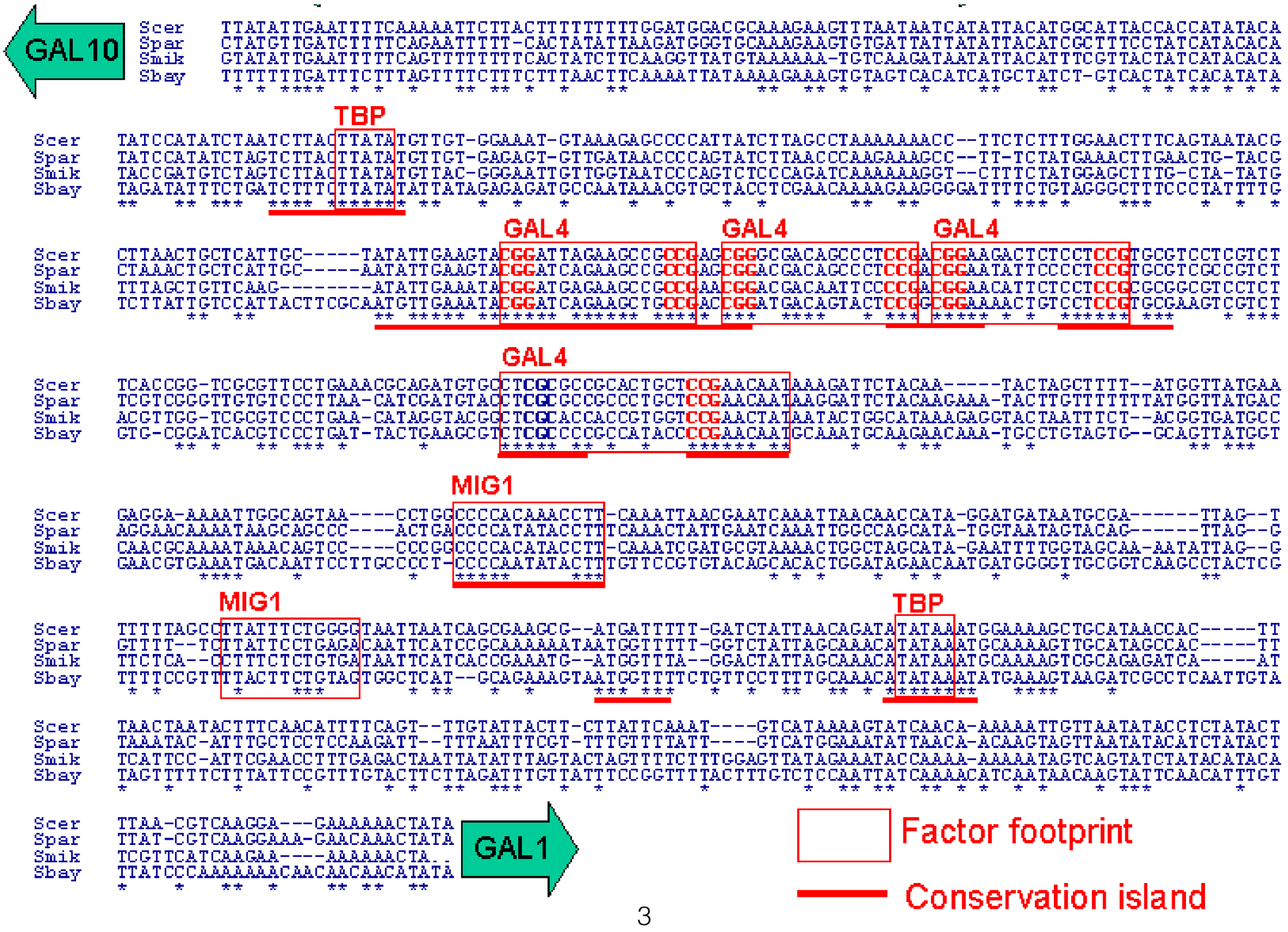
An HMM-based Comparative Genomic Framework for Analyzing Complex Evolutionary Scenarios

Kevin J. Liu
Department of Computer Science
Rice University

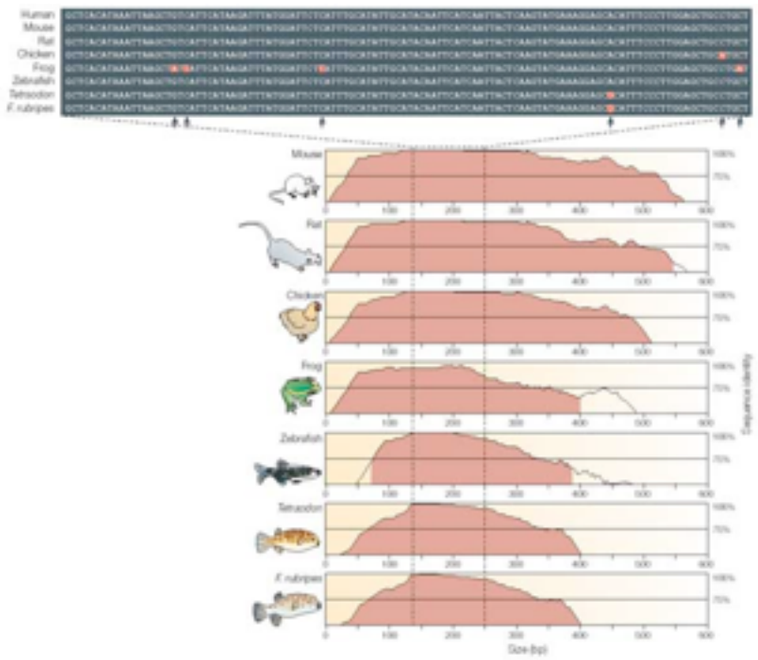
Comparative Genomics: Background

An Example Comparative Genomic Analysis (Nature 423 2003)



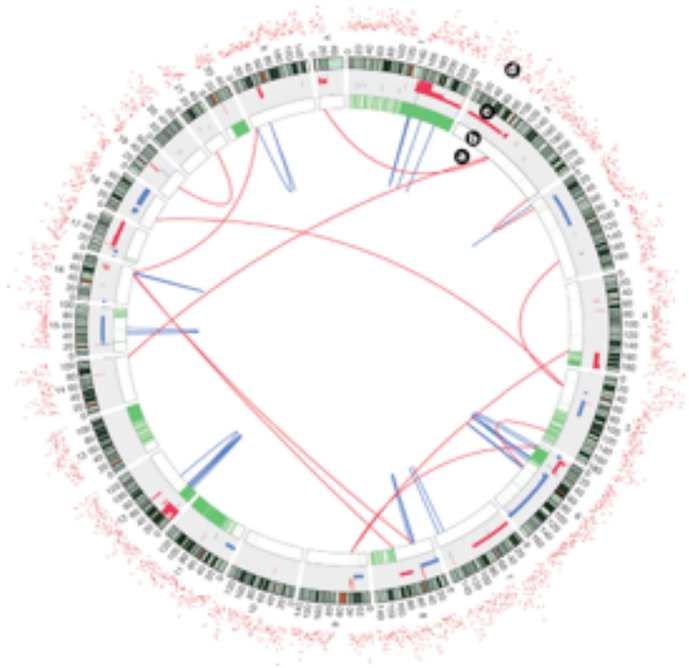
Applications of Comparative Genomics

Detecting regulatory elements



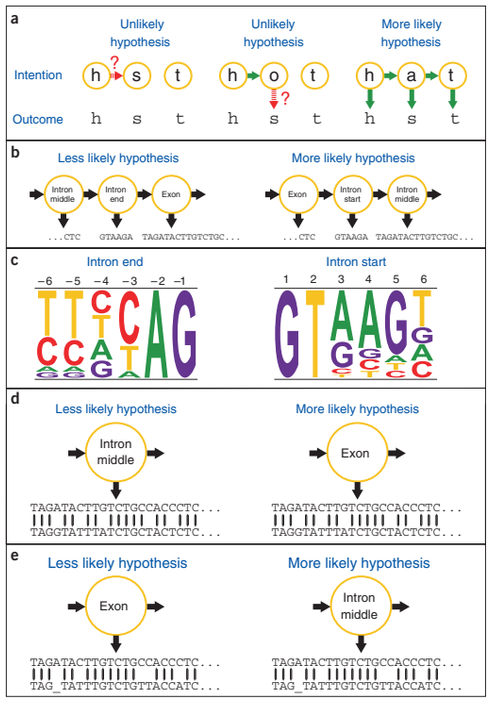
(Nature Reviews Genetics 5, 2004)

Detecting cancer mutations



(Nature 465, 2010)

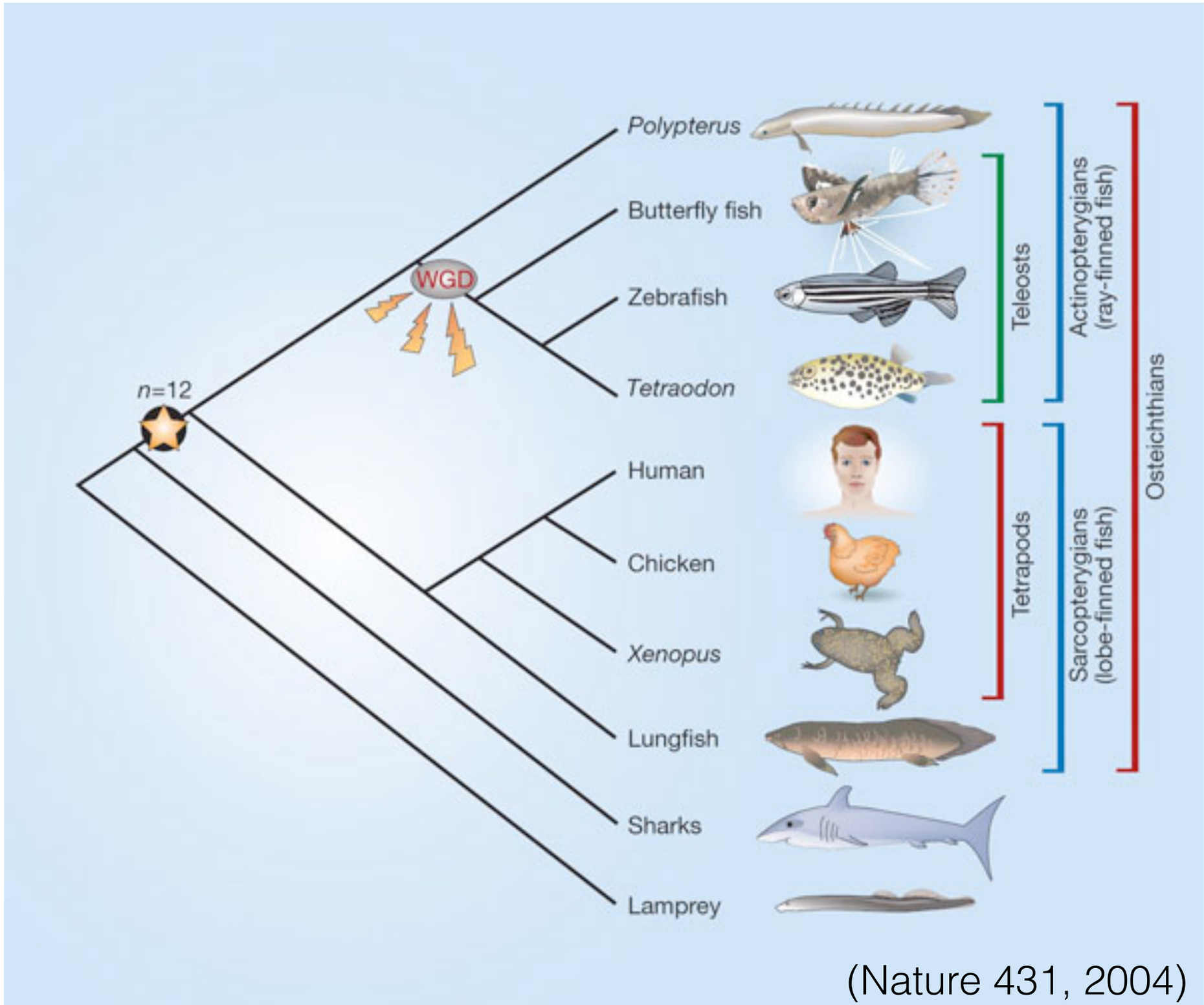
Gene finding



(Nature Biotechnology 25, 2007)

And many, many more …

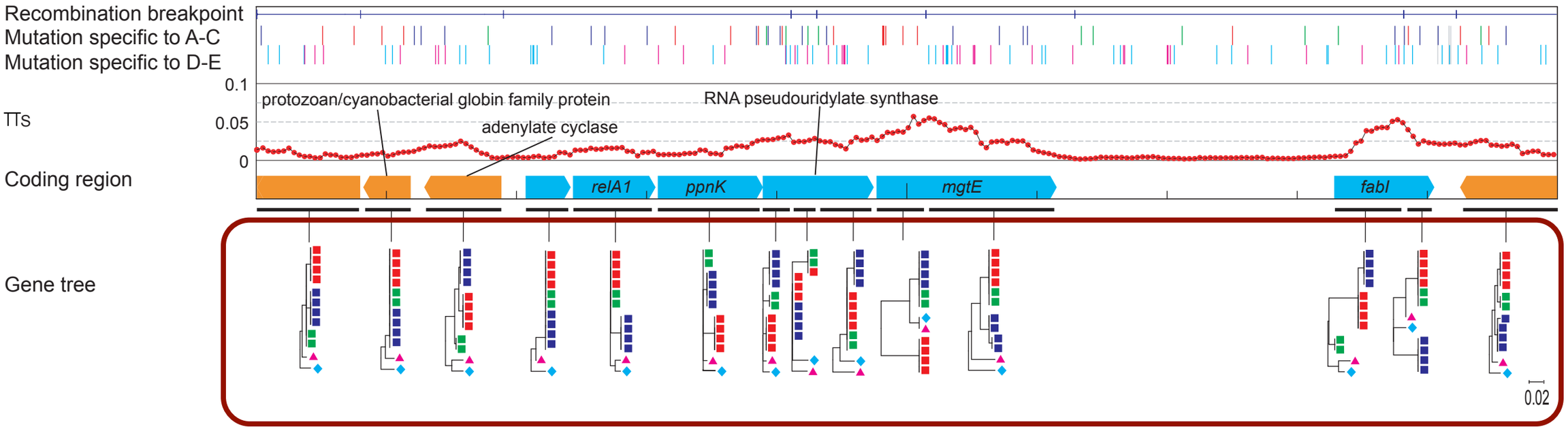
Almost all comparative genomic approaches assume that genomes have evolved down a tree.



(Nature 431, 2004)

- However, it has been shown that:
  - different genomic regions might evolve down different trees, and
  - the set of species might not have evolved in a strictly diverging manner.

(MBE 29, 2013)

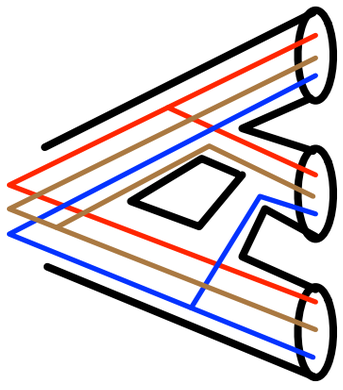


different gene trees for different regions in the Staph aureus genomes, due to horizontal gene transfer!

Comparative Genomics: Going Beyond Trees

A Machine Learning View of Comparative Genomics

Species network
(DAG)
+
gene trees



Genomes

Stochastic
Generative
Model

Observed Data
(Genomic sequences)

Overarching Goal

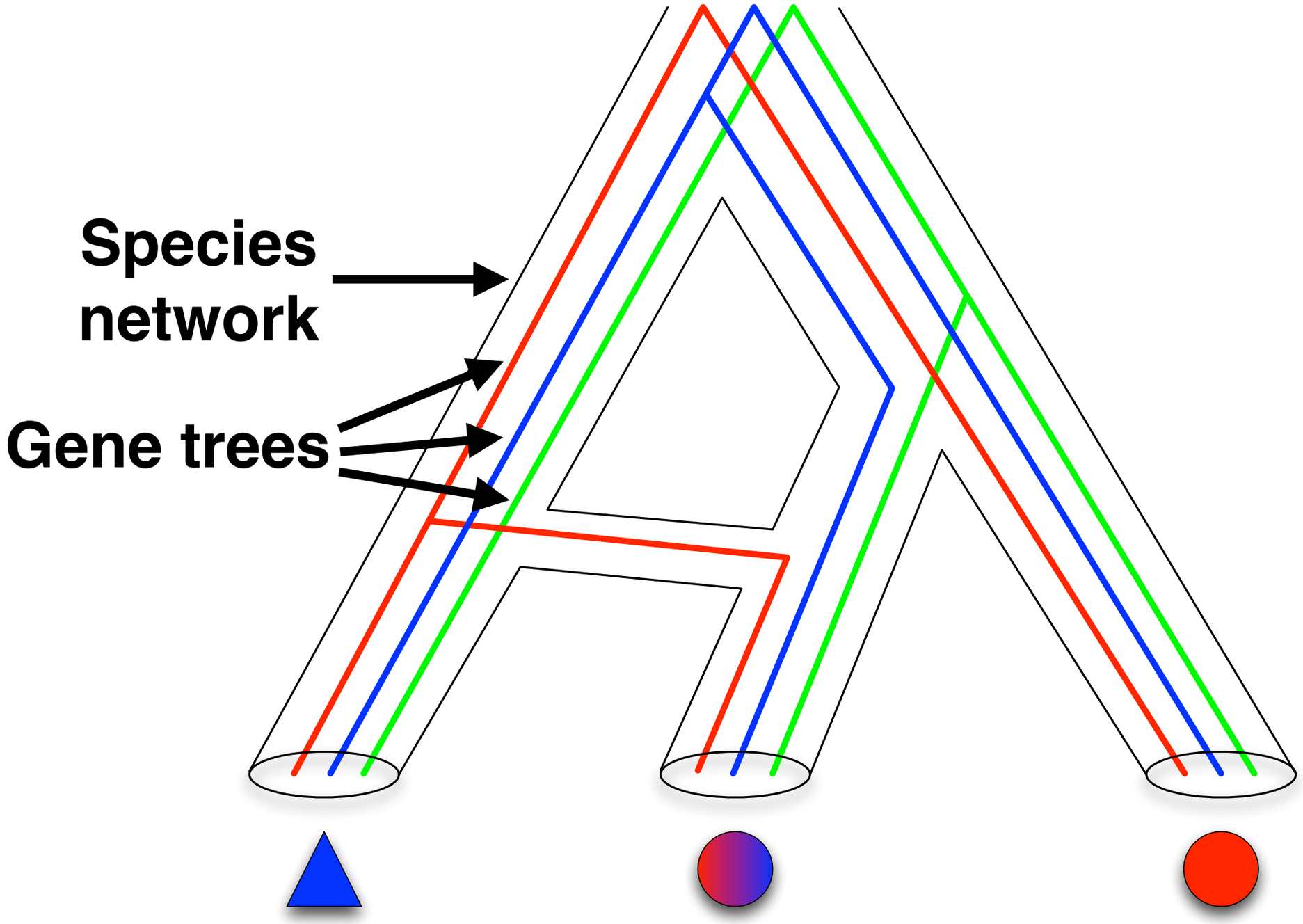
- For every site in the genome, learn:
  - the local gene tree along which the site evolved, and
  - the evolutionary trajectory that the local gene tree took within the species network.
- We also want a confidence measure for the inference.

