Introduction to the multispecies coalescent

Outline

1. Background
   --gene trees vs. species trees
   --coalescence and incomplete lineage sorting

2. Gene tree distributions and anomalous gene trees

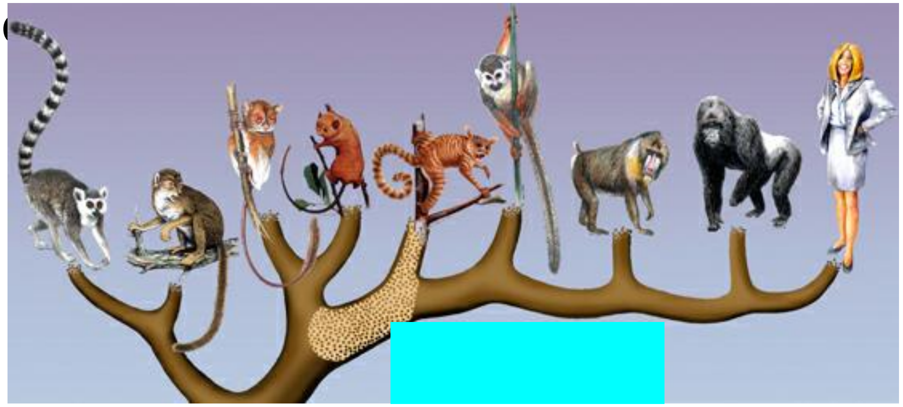
3. Inferring species trees
   a. Concatenation
   b. Consensus trees

4. Conclusions

Population Genetics and Phylogenetics

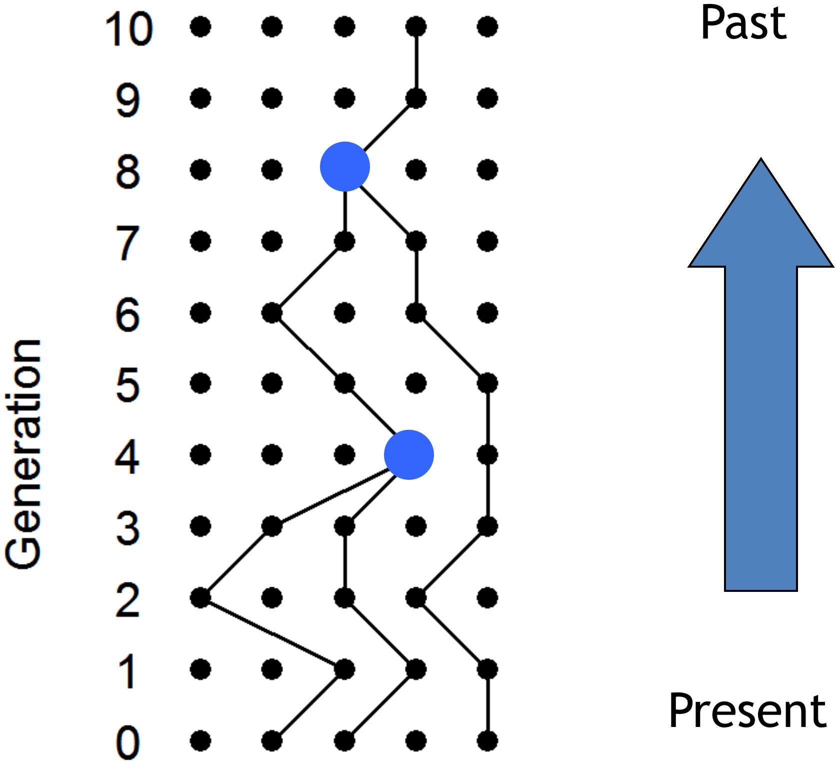
Population genetics: traditionally used to analyze single populations.

Phylogenetics: What is the best way to infer relationships between p
species?



Graphic by Mark A. Klinger, Carnegie Museum of Natural History, Pittsburgh

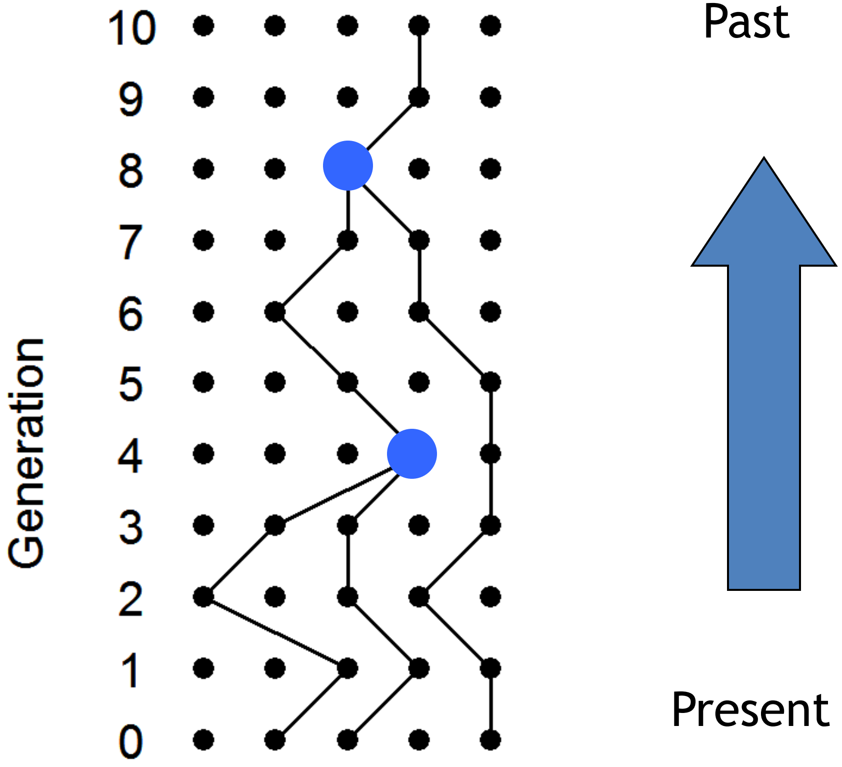
The coalescent process



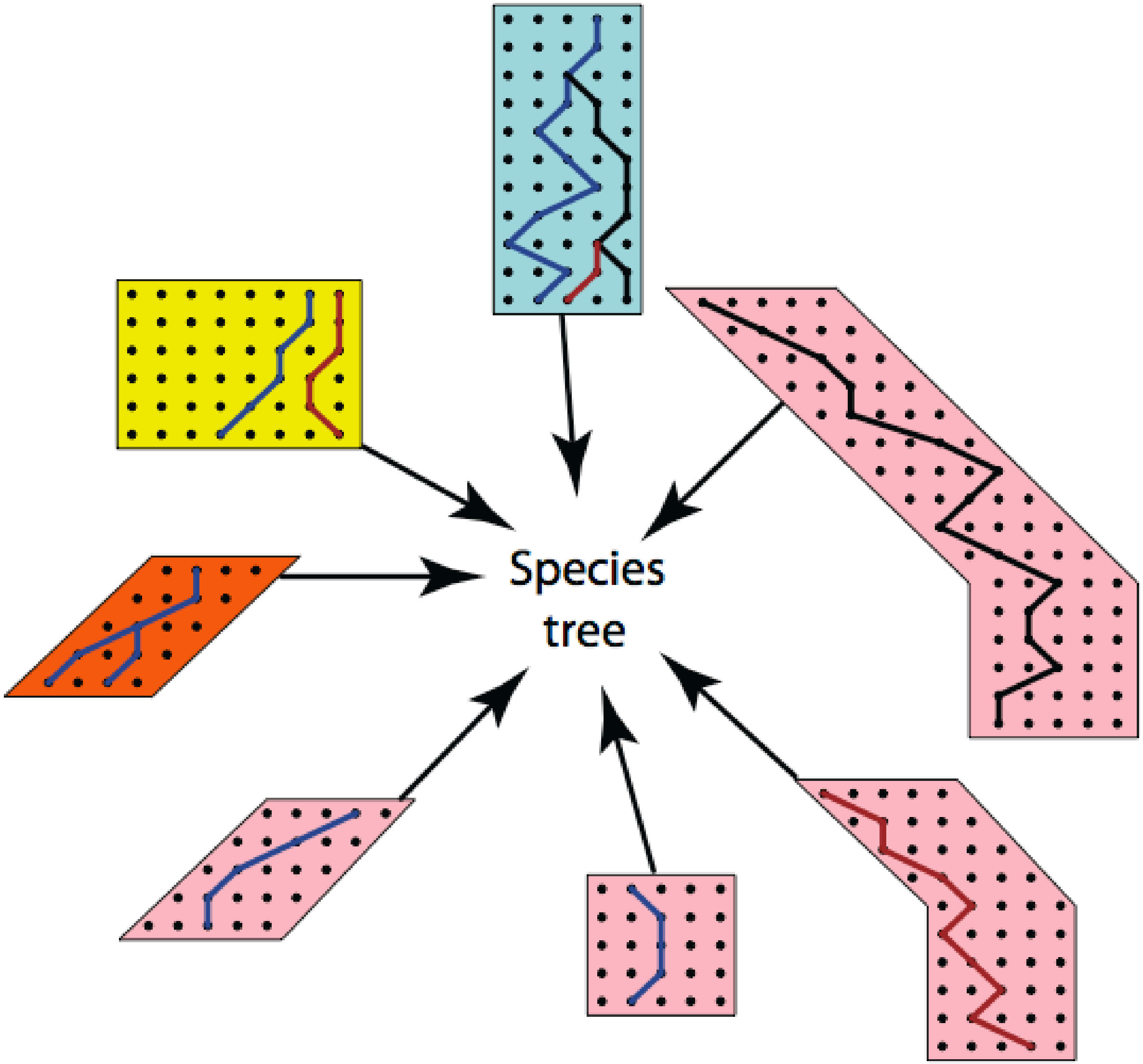
Past

Present

One population

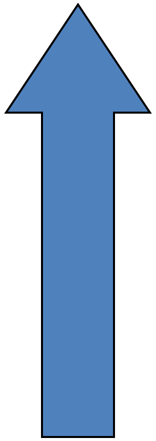


Model for lineages in populations

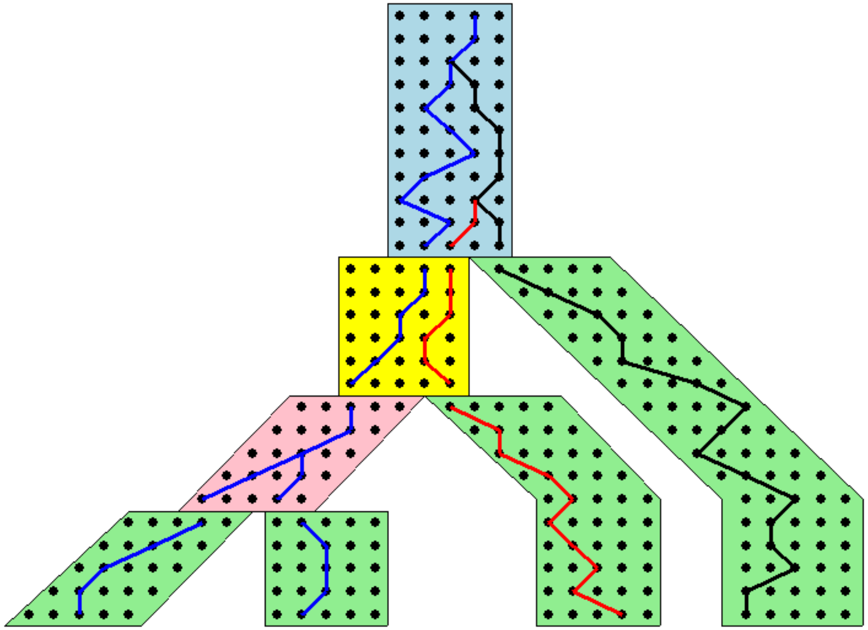


Multiple populations/species

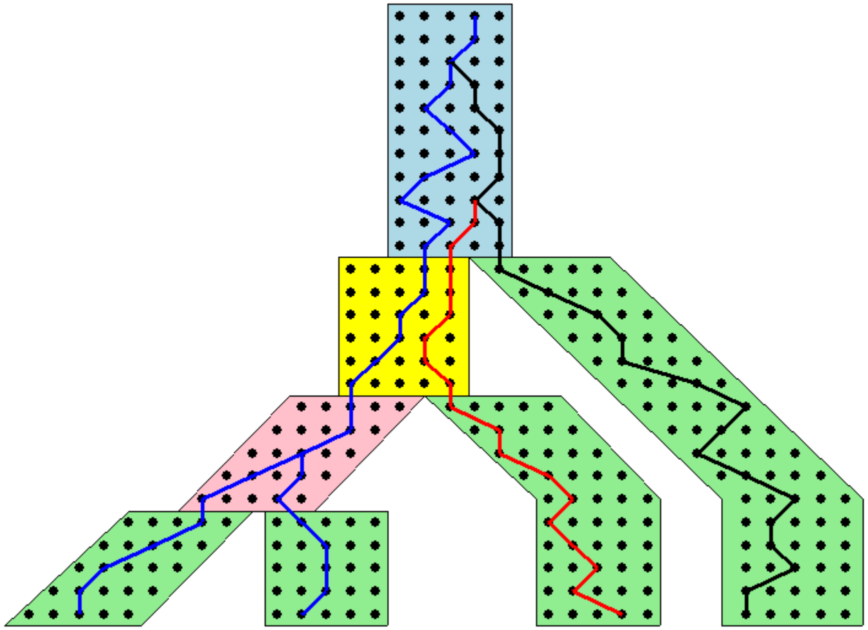
Past



Present

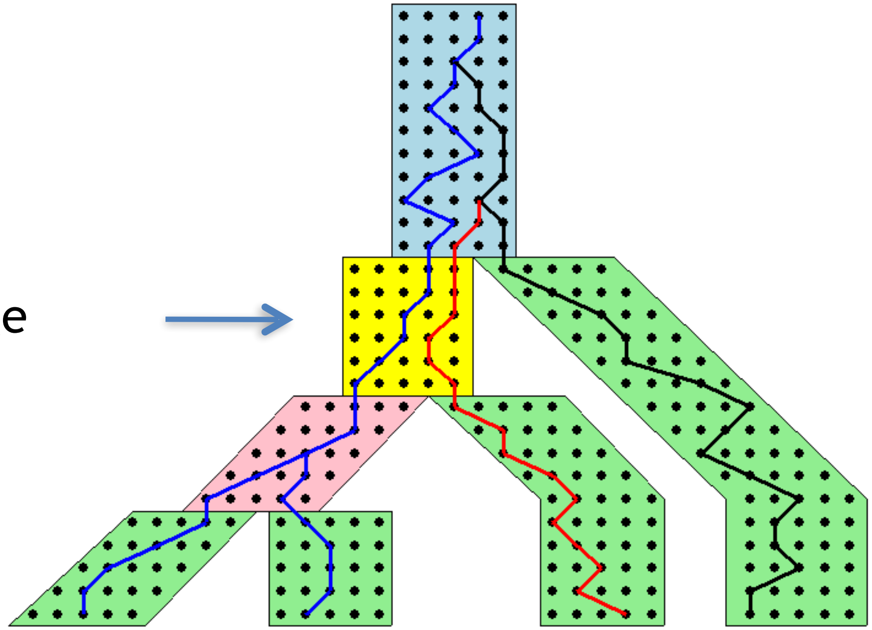


Gene tree in a species tree

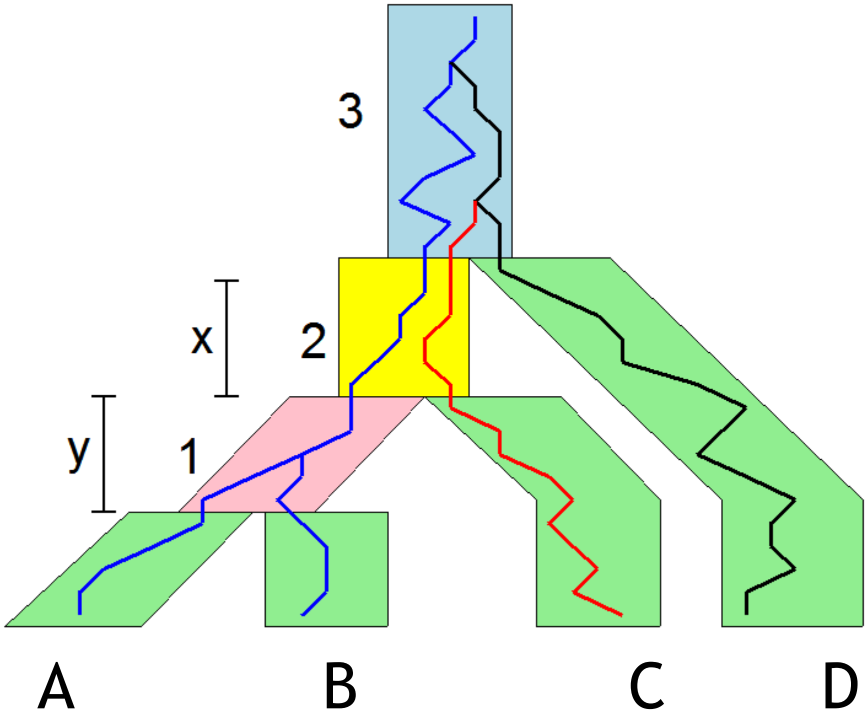


Gene tree in a species tree

Incomplete lineage sorting



Gene tree in a species tree



The gene tree is a random variable. The gene tree distribution is parameterized by the species tree topology and internal branch lengths.

How can we compute probabilities of gene trees given species trees?

-Under a coalescent model, probabilities for gene trees with three species were derived by Nei (1987): $1-(2/3)e^{-T}$

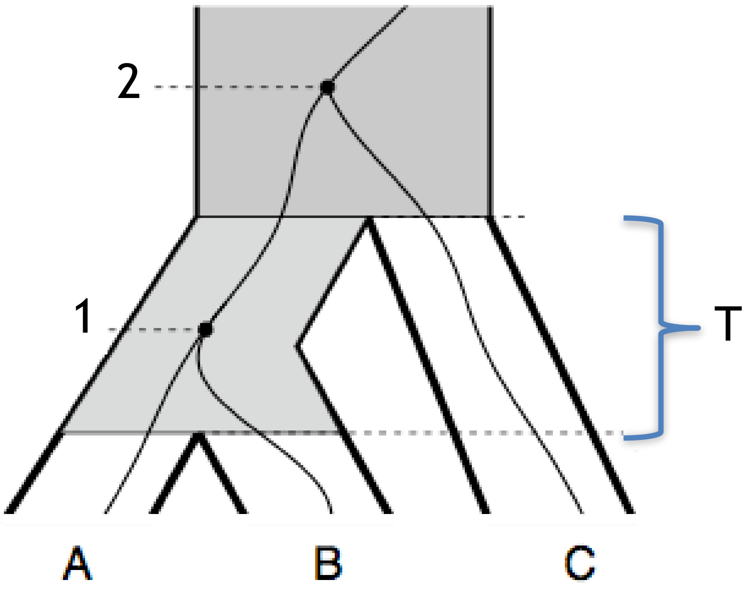
-Probabilities for the gene tree to match the species tree topology for 4 and 5 species given by Pamilo and Nei (1988).

-All 30 species tree/gene tree combinations for 4 species given by Rosenberg (2002).

