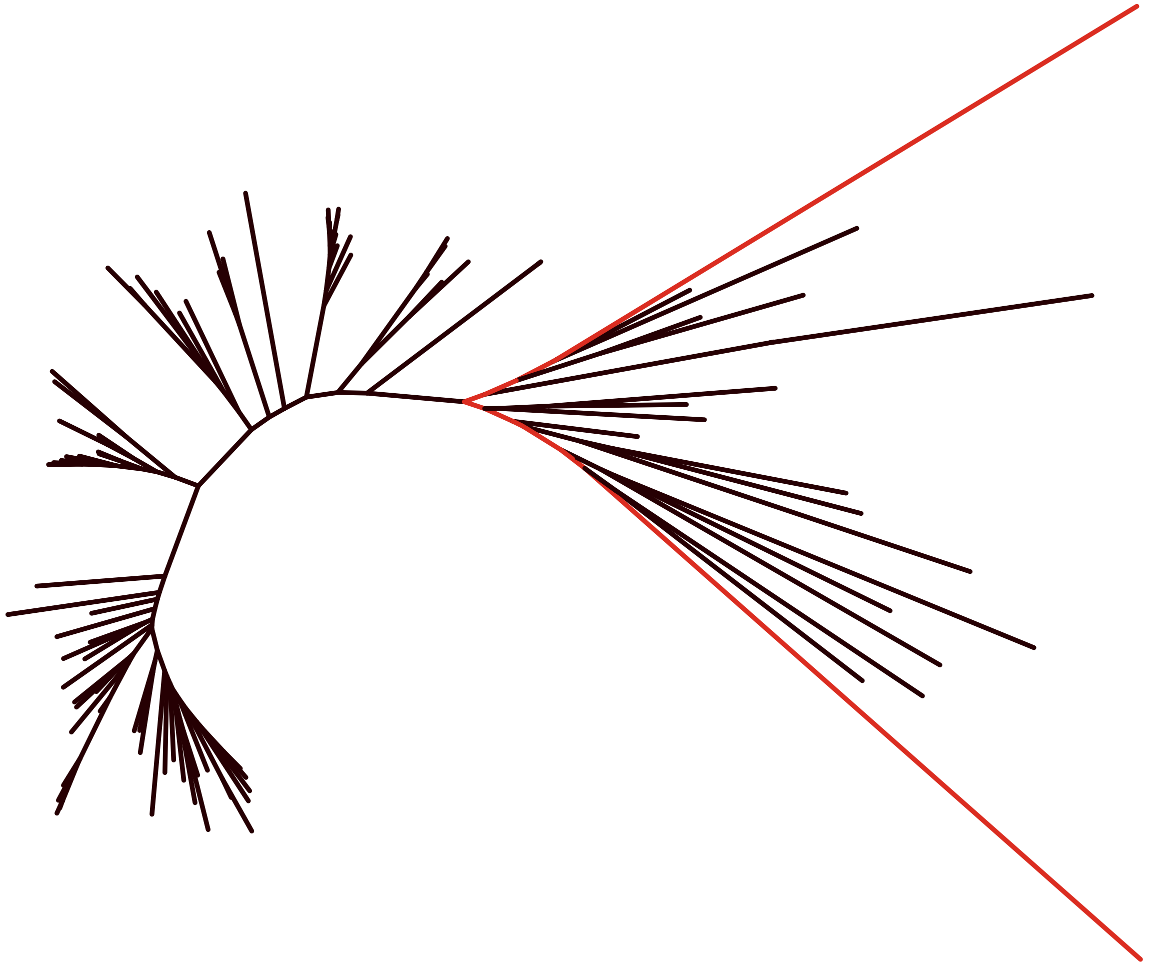


Idea: find errors in the data by building a phylogeny and detecting long branches

A gene tree from the1KP plant dataset (Wicket *et al*, PNAS, 2014)

For unrooted trees?

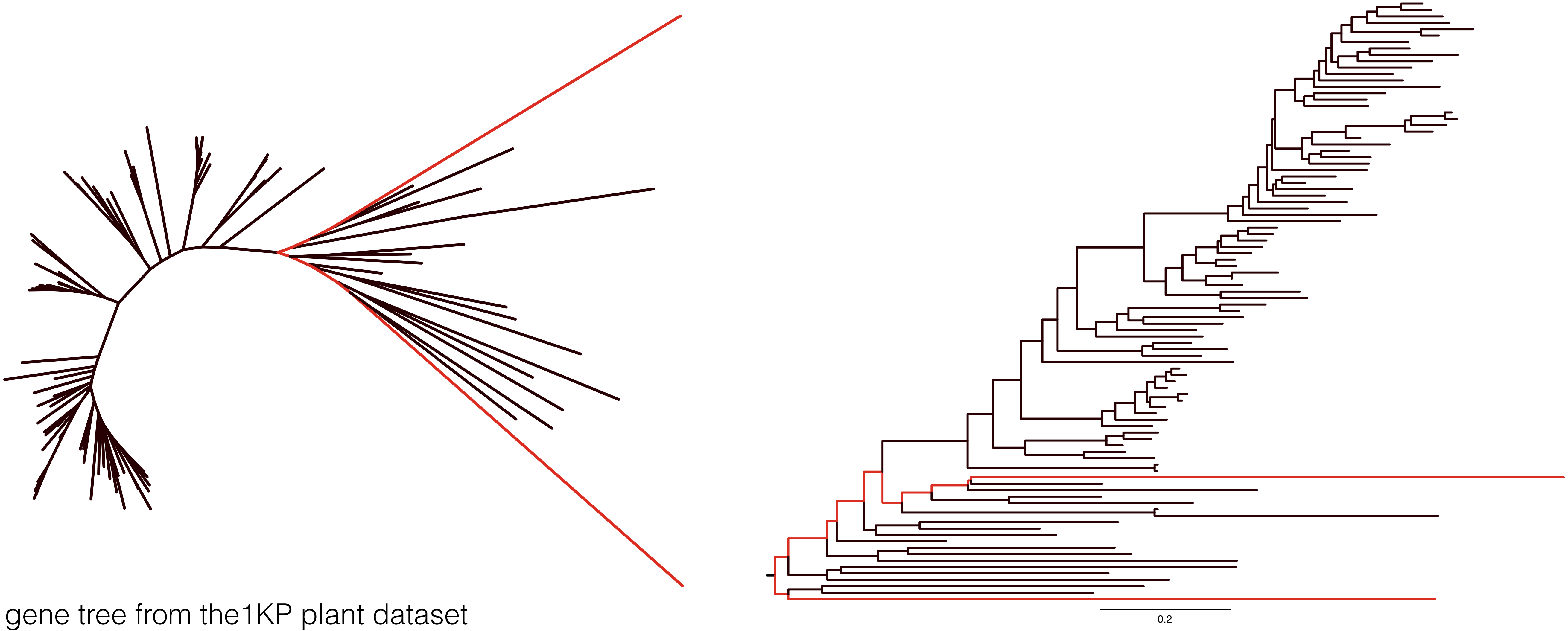
Diameter: the longest path between any two species



A gene tree from the1KP plant dataset
(Wicket *et al,* PNAS, 2014)

For unrooted trees?

Diameter: the longest path between any two species



A gene tree from the1KP plant dataset
(Wicket *et al,* PNAS, 2014)

An optimization problem

The *k*-shrink problem:

- Given:
    - a tree with n leaves and branch lengths
    - some $1 \leq k \leq n$

An optimization problem

The *k*-shrink problem:

- Given:
  - a tree with n leaves and branch lengths
  - some $1 \leq k \leq n$
- Find:
  - for every $1 \leq i \leq k$:
    - the set of i leaves that should be removed to reduce the tree diameter maximally

An optimization problem

The *k*-shrink problem:

- Given:
  - a tree with n leaves and branch lengths
  - s
- Find:
  - for every $1 \leq i \leq k$:
    - the set of i leaves that should be removed to reduce the tree diameter maximally

We have a polynomial time solution

Running Time

- k-shrink can be solved in $O(k^2h+n)$
  where h = the tree height

  - by default, we set $k=O(n^{0.5})$

Running Time

- k-shrink can be solved in $O(k^2h+n)$ where h = the tree height
    - by default, we set $k=O(n^{0.5})$
- Fast enough: processes a tree of n=203,452 leaves with k=2255 in 28 mins

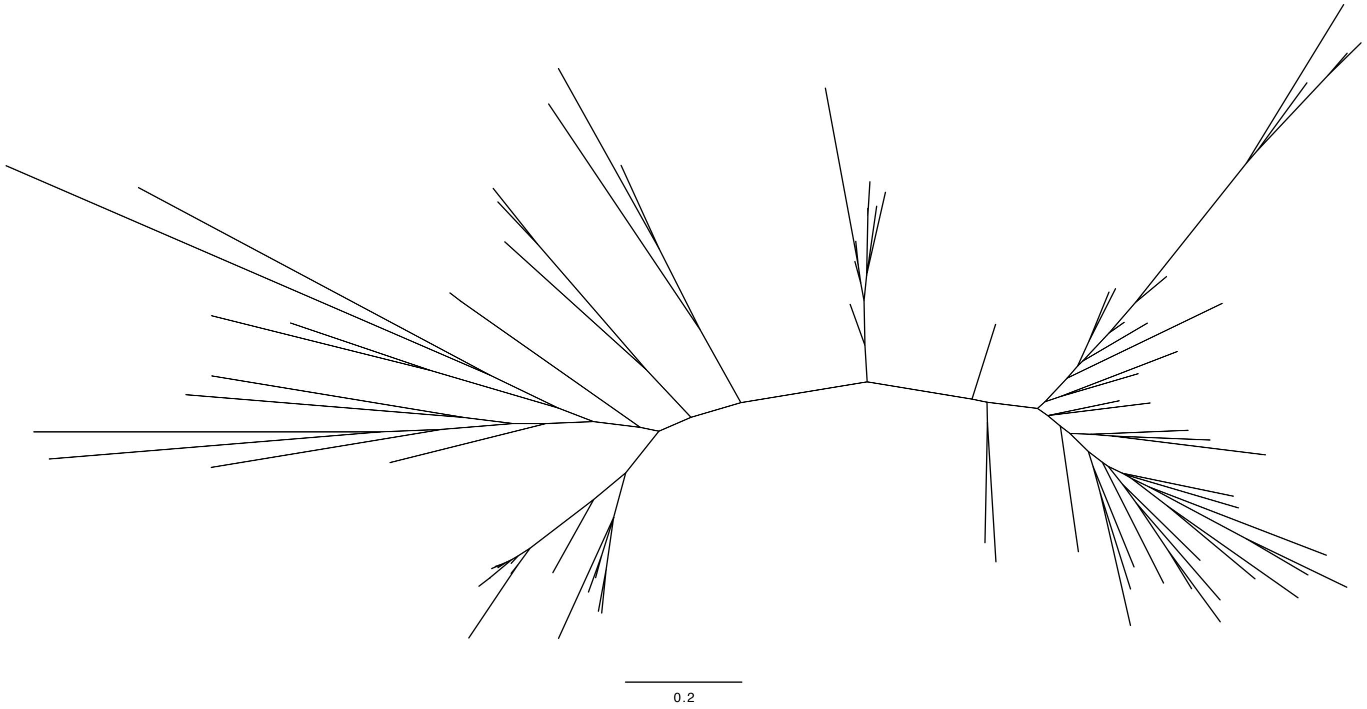
How many do we remove?

- How do we decide how many things to remove?
  - We have the optimal removals for 1 $\leq i \leq k$.
    What i should we use?

How many do we remove?

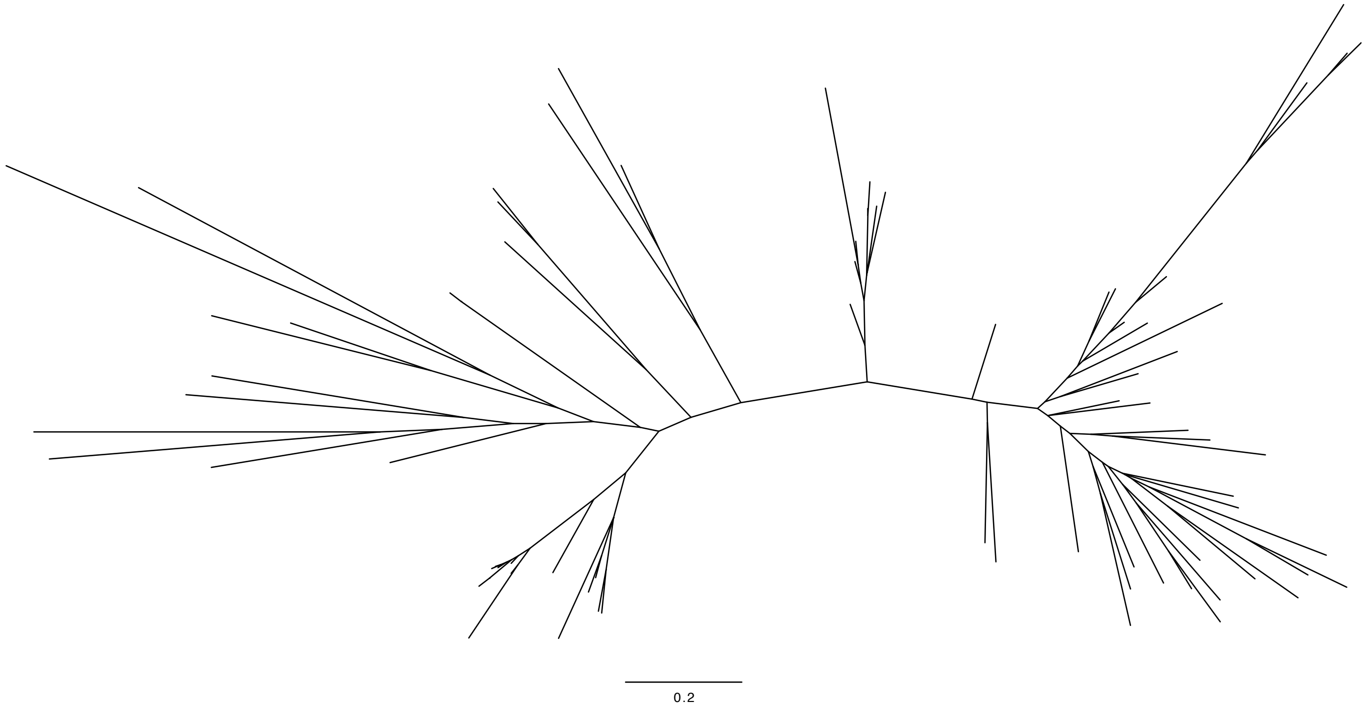
- How do we decide how many things to remove?
  - We have the optimal removals for $1 \leq i \leq k$. What i should we use?
- Find an i where the corresponding reduction in the diameter is unexpectedly high
  - needs statistical tests to find outliers

What to remove?



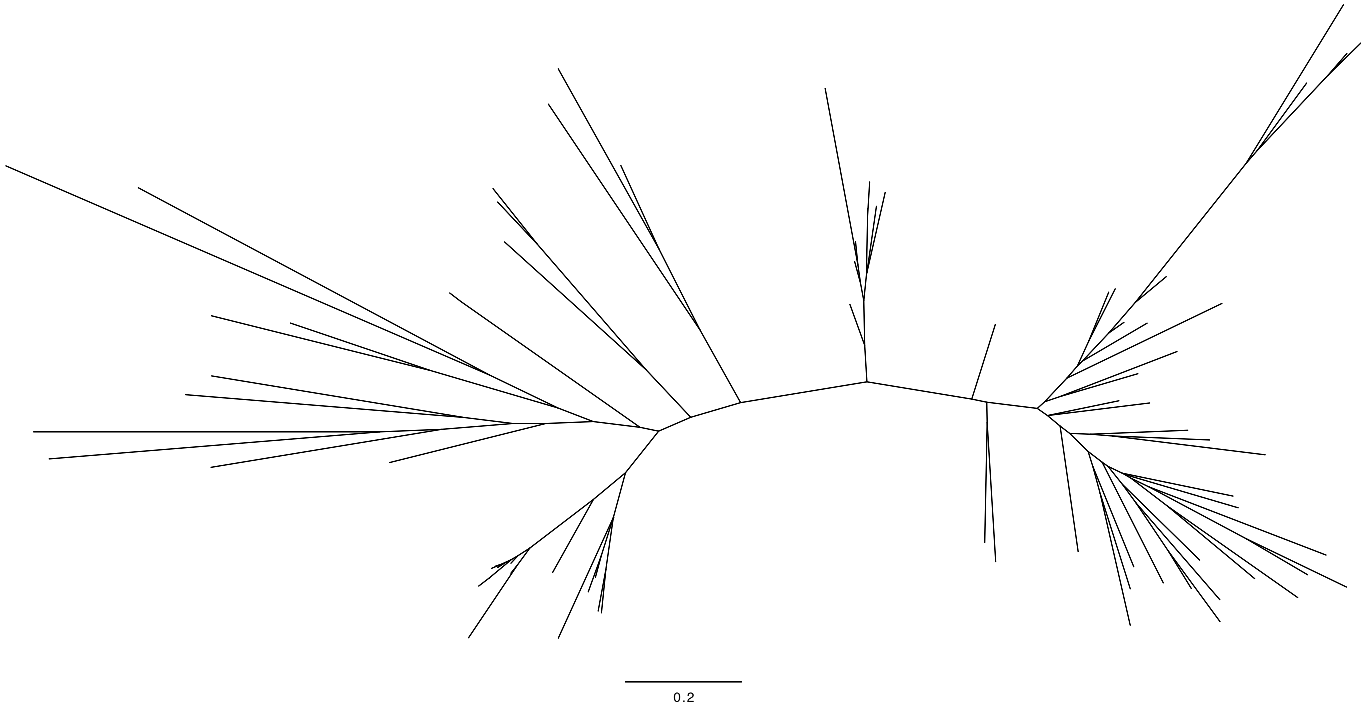
What to remove?