My Approach

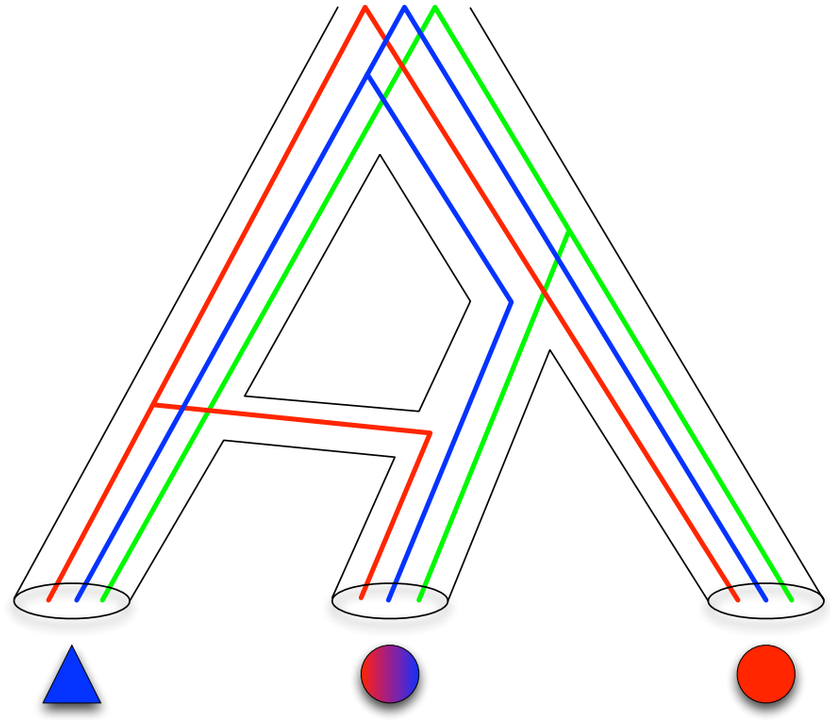
- Modeling: Combine species networks and hidden Markov models into one unified framework, PhyloNet-HMM.

- Inference: Using genomic sequence data, the task is to learn the model.

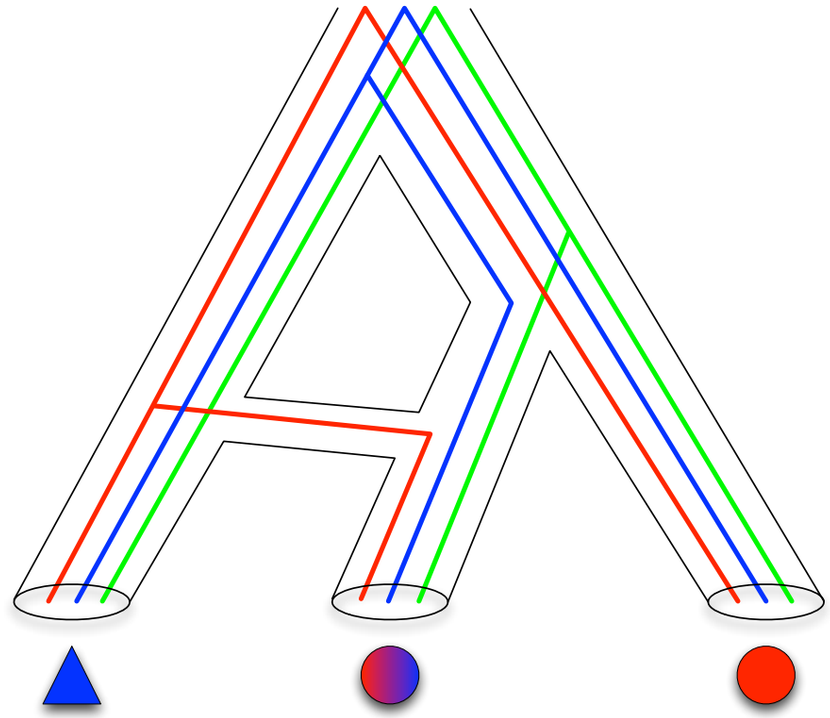
Gene Trees with Different Trajectories in a Species Network



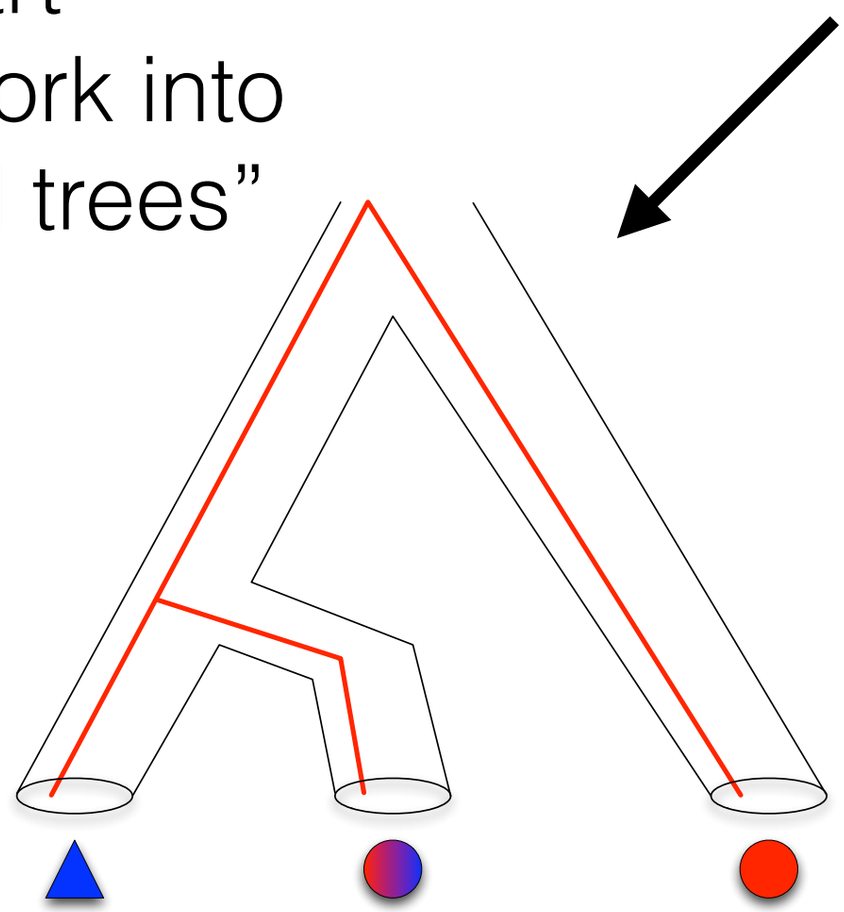
Disentangling Gene Tree Trajectories

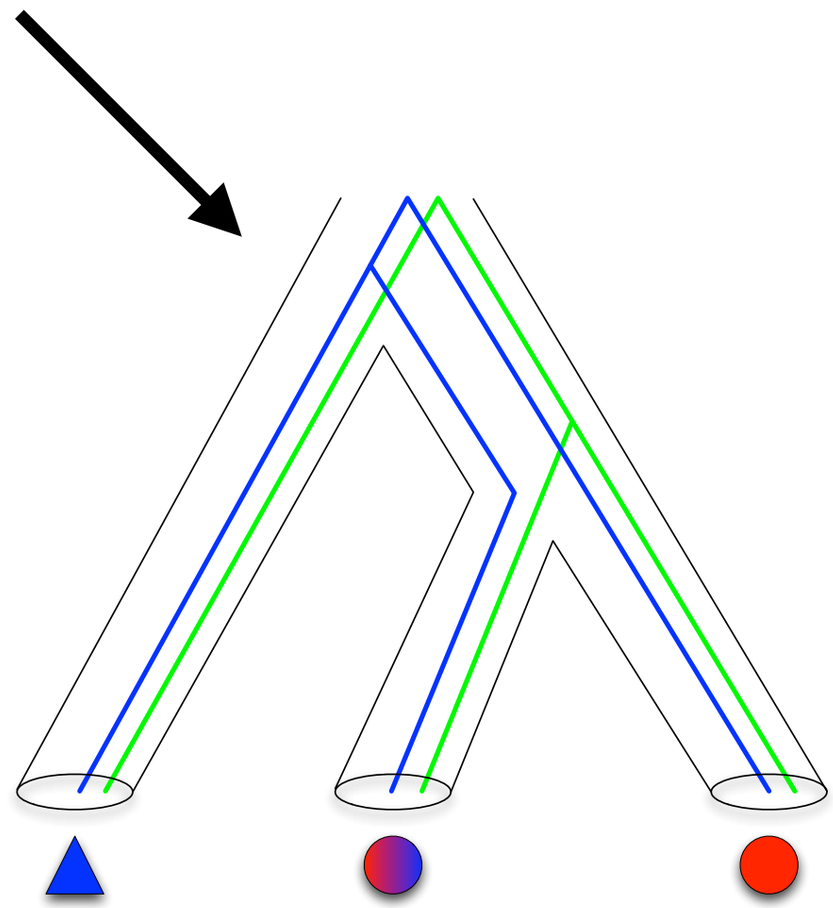


Disentangling Gene Tree Trajectories

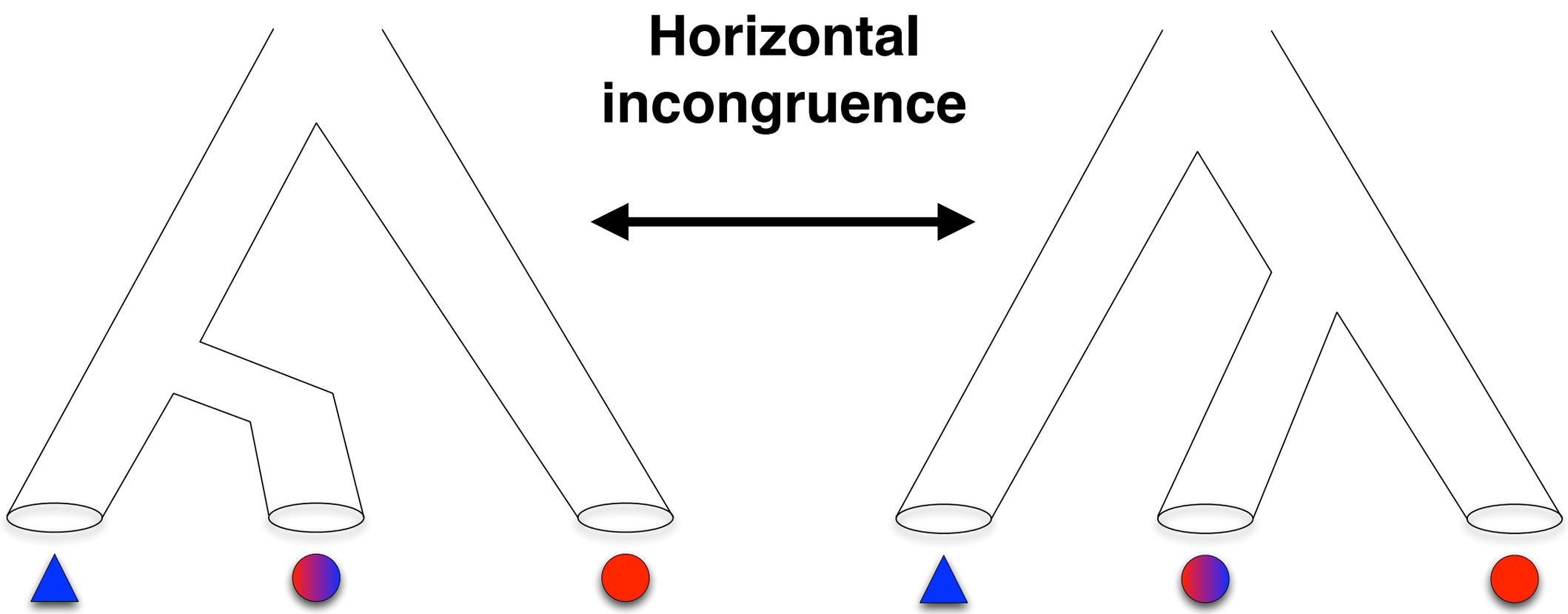


“Pull apart” species network into two “parental trees”