-General case implemented by program phylonet (Nakhleh et al.) and hybrid-coal (Zhu et al., 2017)

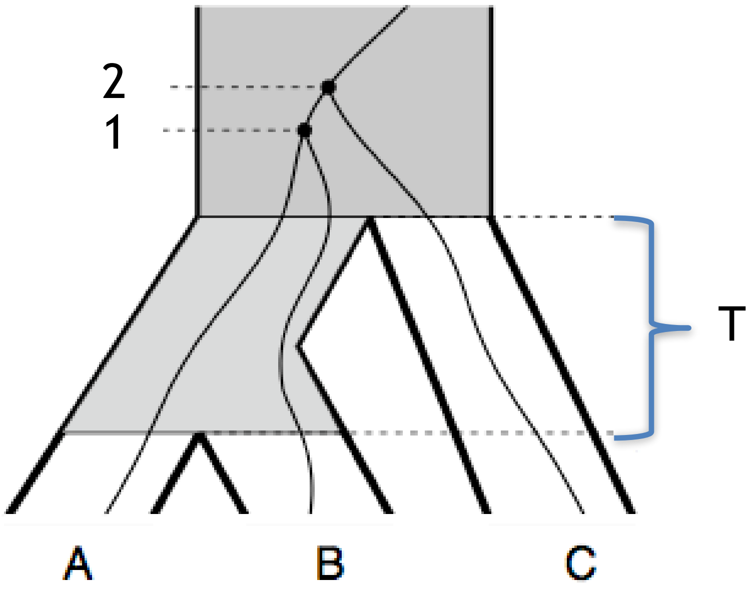
Coalescent histories as cases

Probability that the gene tree matches the species tree for three taxa



Probability:

$$\Pr[X \le T] = 1 - e^{-T}$$



Probability:

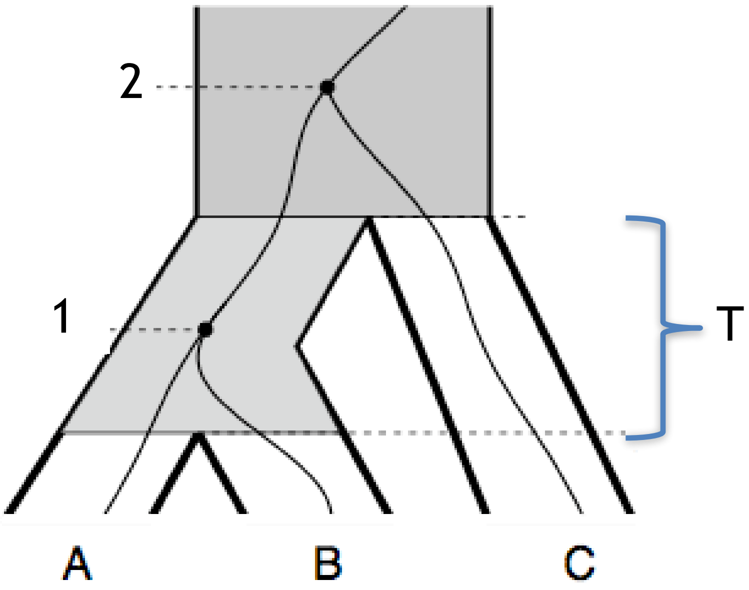
$$(1/3)\Pr[X > T] = (1/3)e^{-T}$$

Coalescent histories as cases

Probability that the gene tree matches the species tree for three taxa

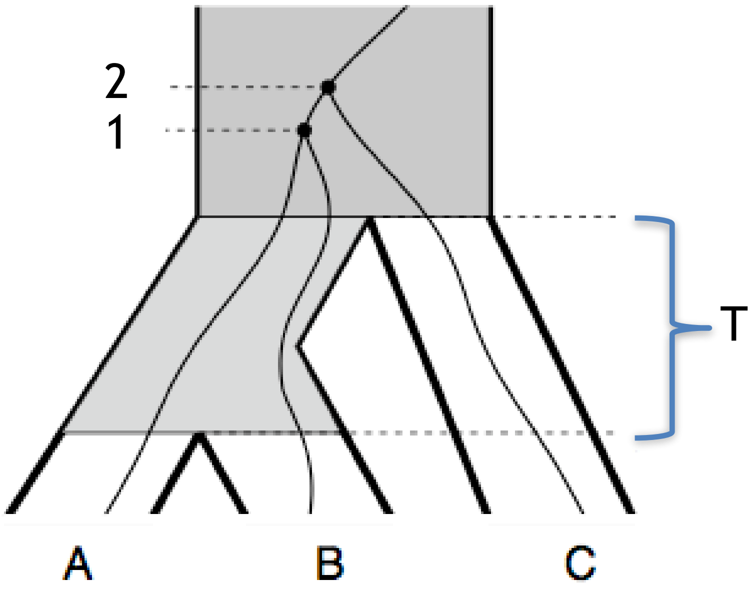
History: (1,2)

History: (2,2)



Pop 2

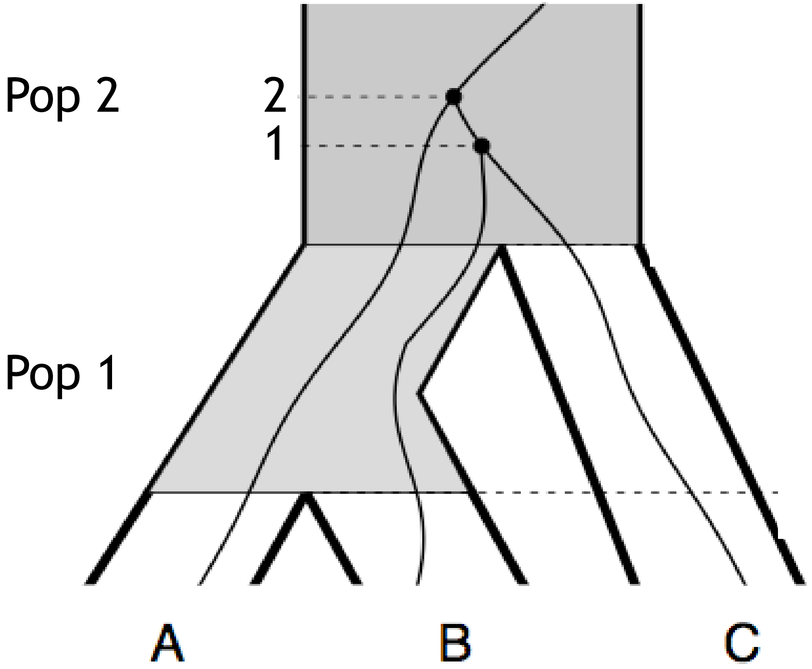
Pop 1



Total probability that the gene tree matches the species tree:

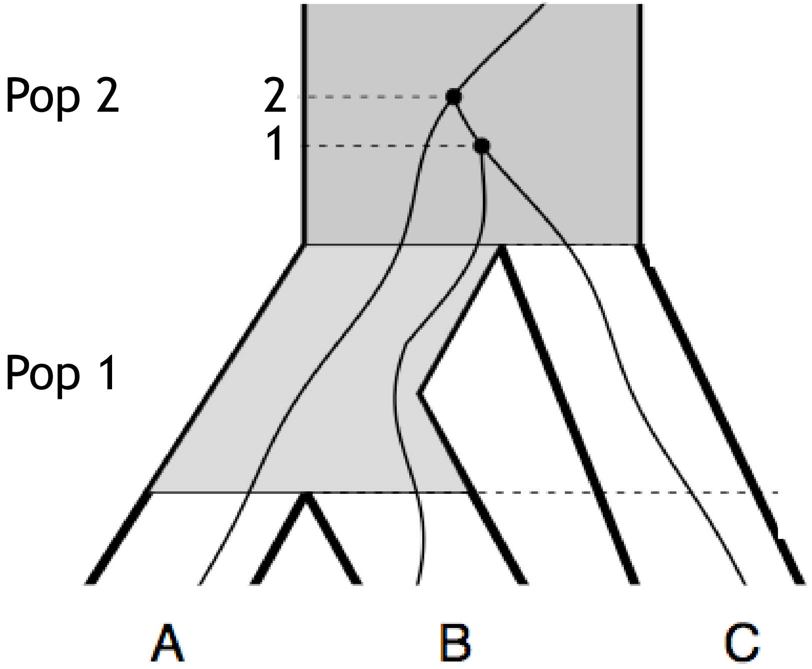
$$1 - e^{-T} + (1/3)e^{-T} = 1 - (2/3)e^{-T} > 1/3$$

Probability of a nonmatching tree: only one coalescent history (2,2)



$$(1/3)\Pr[X > T] = (1/3)e^{-T} < 1/3$$

Probability of a nonmatching tree: only one coalescent history (2,2)



$$(1/3)\Pr[X > T] = (1/3)e^{-T} < 1/3$$

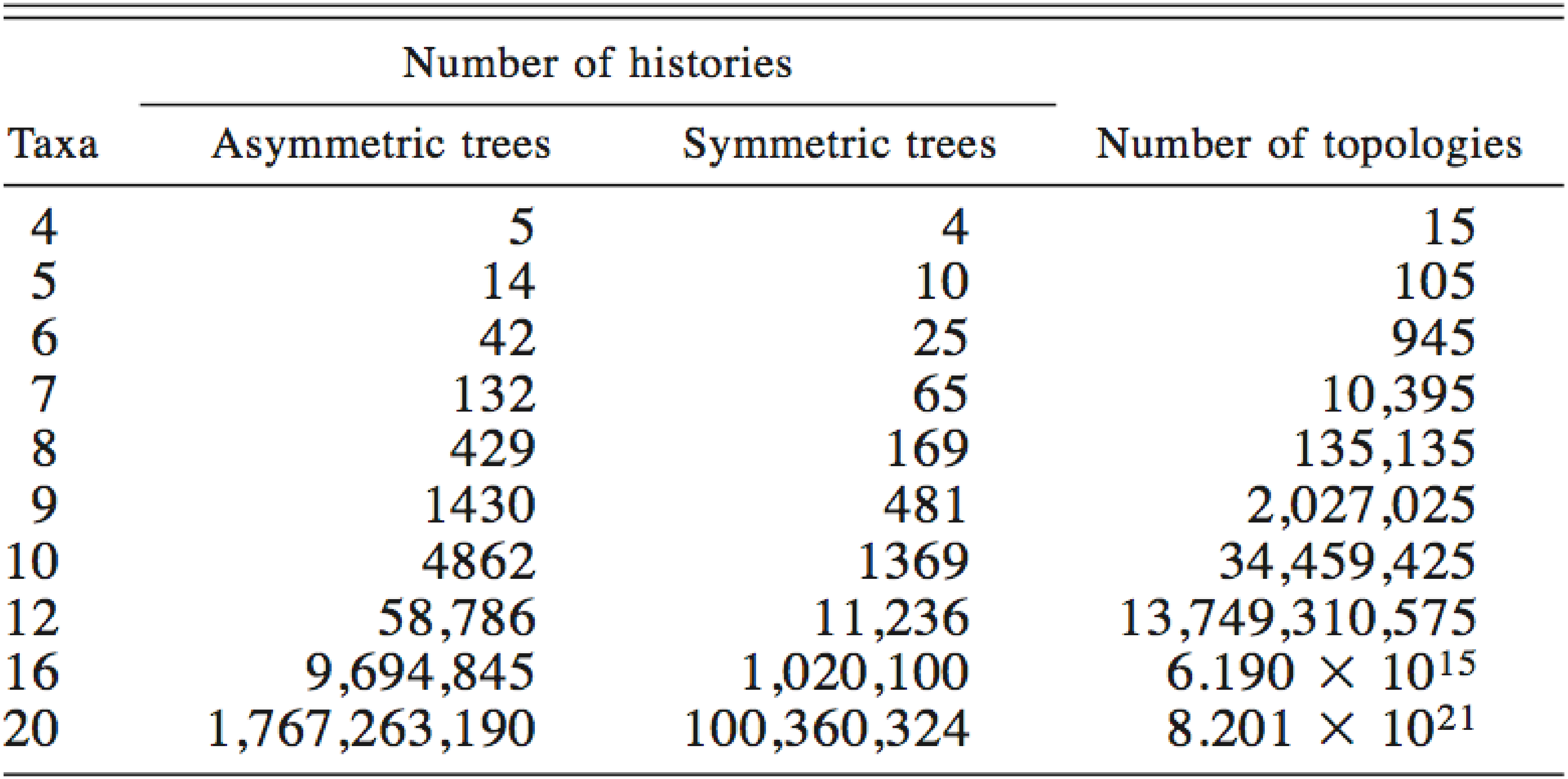
This is always less than the probability of the matching gene tree.

How do we get probabilities of gene trees with more taxa?

Gene tree probabilities

$$\Pr[G \mid S] = \sum_{histories} \Pr[G, histories \mid S]$$

How many coalescent histories?



	Number of histories		
Taxa	Asymmetric trees	Symmetric trees	Number of topologies
4	5	4	15
5	14	10	105
6	42	25	945
7	132	65	10,395
8	429	169	135,135
9	1430	481	2,027,025
10	4862	1369	34,459,425
12	58,786	11,236	13,749,310,575
16	9,694,845	1,020,100	6.190×10^{15}
20	1,767,263,190	100,360,324	8.201×10^{21}

Gene tree probabilities

$$\Pr[G = g \mid S] = \sum_{histories} \Pr[G = g, histories \mid S]$$

$$= \sum_{histories} \prod_{b} w_b \, p_{u(b),v(b)}(T_b)$$

combinatorial enumeration, complexity only known in special cases

u coalesce into v

branch length

internal branches of S

probability coalescences are consistent with g

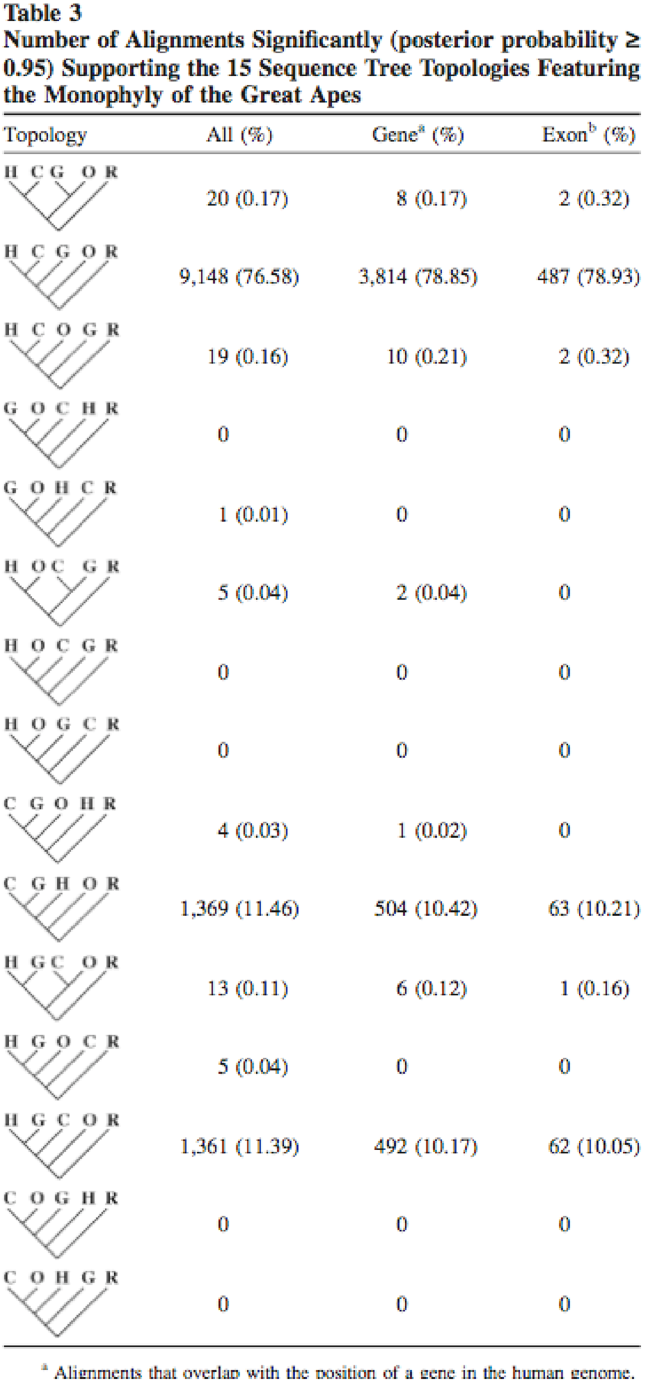


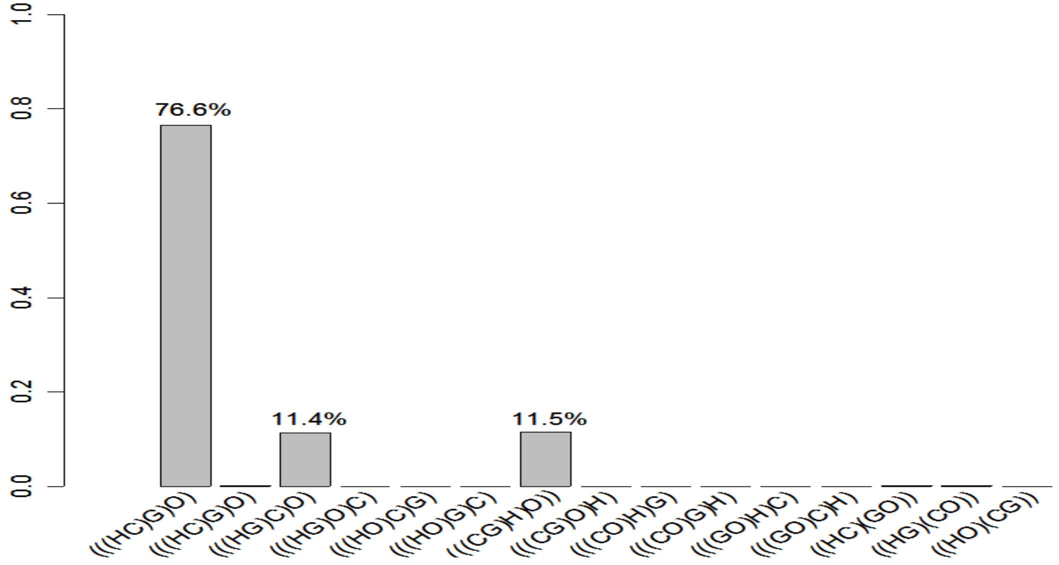
Table 3
Number of Alignments Significantly (posterior probability ≥ 0.95) Supporting the 15 Sequence Tree Topologies Featuring the Monophyly of the Great Apes

Topology	All (%)	Gene[a] (%)	Exon[b] (%)
H C G O R	20 (0.17)	8 (0.17)	2 (0.32)
H C G O R	9,148 (76.58)	3,814 (78.85)	487 (78.93)
H C O G R	19 (0.16)	10 (0.21)	2 (0.32)
G O C H R	0	0	0
G O H C R	1 (0.01)	0	0
H O C G R	5 (0.04)	2 (0.04)	0
H O C G R	0	0	0
H O G C R	0	0	0
C G O H R	4 (0.03)	1 (0.02)	0
C G H O R	1,369 (11.46)	504 (10.42)	63 (10.21)
H G C O R	13 (0.11)	6 (0.12)	1 (0.16)
H G O C R	5 (0.04)	0	0
H G C O R	1,361 (11.39)	492 (10.17)	62 (10.05)
C O G H R	0	0	0
C O H G R	0	0	0

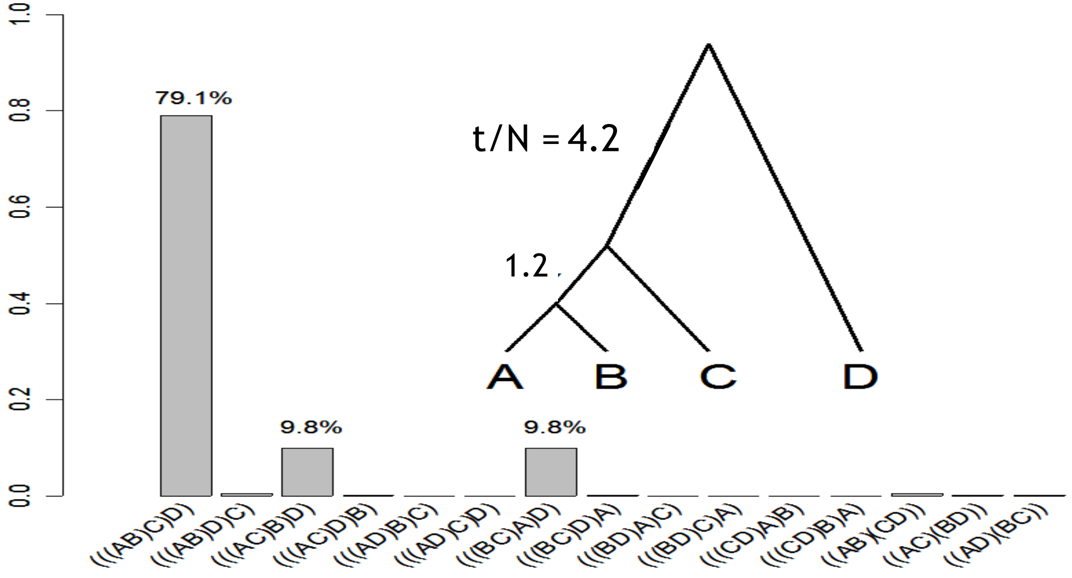
[a] Alignments that overlap with the position of a gene in the human genome.