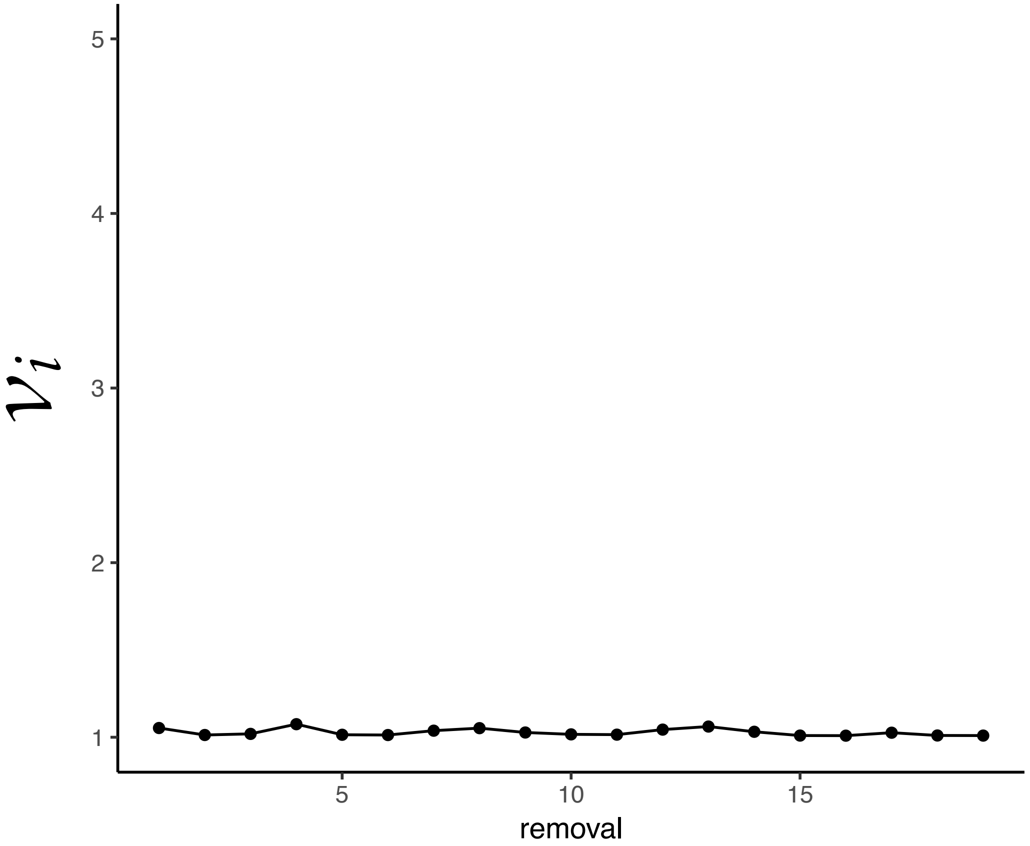
Let $v_i = \dfrac{\text{the diameter after } i\text{-}1 \text{ removals}}{\text{the diameter after } i \text{ removals}}$



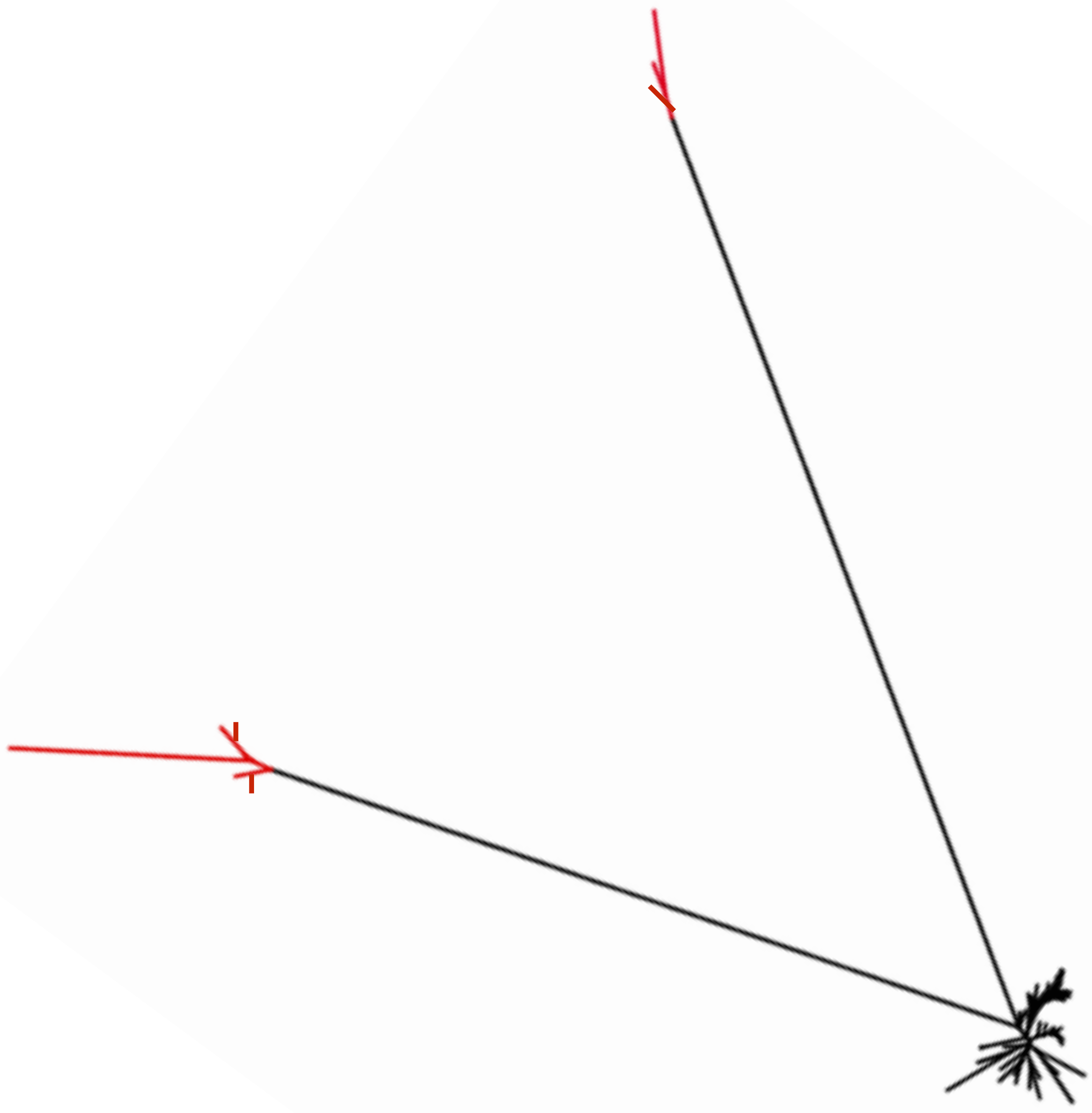
What to remove?

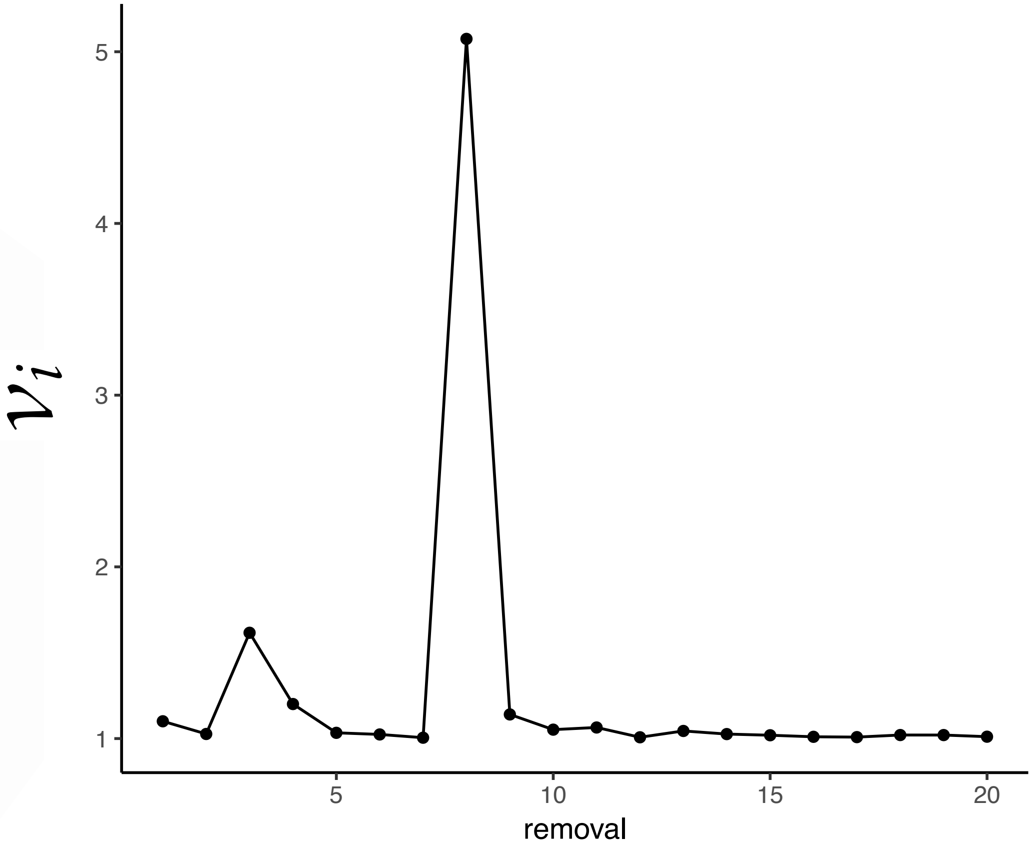
$$\text{Let } v_i = \frac{\text{the diameter after } i\text{-1 removals}}{\text{the diameter after } i \text{ removals}}$$





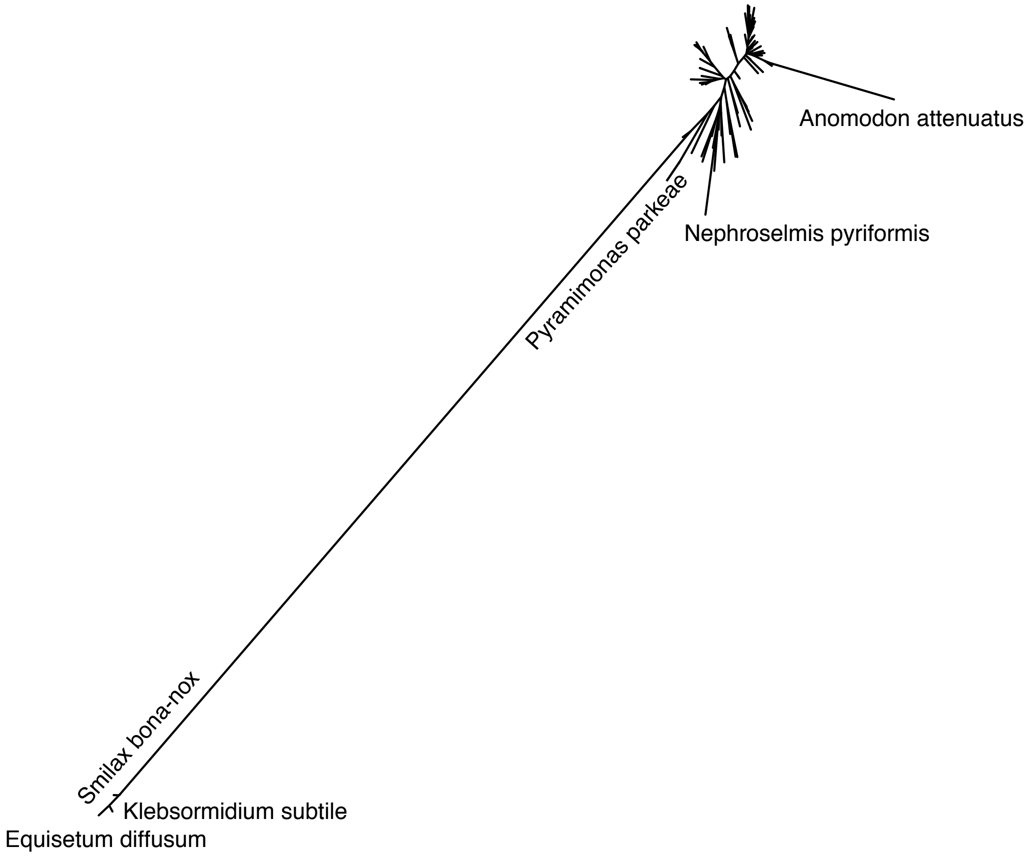
What to remove?