“Horizontal” and “Vertical” Incongruence

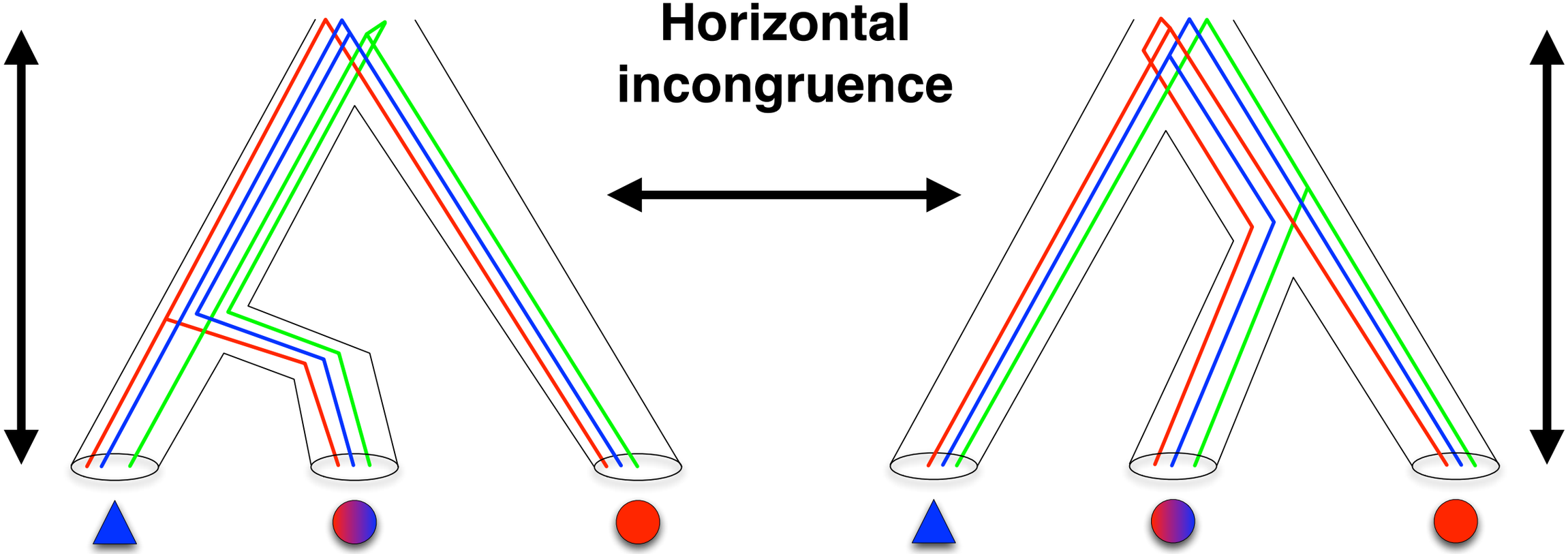


“Horizontal” and “Vertical” Incongruence

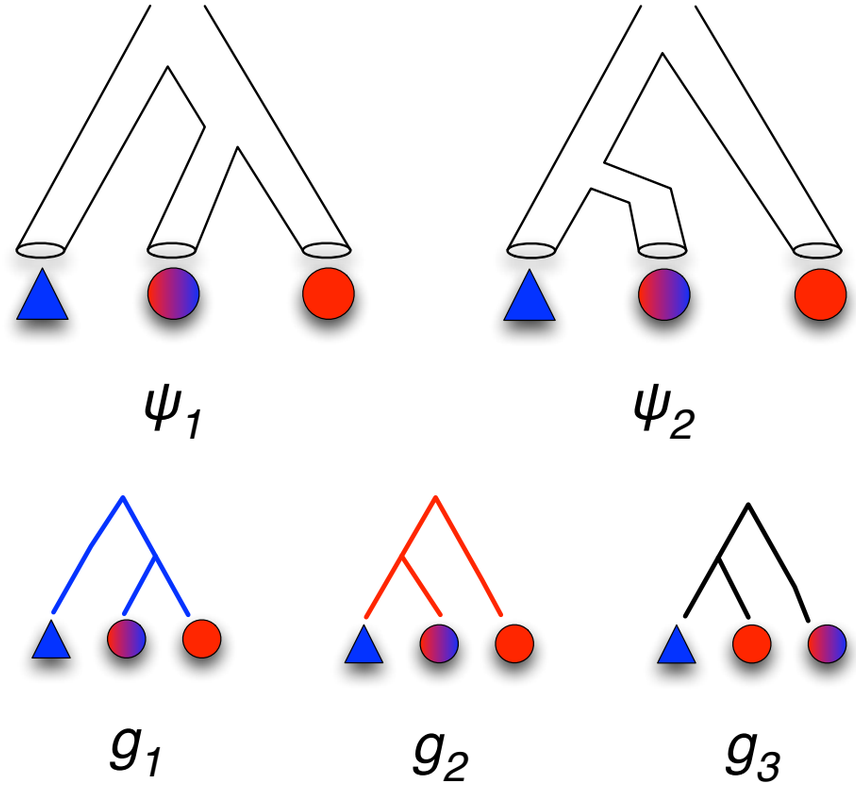
Vertical incongruence

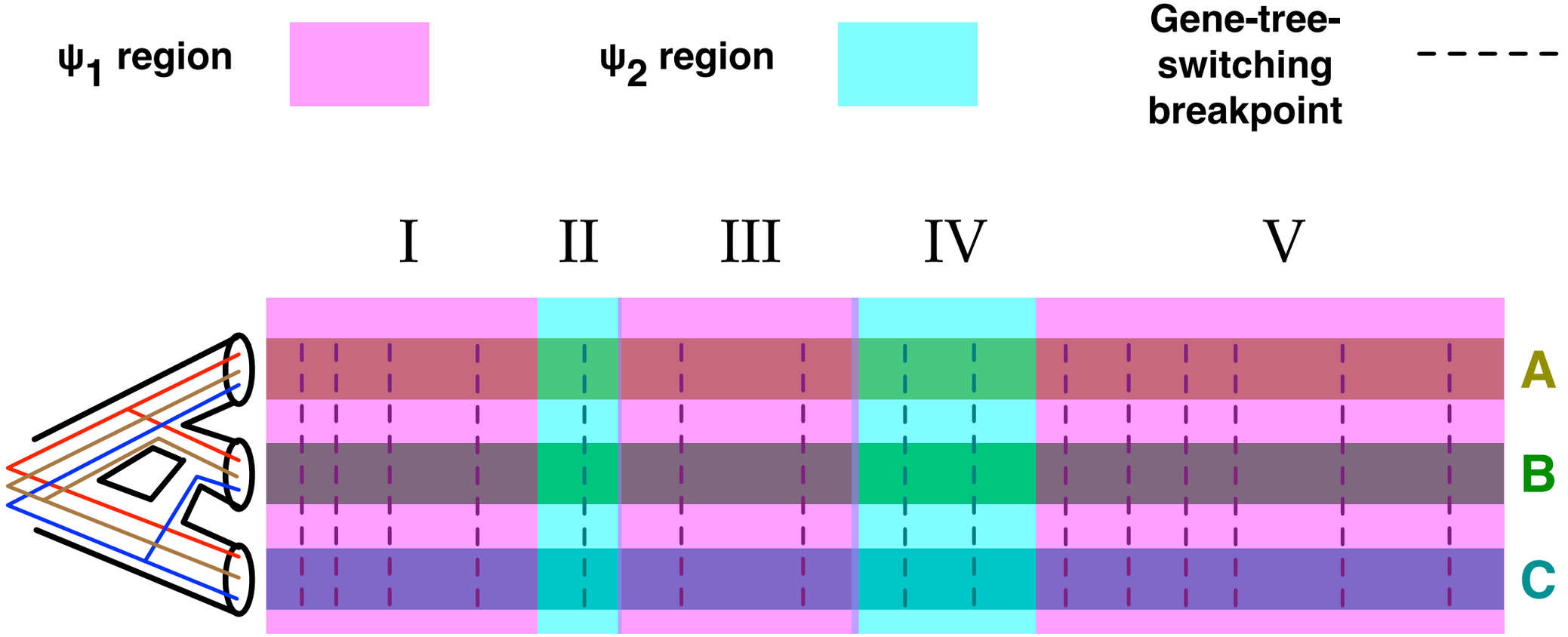
Vertical incongruence

Horizontal incongruence



A Sequence-Level View of Local Incongruence





Insight #1

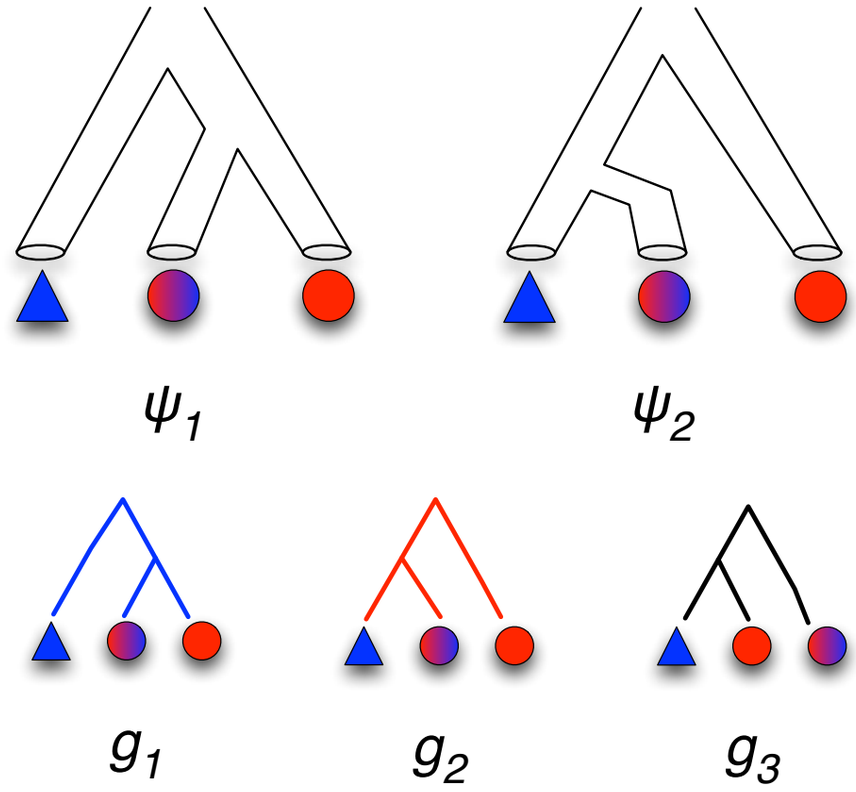
- “Horizontal” and “vertical” incongruence between neighboring gene trees represent two different types of dependence.
- Model the two dependence types using two classes of transitions in a graphical model.

Insight #2

- DNA sequences are observed, not gene trees.
- Under traditional models of DNA sequence evolution, the probability $P(s|g)$ of observing DNA sequences s given a gene tree g can be efficiently calculated using dynamic programming.

Insight #1 + Insight #2 = Use a Hidden Markov Model (HMM)

PhyloNet-HMM: Problem Definition



For each site $1 \leq i \leq k$, let π_i be a random variable that takes a value from the set $(g_x, \psi_y) : g_x \in G(n), \psi_y \in \Psi$.

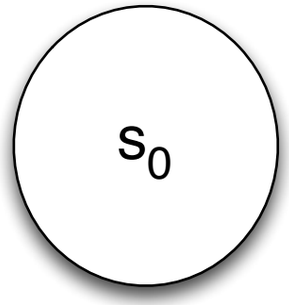
Input: A set S of n aligned genomes, each of length k, and a set Ψ of parental trees corresponding to a species network.

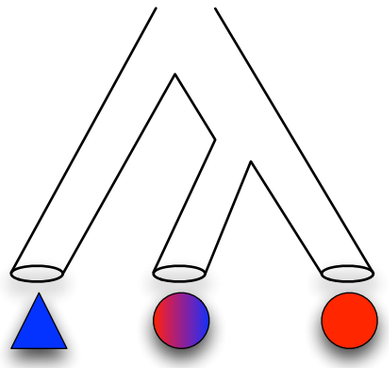
Output: For each site $1 \leq i \leq k$, the probability

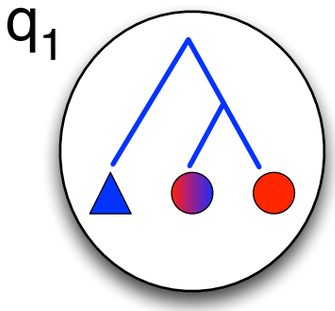
$$\mathbf{P}(\pi_i = (g_x, \psi_y)|S)$$

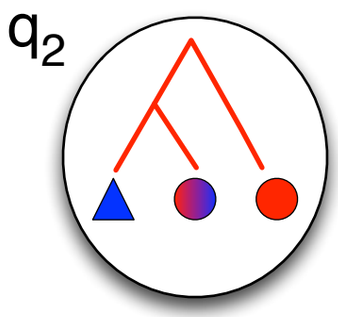
for every $g_x \in G(n)$ and $\psi_y \in \Psi$.

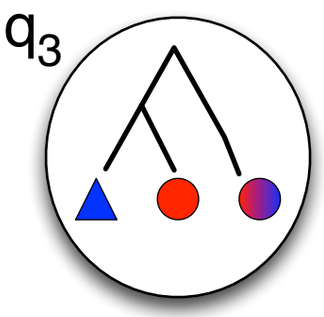
PhyloNet-HMM: Hidden States

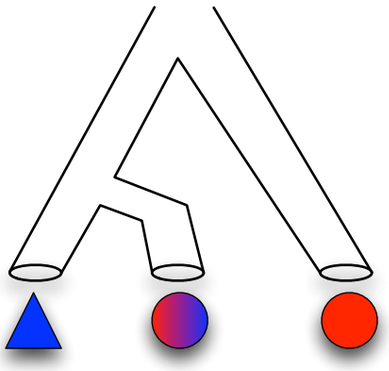


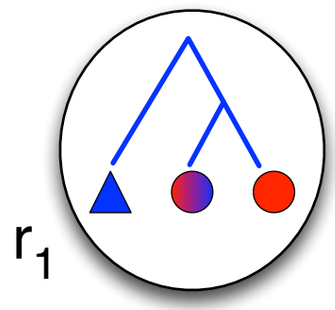


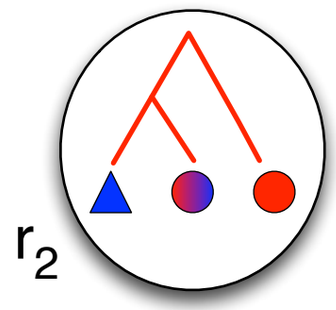


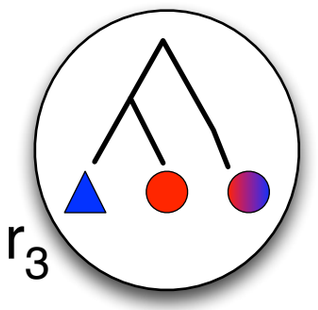




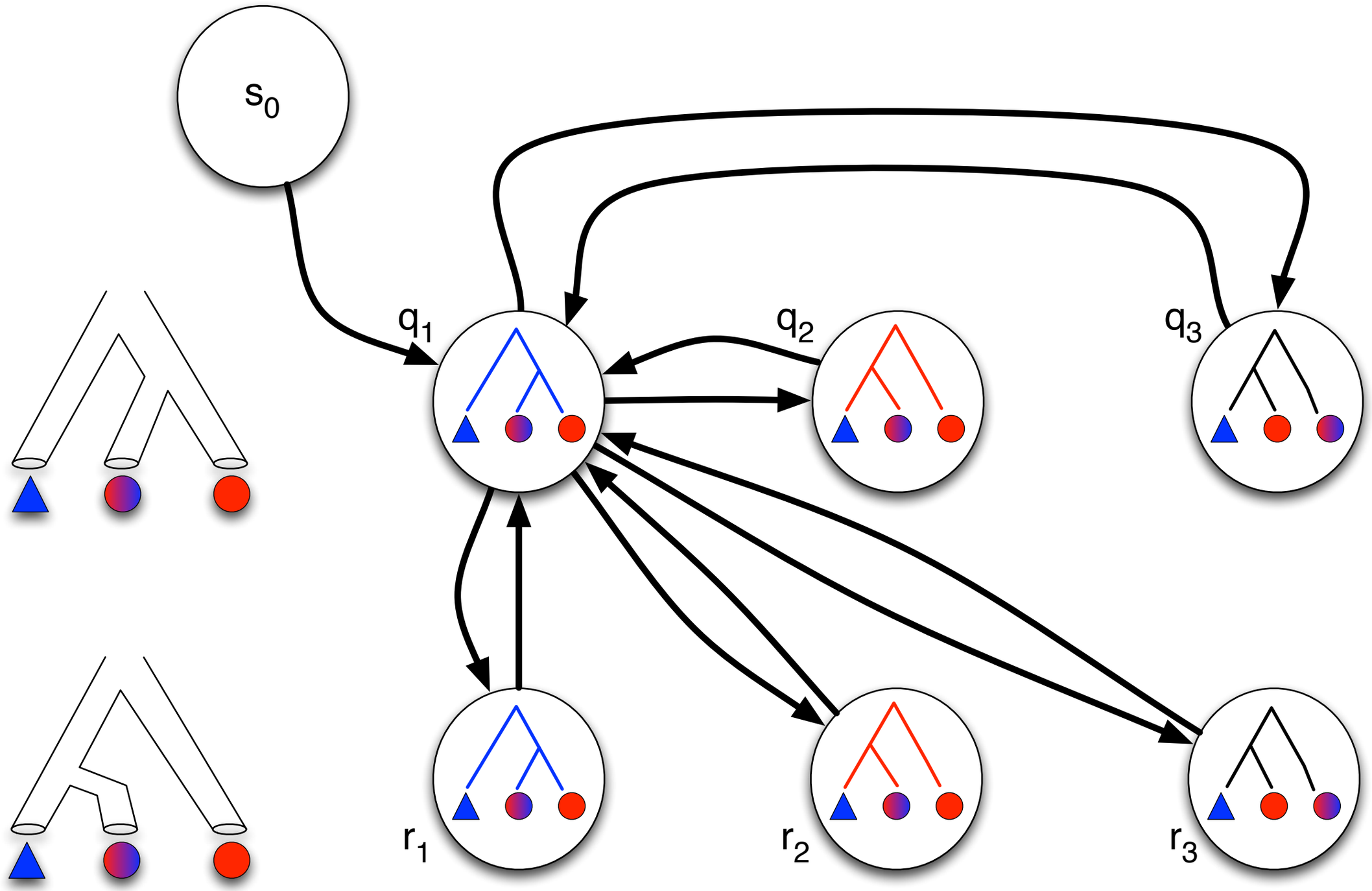








PhyloNet-HMM: Hidden States and Transitions Involving q_1



PhyloNet-HMM

- Each hidden state s_i is associated with a gene tree $g(s_i)$ contained within a "parental" tree $f(s_i)$
- The set of HMM parameters λ consists of
  - The initial state distribution π
  - Transition probabilities

$$a_{ij} = \begin{cases} \mathbf{P}(g(s_i)|f(s_i)) \cdot \gamma & \text{if } s_i \text{ and } s_j \text{ in different rows} \\ \mathbf{P}(g(s_i)|f(s_i)) \cdot (1-\gamma) & \text{if } s_i \text{ and } s_j \text{ in same row} \end{cases}$$

    where γ is the "horizontal" parental tree switching frequency.