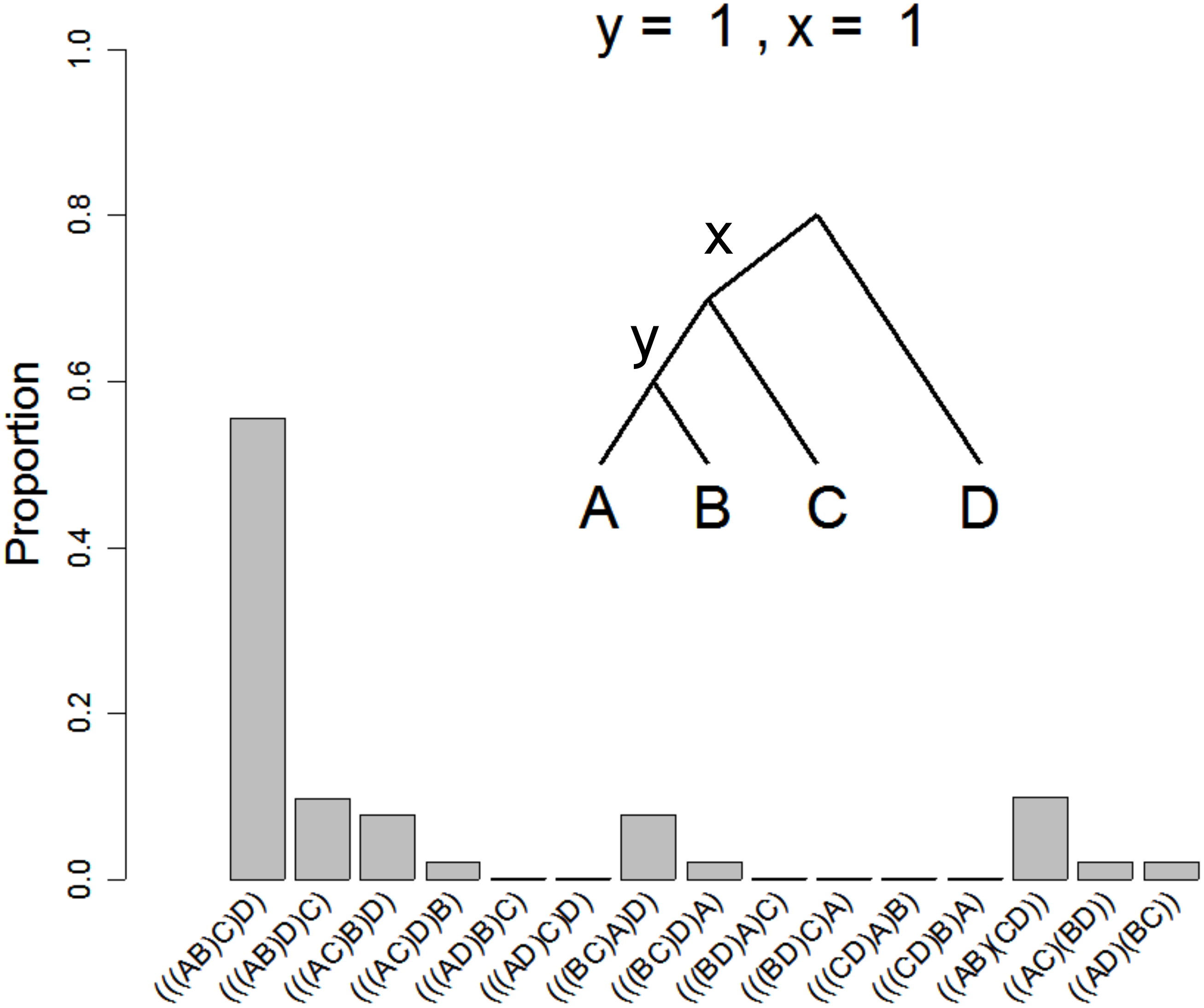
Table from Ebersberger et al. 2007. Mapping human genetic ancestry. MBE 24:2266-2276

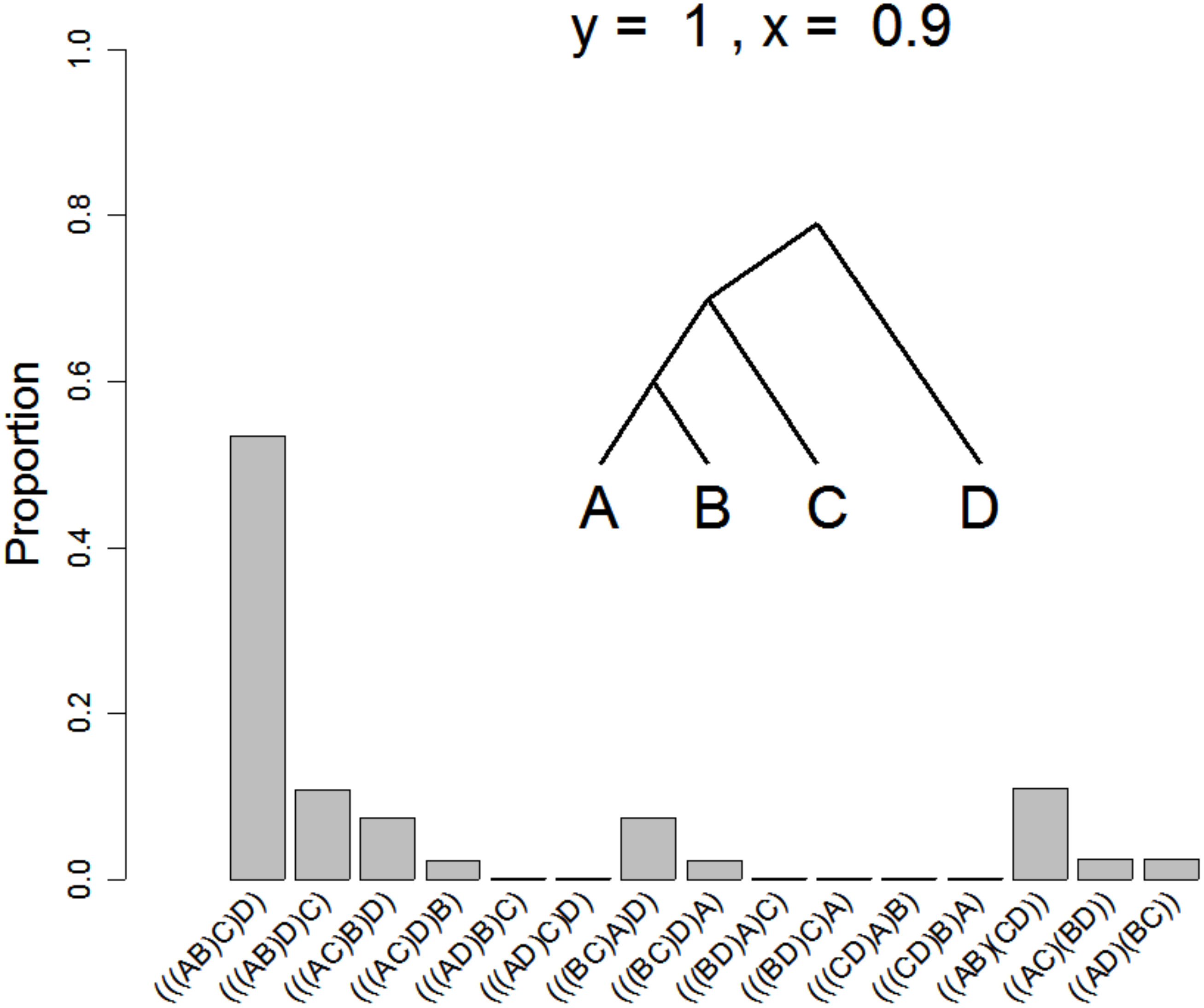
Data from Ebersberger et al. 2007. Mol. Biol. Evol. 24:2266-2276.

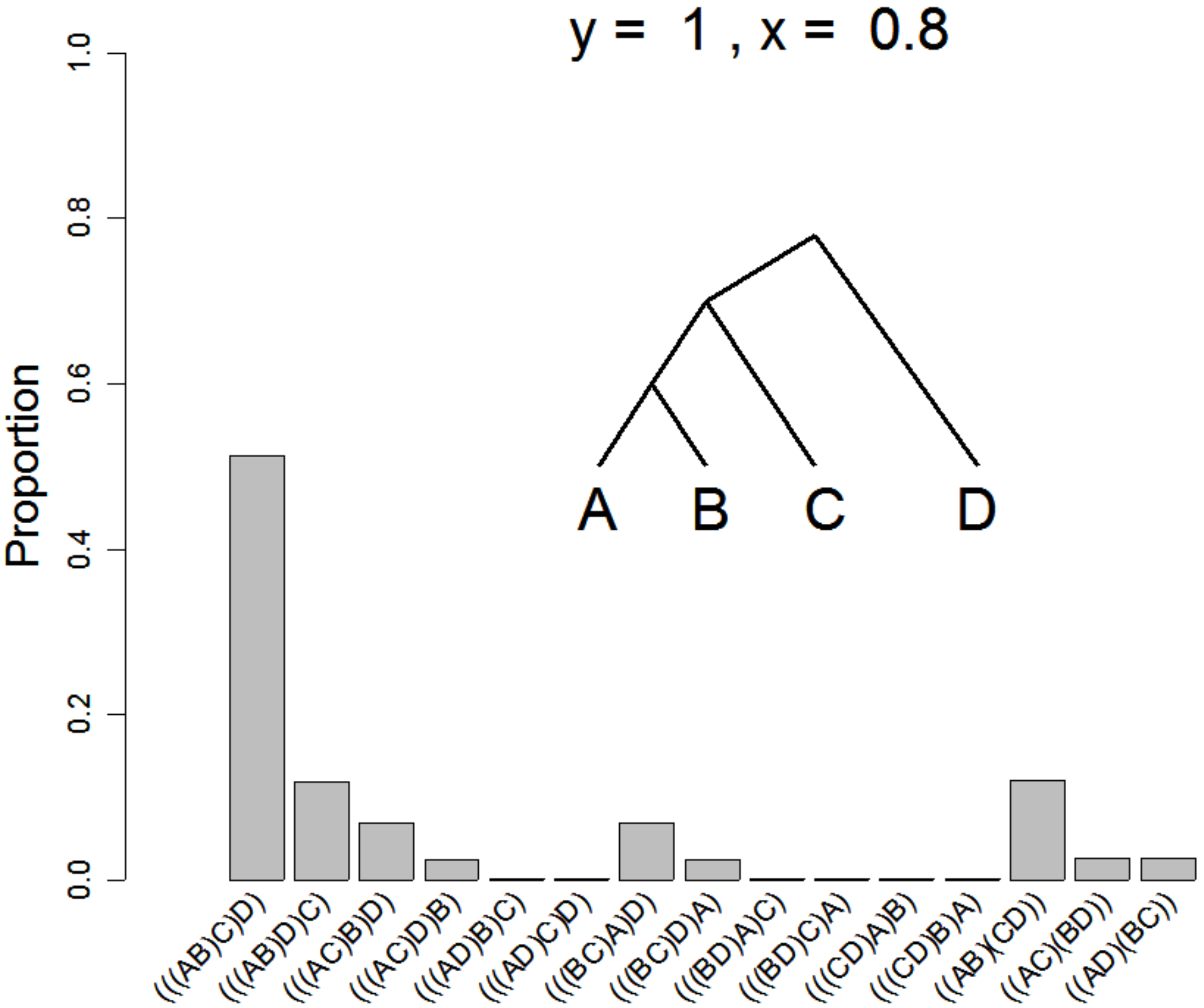


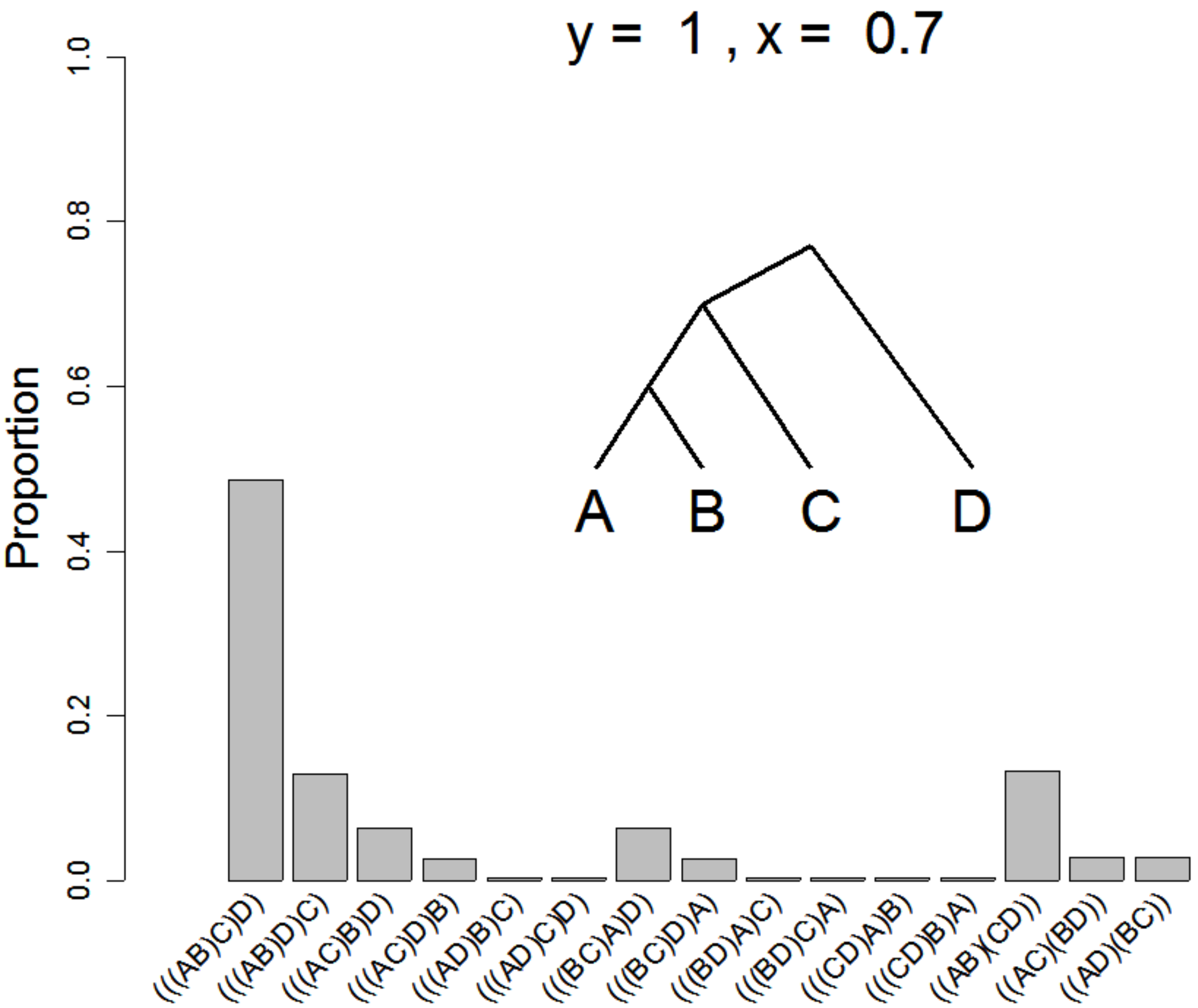
Theoretical distribution based on parameters from Rannala and Yang, 2003. Genetics 164:1645-1656.