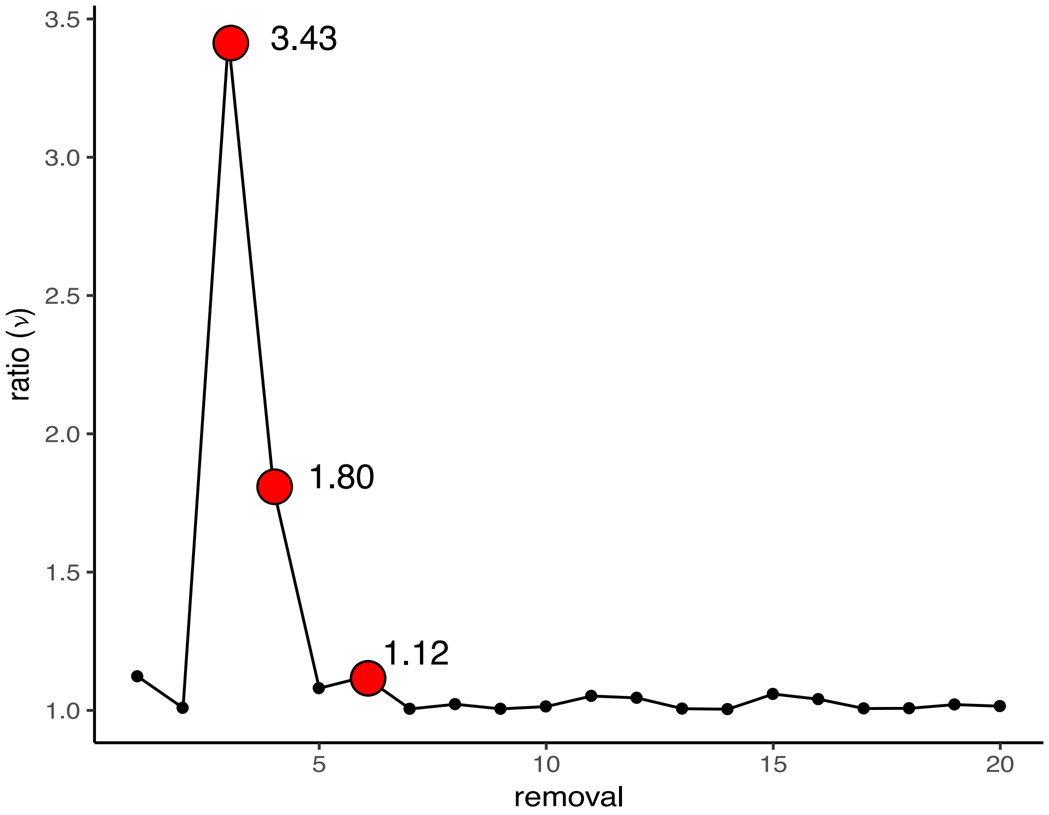




“Signature” of each species

Signature of $x = max\ \log(v_i)$ among all i that remove x

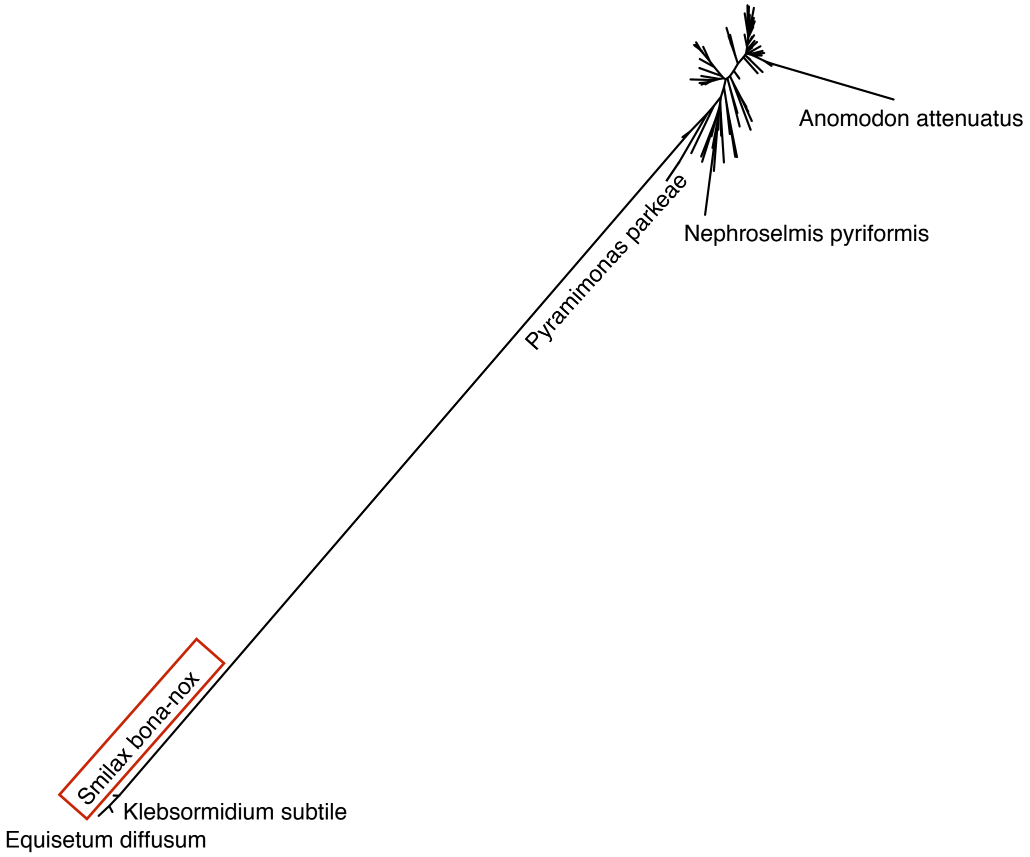


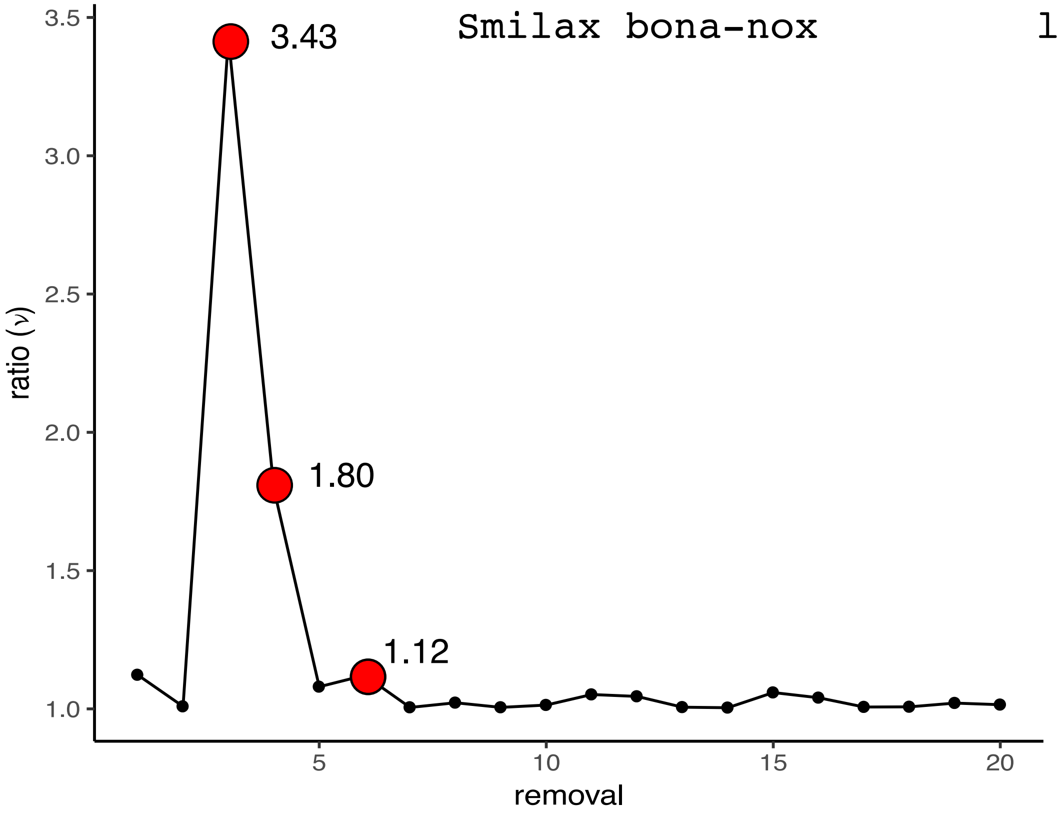


```
Optimal removing sets:
i=1  1.12  Anomodon attenuatus
i=2  1.01  Equisetum diffusum, Anomodon attenuatus
i=3  3.43  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile
i=4  1.80  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Anomodon attenuatus
i=5  1.08  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Nephroselmis pyriformis, Anomodon attenuatus
i=6  1.12  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Nephroselmis pyriformis, Pyramimonas parkeae, Anomodon attenuatus
........
```

“Signature” of each species

Signature of $x = max \log(v_i)$ among all i that remove x



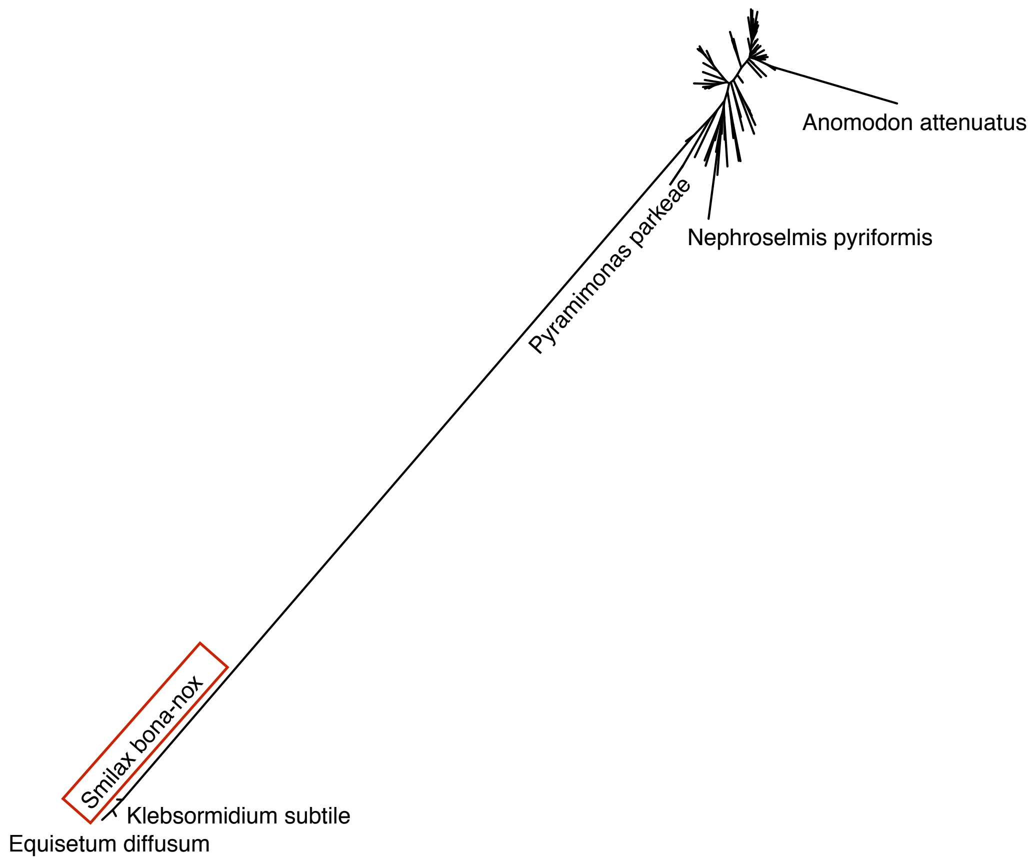


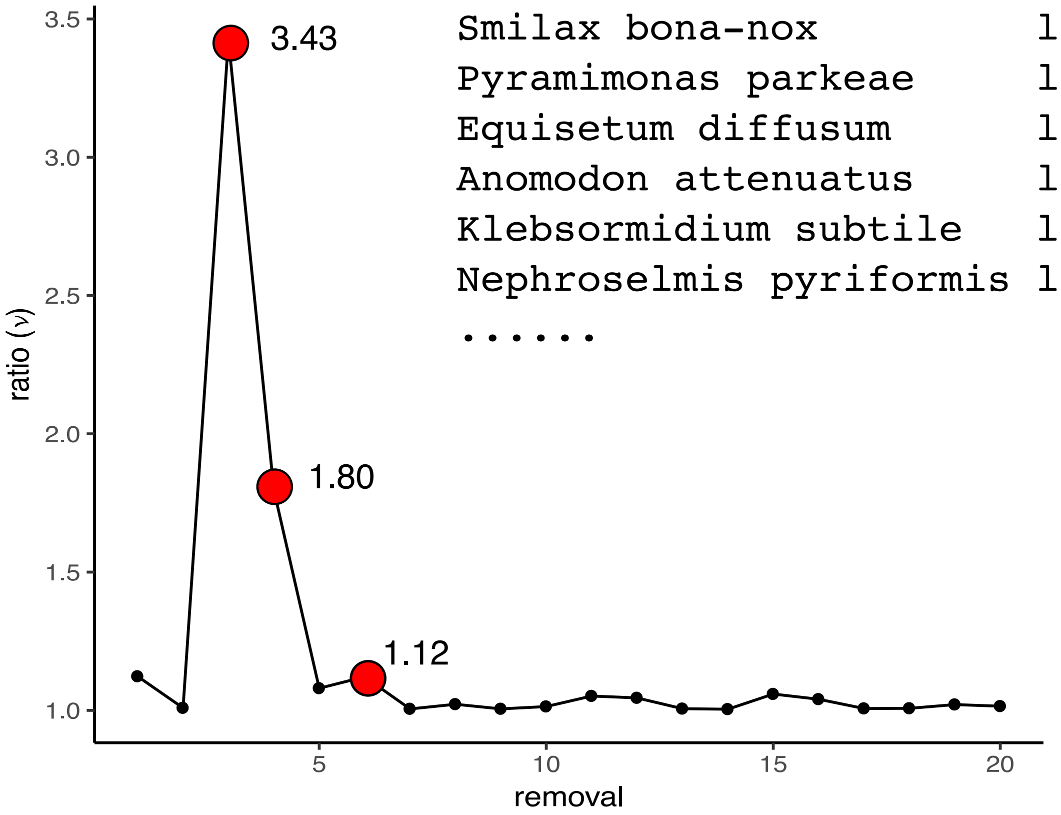
```
Species Signatures:
Smilax bona-nox         log(3.43)
```

```
Optimal removing sets:
i=1  1.12  Anomodon attenuatus
i=2  1.01  Equisetum diffusum, Anomodon attenuatus
i=3  3.43  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile
i=4  1.80  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Anomodon attenuatus
i=5  1.08  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Nephroselmis pyriformis, Anomodon attenuatus
i=6  1.12  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Nephroselmis pyriformis, Pyramimonas parkeae, Anomodon attenuatus
........
```

“Signature” of each species

Signature of $x = max\ \log(v_i)$ among all i that remove x



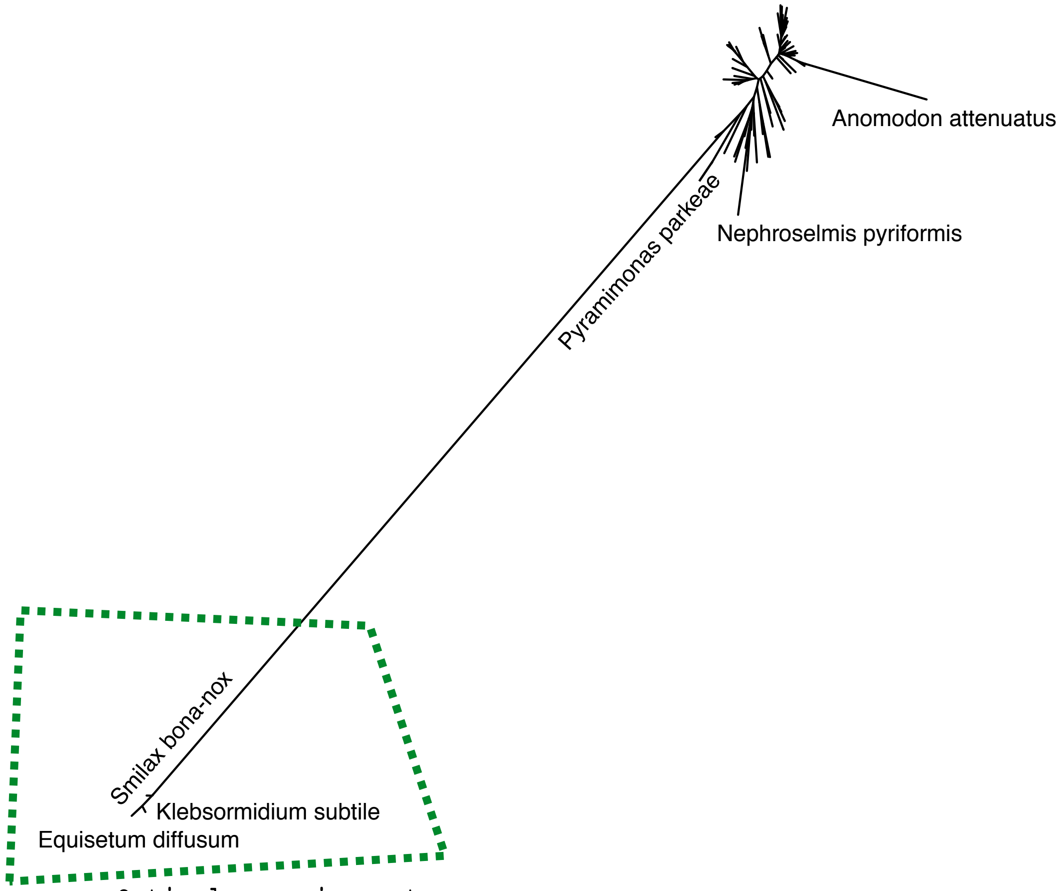


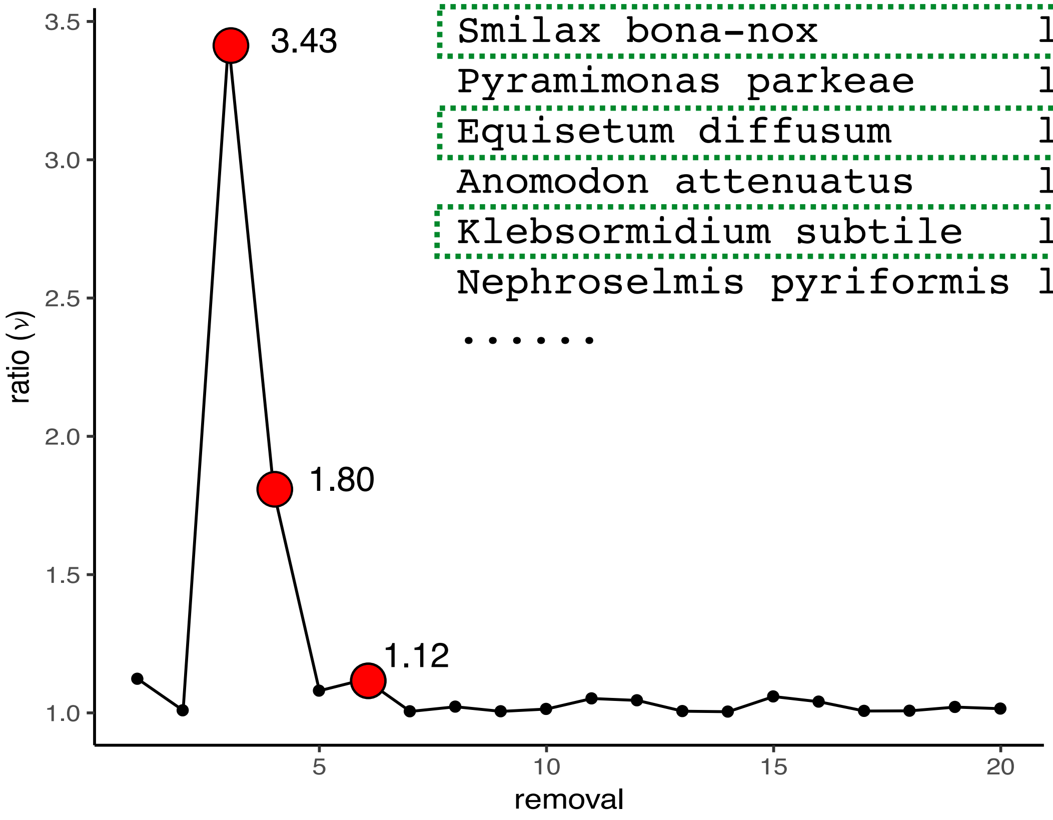
```
Species Signatures:
Smilax bona-nox         log(3.43)
Pyramimonas parkeae     log(1.12)
Equisetum diffusum      log(3.43)
Anomodon attenuatus     log(1.80)
Klebsormidium subtile   log(3.43)
Nephroselmis pyriformis log(1.12)
......
```

```
Optimal removing sets:
i=1  1.12  Anomodon attenuatus
i=2  1.01  Equisetum diffusum, Anomodon attenuatus
i=3  3.43  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile
i=4  1.80  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Anomodon attenuatus
i=5  1.08  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Nephroselmis pyriformis, Anomodon attenuatus
i=6  1.12  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Nephroselmis pyriformis, Pyramimonas parkeae, Anomodon attenuatus
........
```

“Signature” of each species

Signature of $x = \mathit{max}\ \log(v_i)$ among all i that remove x





```
Species Signatures:
Smilax bona-nox           log(3.43)
Pyramimonas parkeae       log(1.12)
Equisetum diffusum        log(3.43)
Anomodon attenuatus       log(1.80)
Klebsormidium subtile     log(3.43)
Nephroselmis pyriformis   log(1.12)
......
```

```
Optimal removing sets:
i=1  1.12  Anomodon attenuatus
i=2  1.01  Equisetum diffusum, Anomodon attenuatus
i=3  3.43  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile
i=4  1.80  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Anomodon attenuatus
i=5  1.08  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Nephroselmis pyriformis, Anomodon attenuatus
i=6  1.12  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Nephroselmis pyriformis, Pyramimonas parkeae, Anomodon attenuatus
........
```

Three statistical tests of TreeShrink

- The “per-gene” test:
  requires only a single tree

- The “all-gene” test:
  requires a collection of gene trees

- The “per-species” test:
  requires a collection of gene trees

Statistical tests

- The “per-gene” test (input: a single tree)
  - Fit a log-normal distribution to the signatures
  - Remove taxa with outlier signatures
  - Outlier: CDF above $1-\alpha$ fore a given α (false positive tolerance)
- The “all-gene” test
- The “per-species” test

Statistical tests

- The “per-gene” test
- The “all-gene” test (input: a collection of gene trees)
    - Combine all signature values *across all genes*
    - Compute a kernel density over the empirical distribution
    - Remove the taxa of the outlier signatures
    - Outlier: CDF above $1-\alpha$ fore a given α
- The “per-species” test

Statistical tests

- The “per-gene” test
- The “all-gene” test
- The “per-species” test (input: a collection of gene trees)
  - Compute a kernel density function *for each species* over its signatures across genes
  - Remove the taxa of the outlier signatures
  - Outlier: CDF above $1-\alpha$ for a given α

Methods

- The three tests of TreeShrink
- Alternative filtering methods
    - *RootedFiltering:* root gene trees and remove taxa X standard deviations more distant to the root than average
    - *RogueNarok:* rogue taxon removal based; finds unstable nodes based on bootstrap replicates
    - *RandomFiltering*: randomly choose what to remove.