  - The emission probabilities $b_i = \mathbf{P}(O_t|g(s_i))$

Three Problems Addressed Using PhyloNet-HMM

1. What is the likelihood of the model given the observed DNA sequences?

   - Forward algorithm calculates prefix probability $\alpha_t(i) = \mathbf{P}(O_1, O_2, \ldots, O_t, q_t = S_i|\lambda)$

   - Backward algorithm calculates suffix probability $\beta_t(i) = \mathbf{P}(O_{t+1}, O_{t+2}, \ldots, O_k|q_t = S_i, \lambda)$

   - Model likelihood is $\mathbf{P}(O|\lambda) = \sum_{i=1}^{N} \alpha_k(i)$

2. Which sequence of hidden states best explains the observed DNA sequences?

   - Posterior decoding probability $\gamma_t(i)$ is the probability that HMM is in state s_i at time t, calculated as:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\mathbf{P}(O|\lambda)}$$

3. How do we choose parameter values that maximize the model likelihood?

   - Apply hill-climbing to optimize $\arg\max_{\lambda} \mathbf{P}(O|\lambda)$

Related Methods

1. Methods that work for at most three genomes, including:

   - D-statistic (Durand *et al.* 2012)
   - CoalHMM (Mailund *et al.* 2012)

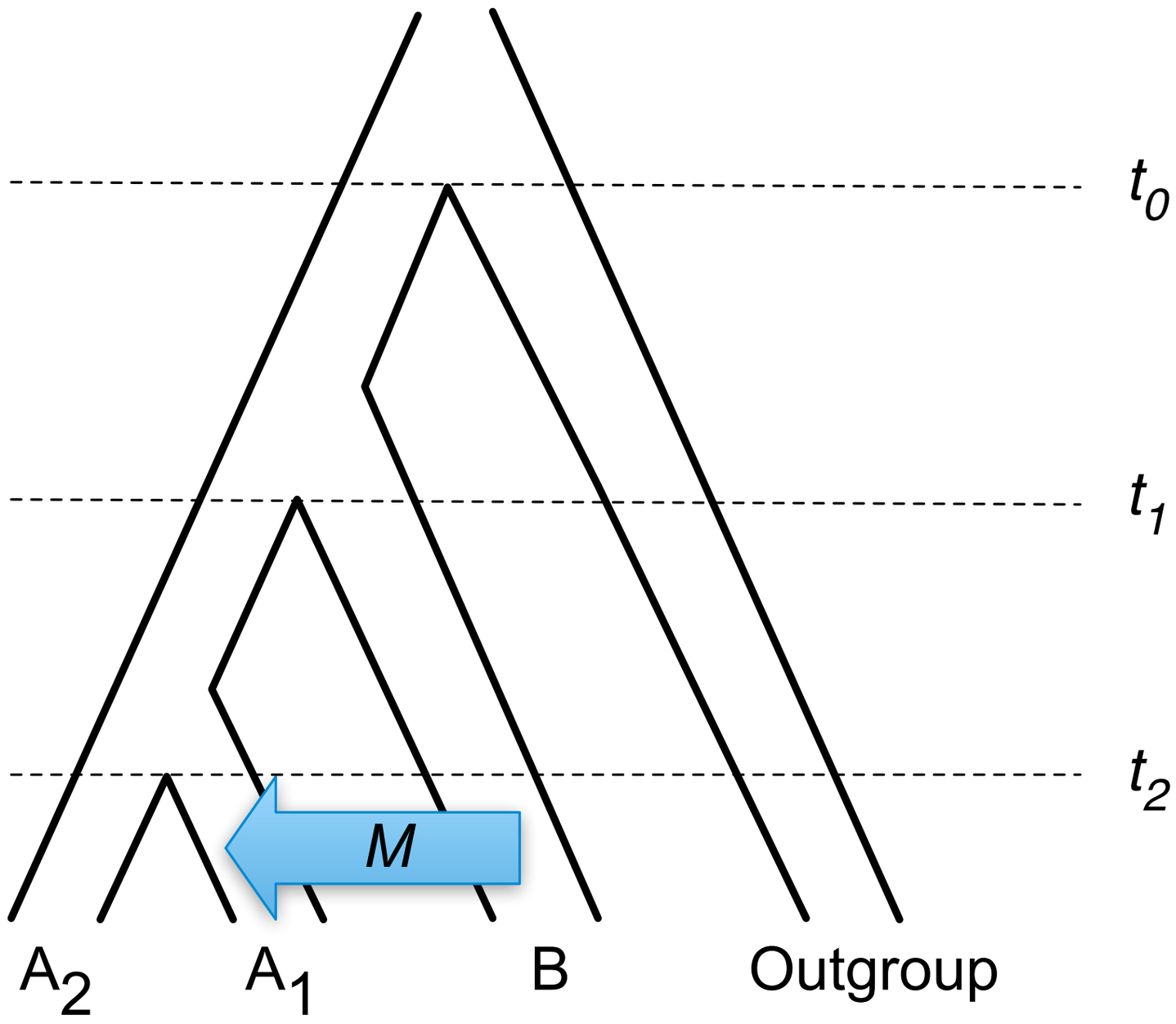
2. Methods that consider vertical incongruence or horizontal incongruence but not both, including:

   - CoalHMM (Hobolth *et al.* 2007, Schierup *et al.* 2009)
   - RecHMM (Westesson and Holmes 2009)

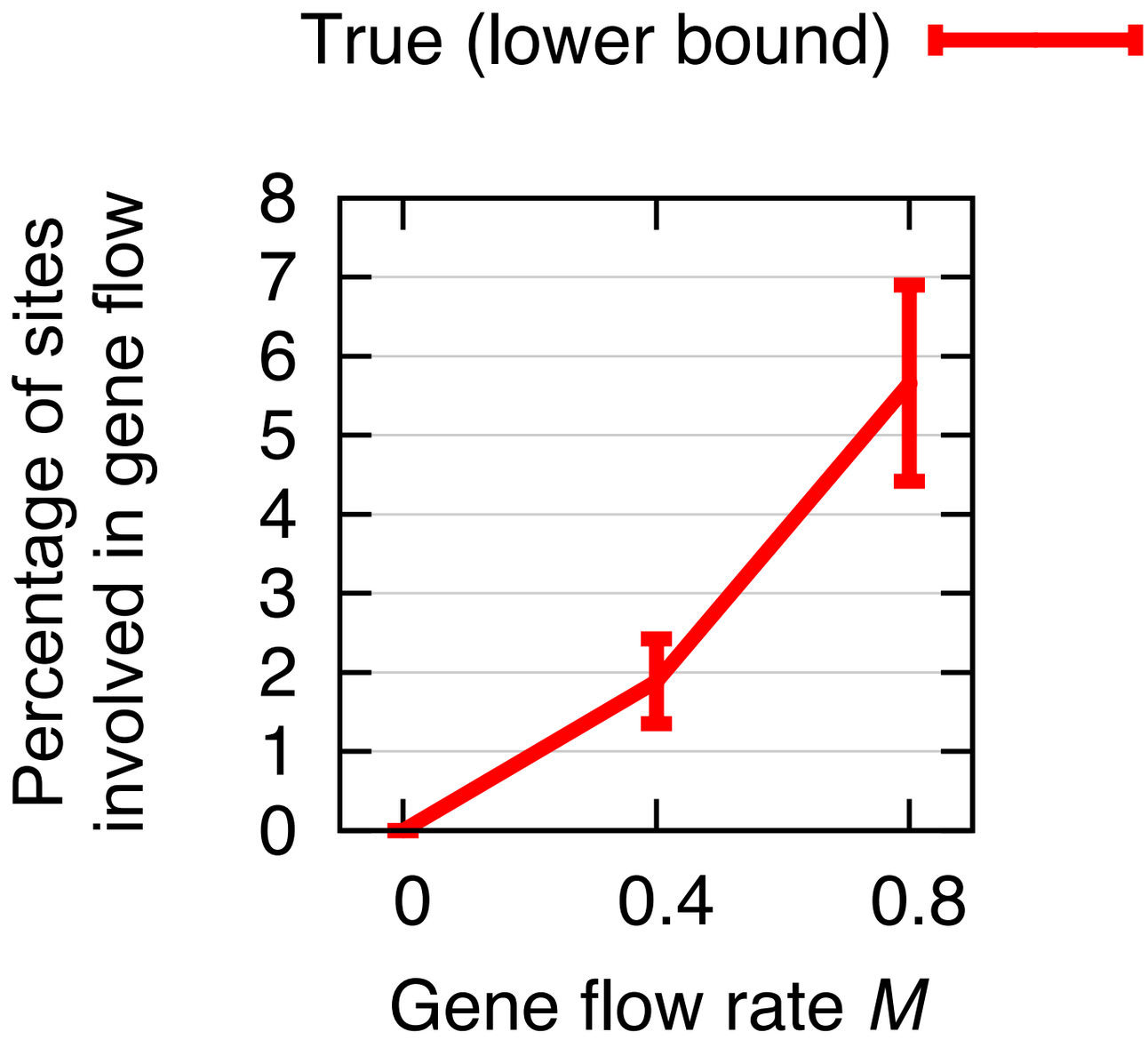
Evaluating PhyloNet-HMM

- Simulation study using:
  - Species tree model
  - Species network model
- Empirical study of different sets of mouse genomes:
  - Controls: lab mice, wild mice from populations that lacked gene flow
  - Additional wild mice from populations where gene flow was suspected

Simulation Model

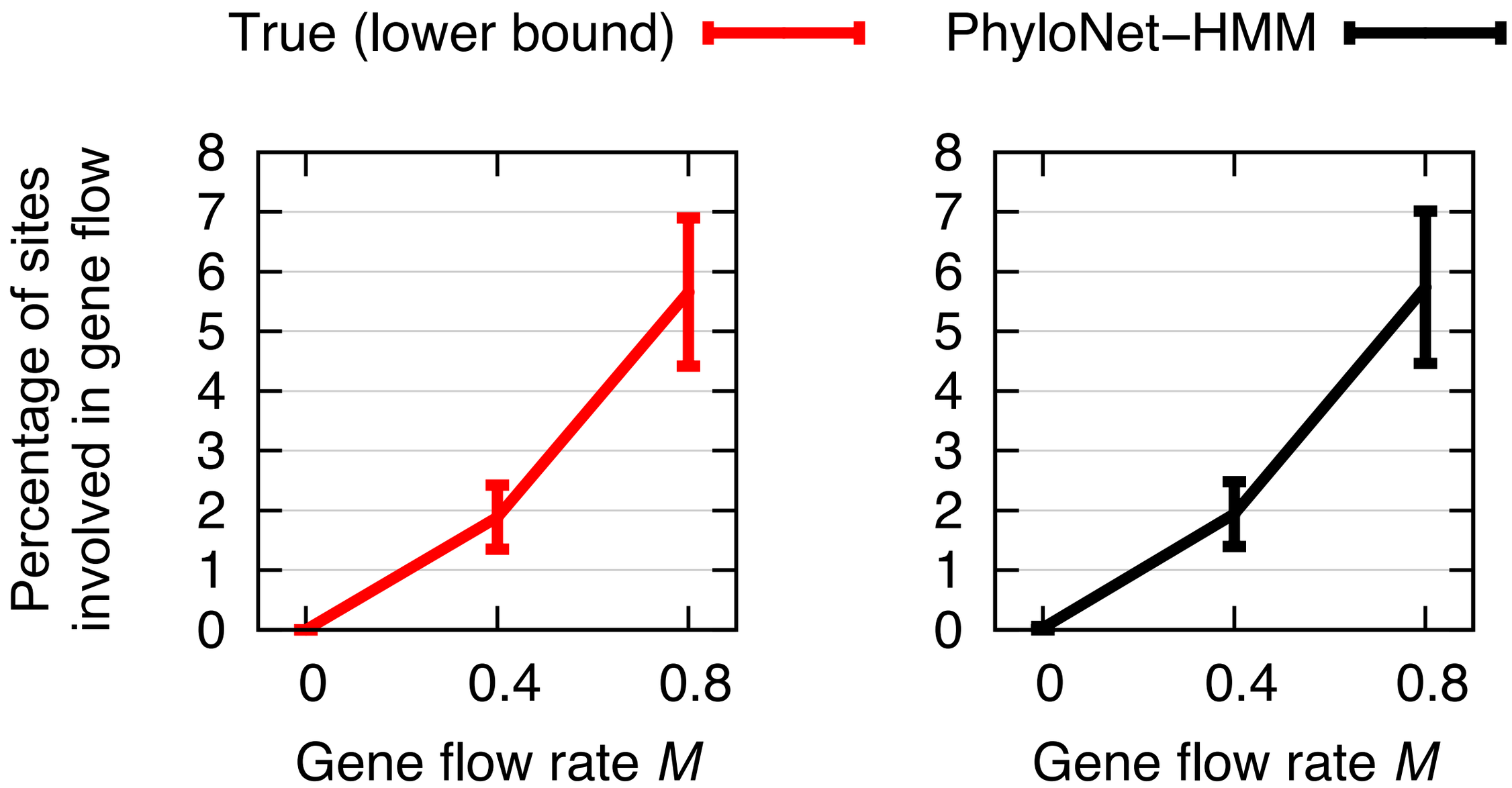


Simulation Study Results



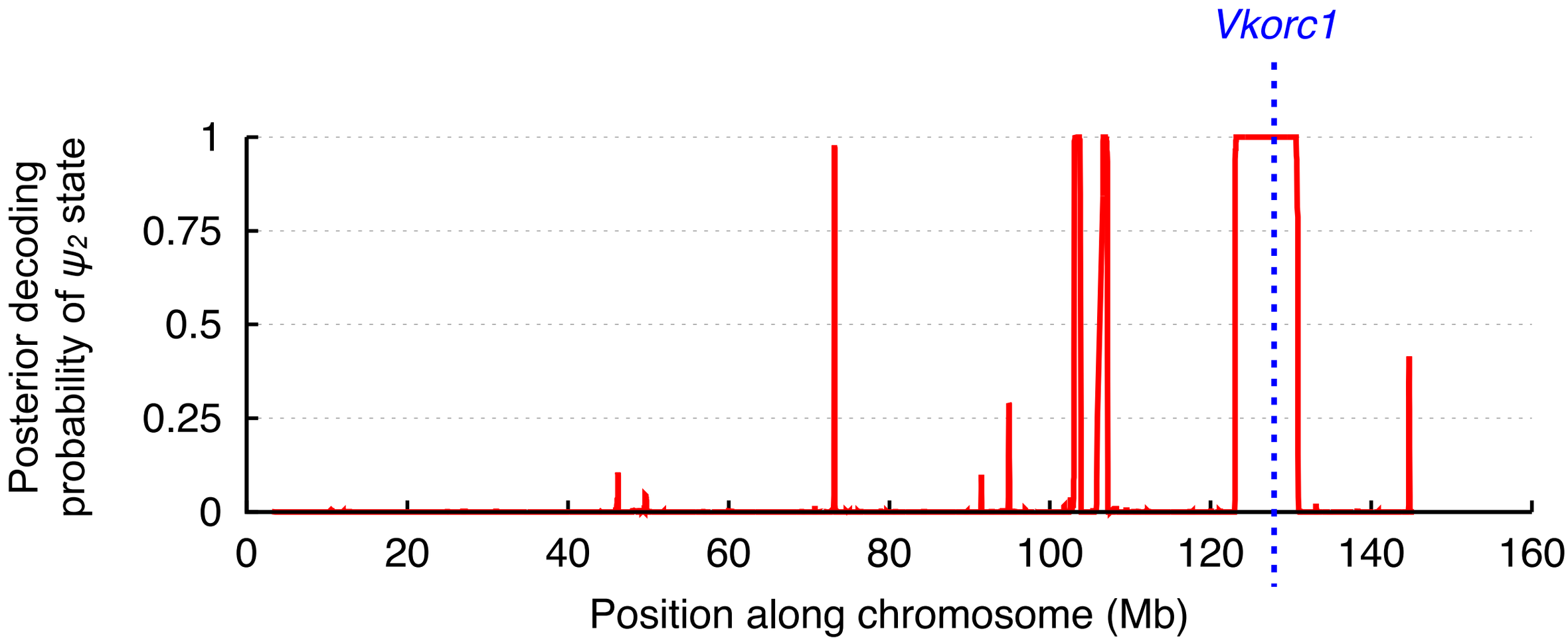
Liu *et al.*, to appear in PLoS Computational Biology.

Simulation Study Results



Liu *et al.*, to appear in PLoS Computational Biology.

Empirical Study: Non-control Mice (Chromosome 7)



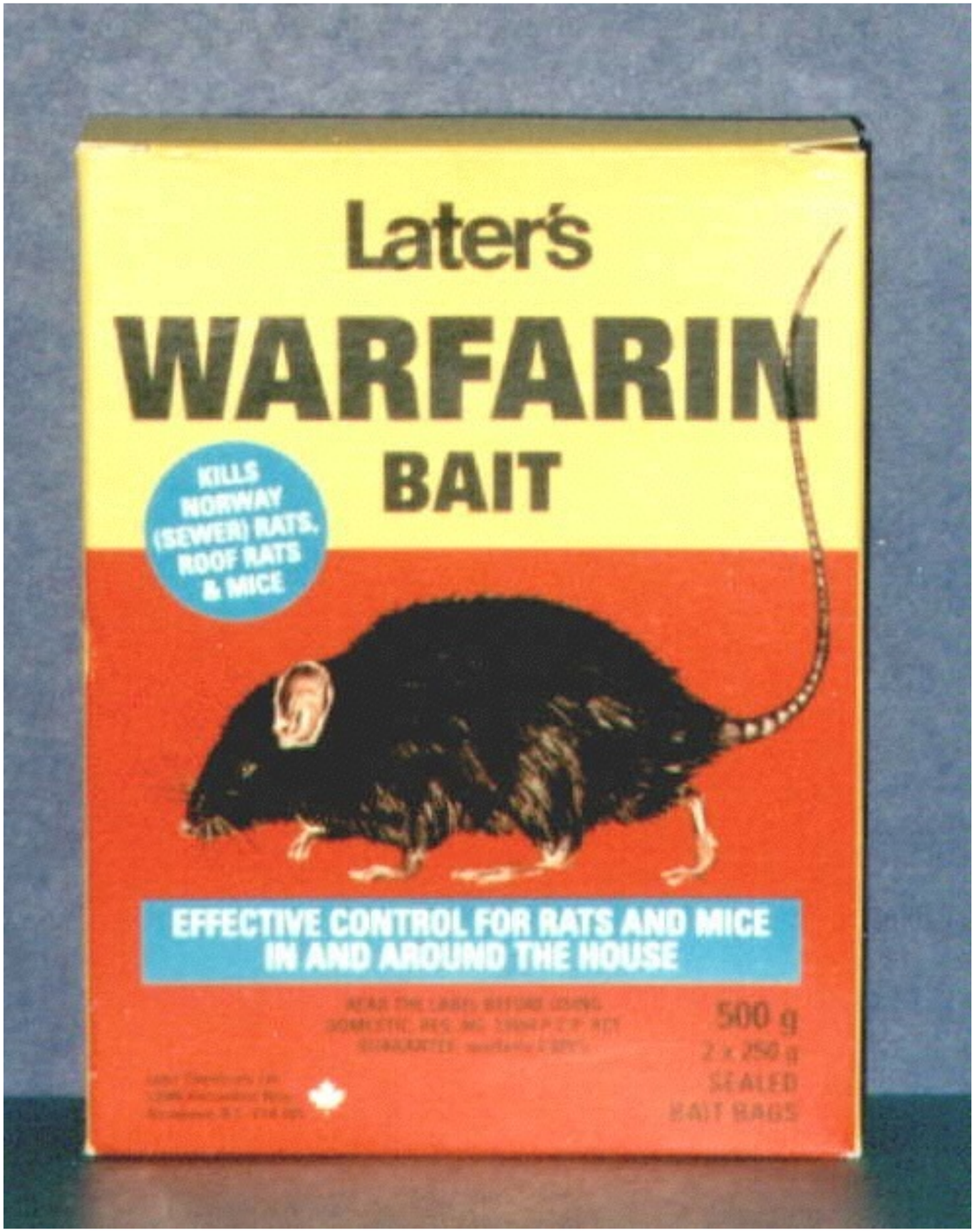
Liu *et al.*, under review by PNAS.