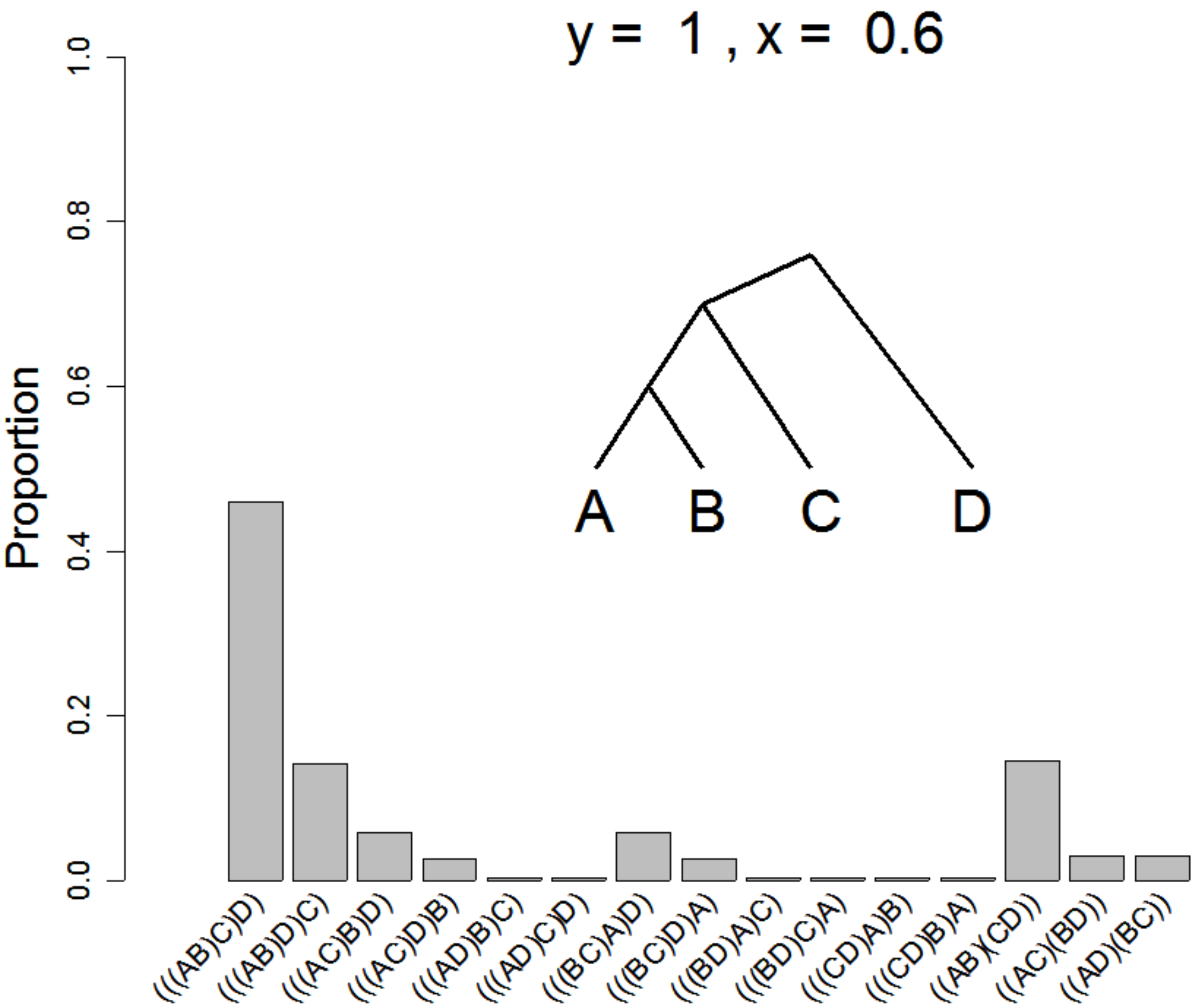


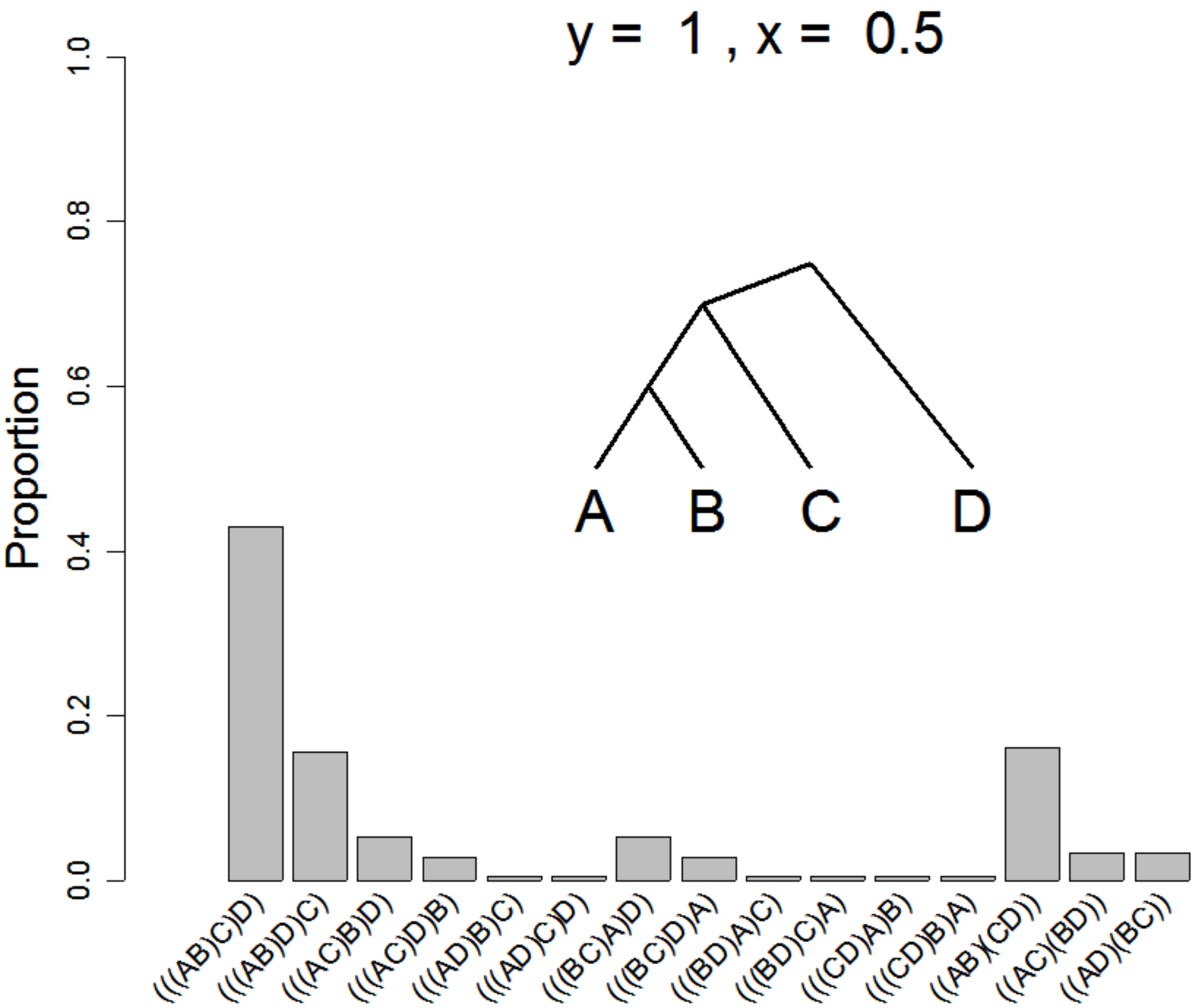

y = 1 , x = 1
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
x
y
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