Measurements

- Effects of filtering on taxon occupancy
    - Proportion of data retained for each species
- Effects of filtering on gene tree discordance
    - Reduction in pairwise MS distance of gene trees on controlled amount of filtering

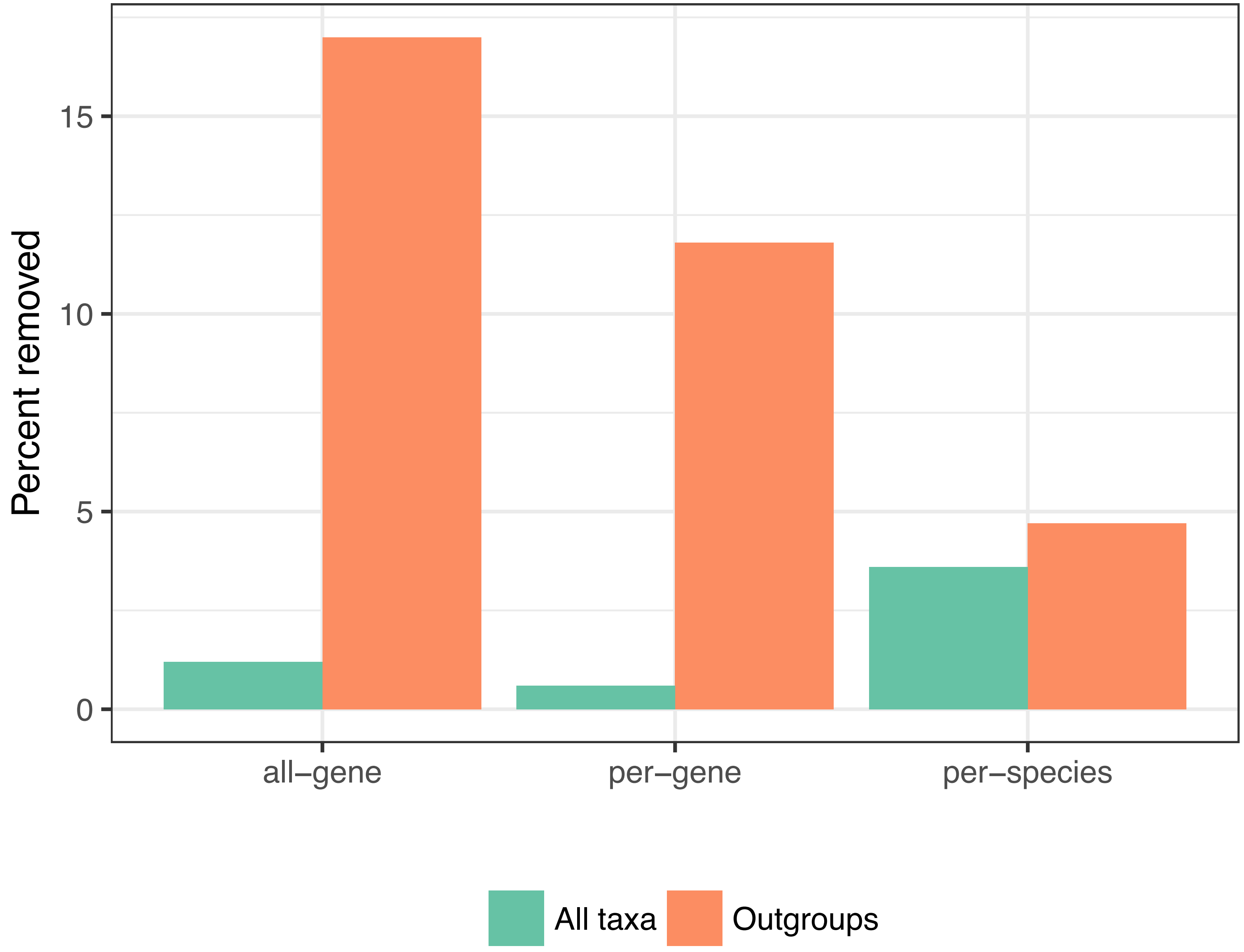
Datasets

- 6 phylogenomic datasets
- Gene number: 95 - 1478
- Species number: 26 - 164

	Genes	Species
Plants	852	104
Insects	1478	144
Mammals	424	37
Frogs	95	164
Metazoa-Cannon	213	78
Metazoa-Rouse	393	26

Results: outgroup removal

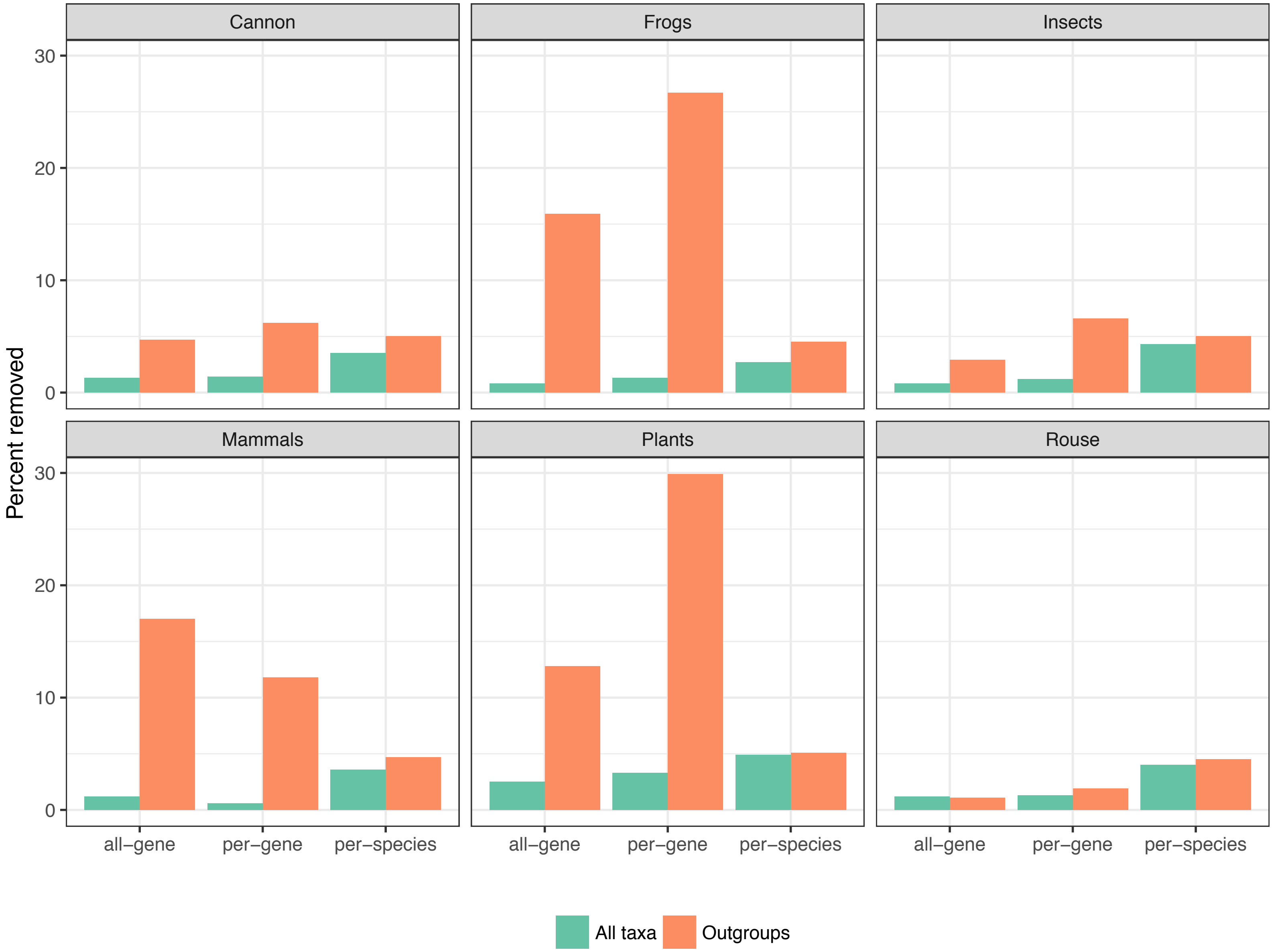
- Percent of the data removed for $\alpha=0.05$ for
  - All species
  - Outgroups