The *Vkorc1* Gene and Personalized Warfarin Therapy

- Mutant *Vkorc1* gene contributes to warfarin resistance
- Warfarin resistant individuals require larger-than-normal dose to prevent clotting complications (like stroke)

Rost *et al.* Nature 427, 537-541 2004.

Warfarin is Really Glorified Rodent Poison



Reproduced from UTMB.

The Spread of Warfarin Resistance in Wild Mice

- Humans inadvertently started a gigantic drug trial by giving warfarin to mice in the wild
- Mice shared genes (including one that confers warfarin resistance) to survive (Song *et al.* 2011)
    - Gene sharing occurred between two different species (introgression)
- To find out results from the drug trial, we just need to analyze the genomes of introgressed mice and locate the introgressed genes

Summary and Future Directions

Summary

- PhyloNet-HMM generalizes the basic coalescent model, one of the most widely used models in population genetics, by using a DAG in place of a tree
- Simulated and empirical data sets with tree-like and non-tree-like evolution were used to validate PhyloNet-HMM
- PhyloNet-HMM found non-tree-like evolution in multiple mouse chromosomes
  - Introgressed mouse genes confer warfarin resistance, many with related human genes
  - New candidate genes to target for improved personalization of warfarin therapy
- Study of non-tree-like evolution is a fundamentally important research topic in biology

Future Directions

- Future directions include:
  - Incorporating network search,
  - Detecting adaptive gene flow, and
  - Expanding the model and method to account for other evolutionary events (e.g., sequence insertion/deletion).
- Additional biological systems of interest include:
  - Bacterial species, where horizontal gene transfer plays an important role in the spread of antibiotic resistance,
  - Hybrid plant species, and
  - Other introgressed animal species.

Acknowledgments

Luay Nakhleh
CS

Michael Kohn
Biology

Tandy Warnow
CS (UIUC)

- Supported in part by:
  - A training fellowship from the Keck Center of the Gulf Coast Consortia, on Rice University's NLM Training Program in Biomedical Informatics (Grant No. T15LM007093).
  - NLM (Grant No. R01LM00949405 to Luay Nakhleh)
  - NHLBI (Grant No. R01HL09100704 to Michael Kohn)

Questions?

- My website: http://www.cs.rice.edu/~kl23

- Nakhleh lab website: http://bioinfo.cs.rice.edu

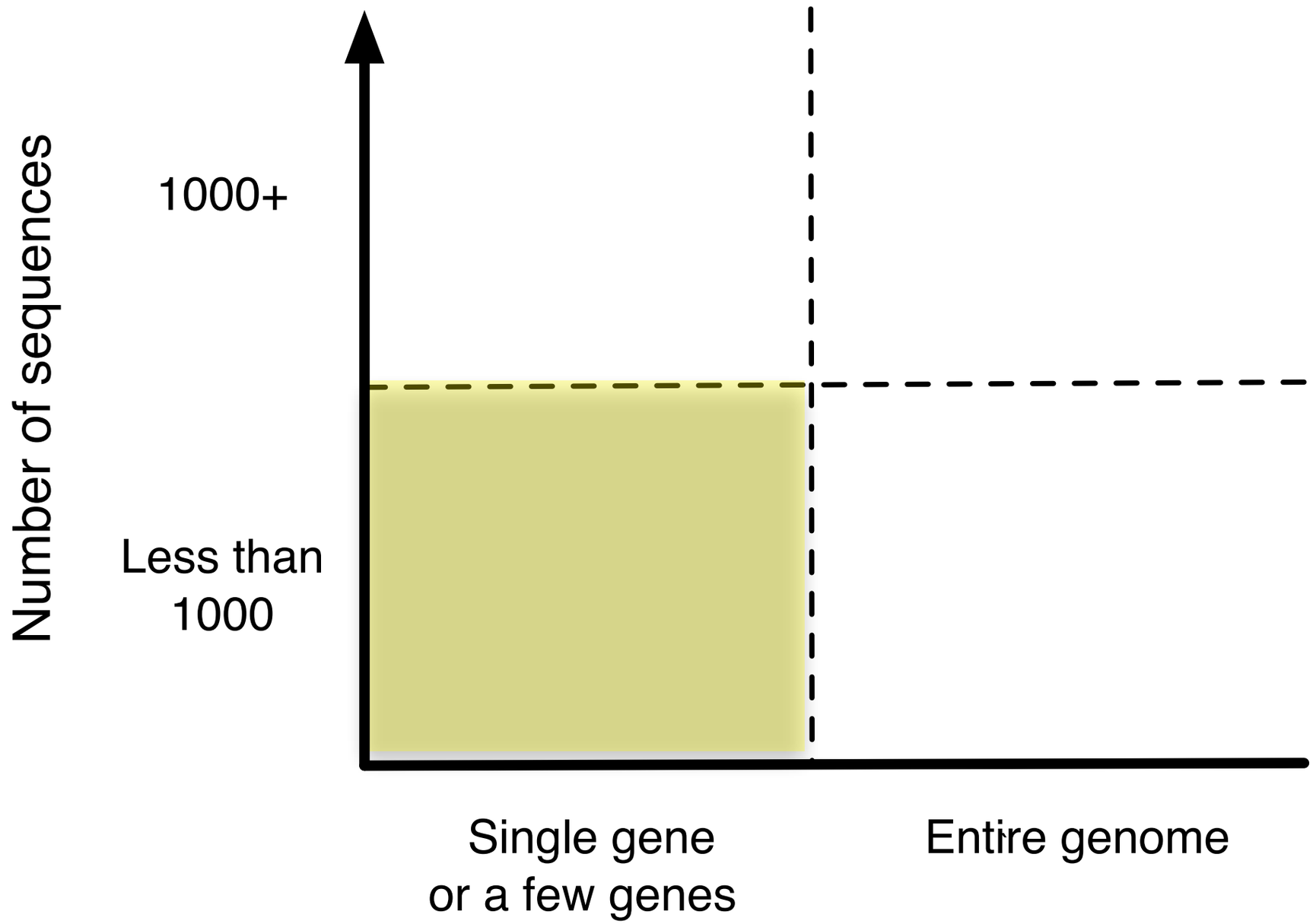
- Warnow lab website: http://www.cs.utexas.edu/~phylo

Evolution: Unifying Theme #1

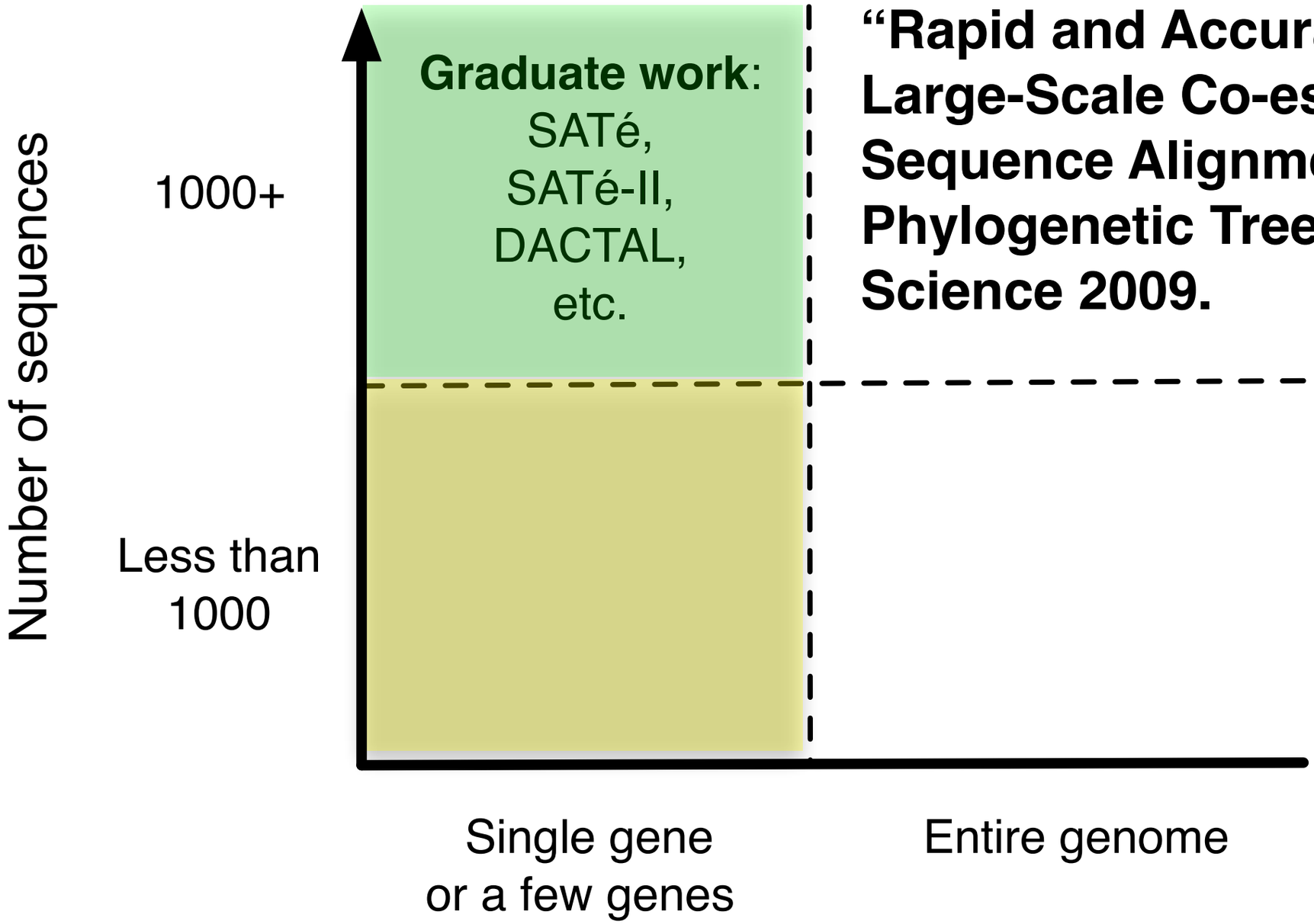
- “Nothing In Biology Makes Sense Except in the Light of Evolution” – 1973 essay by T. Dobzhansky, a famous biologist

- My primary goal: use evolutionary principles to
  - Create computational methods to analyze heterogeneous large-scale biological data,
  - Then apply them to obtain new biological and biomedical discoveries

The Pre-genomic Era

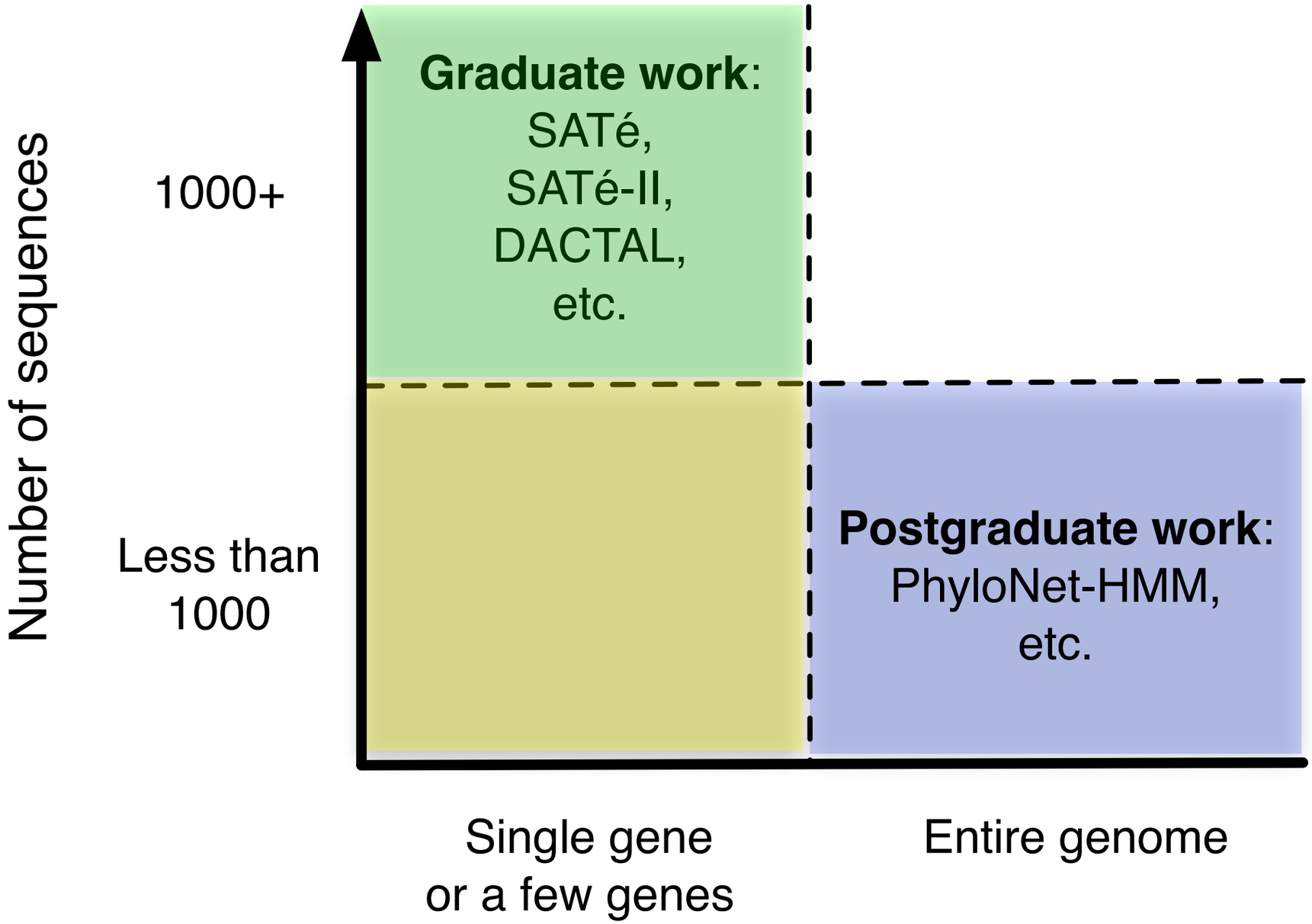


My Contributions

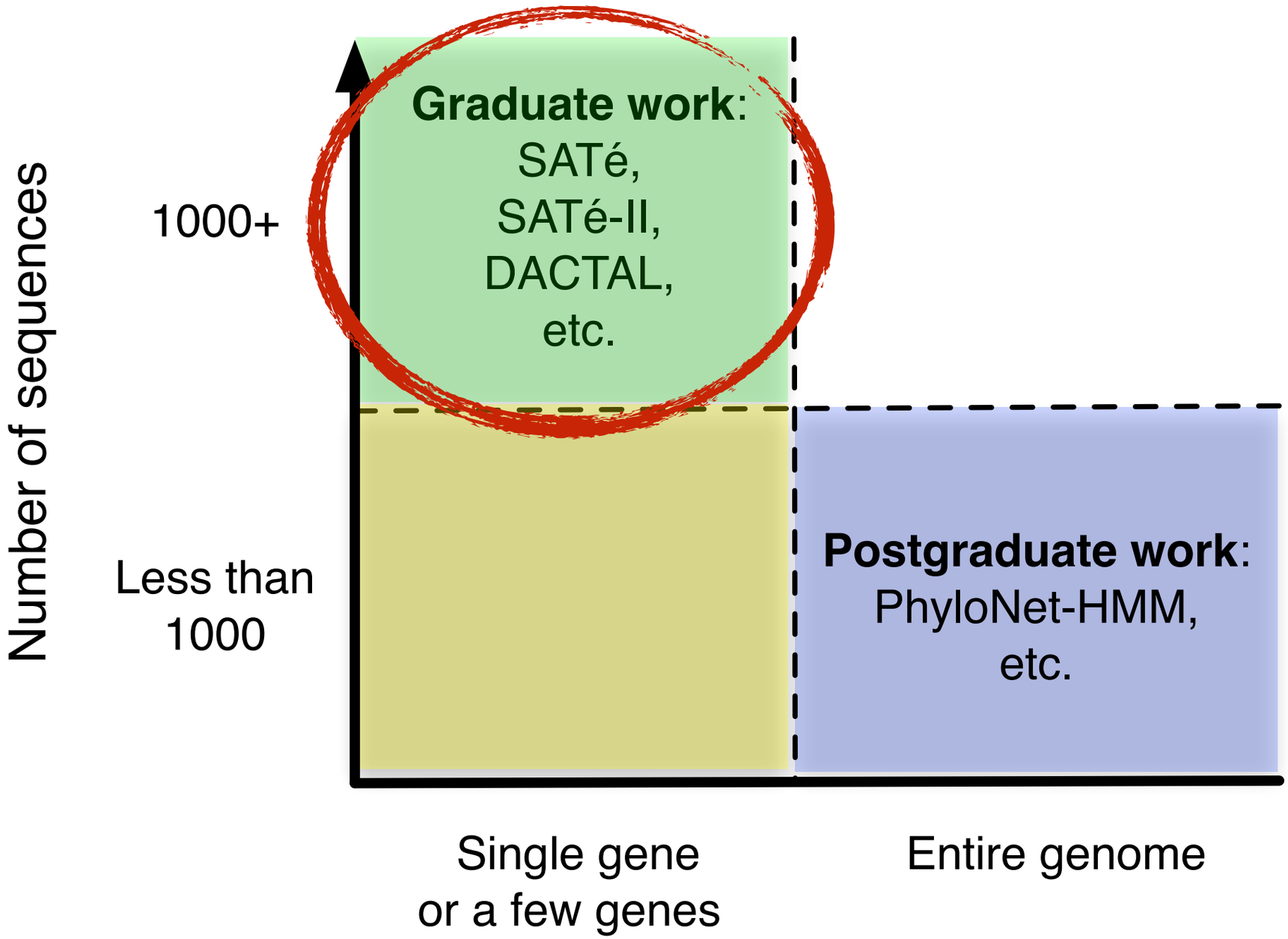


Highlight: Liu *et al.* "Rapid and Accurate Large-Scale Co-estimation of Sequence Alignments and Phylogenetic Trees," Science 2009.