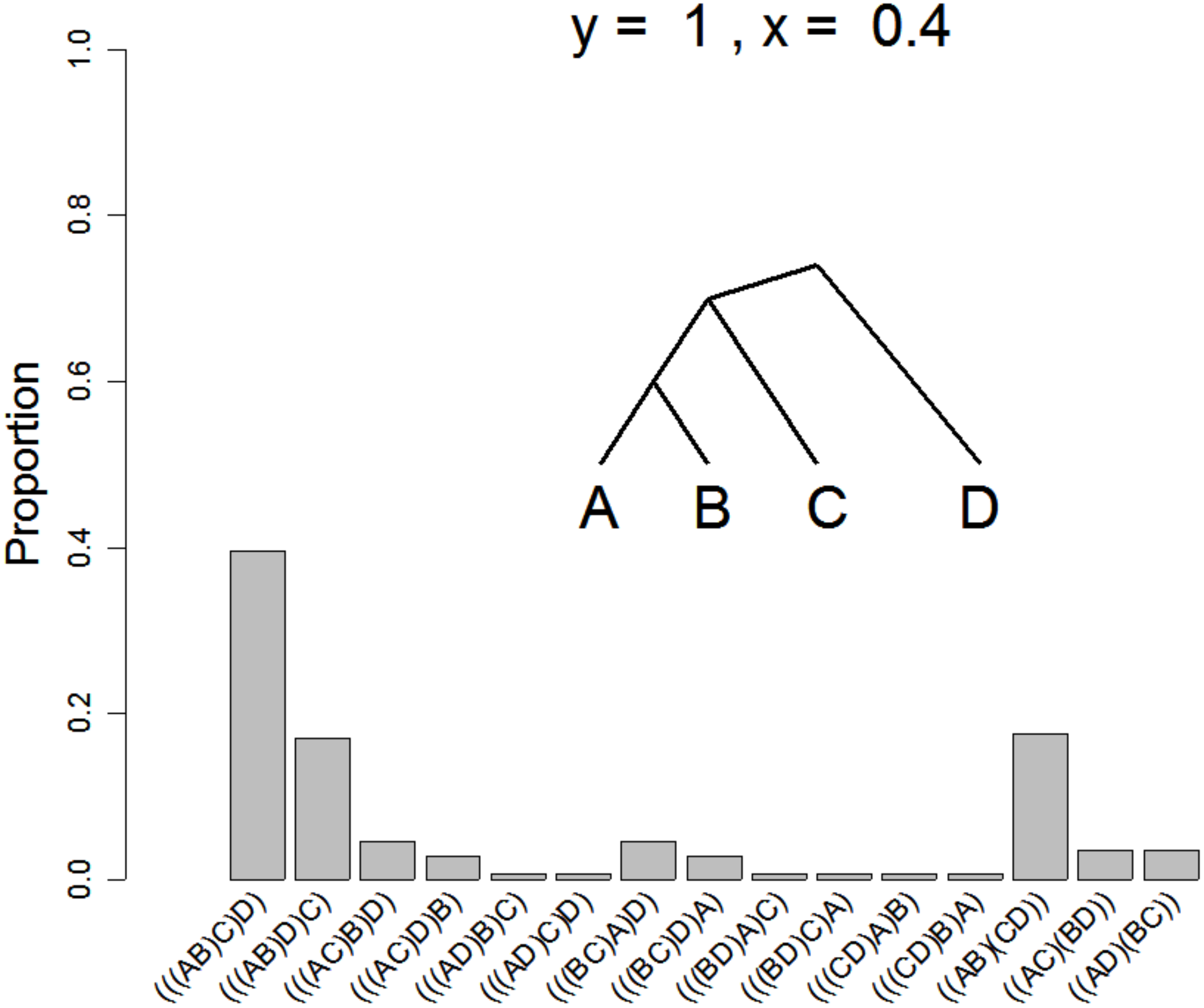

y = 1 , x = 0.9
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

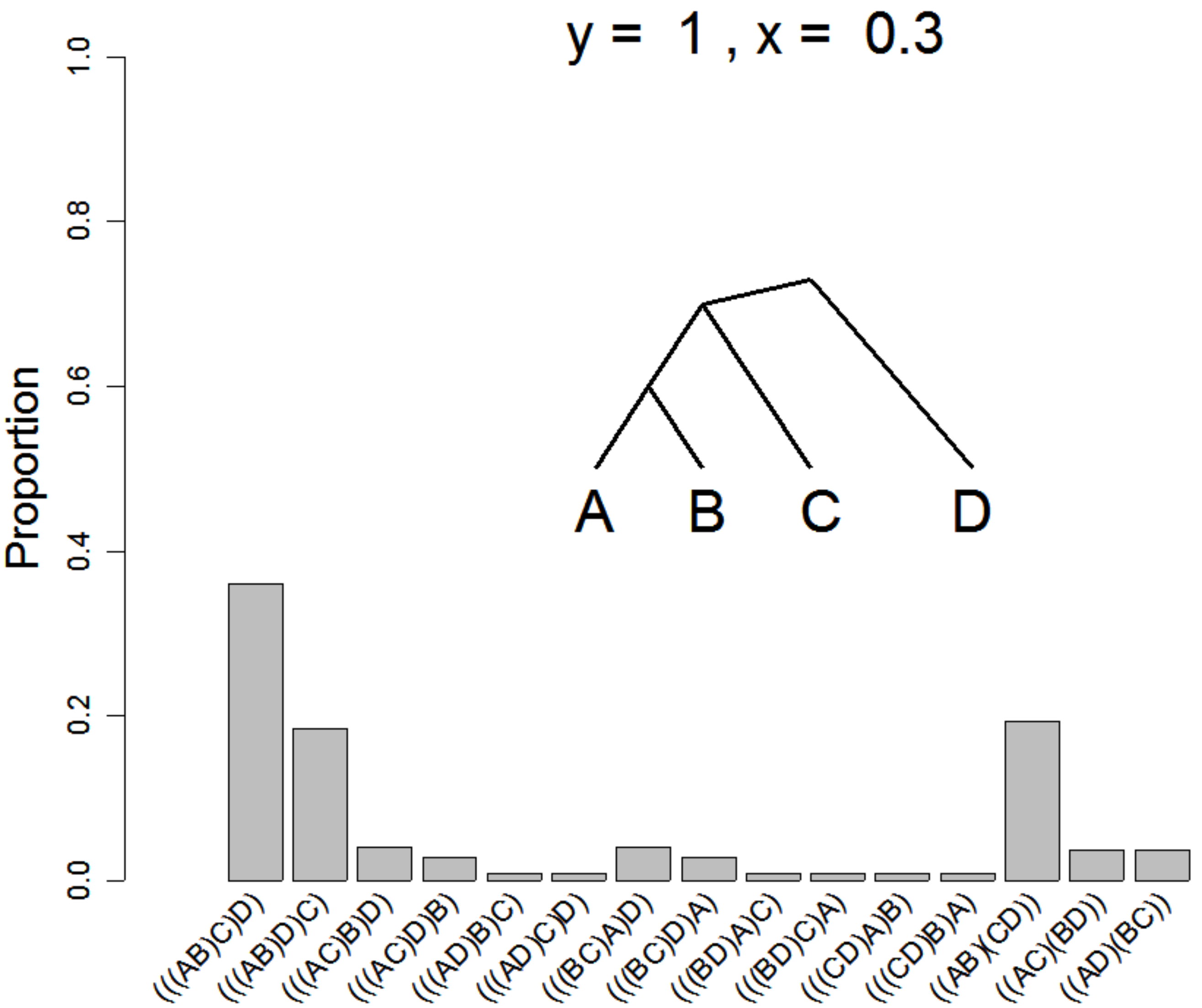

y = 1 , x = 0.8
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

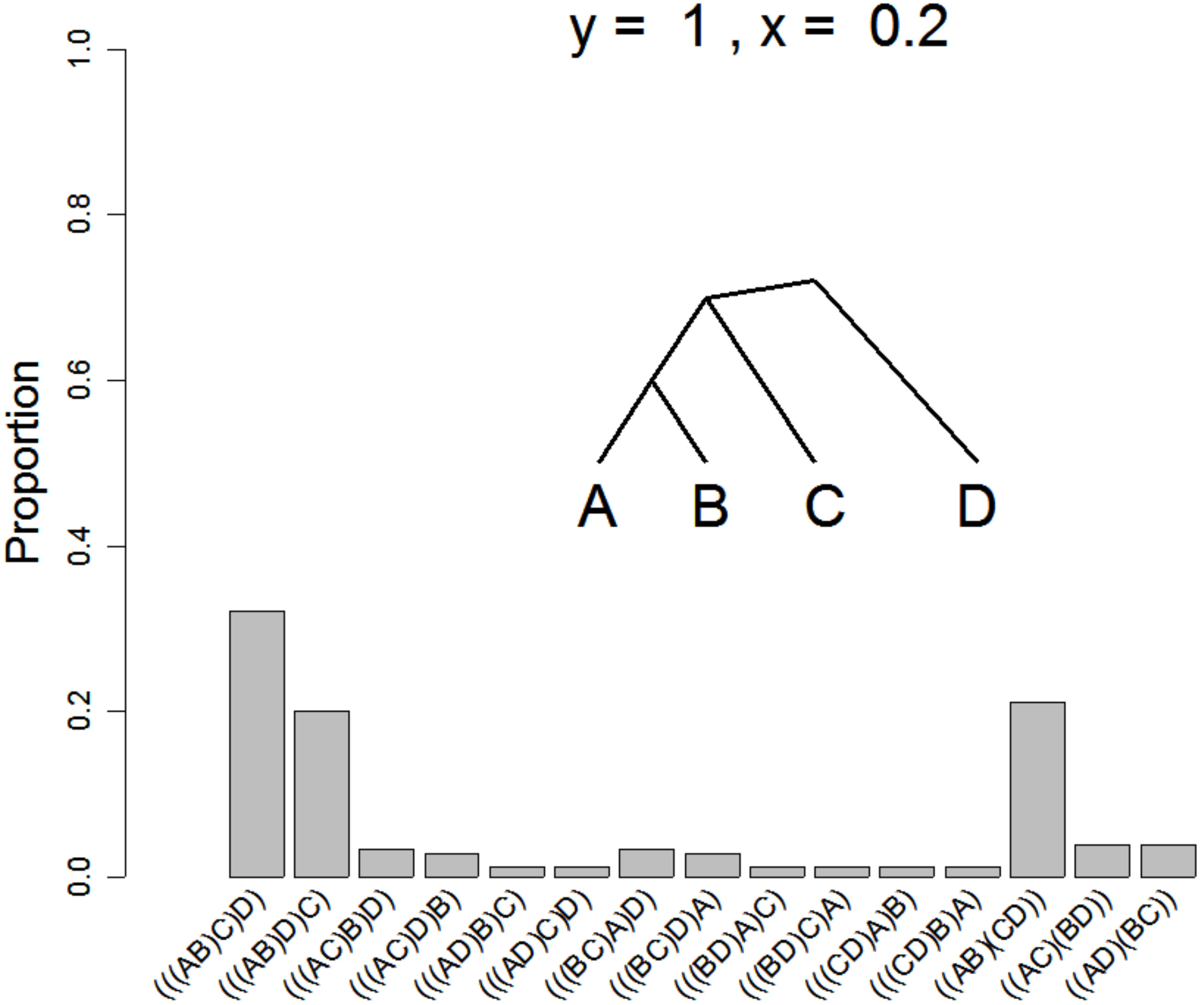

y = 1 , x = 0.7
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

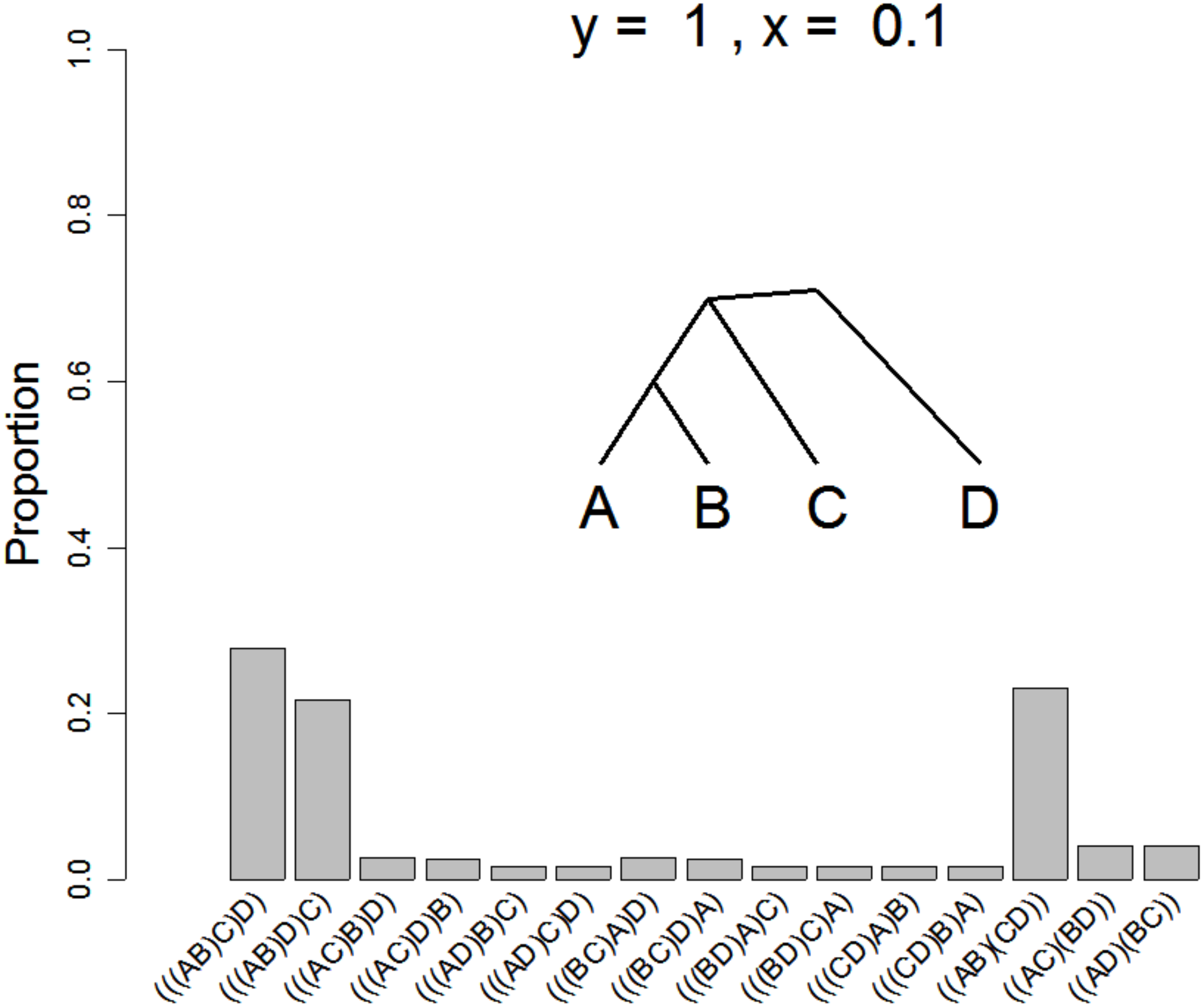

y = 1 , x = 0.6
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

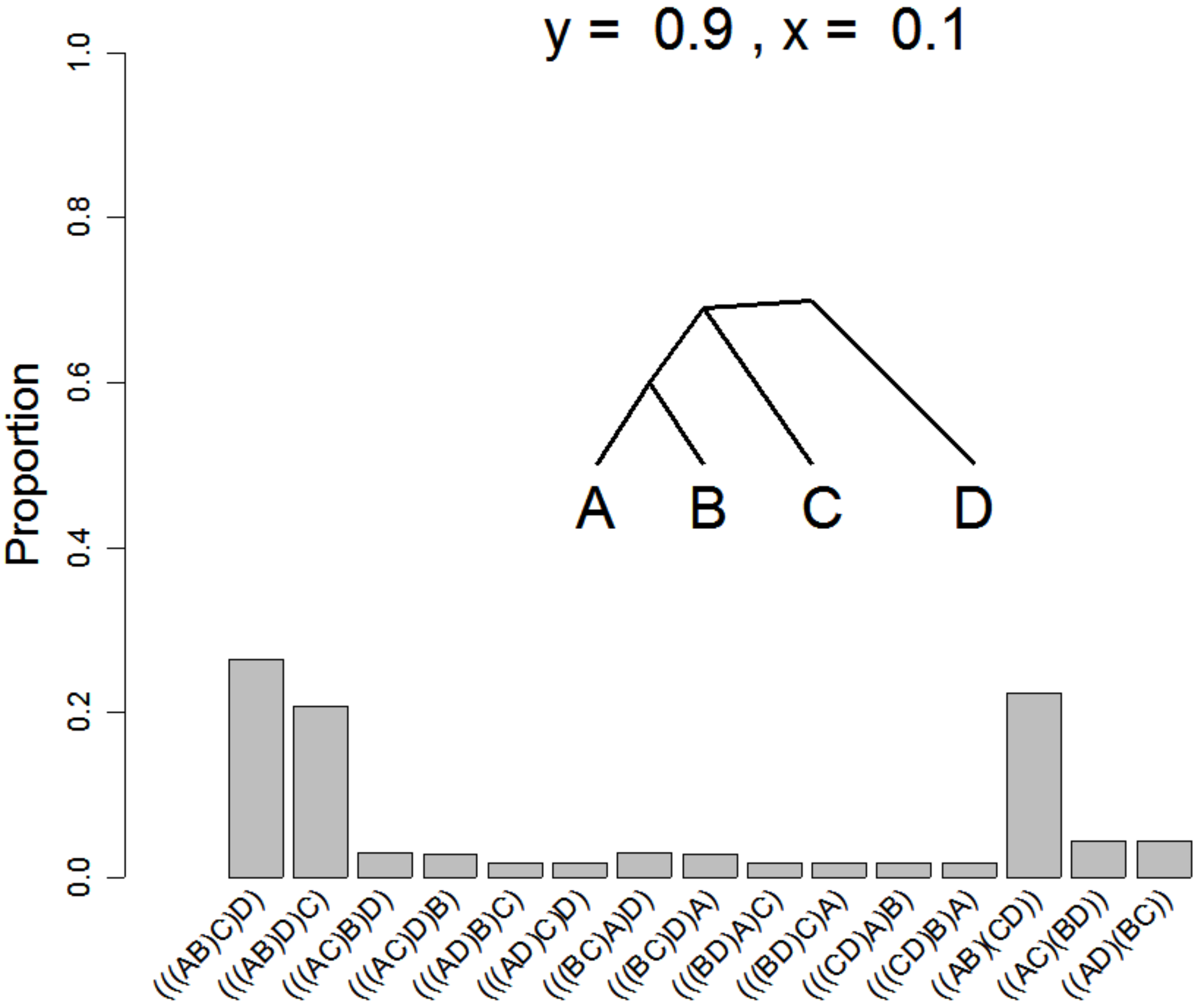

y = 1 , x = 0.5
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

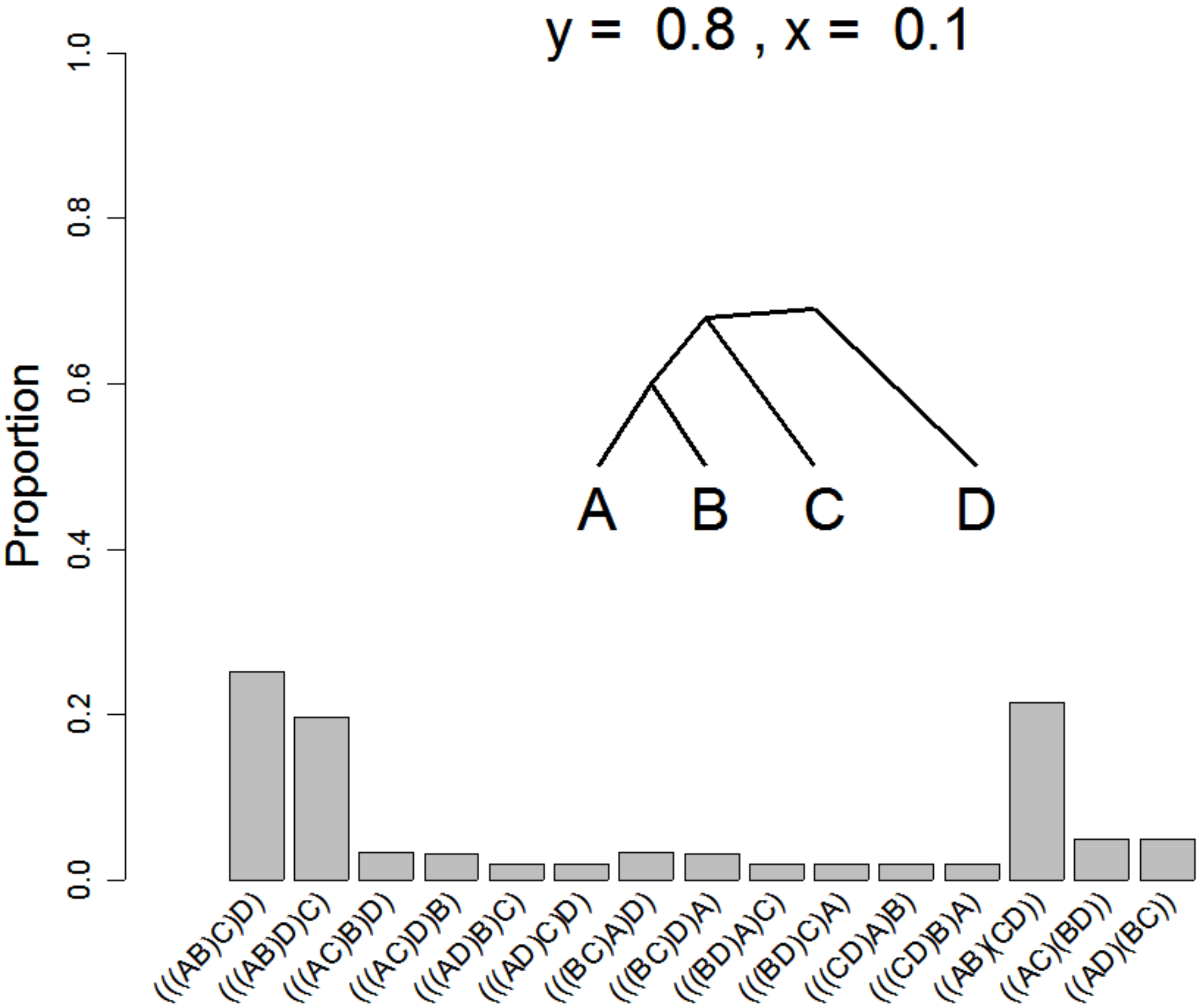

y = 1 , x = 0.4
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

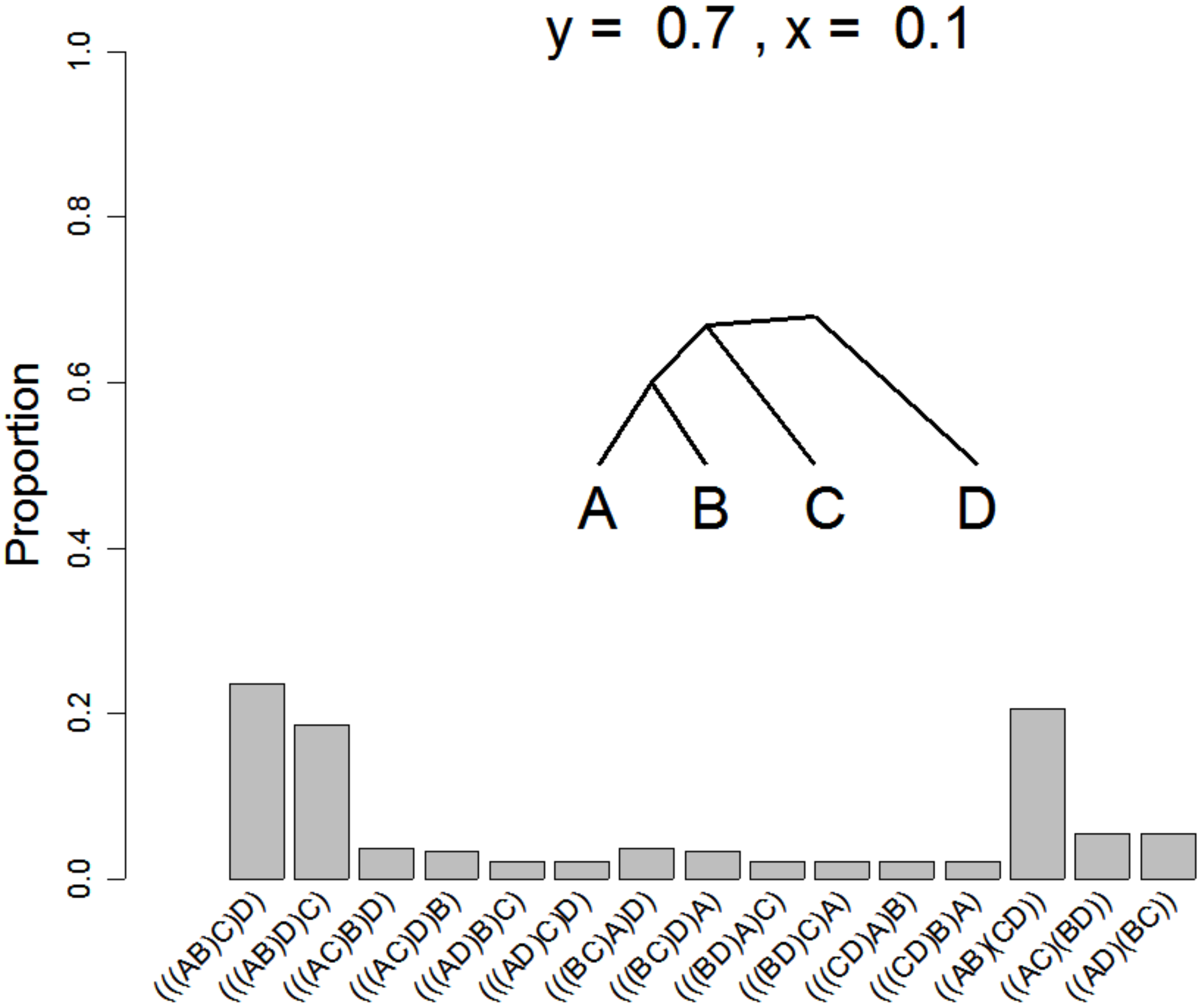

y = 1 , x = 0.3
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

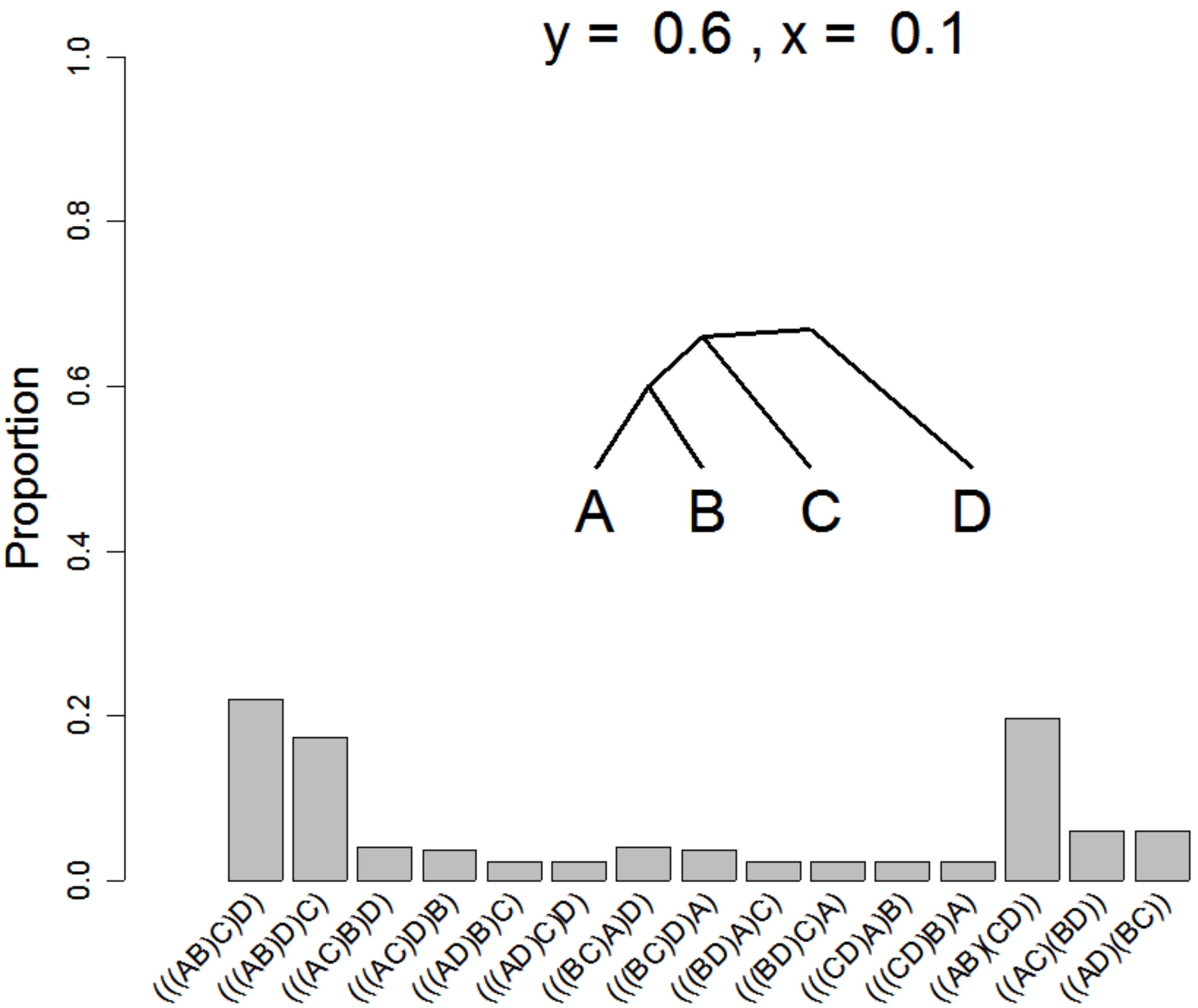

y = 1 , x = 0.2
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

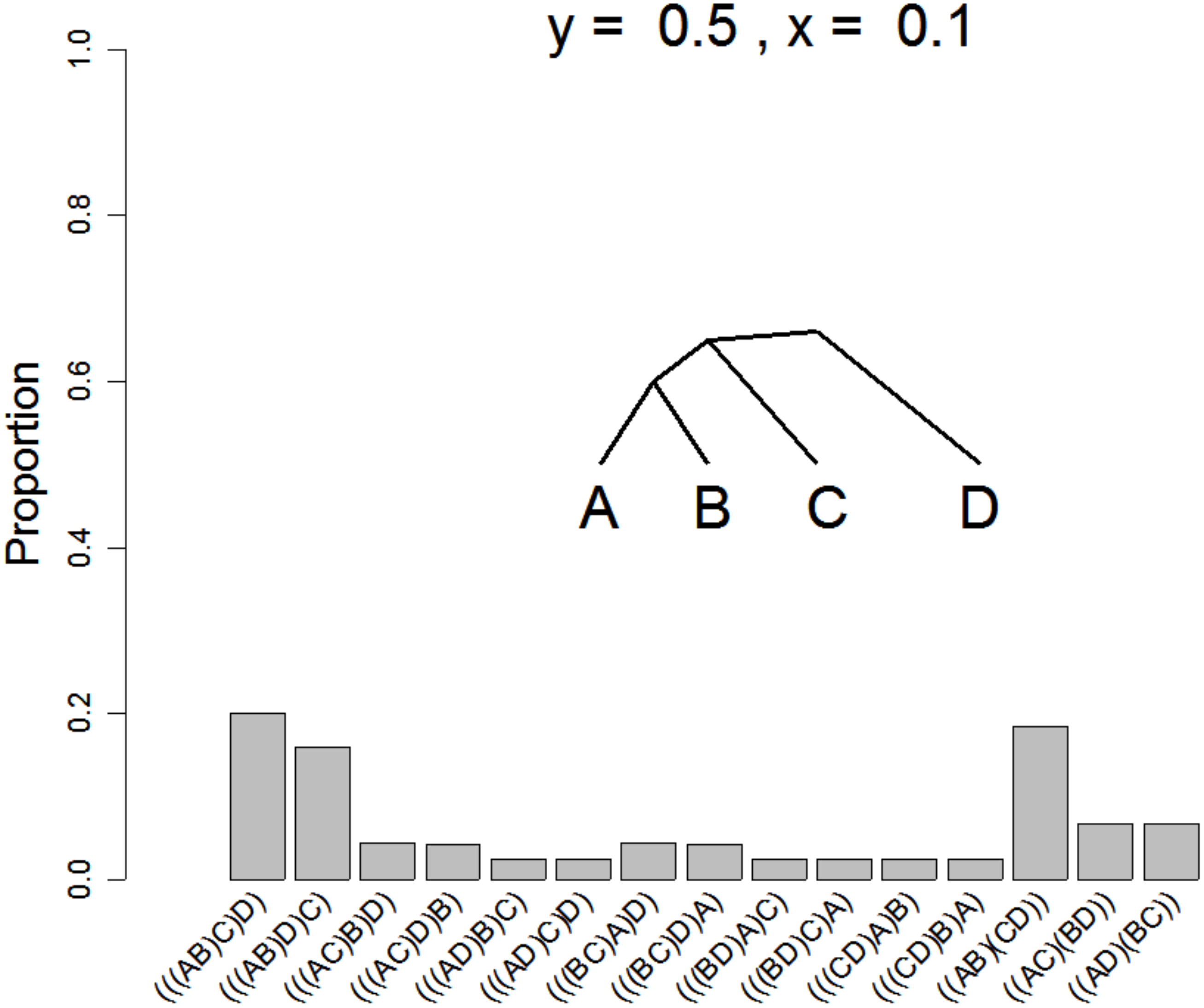

y = 1 , x = 0.1
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

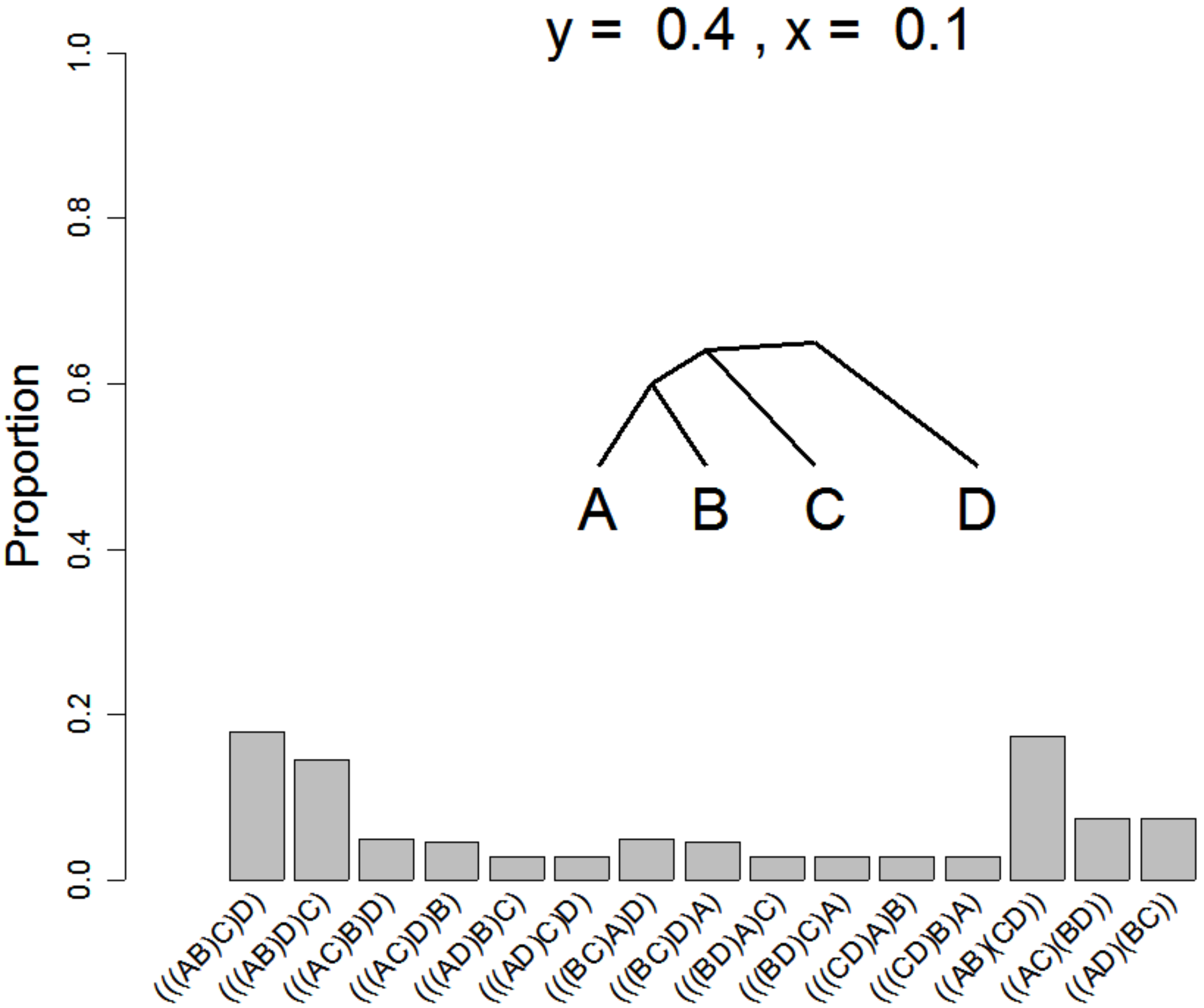

y = 0.9 , x = 0.1
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))