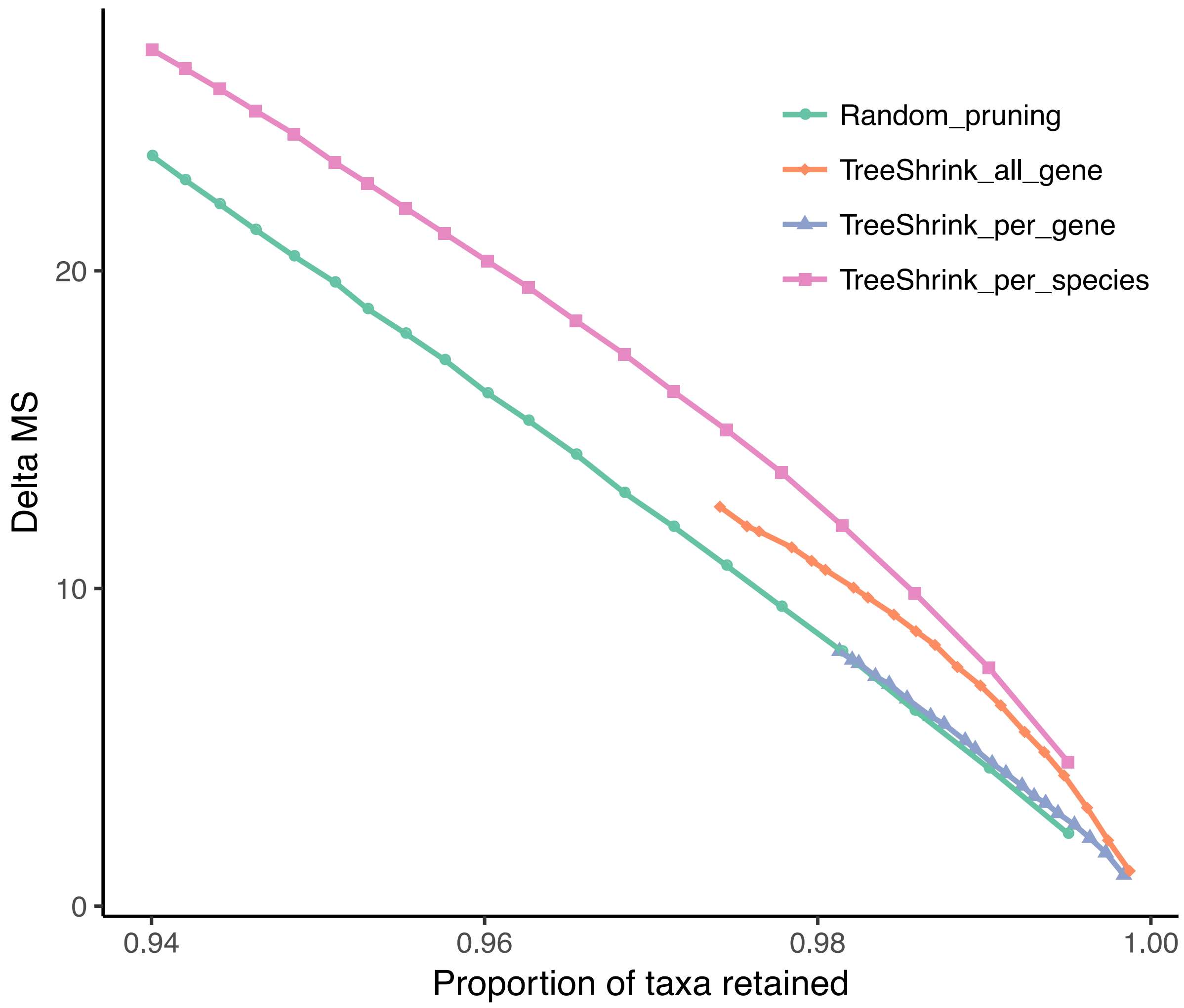
Mammalian dataset

Results: outgroup removal

- Percent of the data removed for α=*0.05* for
  - All species
  - Outgroups

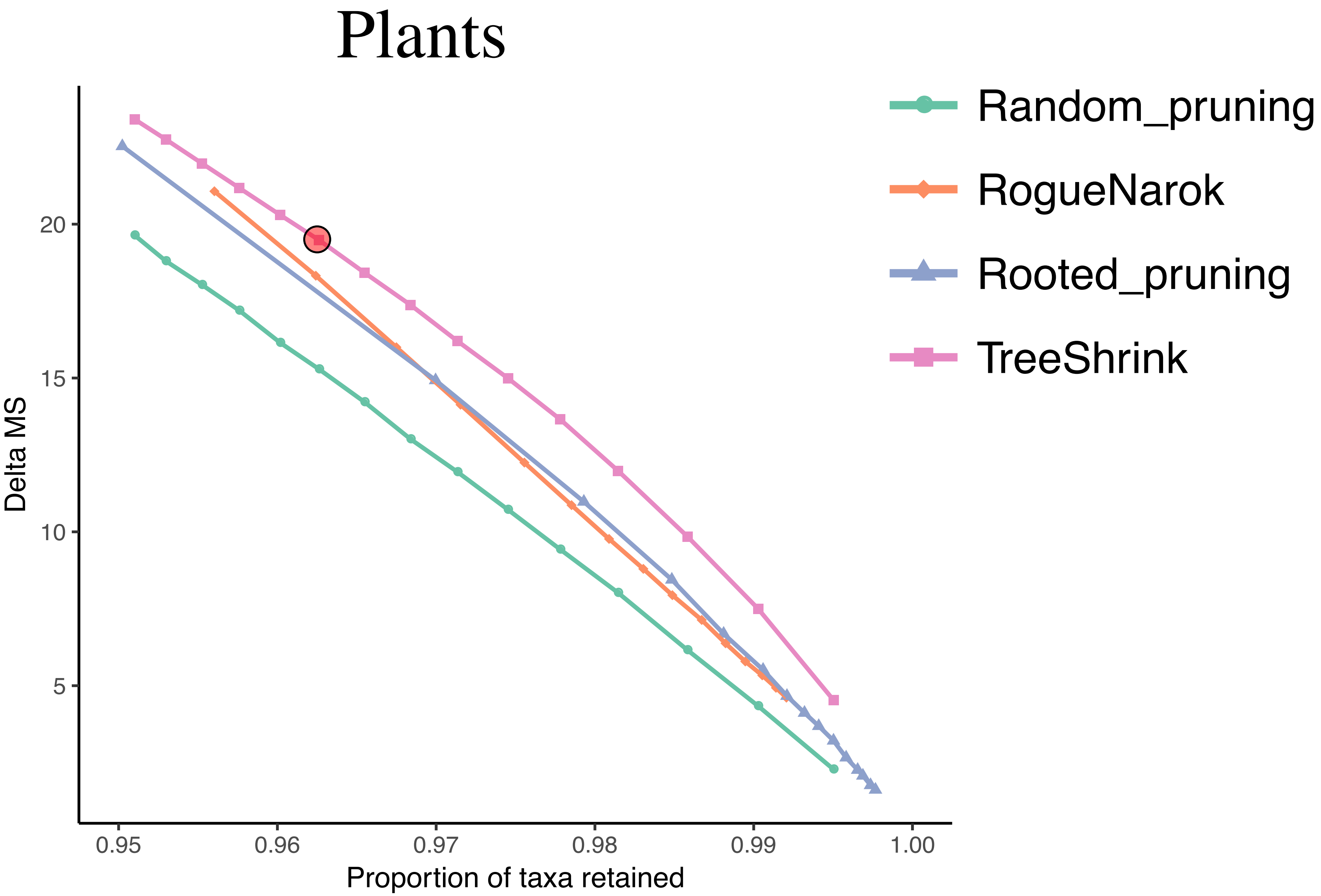


Impact of filtering on discordance

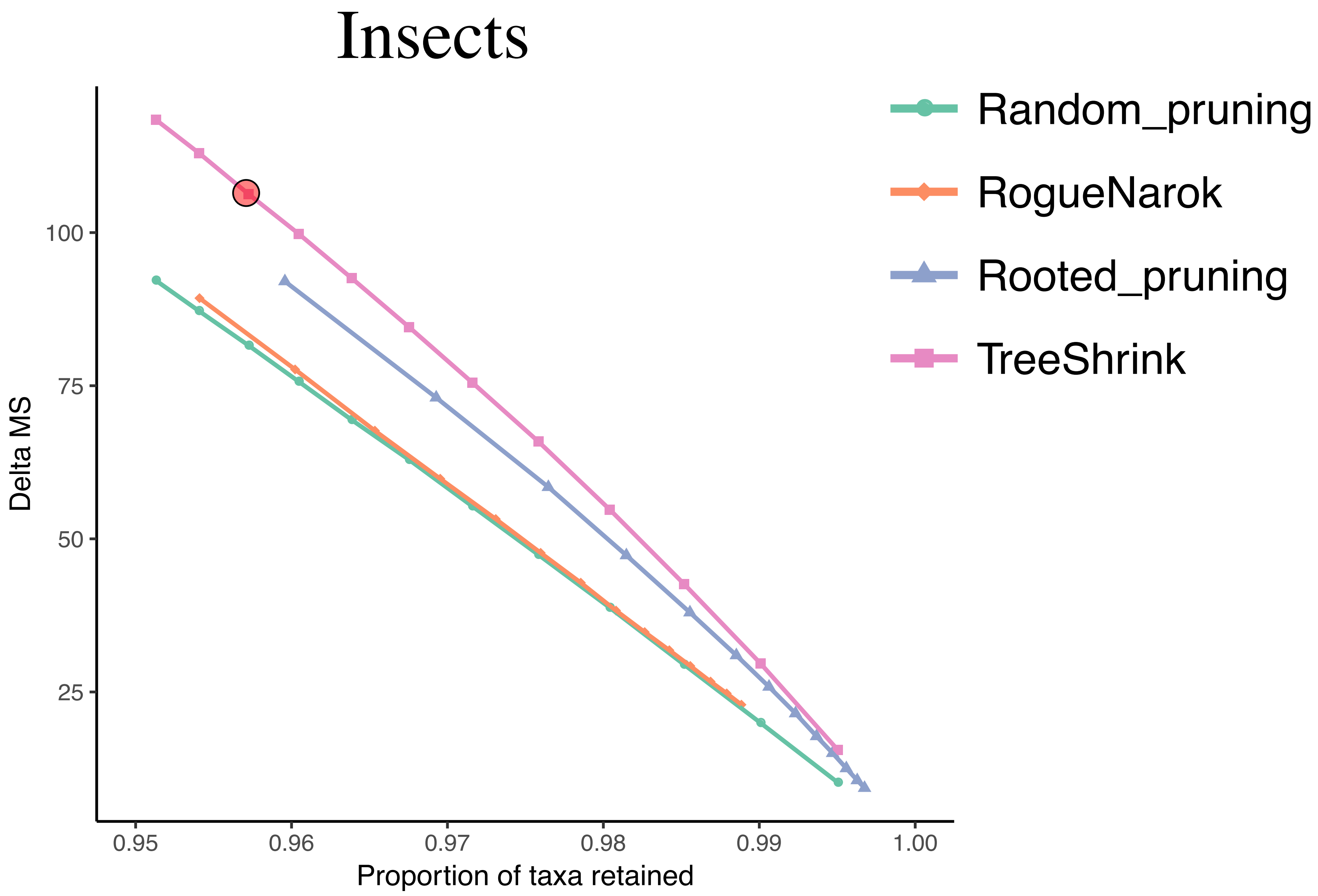


Plant dataset

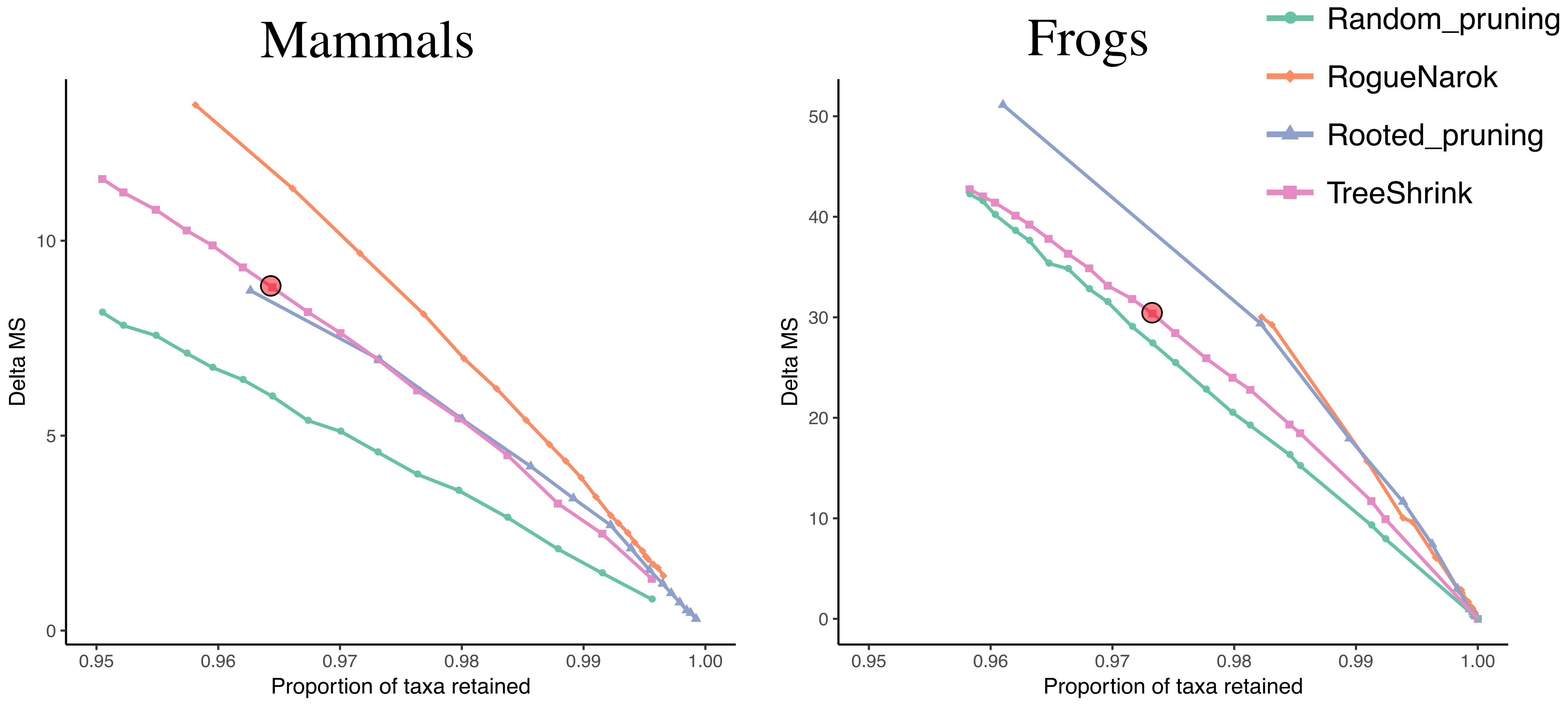
TreeShrink versus alternative methods (discordance)



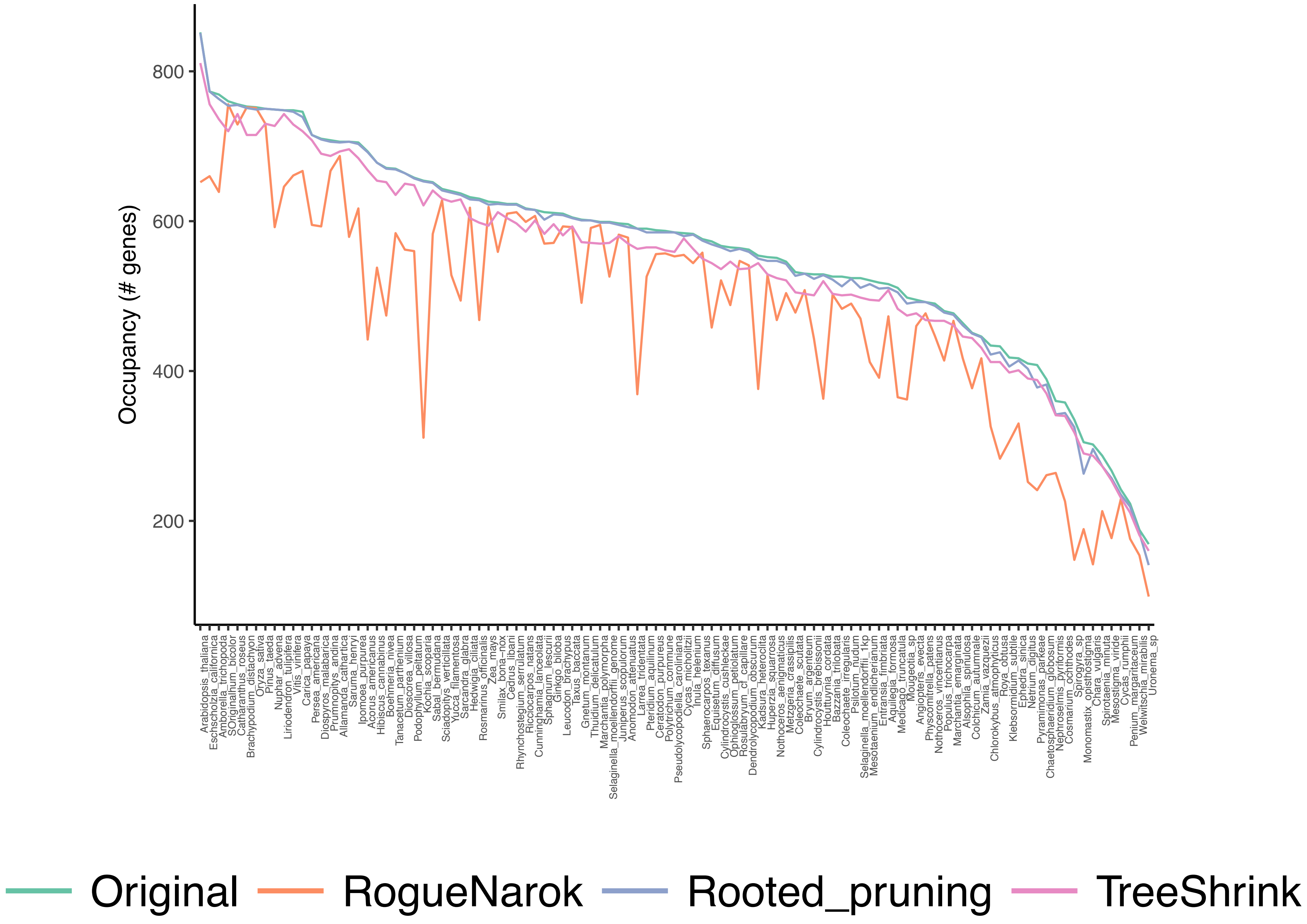
TreeShrink versus alternative methods (discordance)

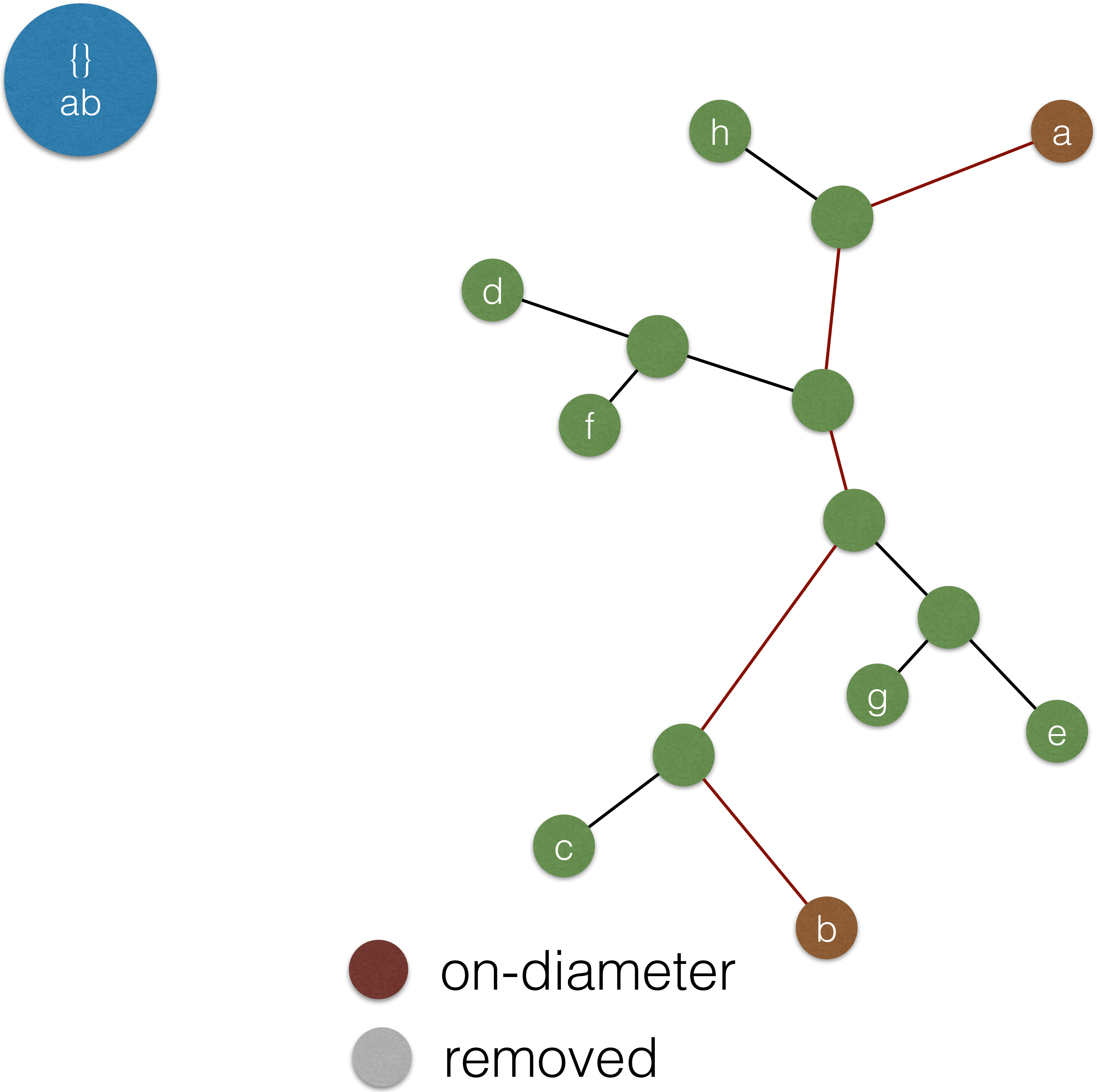


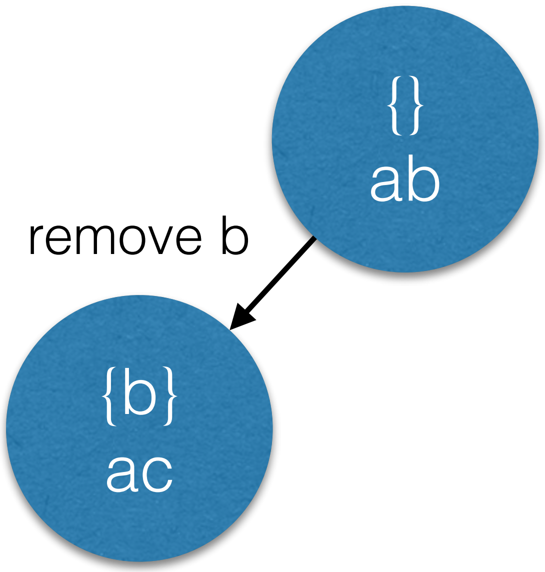
Results: TreeShrink versus Alternative Methods