My Contributions



Outline for Today's Talk



Part I: Fast and Accurate Alignment and Tree Estimation on Large-Scale Datasets

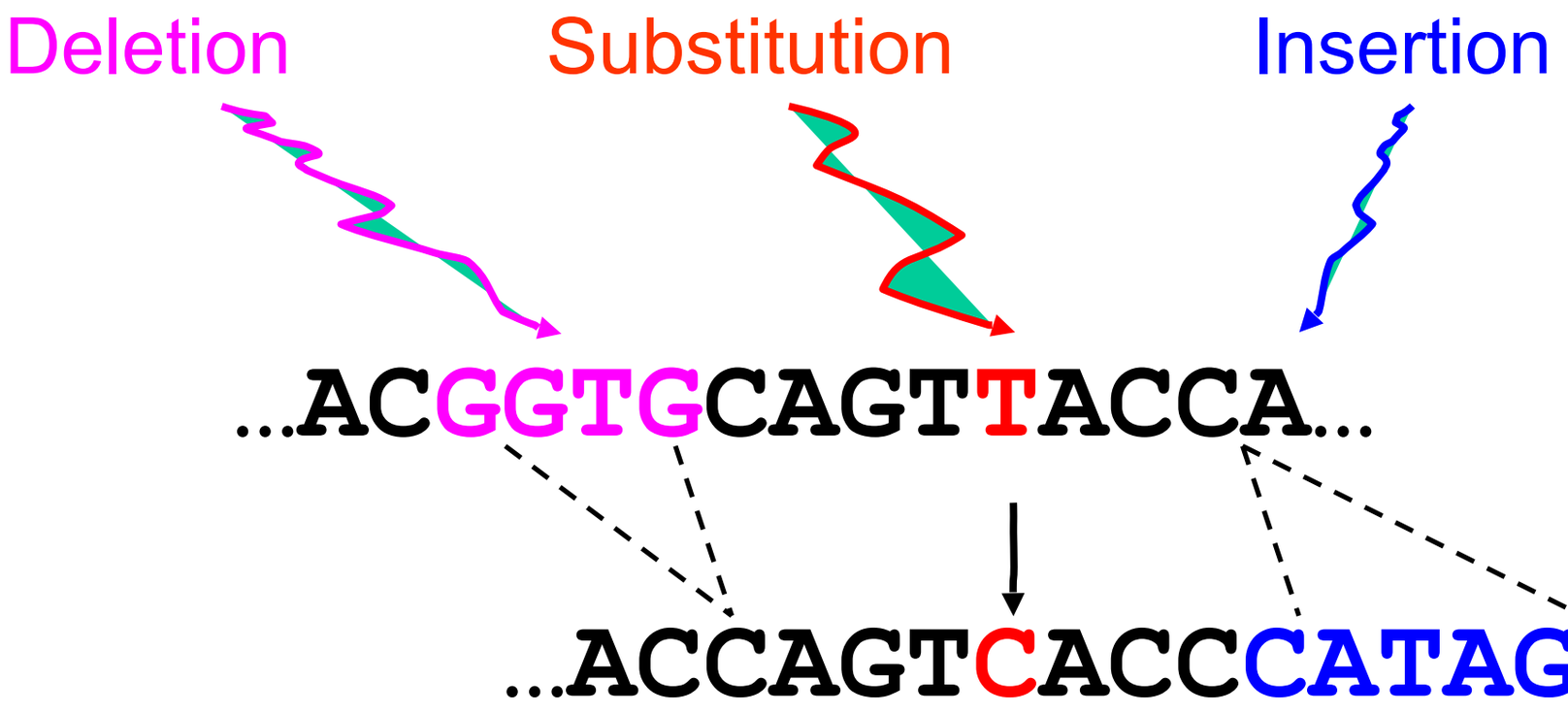
SATé: Simultaneous Alignment and Tree estimation (Liu *et al.* Science 2009)

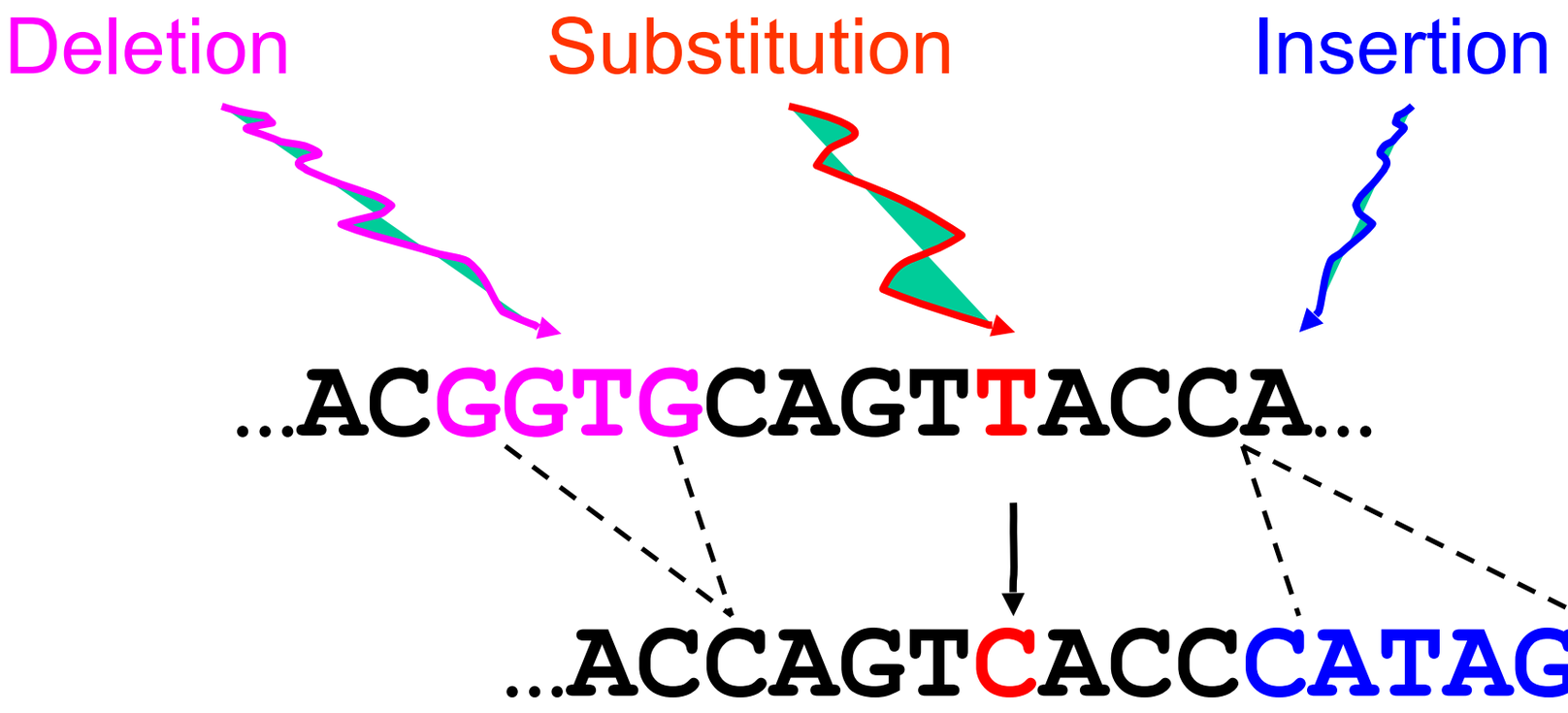
- Standard methods for alignment and tree estimation have unacceptably high error and/or cannot analyze large datasets

- SATé has equal or typically better accuracy than all existing methods on datasets with up to thousands of sequences

- 24 hour analyses using standard desktop computer

- SATé-II (Liu *et al.* Systematic Biology 2012) is more accurate and faster than SATé on datasets with up to tens of thousands of taxa


Deletion
Substitution
Insertion
...ACGGTGCAGTTACCA...
...ACCAGTCACCCATAGA...



The true alignment is:

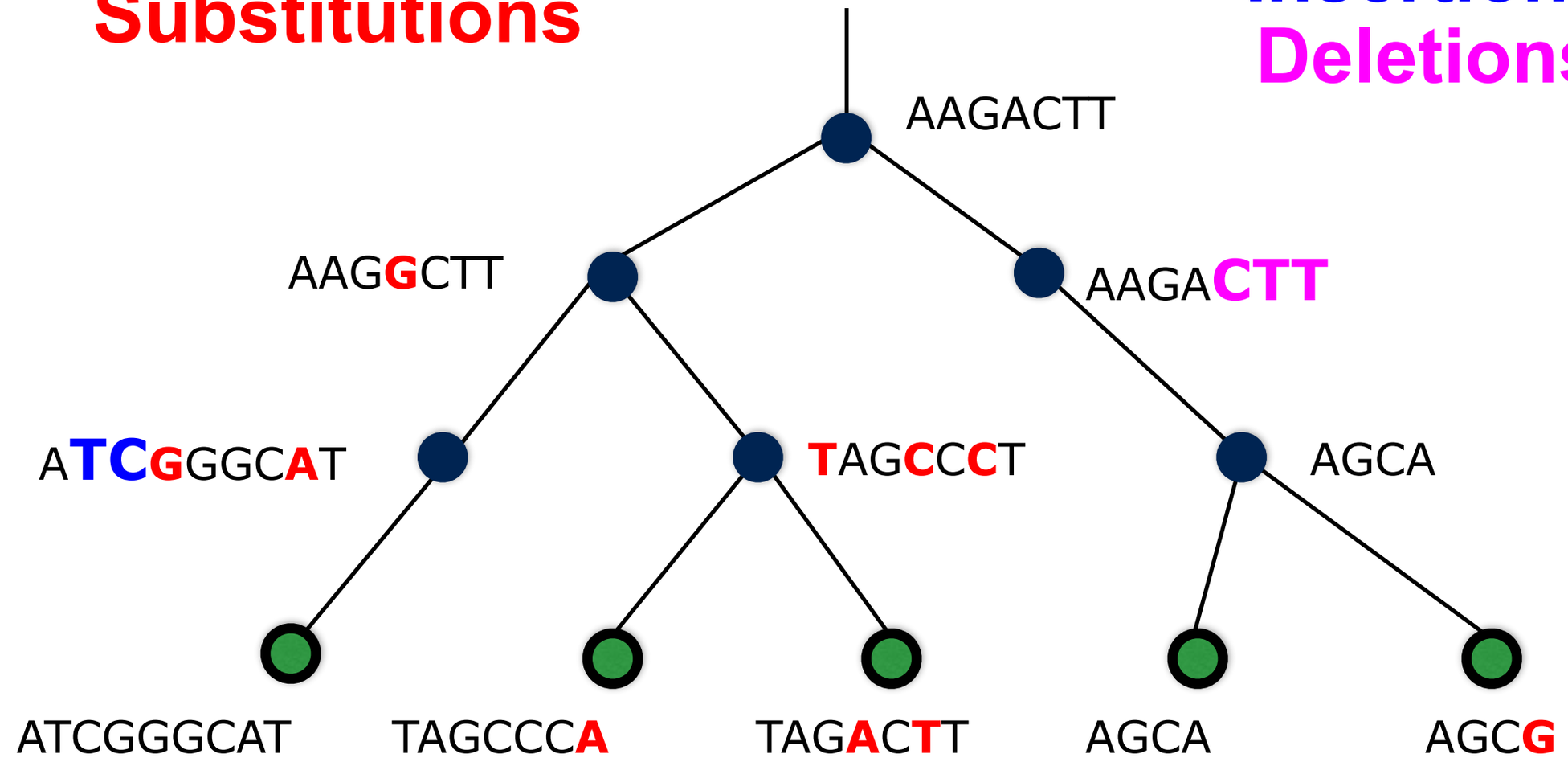
…ACGGTGCAGTTACC-----A…

…AC-----CAGTCACCCATAGA…

DNA Sequence Evolution (Example)