y = 0.8 , x = 0.1
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))


y = 0.7 , x = 0.1
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

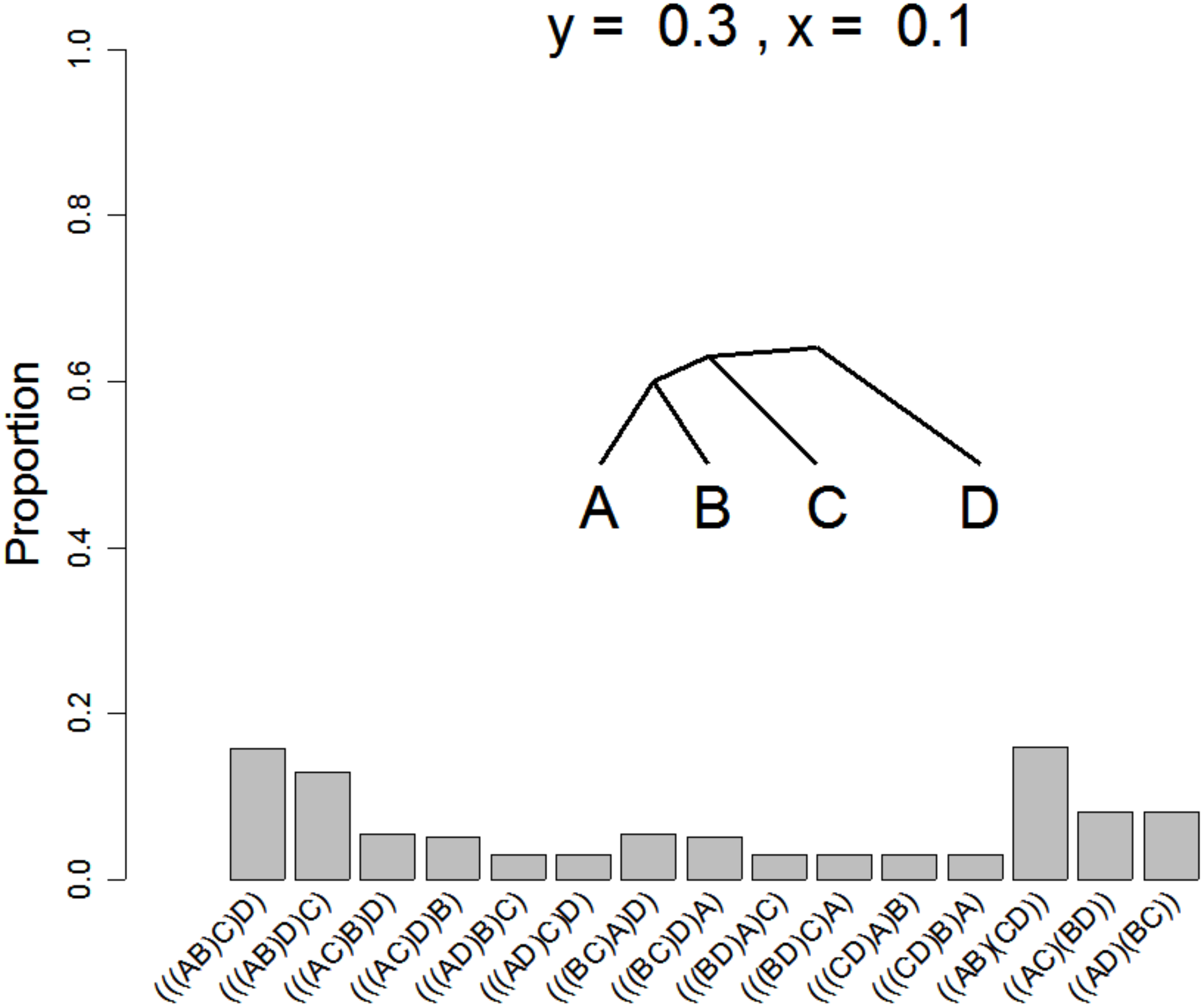

y = 0.6 , x = 0.1
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

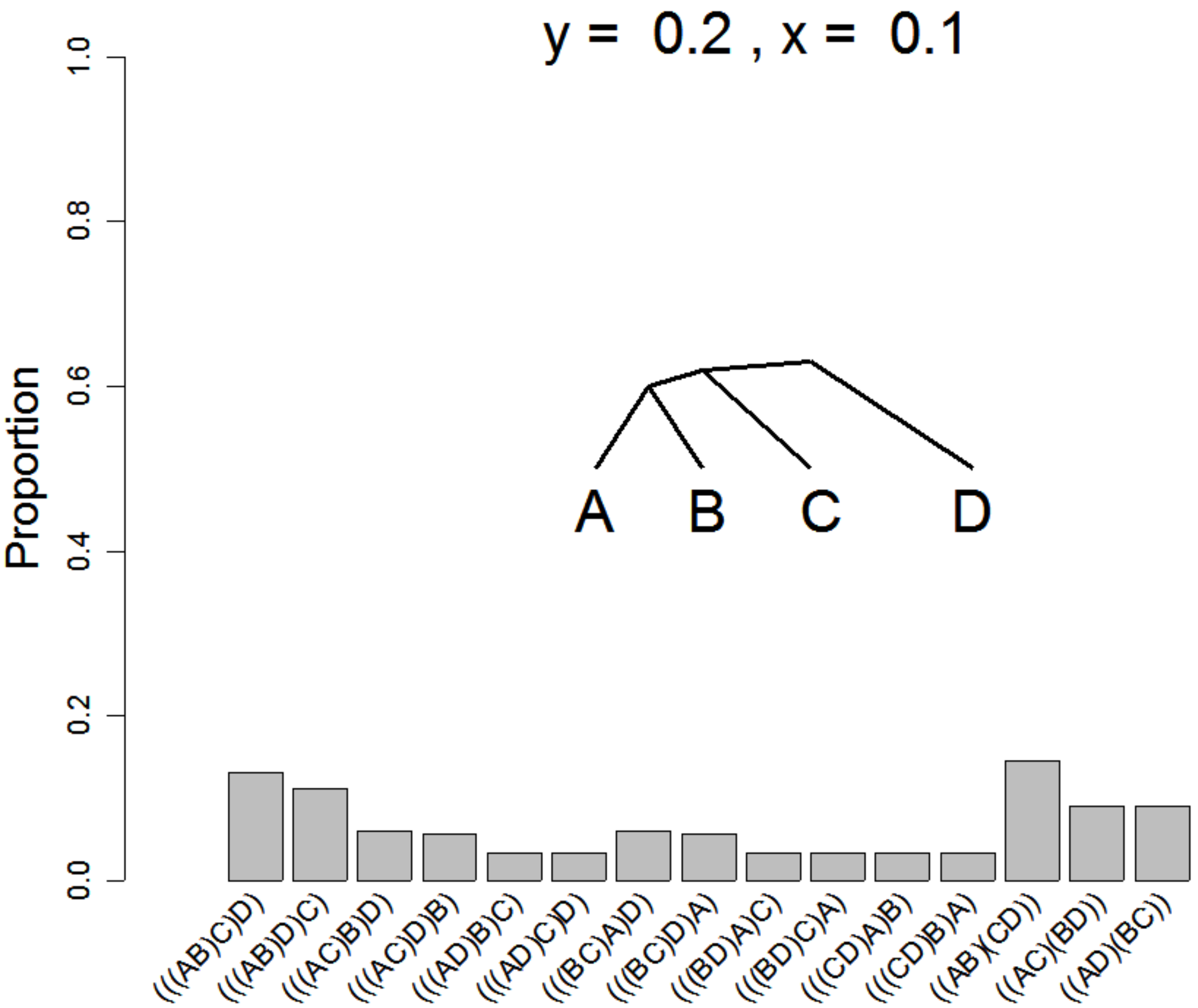

y = 0.5 , x = 0.1
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

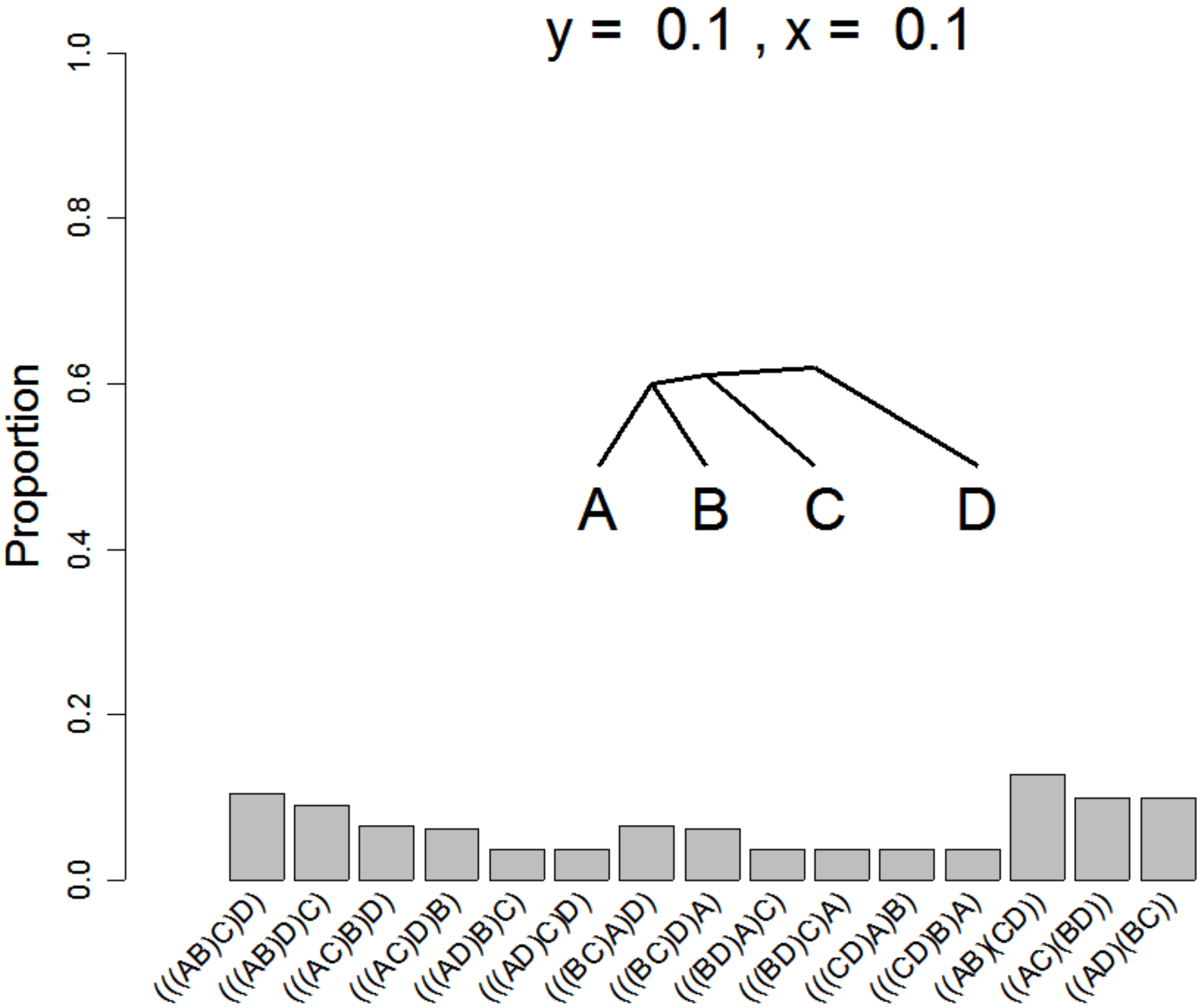


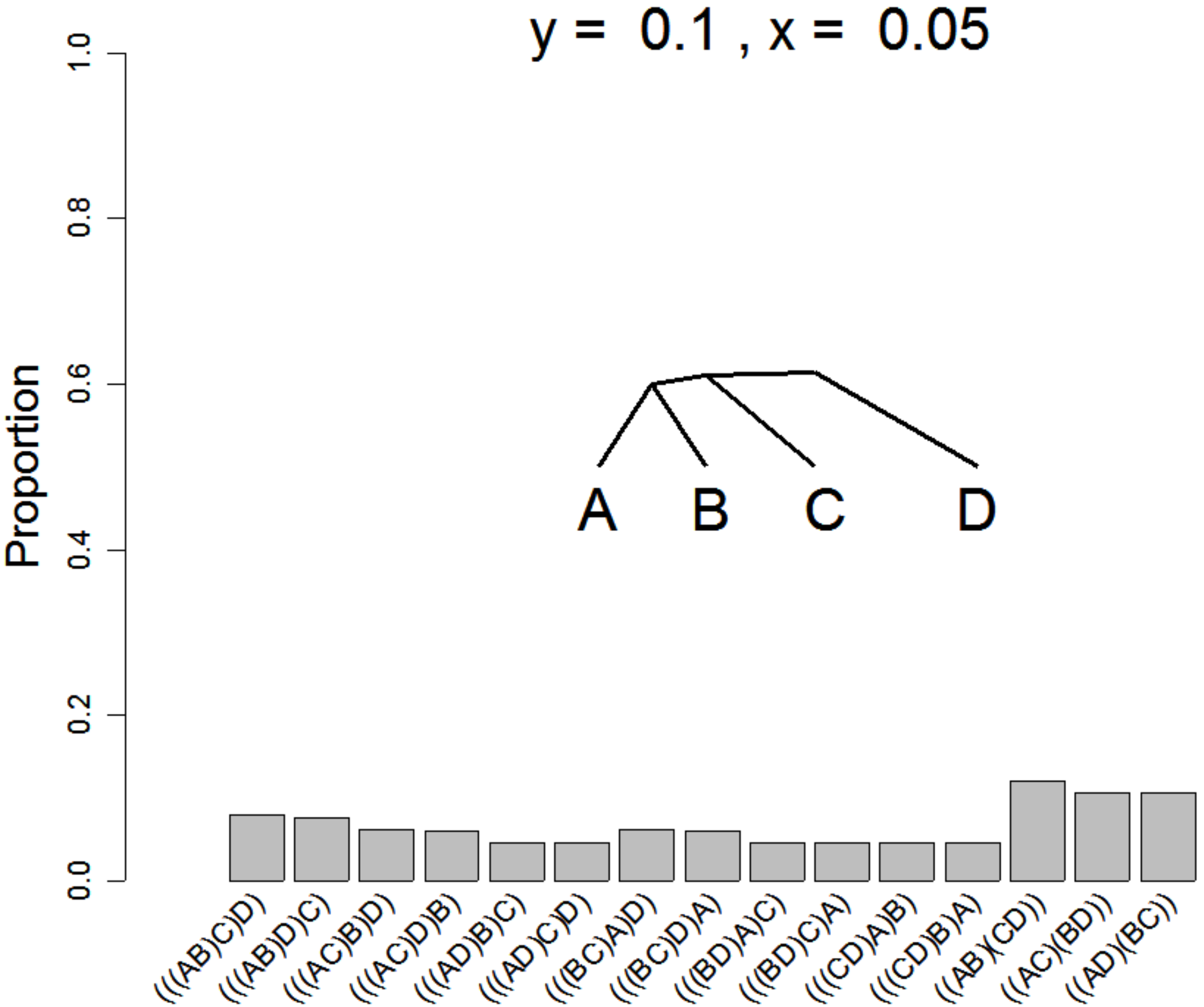
y = 0.4 , x = 0.1

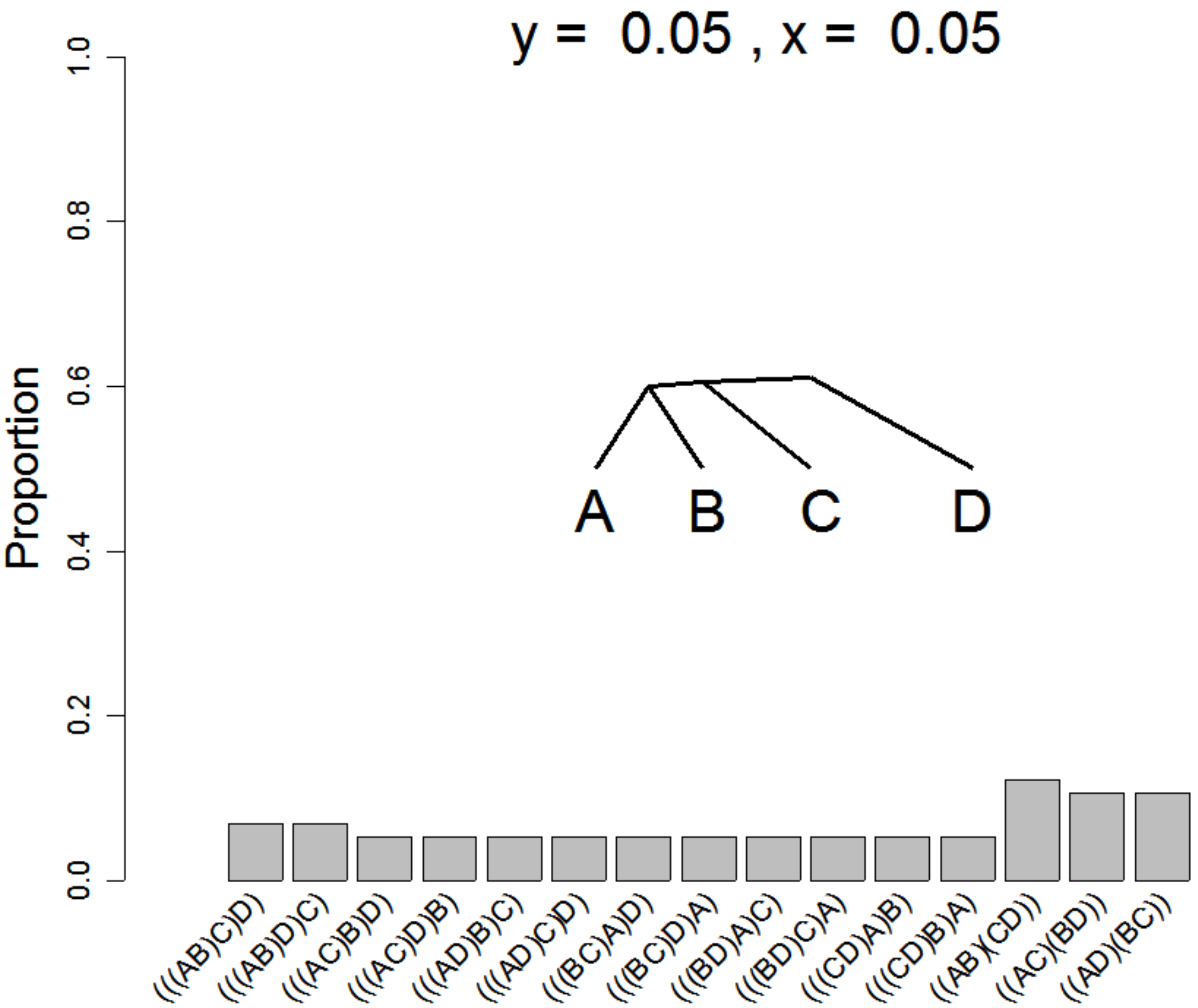
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))