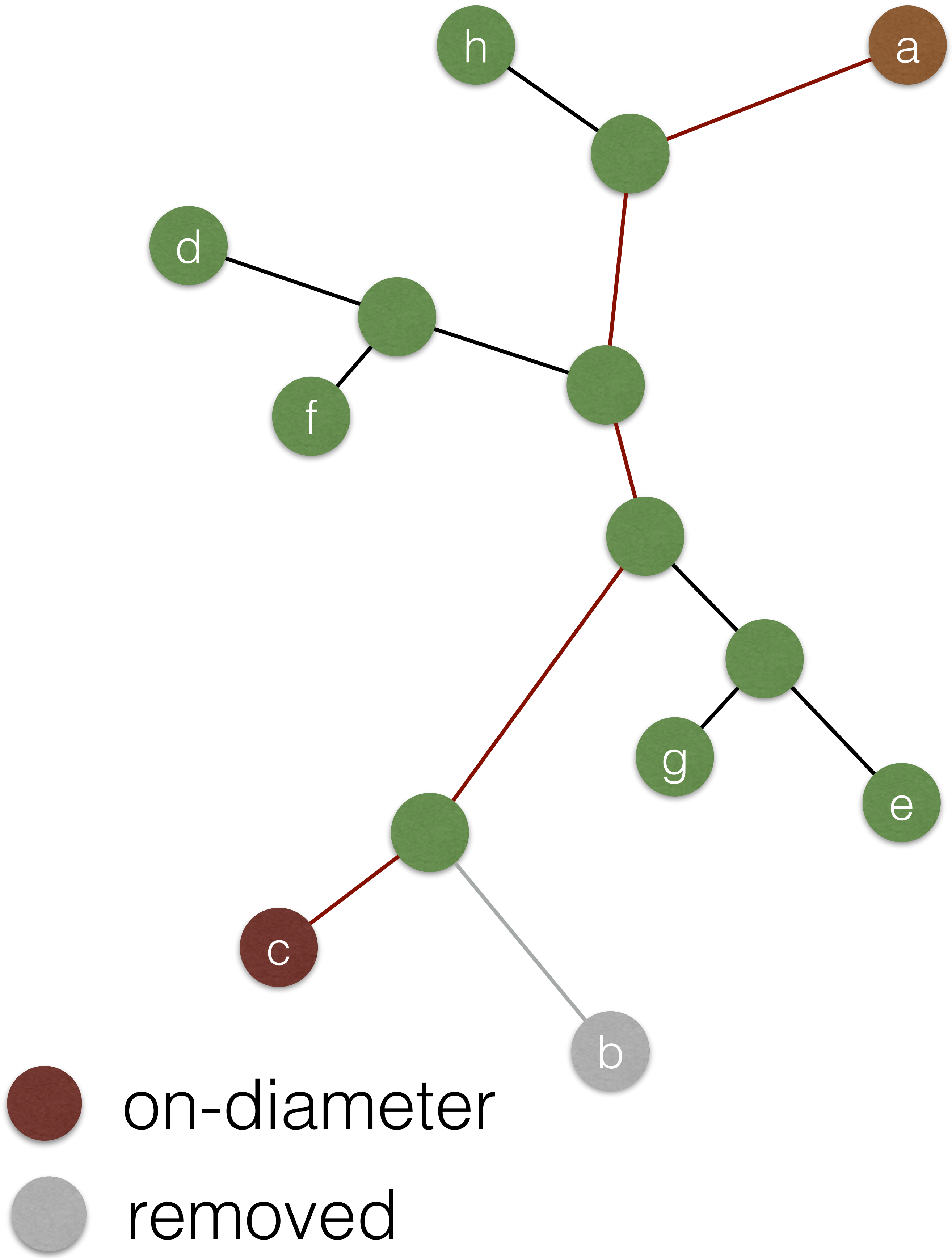


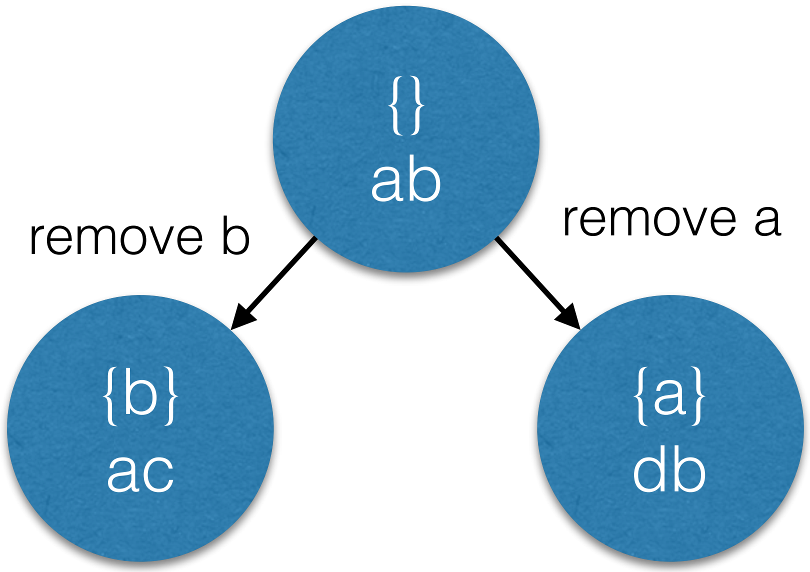
TreeShrink versus alternative methods (occupancy)