Substitutions

Insertions
Deletions



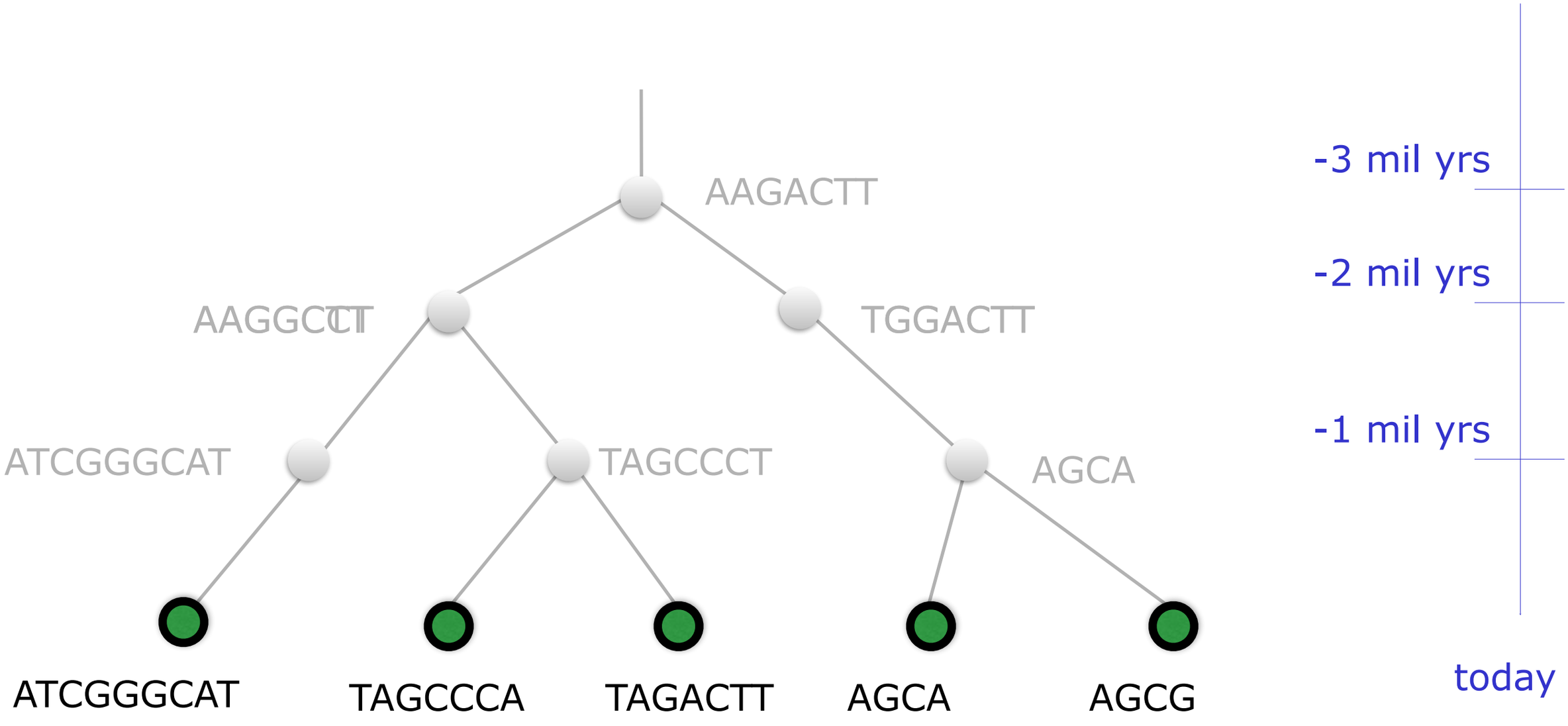
-3 mil yrs

-2 mil yrs

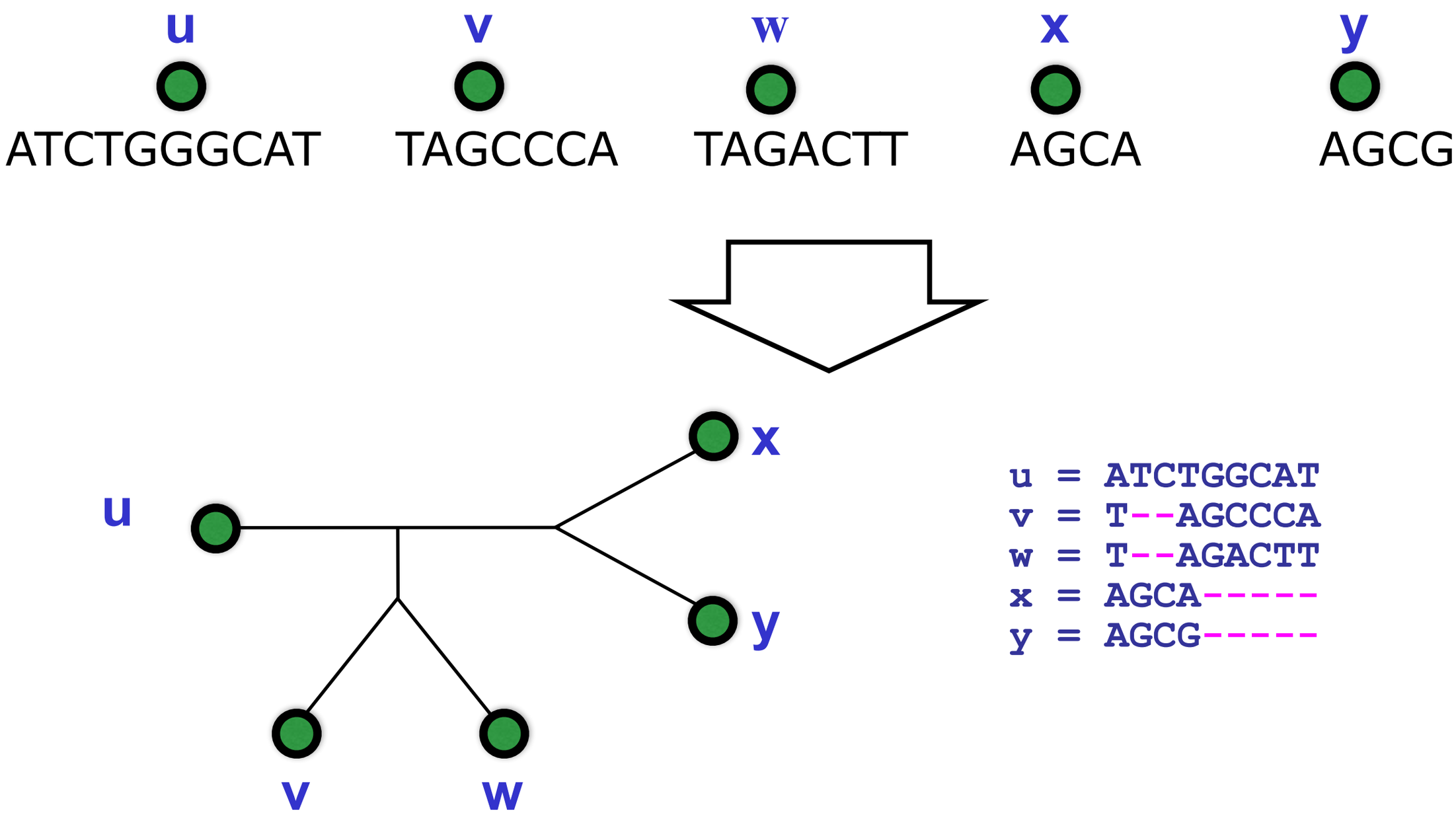
-1 mil yrs

today

DNA Sequence Evolution (Example)



Tree and Alignment Estimation Problem (Example)



Many Trees and Many Alignments

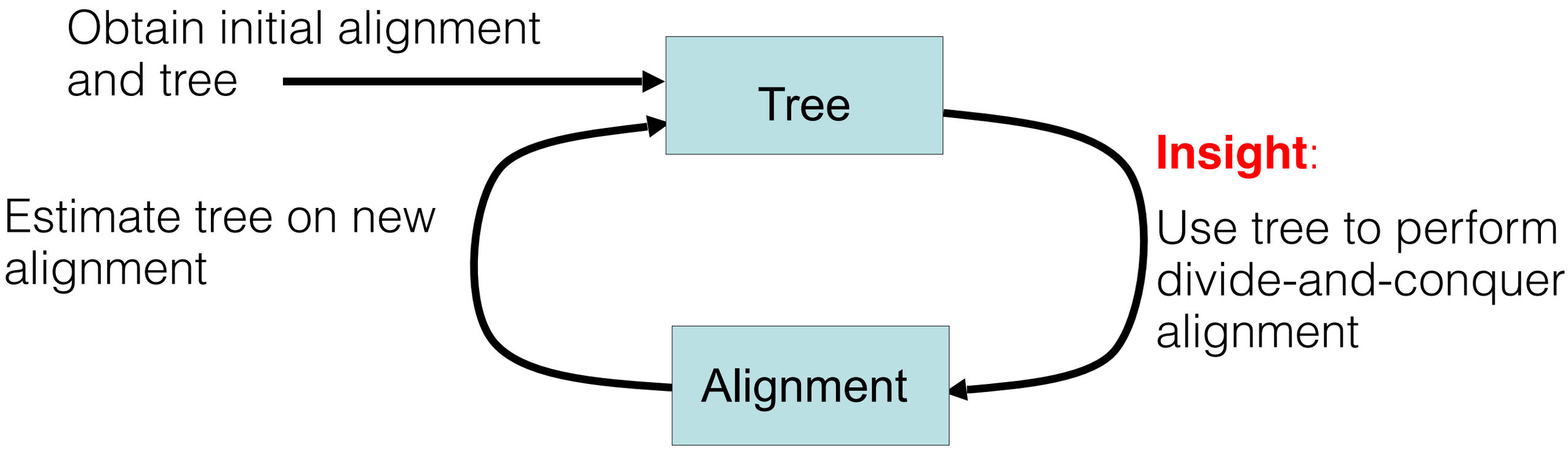
- Number of trees |***T***| grows exponentially in n, the number of leaves:

$$|T| = (2n - 5)!!$$

- The number of alignments |*A*| also grows exponentially in *n* and the length of the longest unaligned sequence.

- All of the common and useful optimization problems are NP-hard.

SATé Algorithm

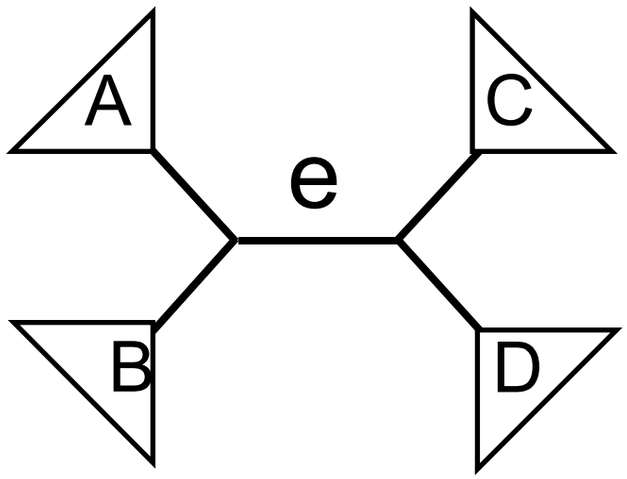


Insight: iterate - use a moderately accurate tree to obtain a more accurate tree

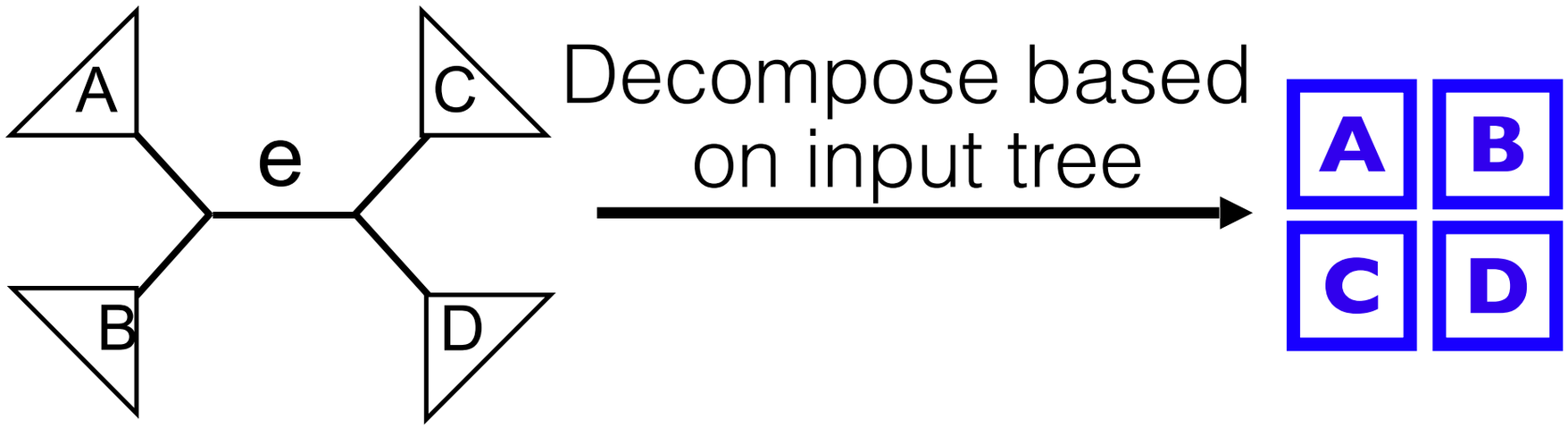
If new alignment/tree pair has worse likelihood, realign using a different decomposition

Repeat until convergence under the maximum likelihood optimization criterion

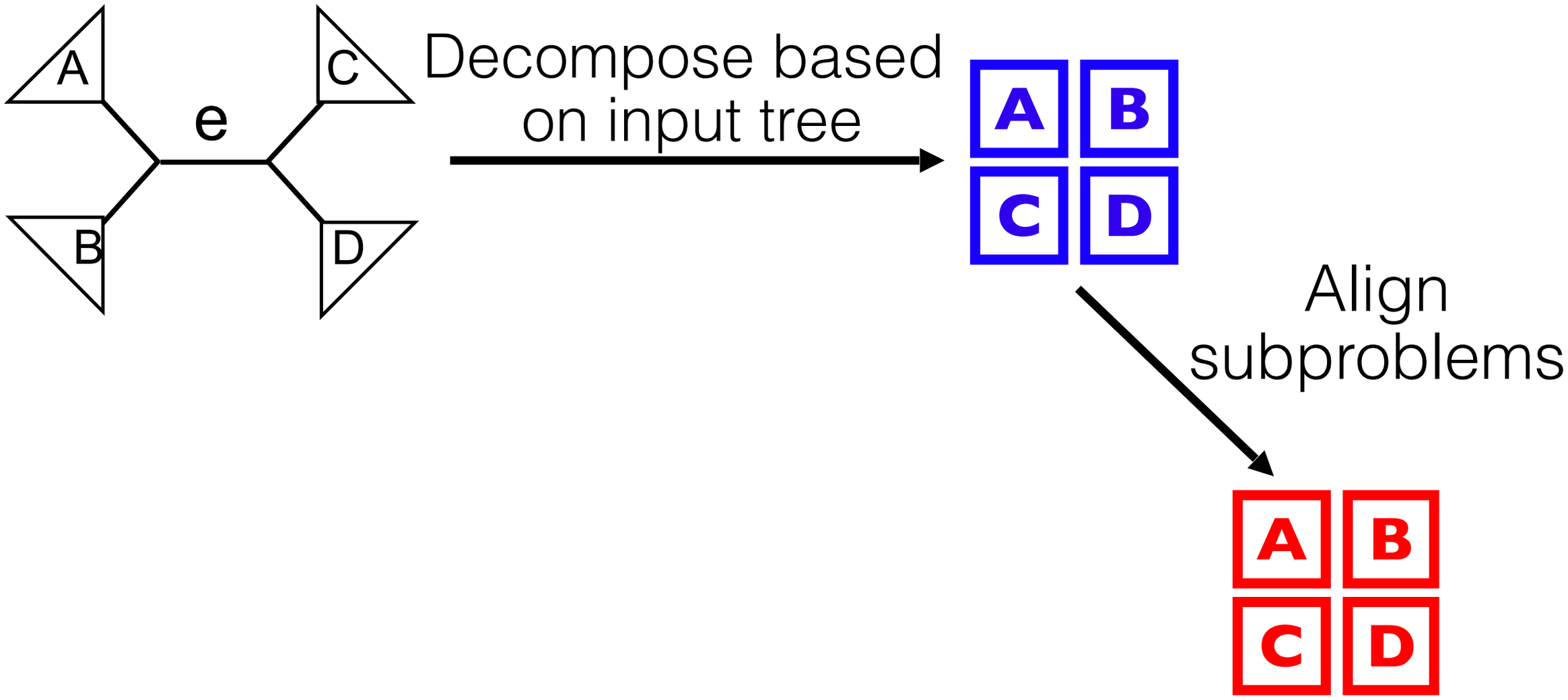
SATé iteration
(Actual decomposition size is configurable)



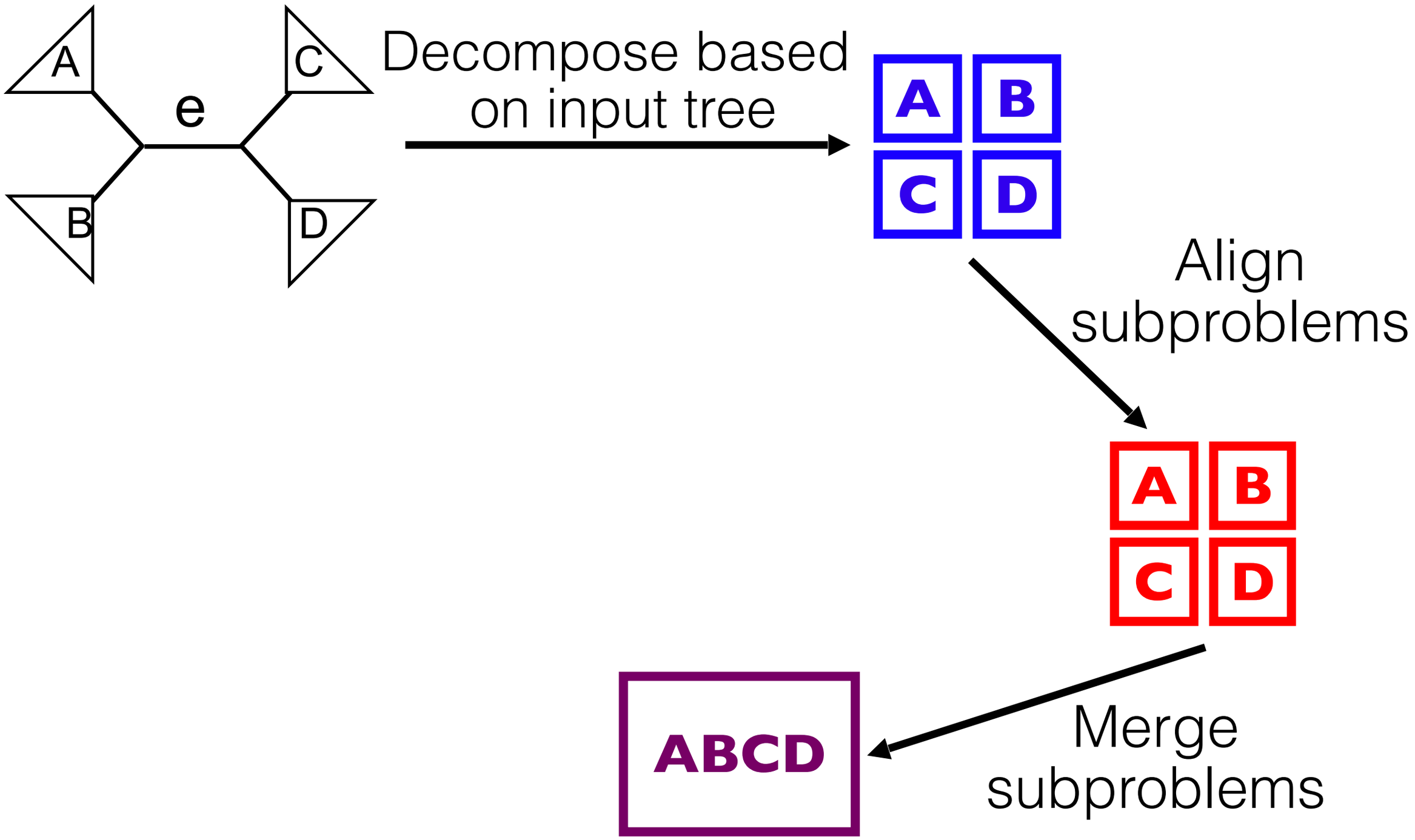
SATé iteration
(Actual decomposition size is configurable)



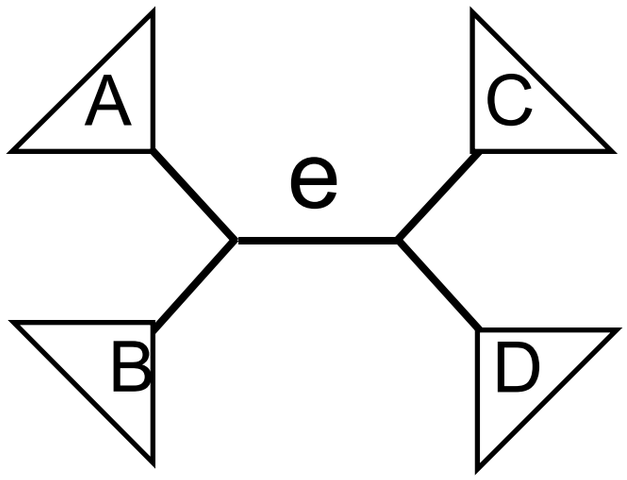
SATé iteration
(Actual decomposition size is configurable)