y = 0.3 , x = 0.1
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

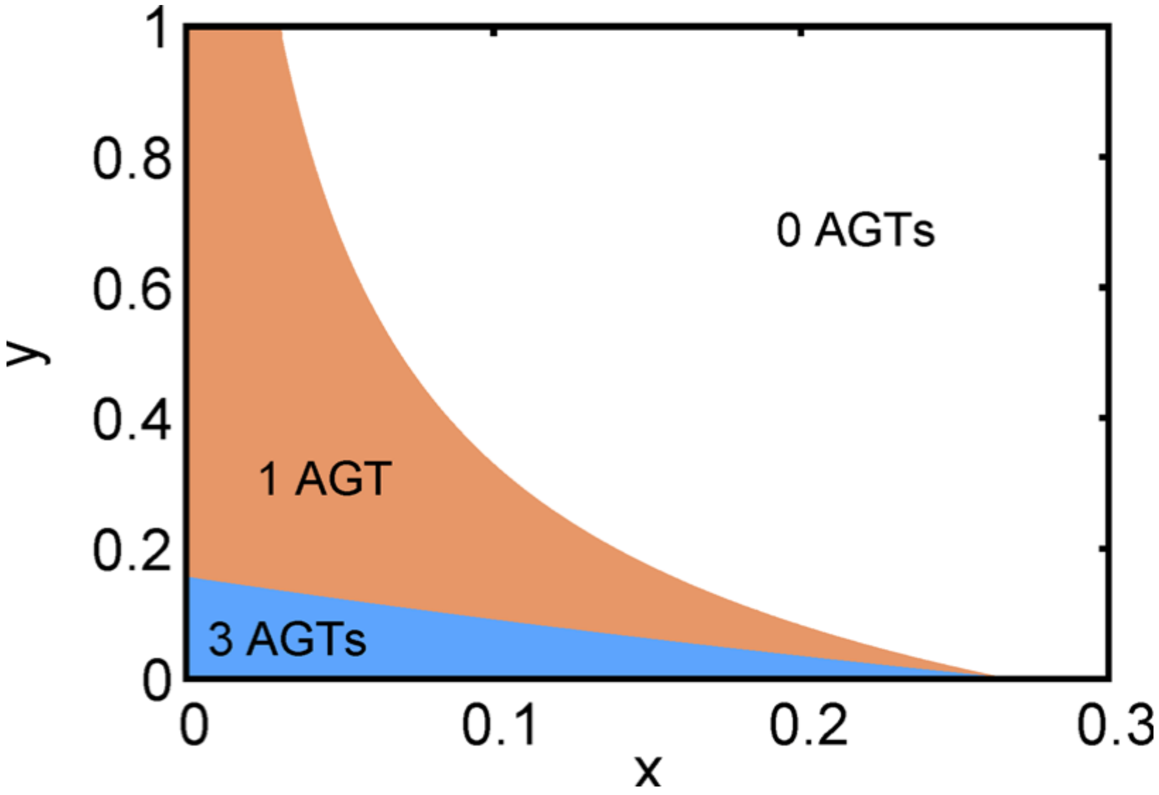

y = 0.2 , x = 0.1
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))


y = 0.1 , x = 0.1
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))


y = 0.1 , x = 0.05
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))


y = 0.05 , x = 0.05
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
(((AB)C)D)
(((AB)D)C)
(((AC)B)D)
(((AC)D)B)
(((AD)B)C)
(((AD)C)D)
(((BC)A)D)
(((BC)D)A)
(((BD)A)C)
(((BD)C)A)
(((CD)A)B)
(((CD)B)A)
((AB)(CD))
((AC)(BD))
((AD)(BC))

Definition: a gene tree which is more probable than the gene tree matching the species tree is called an ***anomalous gene tree*** (Degnan and Rosenberg, 2006).



Theorem. For the asymmetric species tree topology with four species and for any species tree topology with more than four species, there exist branch lengths such that at least one gene tree is anomalous (Degnan and Rosenberg, 2006).

Why can AGTs occur?

If branches are very short, most coalescences occur more anciently than the root of the species tree.

Why can AGTs occur?

If branches are very short, most coalescences occur more anciently than the root of the species tree.

In this case, we assume every sequence of coalescences is equally likely.

Why can AGTs occur?

If branches are very short, most coalescences occur more anciently than the root of the species tree.

In this case, we assume every sequence of coalescences is equally likely.

Gene trees with more symmetry are compatible with more sequences of coalescences. ((AB)(CD)) can have either (AB) first or (CD) first. (((AB)C)D) must have (AB) before ((AB)C).

Why can AGTs occur?

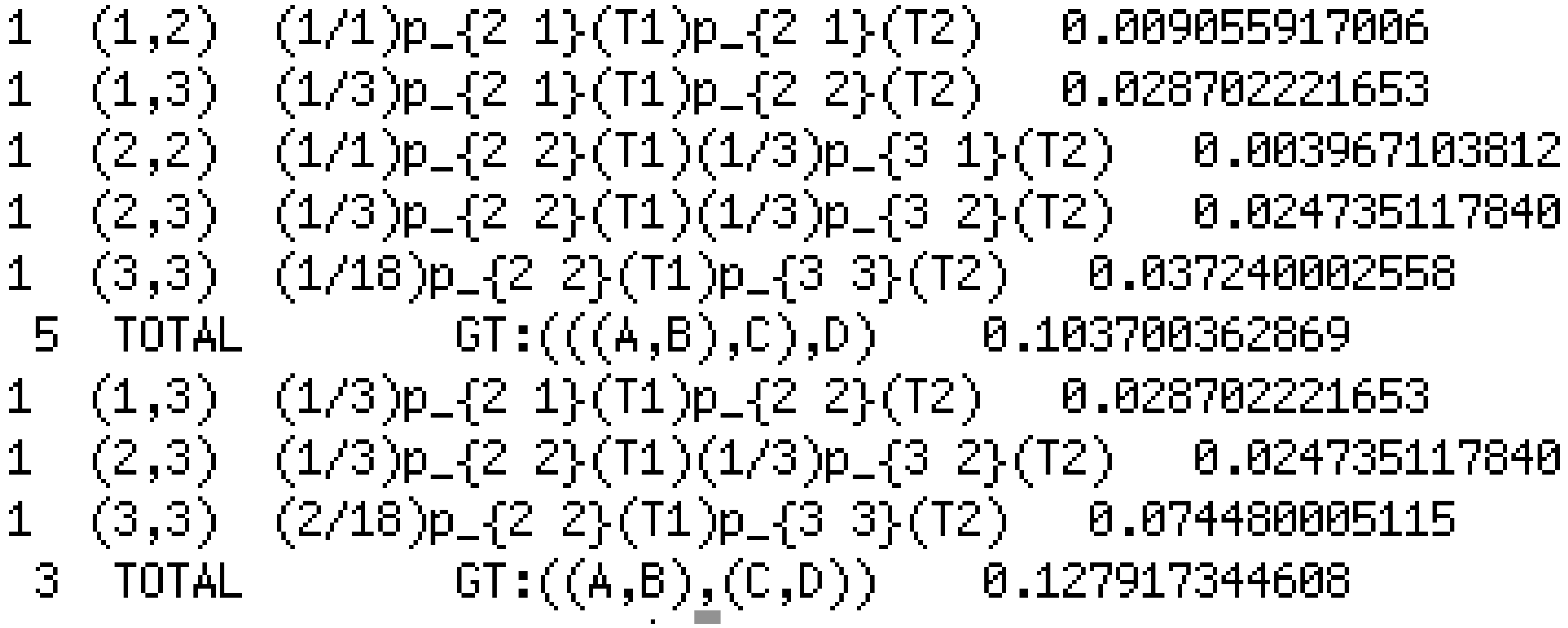
If branches are very short, most coalescences occur more anciently than the root of the species tree.

In this case, we assume every sequence of coalescences is equally likely.

Gene trees with more symmetry are compatible with more sequences of coalescences. ((AB)(CD)) can have either (AB) first or (CD) first. (((AB)C)D) must have (AB) before ((AB)C).

Thus gene trees with more symmetry can have higher probability than gene trees with less symmetry, regardless of the species tree.

COAL output ST (((A:1.0,B:1.0):0.1,C:1.1):0.1,D:1.2)



```
1  (1,2)  (1/1)p_{2 1}(T1)p_{2 1}(T2)   0.009055917006
1  (1,3)  (1/3)p_{2 1}(T1)p_{2 2}(T2)   0.028702221653
1  (2,2)  (1/1)p_{2 2}(T1)(1/3)p_{3 1}(T2)   0.003967103812
1  (2,3)  (1/3)p_{2 2}(T1)(1/3)p_{3 2}(T2)   0.024735117840
1  (3,3)  (1/18)p_{2 2}(T1)p_{3 3}(T2)   0.03724002558
 5  TOTAL         GT:(((A,B),C),D)    0.103700362869
1  (1,3)  (1/3)p_{2 1}(T1)p_{2 2}(T2)   0.028702221653
1  (2,3)  (1/3)p_{2 2}(T1)(1/3)p_{3 2}(T2)   0.024735117840
1  (3,3)  (2/18)p_{2 2}(T1)p_{3 3}(T2)   0.07448005115
 3  TOTAL         GT:((A,B),(C,D))    0.12791734608
```

Some limitations on AGTs

--If a gene tree doesn't match the species tree, its probability must be $<$ 1/3.

Some limitations on AGTs

--If a gene tree doesn’t match the species tree, its probability must be < 1/3.

--There are limits to how much more likely an AGT can be than a matching gene tree. For the (((AB)C)D) species tree,

$$\Pr[((AB)(CD))] – \Pr[(((AB)C)D)] < 1/18$$

Some limitations on AGTs

--If a gene tree doesn't match the species tree, its probability must be < 1/3.

--There are limits to how much more likely an AGT can be than a matching gene tree. For the (((AB)C)D) species tree,

$$\Pr[((AB)(CD))] - \Pr[(((AB)C)D)] < 1/18$$

--If a gene tree is a "caterpillar" (pectinate), then it is not an AGT ("There are no caterpillars in a wicked forest").

Some limitations on AGTs

--If a gene tree doesn't match the species tree, its probability must be < 1/3.

--There are limits to how much more likely an AGT can be than a matching gene tree. For the (((AB)C)D) species tree,

$$\Pr[((AB)(CD))] - \Pr[(((AB)C)D)] < 1/18$$

--If a gene tree is a "caterpillar" (pectinate), then it is not an AGT ("There are no caterpillars in a wicked forest").

--There must be at least one or two very short branches in the species tree in order for there to be an AGT. For four taxa, $x < 0.1568$ or $y < 0.1568$. If N=100,000, x=0.1 is 10,000 generations.

What implications do AGTs have?

Methods of species tree inference might be statistically inconsistent and can infer AGTs

How do concatenation and consensus methods perform when there are AGTs?

Species Tree inference—concatenation

Species Trees are often estimated by concatenating several gene sequences and analyzing as one (data from Chen and Li, 2001).

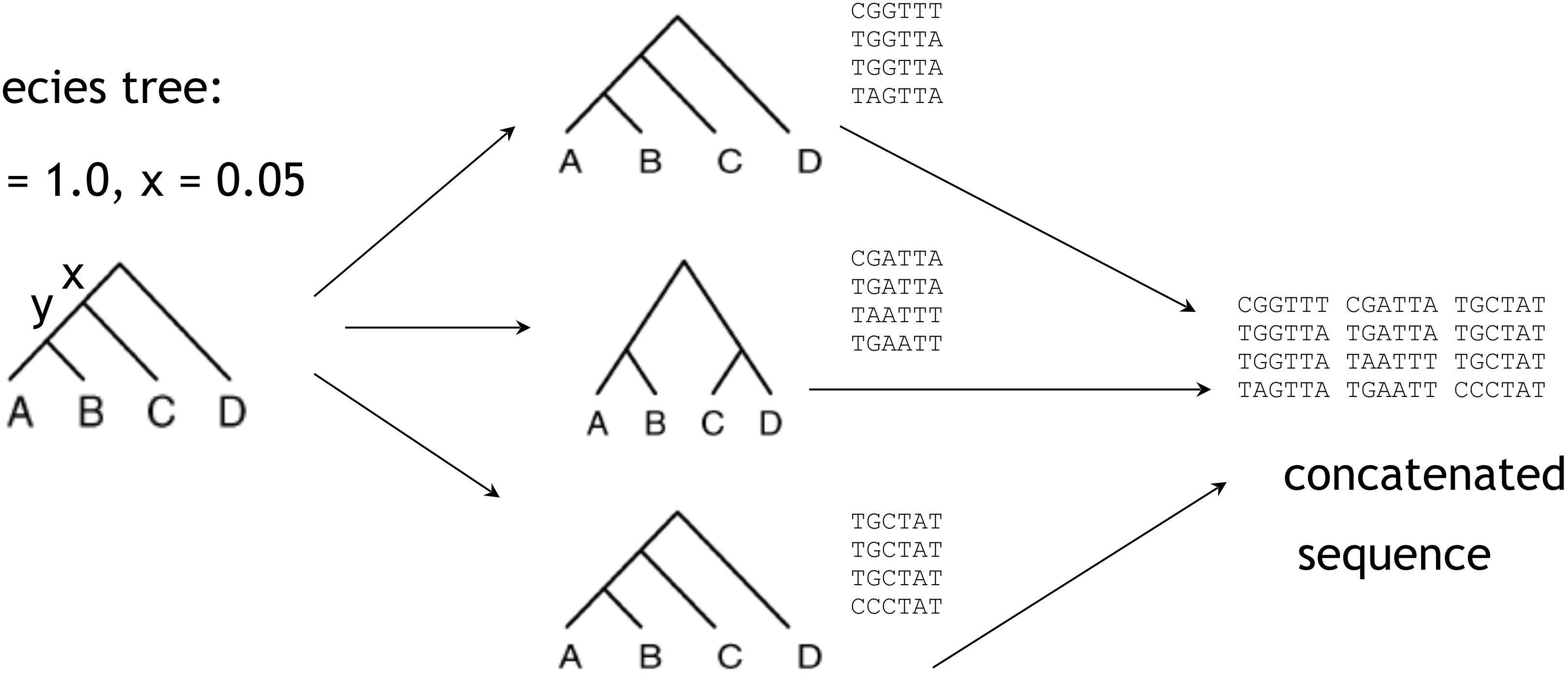
	Gene 1	**Gene 2**	**Gene 3**
Human	CTTGAATAATTTTTAC	TAGAGTTTCCTTGTGGTG	CGGTTT
Chimp	CTTCAATAATTTTTAC	TAGAGTTTCCTTGTGGTA	TGGTTT
Gorilla	TTTGAATAATTTTTAC	TAGAGTTTCCTTGTGGTA	TGGTTT
Orang	CTTGAATAATTTTTAT	CAGAGTTTCCTTGTGGTC	CRGTTT

Concatenation and gene tree discordance

How does concatenation perform when sequences are generated from different topologies?

Species tree:

y = 1.0, x = 0.05



```
CGGTTT
TGGTTA
TGGTTA
TAGTTA
```

```
CGATTA
TGATTA
TAATTT
TGAATT
```

```
TGCTAT
TGCTAT
TGCTAT
CCCTAT
```

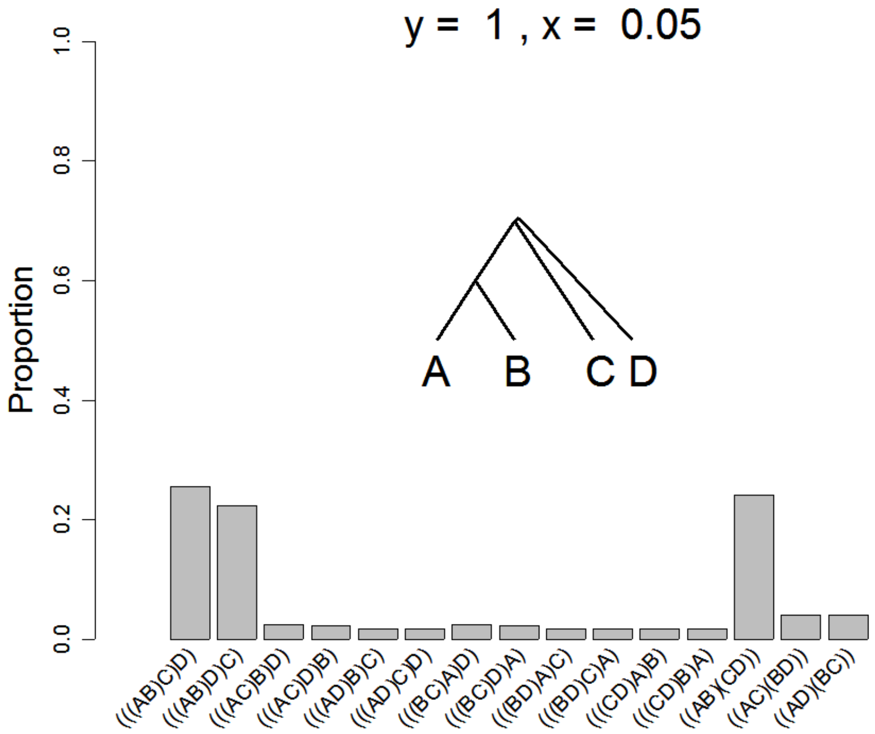
```
CGGTTT CGATTA TGCTAT
TGGTTA TGATTA TGCTAT
TGGTTA TAATTT TGCTAT
TAGTTA TGAATT CCCTAT
```