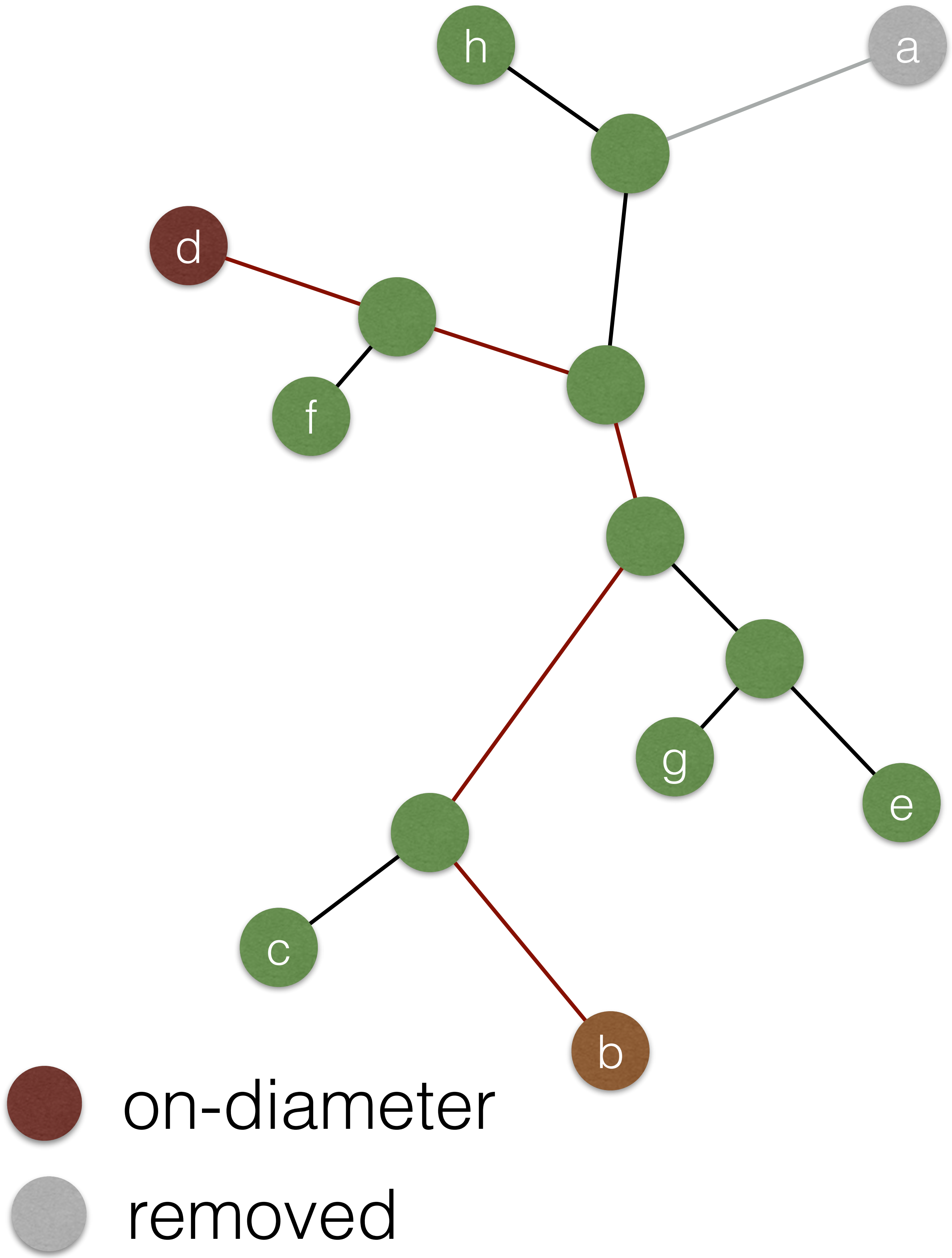


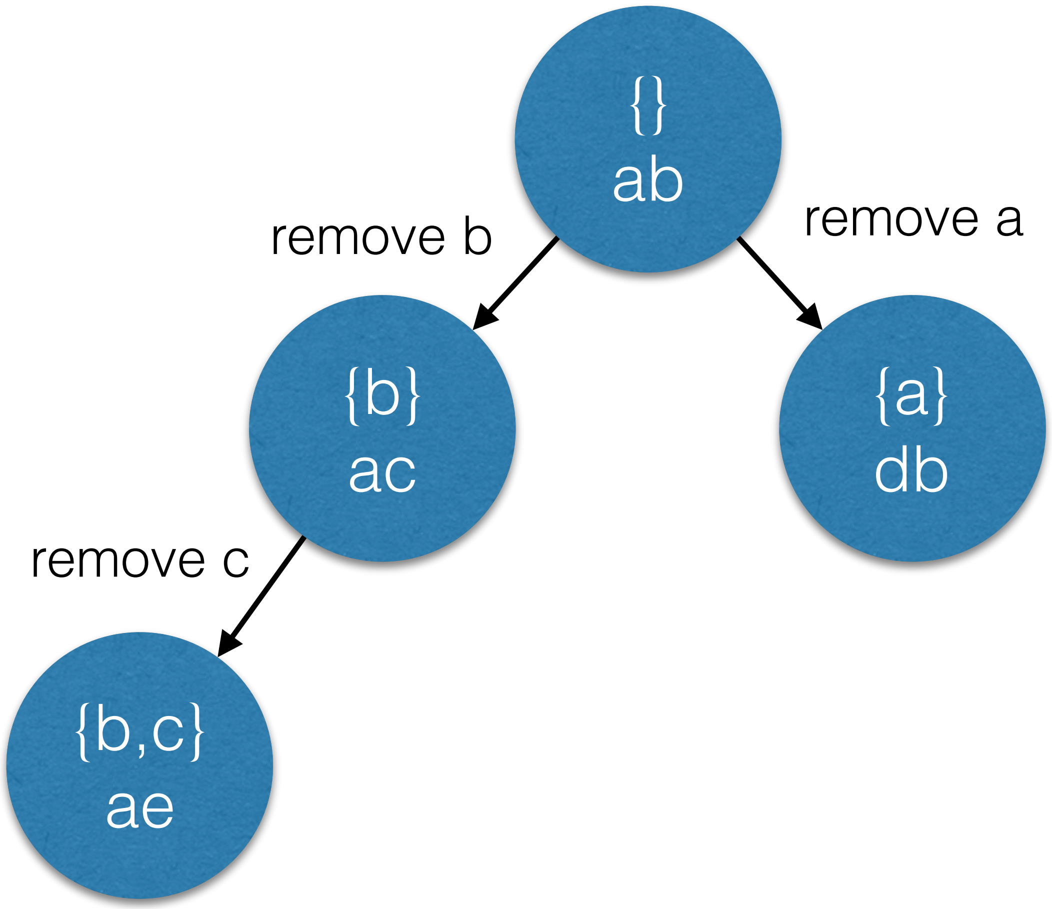

{}
ab
h
a
d
f
g
e
c
b
on-diameter
removed

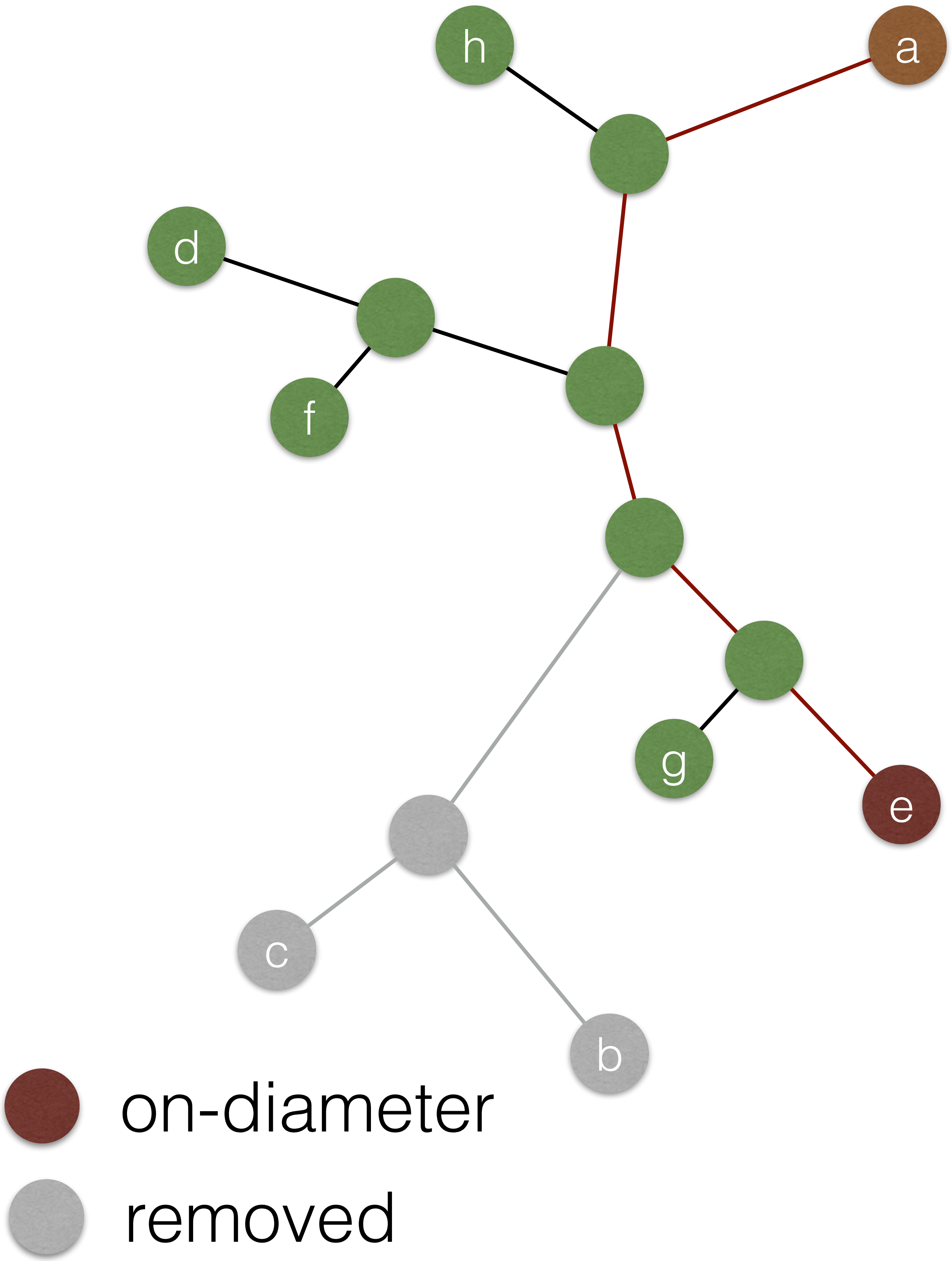

{}
ab
remove b
{b}
ac

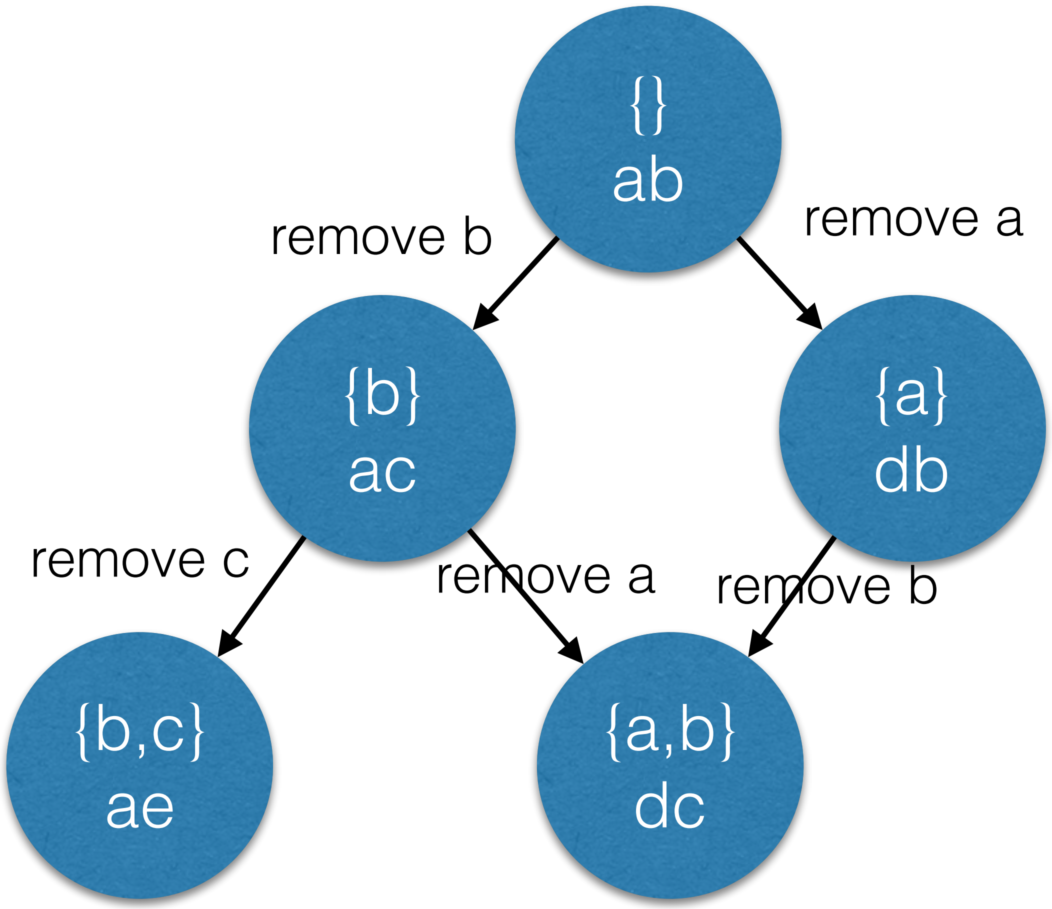

h
a
d
f
g
e
c
b
on-diameter
removed

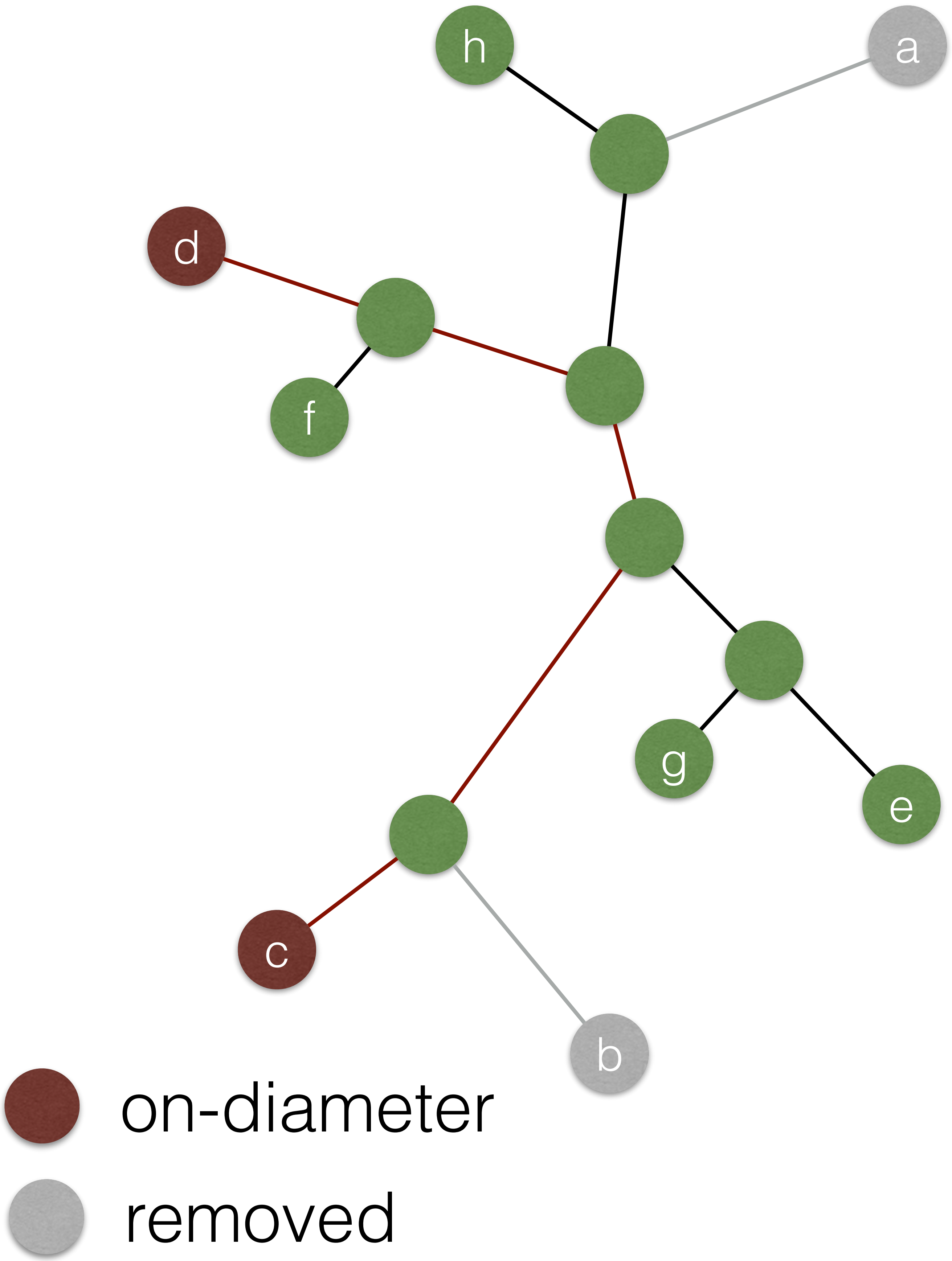

{}
ab
remove b
remove a
{b}
ac
{a}
db

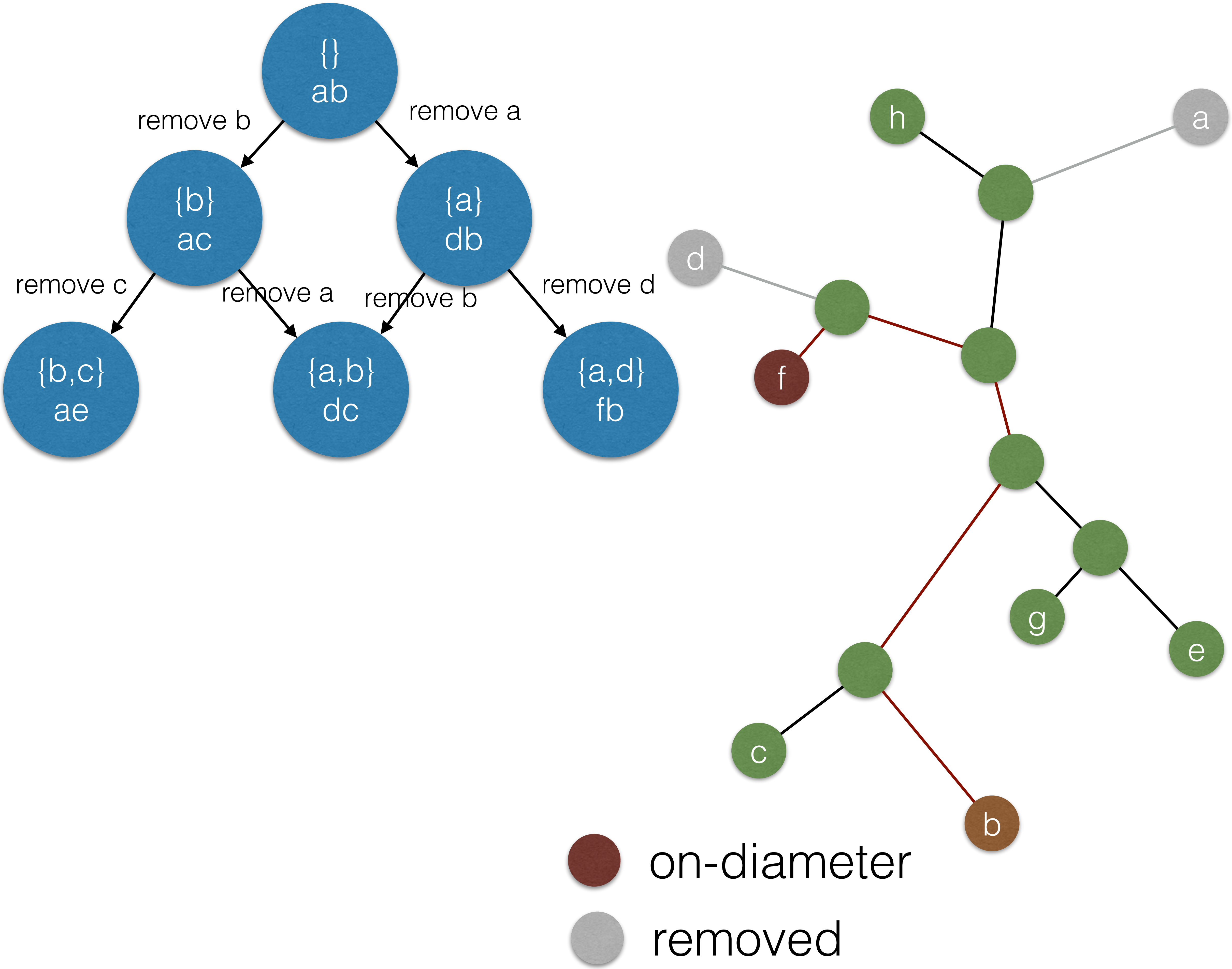

h
a
d
f
g
e
c
b
● on-diameter
● removed

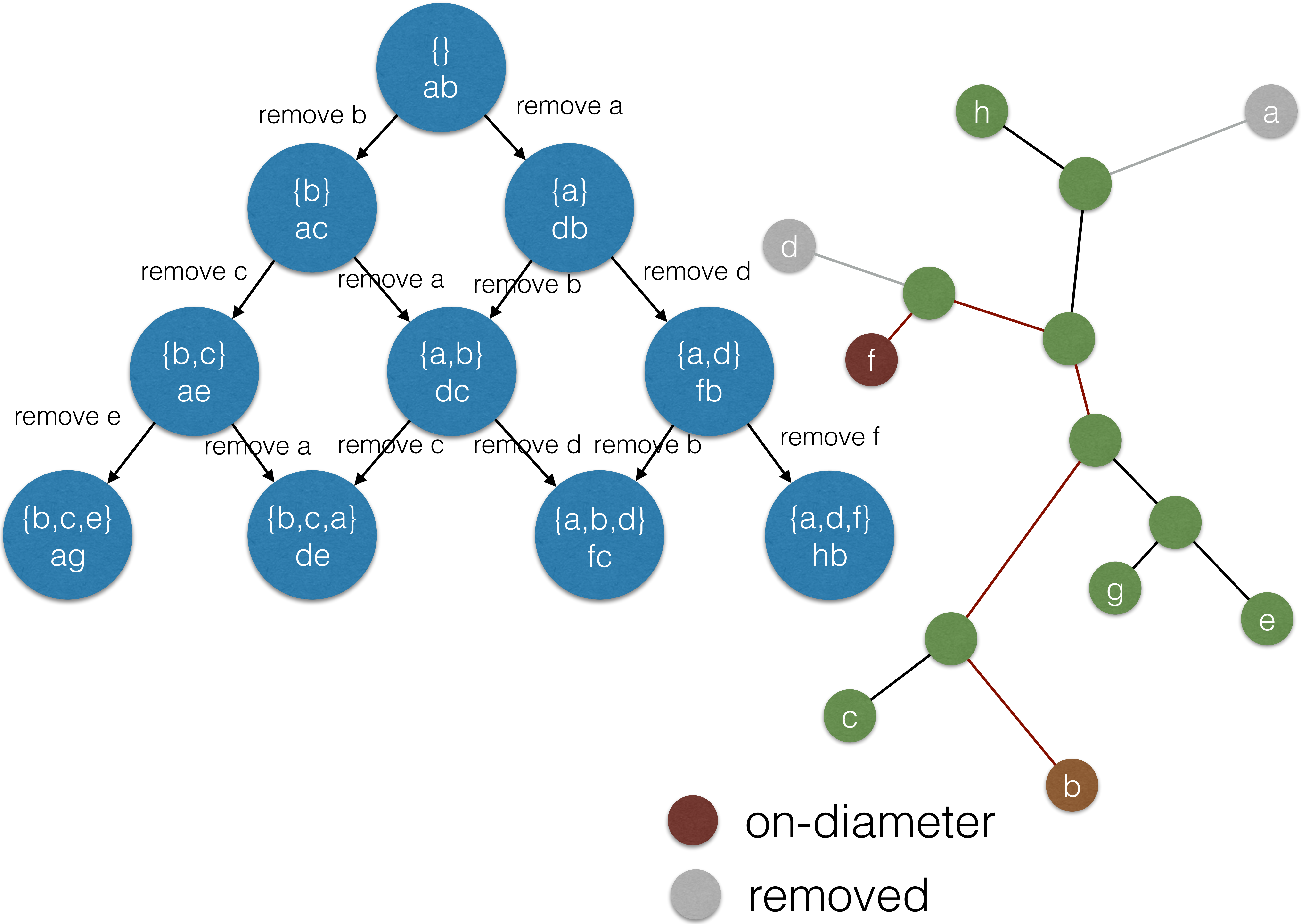

{}
ab
remove b
remove a
{b}
ac
{a}
db
remove c
{b,c}
ae


h
a
d
f
g
e
c
b
on-diameter
removed


{}
ab
remove b
remove a
{b}
ac
{a}
db
remove c
remove a
remove b
{b,c}
ae
{a,b}
dc


h
a
d
f
g
e
c
b
on-diameter
removed


{}
ab
remove b
remove a
{b}
ac
{a}
db
remove c
remove a
remove b
remove d
{b,c}
ae
{a,b}
dc
{a,d}
fb
h
a
d
f
g
e
c
b
● on-diameter
● removed


{}
ab
remove b
remove a
{b}
ac
{a}
db
remove c
remove a
remove b
remove d
{b,c}
ae
{a,b}
dc
{a,d}
fb
remove e
remove a
remove c
remove d
remove b
remove f
{b,c,e}
ag
{b,c,a}
de
{a,b,d}
fc
{a,d,f}
hb
h
a
d
f
g
e
c
b
● on-diameter
● removed

Solution space

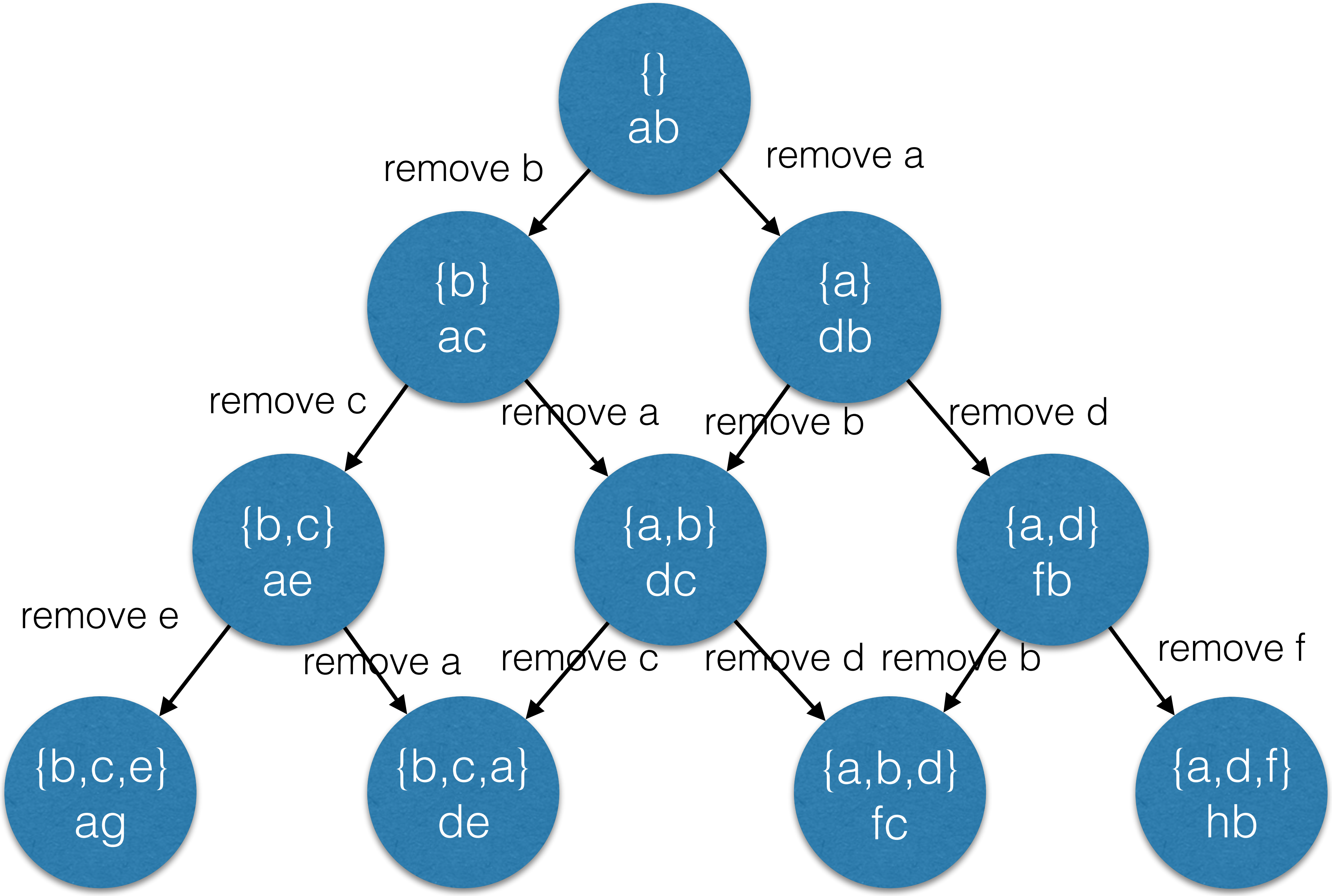
k = 3

i = 0

i = 1

i = 2

i = 3

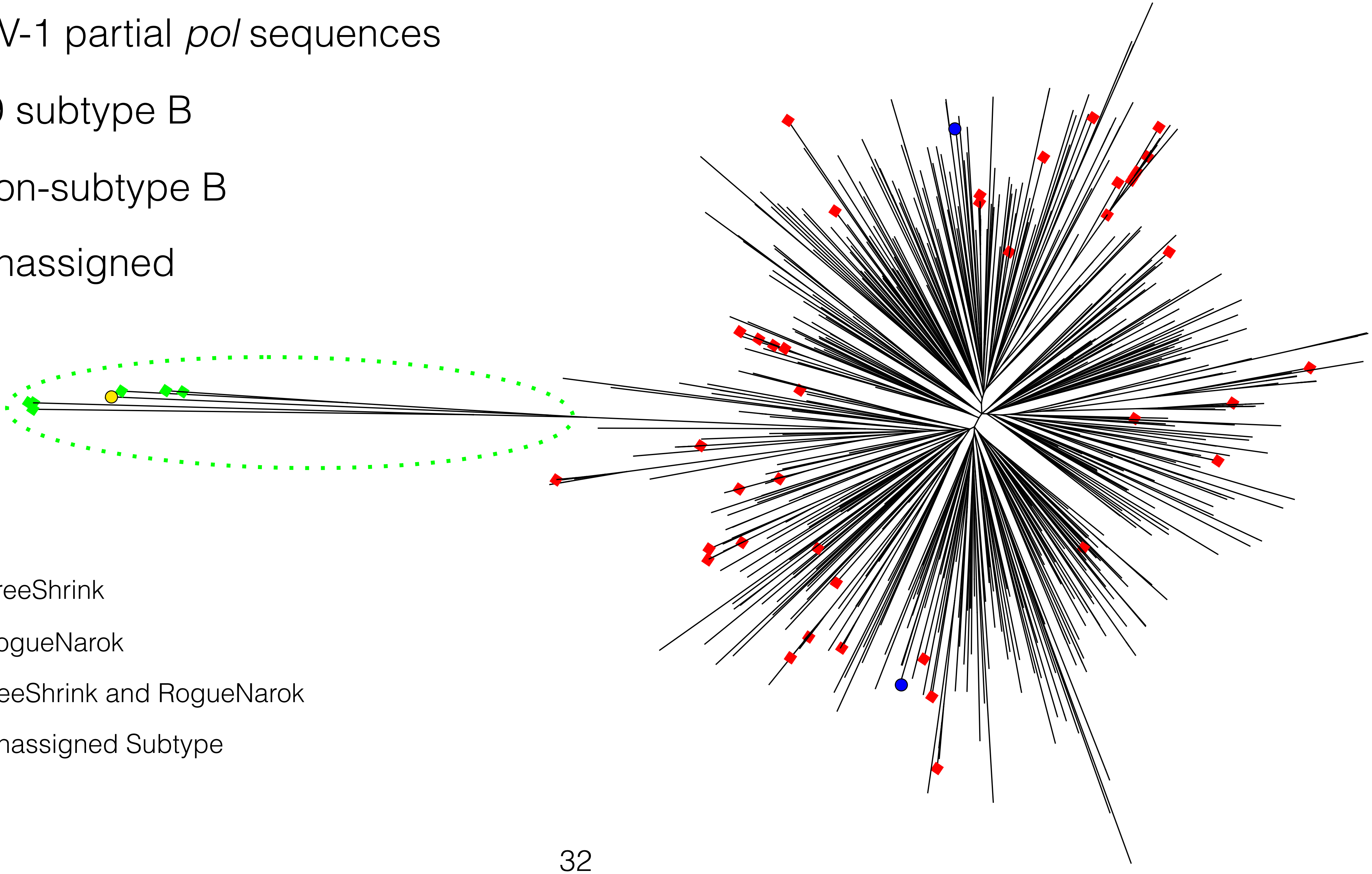


The TreeShrink tool is publicly available
https://github.com/uym2/TreeShrink

Uyen Mai

A single HIV tree

- 648 HIV-1 partial *pol* sequences
    - 639 subtype B
    - 7 non-subtype B
    - 2 unassigned



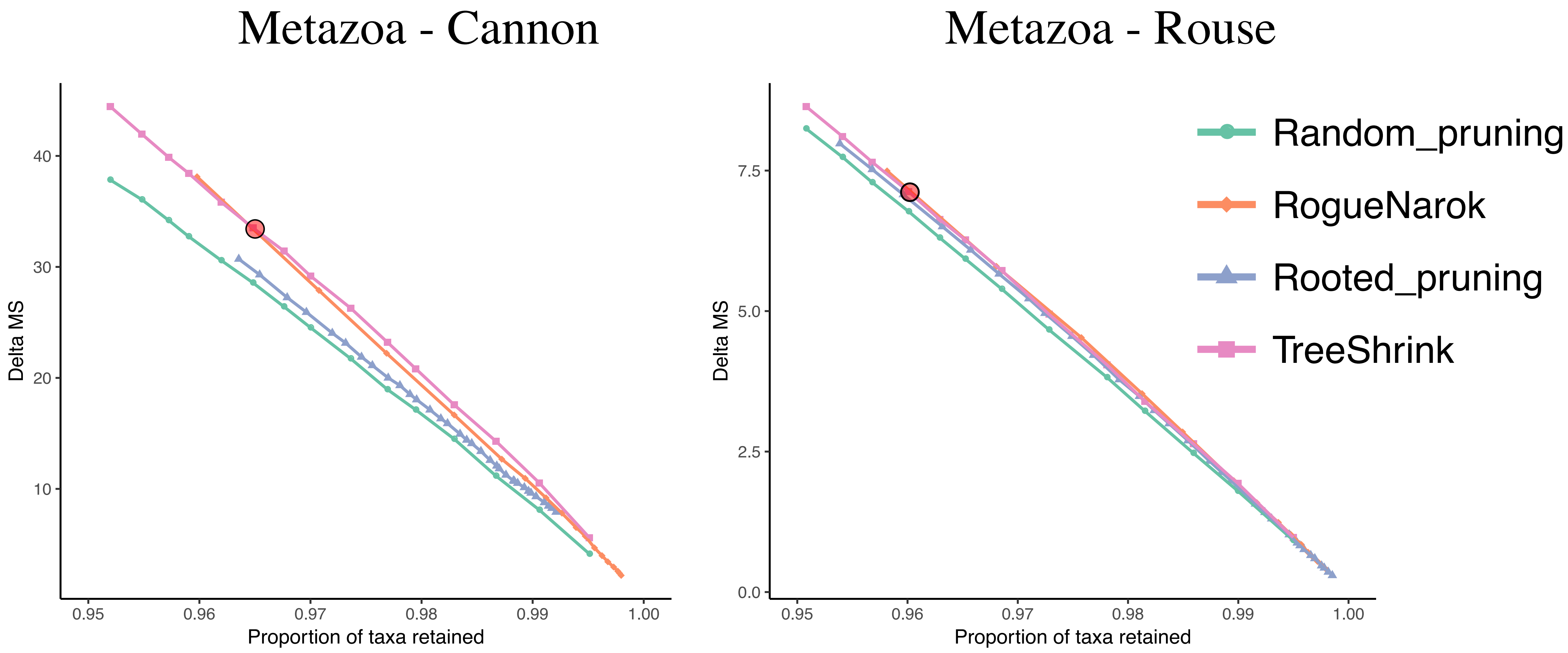
🟩 TreeShrink

🟥 RogueNarok

🟡 TreeShrink and RogueNarok

🔵 Unassigned Subtype

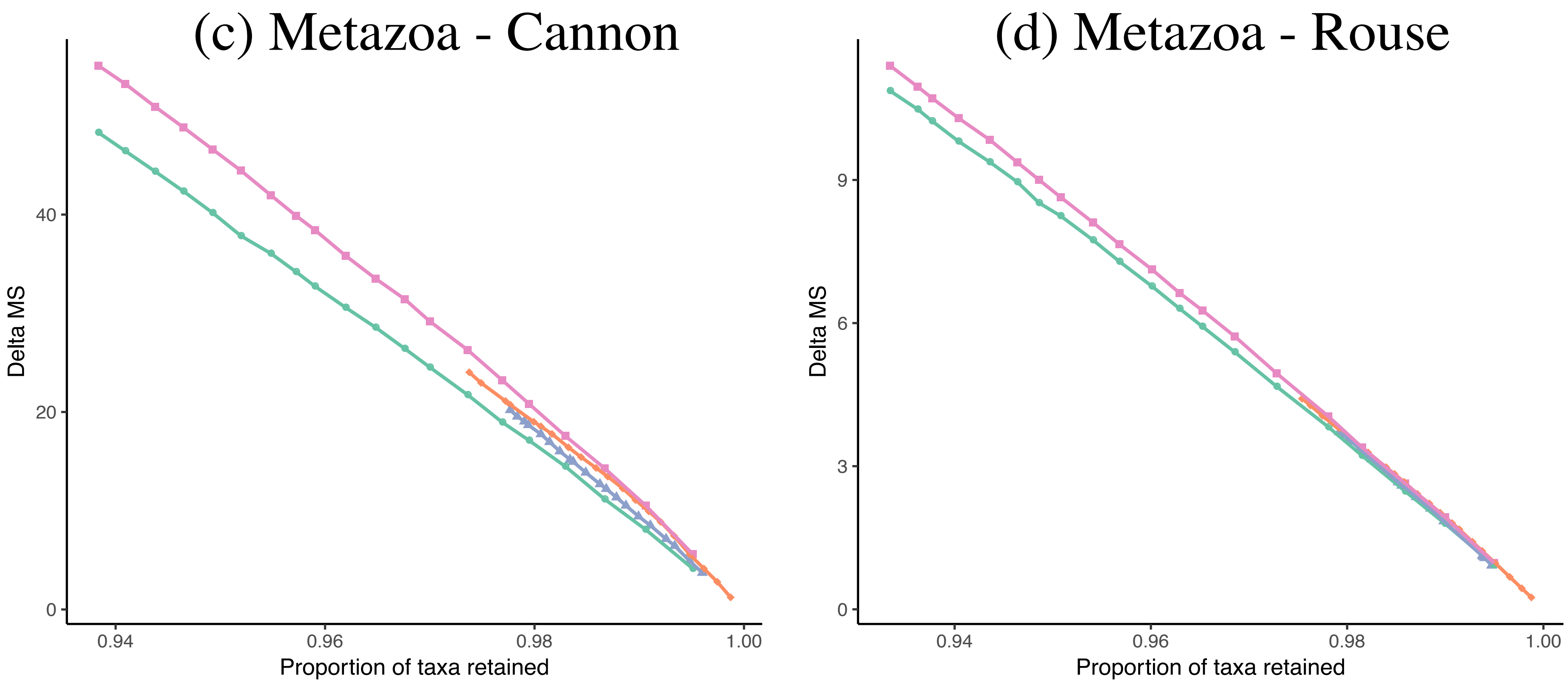
Results: TreeShrink versus Alternative Methods



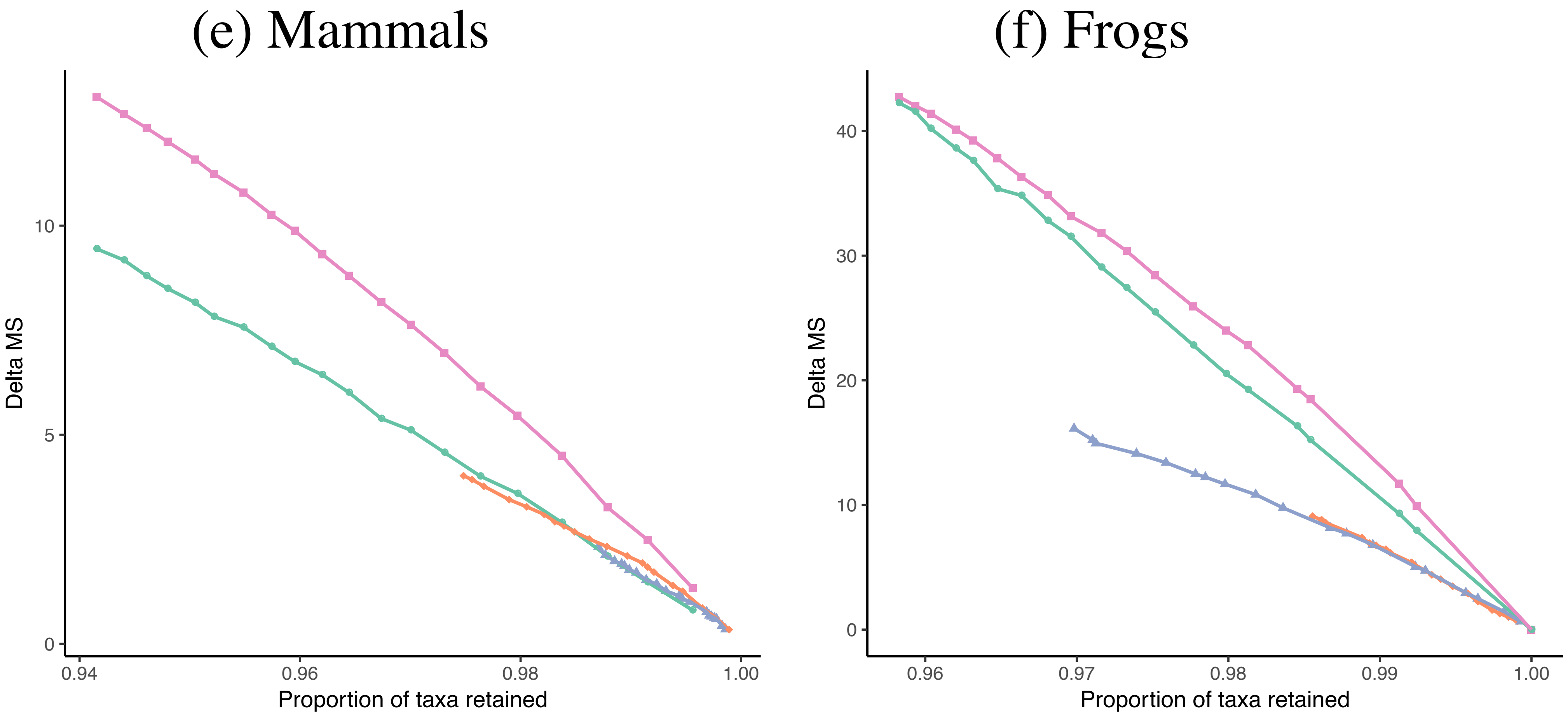
Metazoa - Cannon

Metazoa - Rouse

Results: The 3 Tests of TreeShrink

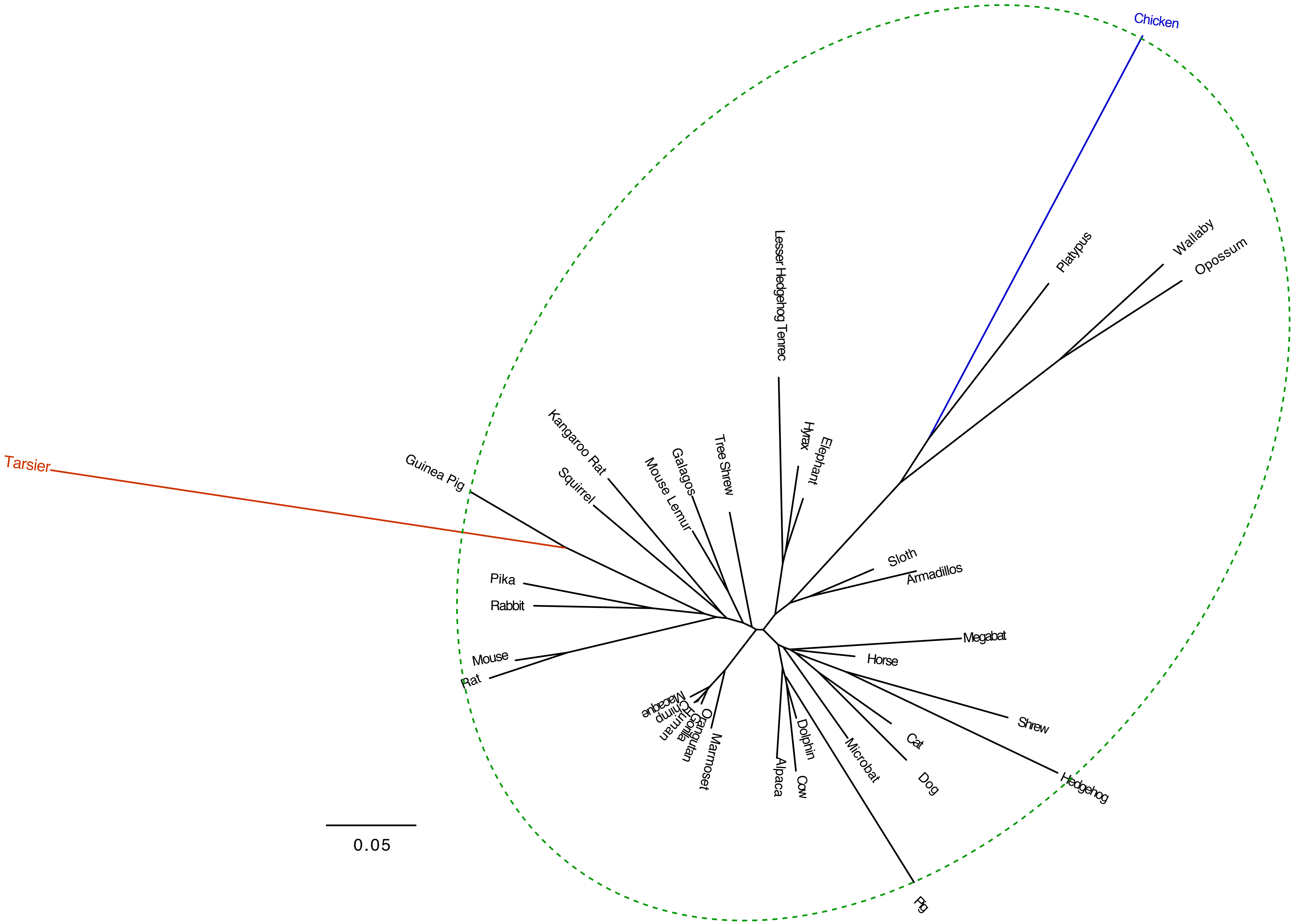


Results: The 3 Tests of TreeShrink



- Can be done in other ways too (e.g., $O(n.k+k^2 logk)$), but harder to implement

Can be just outgroups


Chicken
Platypus
Wallaby
Opossum
Lesser Hedgehog Tenrec
Hyrax
Elephant
Tree Shrew
Galagos
Mouse Lemur
Kangaroo Rat
Squirrel
Guinea Pig
Tarsier
Pika
Rabbit
Mouse
Rat
Sloth
Armadillos
Megabat
Horse
Shrew
Hedgehog
Cat
Dog
Microbat
Dolphin
Cow
Alpaca
Pig
Marmoset
Orangutan
Gorilla
Human
Chimp
Macaque
0.05