concatenated

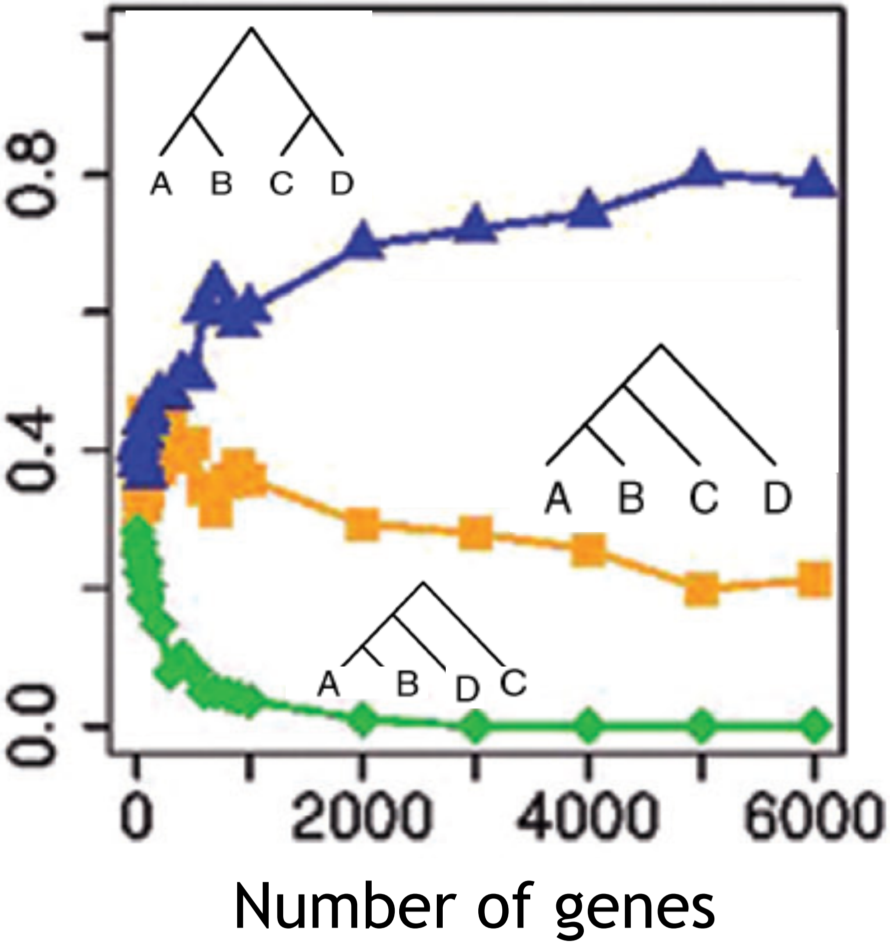
sequence

Simulated gene trees

Trees inferred from concatenated sequences using ML

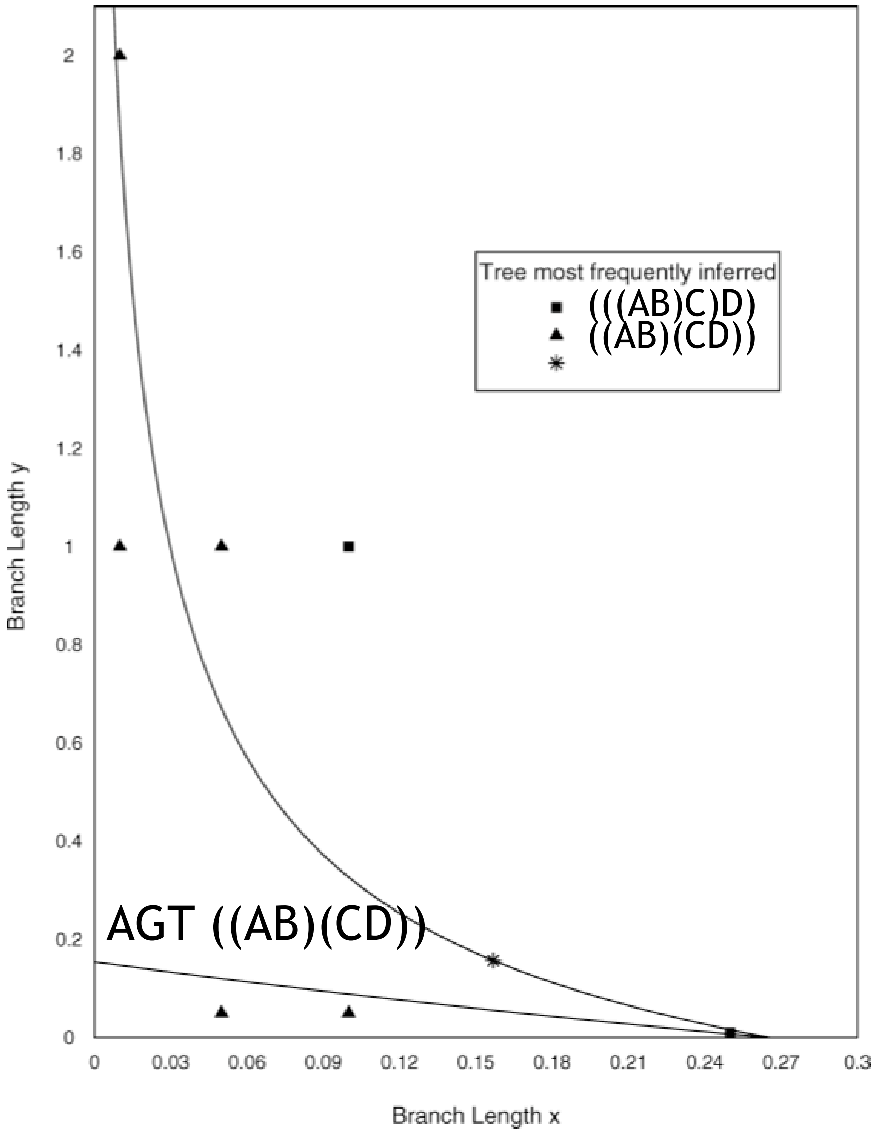


y = 1.0, x = 0.05

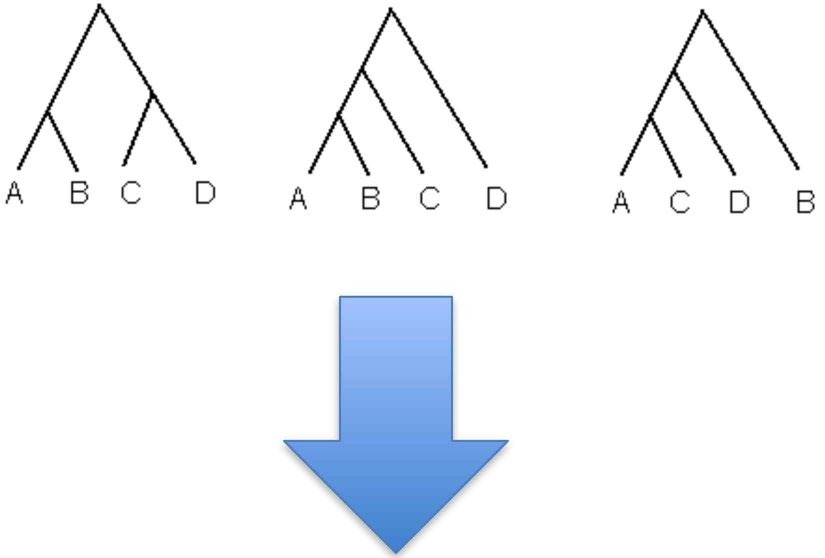


Concatenation and the anomaly zone

ST (((AB):y,C):x,D)

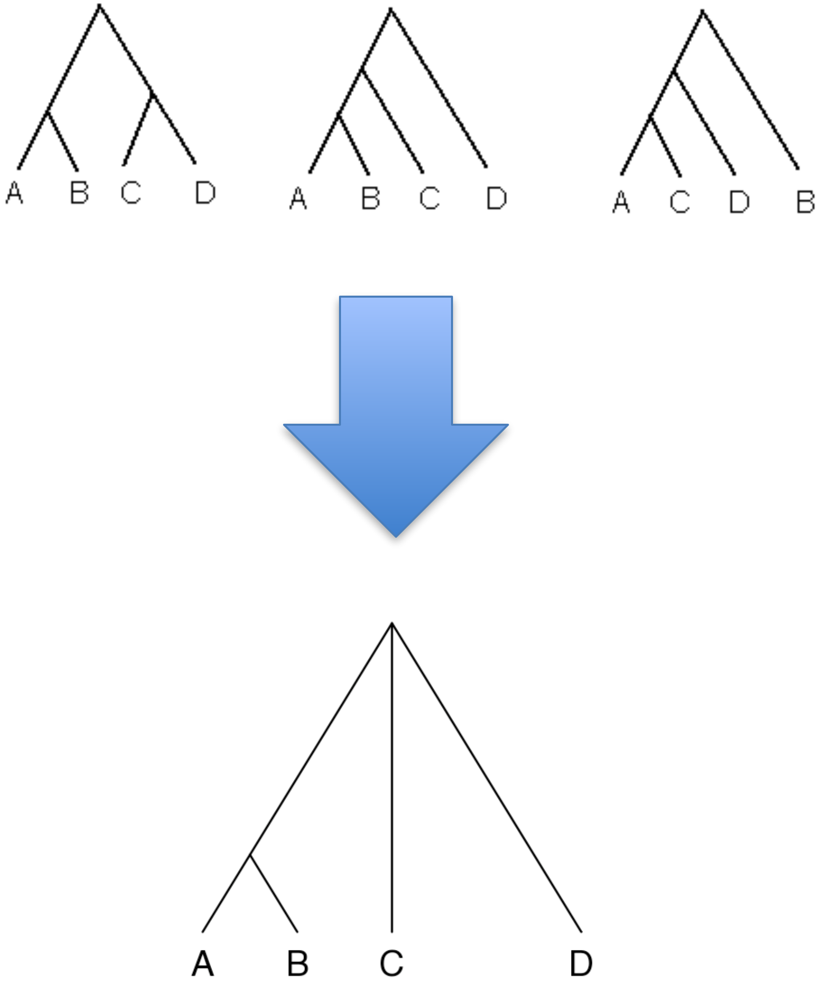


Consensus methods



?

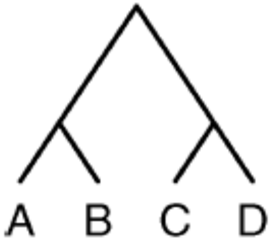
Consensus methods



Majority rule—consensus tree has all clades that were observed in > 50% of trees.

Greedy—sort clades by their proportions. Accept the most frequently observed clades one at a time that are compatible with already accepted clades. Do this until you have a fully resolved tree.

R*—for each set of 3 taxa, find the most commonly occurring triple e.g., (AB)C, (AC)B or (BC)A. Build the tree from the most commonly occurring triple.



(AB)D, (CD)B are two rooted triples

ASTRAL—a median tree. Find the tree that minimizes the sum of distances to input gene trees. For ASTRAL, a quartet distance is used.

Asymptotic consensus trees

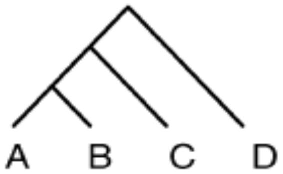
Consensus trees are usually *statistics*, functions of data like x-bar.

Definition: an ***asymptotic consensus tree*** is the tree that is obtained by computing the consensus tree using topology probabilities from the multispecies coalescent model.

Motivation: if there are a large number of independent loci, observed gene tree, clade, and rooted triple proportions should approximate their theoretical probabilities.

Gene tree	Probability	(.1, .1)
(((AB)C)D)	p_1	.104
(((AB)D)C)	p_2	.091
(((AC)B)D)	p_3	.066
(((AC)D)B)	p_4	.062
(((AD)B)C)	p_5	.037
(((AD)C)B)	p_6	.037
(((BC)A)D)	p_7	.066
(((BC)D)A)	p_8	.062
(((BD)A)C)	p_9	.037
(((BD)C)A)	p_{10}	.037
(((CD)A)B)	p_{11}	.037
(((CD)B)A)	p_{12}	.037
((AB)(CD))	p_{13}	.128
((AC)(BD))	p_{14}	.099
((AD)(BC))	p_{15}	.099

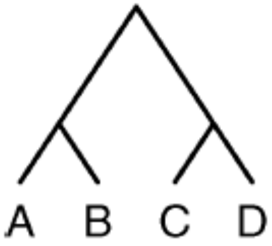
Clade			
{AB}	$p_1 + p_2 + p_{13}$	.322	◀
~~{AC}~~	~~$p_3 + p_4 + p_{14}$~~	~~.227~~	
~~{AD}~~	~~$p_5 + p_6 + p_{15}$~~	~~.174~~	
~~{BC}~~	~~$p_7 + p_8 + p_{15}$~~	~~.227~~	
~~{BD}~~	~~$p_9 + p_{10} + p_{14}$~~	~~.174~~	
~~{CD}~~	~~$p_{11} + p_{12} + p_{13}$~~	~~.202~~	
{ABC}	$p_1 + p_3 + p_7$	.236	◀
~~{ABD}~~	~~$p_2 + p_5 + p_9$~~	~~.165~~	
~~{ACD}~~	~~$p_4 + p_6 + p_{11}$~~	~~.136~~	
~~{BCD}~~	~~$p_8 + p_{10} + p_{12}$~~	~~.136~~	



Greedy consensus tree

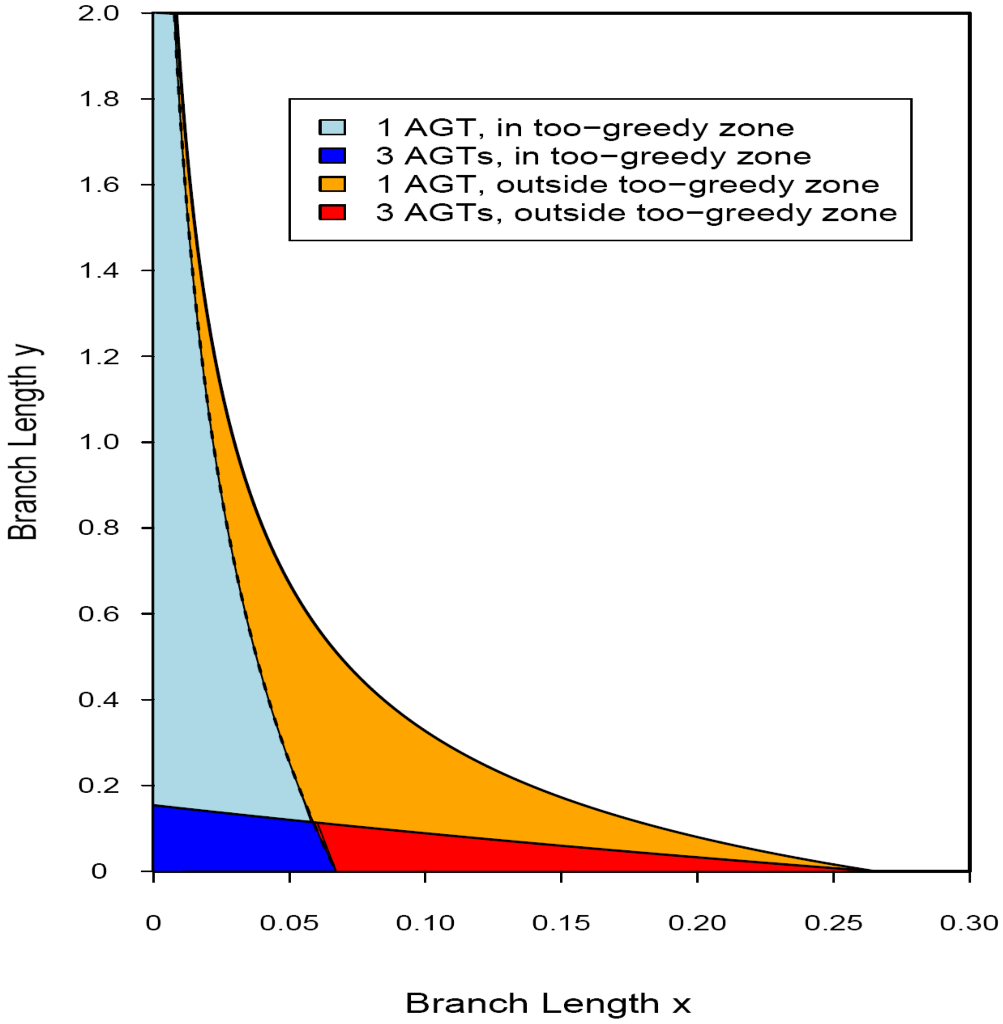
Gene tree	Probability	(.1, .1)	(.05, .05)
(((AB)C)D)	p_1	.104	.079
(((AB)D)C)	p_2	.091	.075
(((AC)B)D)	p_3	.066	.061
(((AC)D)B)	p_4	.062	.060
(((AD)B)C)	p_5	.037	.045
(((AD)C)B)	p_6	.037	.045
(((BC)A)D)	p_7	.066	.061
(((BC)D)A)	p_8	.062	.060
(((BD)A)C)	p_9	.037	.045
(((BD)C)A)	p_{10}	.037	.045
(((CD)A)B)	p_{11}	.037	.045
(((CD)B)A)	p_{12}	.037	.045
((AB)(CD))	p_{13}	.128	.121
((AC)(BD))	p_{14}	.099	.105
((AD)(BC))	p_{15}	.099	.105

Clade			
{AB}	$p_1 + p_2 + p_{13}$	.322 ◀	.275 ◀
~~{AC}~~	~~$p_3 + p_4 + p_{14}$~~	~~.227~~	~~.226~~
~~{AD}~~	~~$p_5 + p_6 + p_{15}$~~	~~.174~~	~~.196~~
~~{BC}~~	~~$p_7 + p_8 + p_{15}$~~	~~.227~~	~~.226~~
~~{BD}~~	~~$p_9 + p_{10} + p_{14}$~~	~~.174~~	~~.196~~
{CD}	$p_{11} + p_{12} + p_{13}$	.202	.212 ◀
~~{ABC}~~	~~$p_1 + p_3 + p_7$~~	~~.236~~ ◀	~~.201~~
~~{ABD}~~	~~$p_2 + p_5 + p_9$~~	~~.165~~	~~.166~~
~~{ACD}~~	~~$p_4 + p_6 + p_{11}$~~	~~.136~~	~~.151~~
~~{BCD}~~	~~$p_8 + p_{10} + p_{12}$~~	~~.136~~	~~.151~~

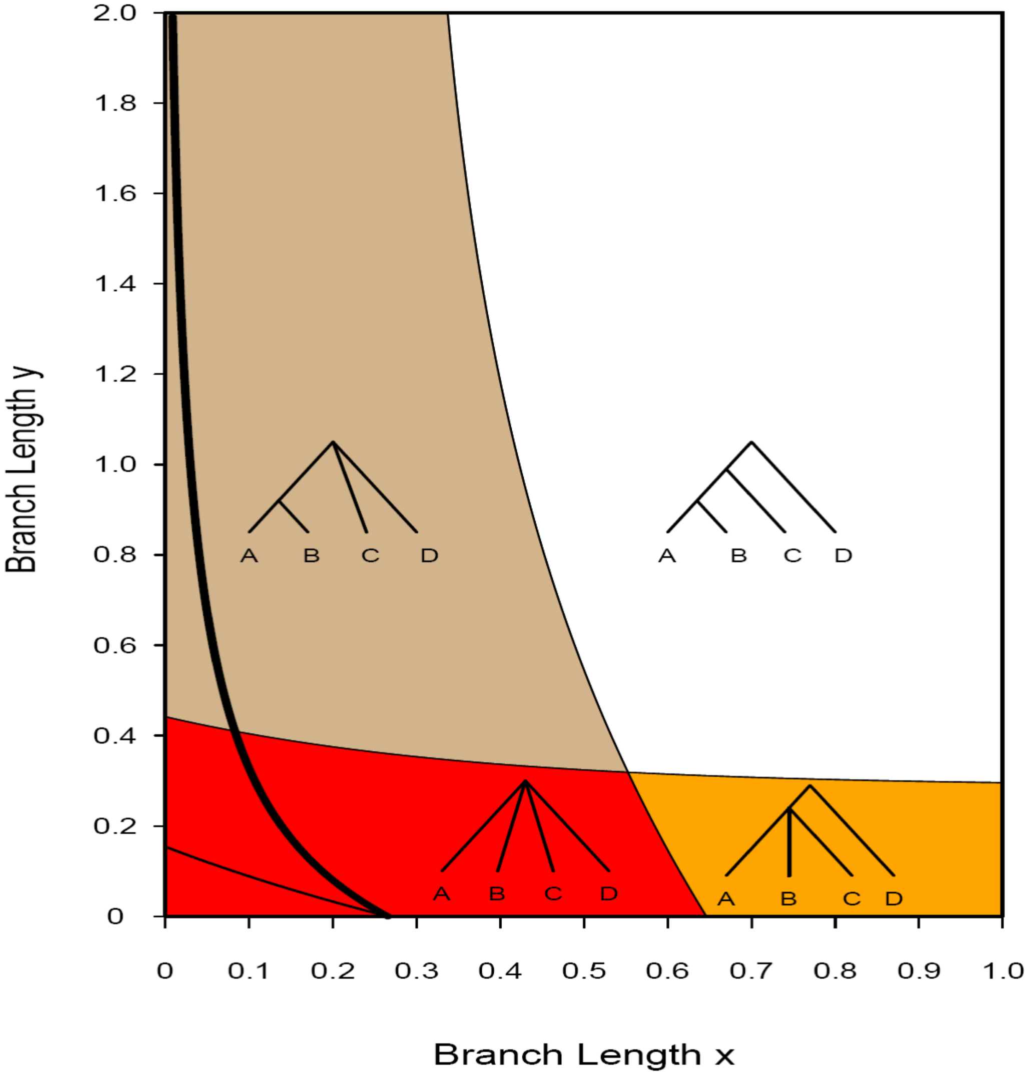


Greedy consensus tree

Inconsistency of greedy consensus



Majority-rule: unresolved zone



Are consensus trees inconsistent estimators of species trees?

Majority Rule. (i) Majority-rule asymptotic consensus trees (MACTs) do not have any clade not on the species tree. (ii) Majority-rule unresolved zones exist for any species tree topology with $n \geq 3$ species.

Greedy Consensus. Greedy asymptotic consensus trees (GACTs) can be misleading estimators of species trees for the 4-species asymmetric tree and for any species tree with $n > 4$ species.

R* Consensus. R* asymptotic consensus trees (RACTs) always match the species tree.

ASTRAL. Yes. An intuitive explanation is that the most likely unrooted four-taxon gene tree matches the unrooted species tree.

Notes

Many methods have been developed for inferring species trees from gene trees and/or multilocus sequence data. ASTRAL seems to be the most popular right now of the two-stage methods (first gene trees, then species trees).

Other methods use sequence data directly, for example *BEAST uses a Bayesian approach to jointly model infer posterior probabilities for the gene trees and species tree with parameters (ancestral divergence times and population sizes). These methods tend to be computationally intensive, making ASTRAL still a popular method.

A different method is SVDquartets (Chifman and Kubatko), which uses invariants in the site patterns. These is often compared to ASTRAL and does well, but not quite as well.

For two-stage methods, there has been a trend from rooted methods to unrooted methods, where the root is then inferred from an outgroup.

These methods (both from sequence data and from gene trees) have been extended to model ancestral hybridization.