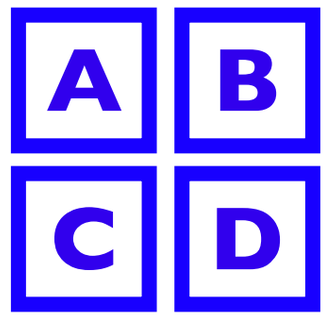
SATé iteration
(Actual decomposition size is configurable)



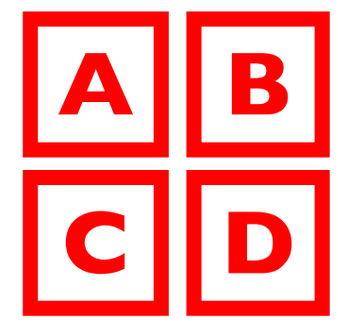
SATé iteration
(Actual decomposition size is configurable)



Decompose based on input tree



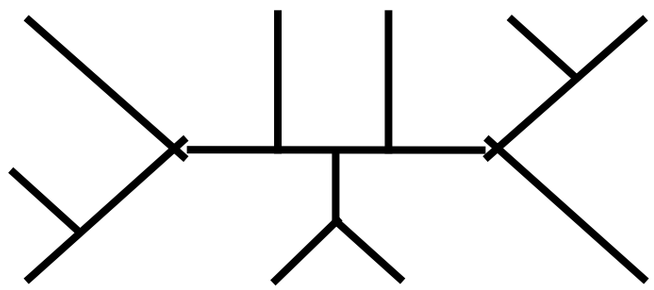
Align subproblems



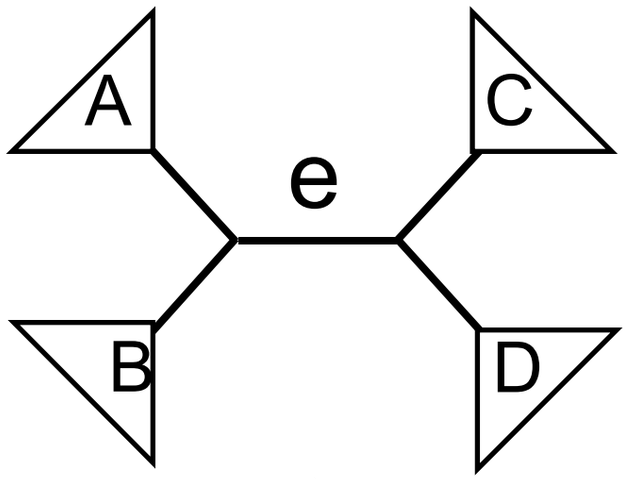
Merge subproblems

ABCD

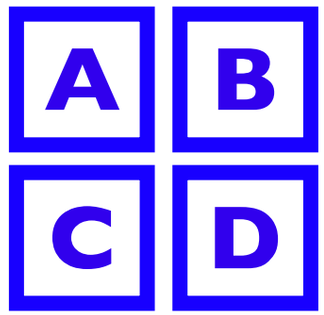
Estimate tree on merged alignment



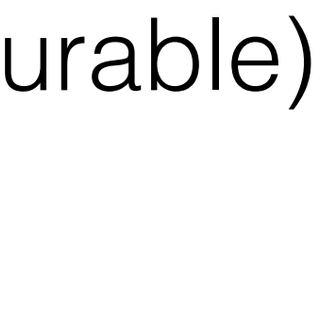
SATé iteration
(Actual decomposition size is configurable)



Decompose based on input tree

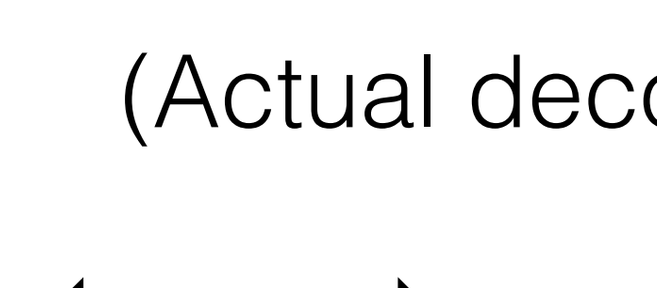


Align subproblems

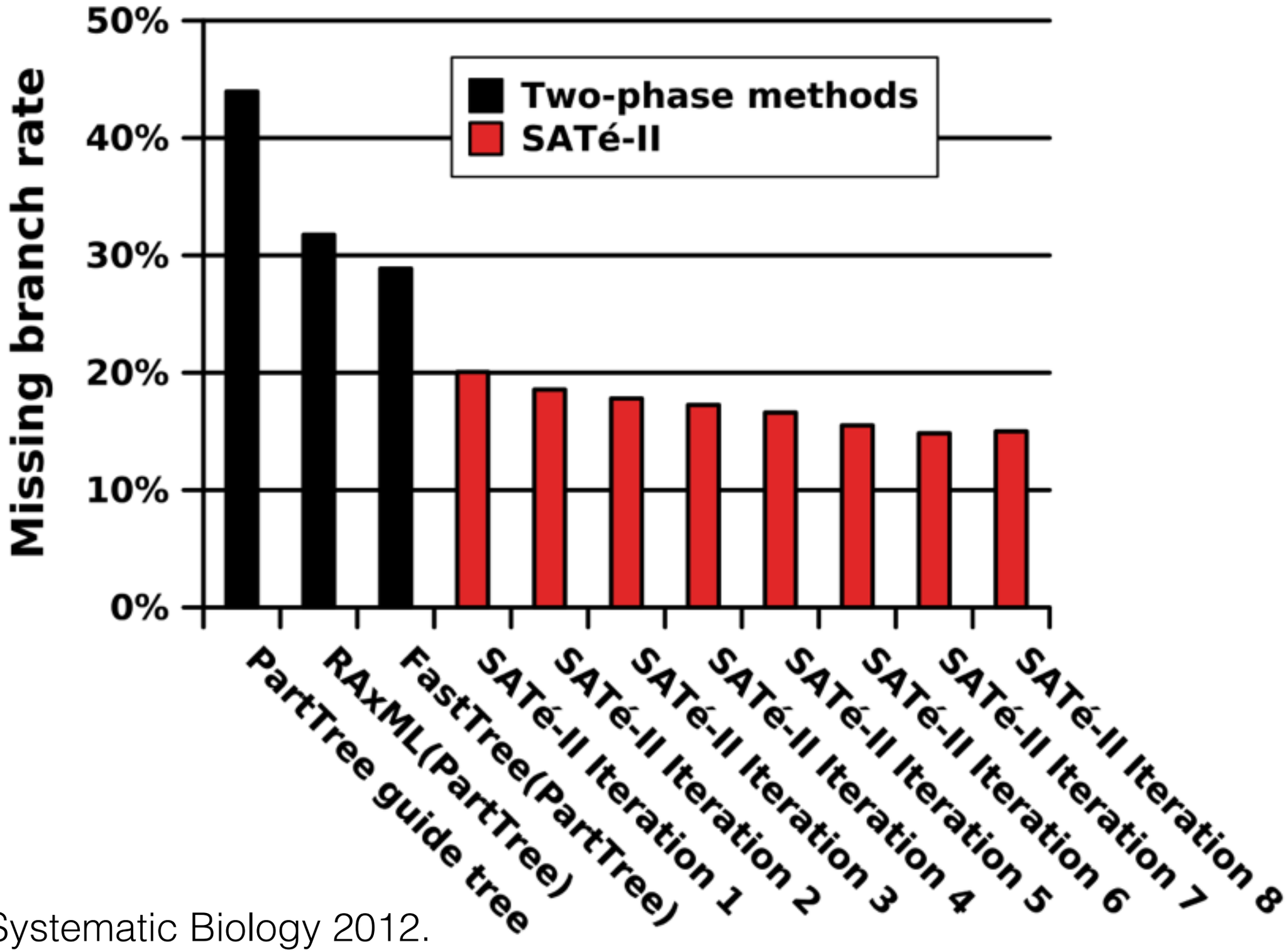


Merge subproblems

Estimate tree on merged alignment



Results on a Dataset with 27,000 Sequences

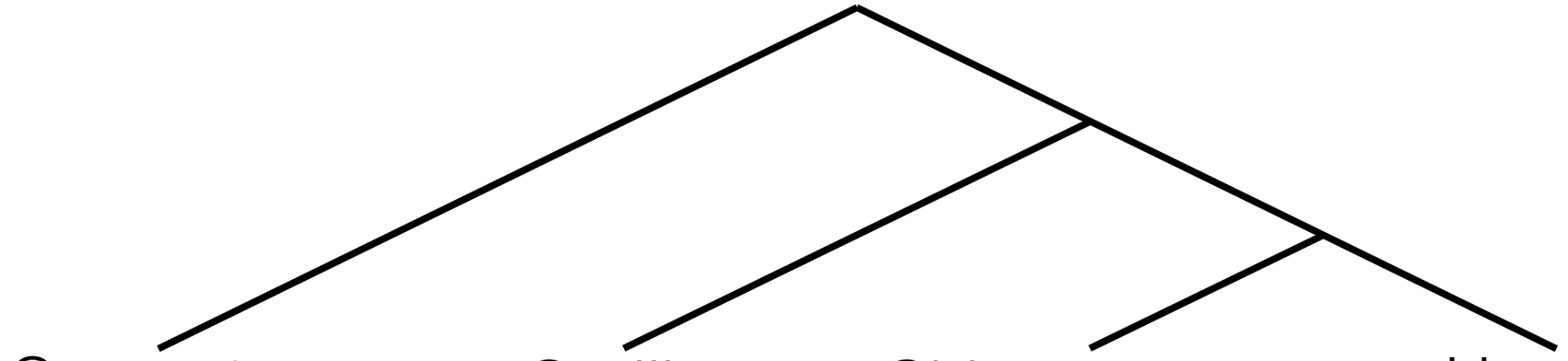


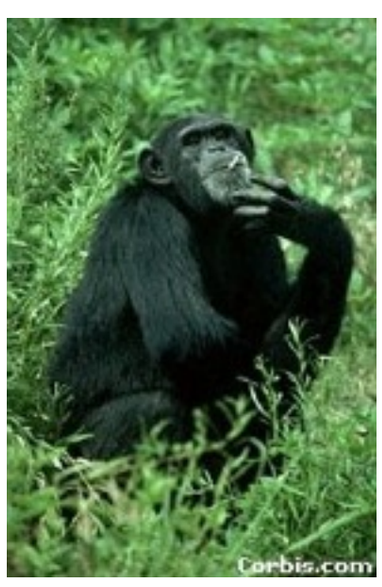
Liu *et al.* Systematic Biology 2012.

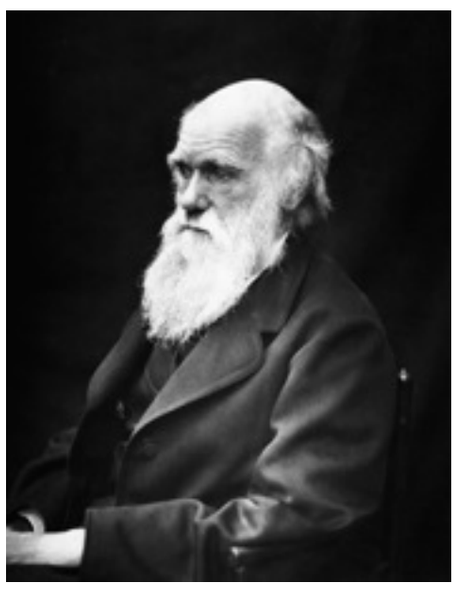
Summary of Part I

- Created novel tree-based divide-and-conquer techniques for simultaneous alignment and tree estimation, enabling:
  - Scalability to thousands of sequences or more
  - High accuracy
- Family of algorithms included:
  - SATé (Liu *et al.* Science 2009)
  - SATé-II (Liu *et al.* Systematic Biology 2012)
  - and others

A Phylogeny, or Evolutionary Tree







Images from the Tree of the Life Website,
University of Arizona, and Wikimedia

Evolutionary History

- Phylogenetics is the study of evolutionary history
- Helps us:
  - Predict gene function
  - Develop drugs and vaccines
  - Understand disease epidemics
  - Study the Tree of Life
  - Etc.

Source: www.tolweb.org

This Talk

- **SATé (Simultaneous Alignment and Tree estimation), Liu et al. Science 2009**
  - Standard phylogenetic methods have unacceptably high error and/or cannot analyze large datasets
  - SATé is more accurate than all existing methods on datasets with up to thousands of taxa
  - 24 hour analyses using standard desktop computer
- **SATé-II, Liu et al. Systematic Biology, in press, 2011**
  - More accurate and faster than SATé on datasets with up to tens of thousands of taxa using a standard desktop computer

Many Trees and Many Alignments

- Number of trees |***T***| grows exponentially in n, the number of leaves:

$$|T| = (2n-5)!!$$

- The number of alignments |*A*| also grows exponentially in *n* and $\max_j k_j$, where k_j is the sequence length of the *j*th sequence (Slowinski MPE 1998):

$$|A| = \sum_{N=\max k_j}^{\sum_j k_j} \sum_{i=0}^{N} (-1)^i \binom{N}{i} \prod_{j=1}^{n} \binom{N-i}{N-k_j-i}$$

- NP-hard optimization problems

Counting Alignments

$$f(k_1, k_2) = f(k_1 - 1, k_2) + f(k_1 - 1, k_2 - 1) + f(k_1, k_2 - 1)$$

$$f(1, 1) = f(1, 0) = 0$$

$$f(k_1, k_2) = \sum_{i=0}^{k_1} \binom{k_1}{i} \binom{k_2 + i}{k_1}$$

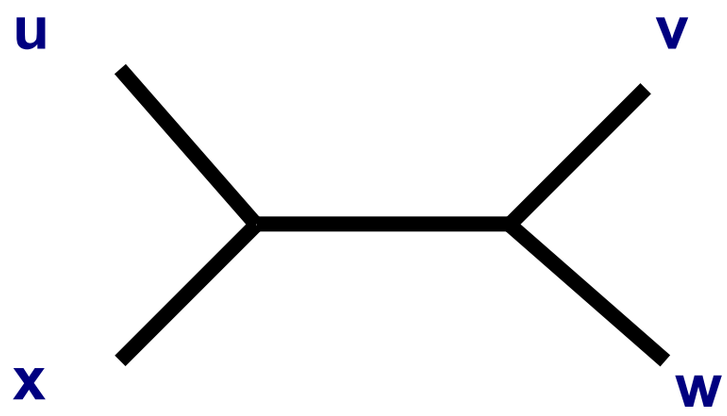
Two-phase Methods

```
u = AGGCTATCACCTGACCTCCA
v = TAGCTATCACGACCGC
w = TAGCTGACCGC
x = TCACGACCGACA
```

Phase 1: Align

```
u = -AGGCTATCACCTGACCTCCA
v = TAG-CTATCAC--GACCGC--
w = TAG-CT-------GACCGC--
x = -------TCAC--GACCGACA
```

Phase 2: Estimate Tree



Many Methods

Alignment method

- **ClustalW**
- **MAFFT**
- **Muscle**
- **Prank**
- Opal
- Probcons (and Probtree)
- Di-align
- T-Coffee
- Etc.

Many Methods

Alignment method

- **ClustalW**
- **MAFFT**
- **Muscle**
- **Prank**
- Opal
- Probcons (and Probtree)
- Di-align
- T-Coffee
- Etc.

Phylogeny method

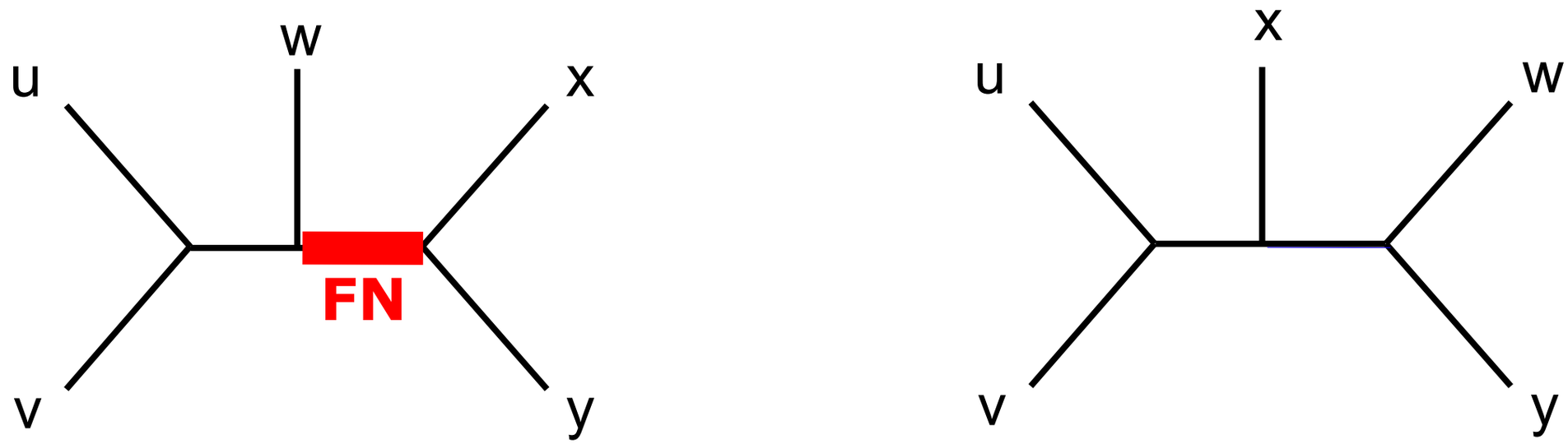
- **Maximum likelihood (ML)**
  - **RAxML**
- Bayesian MCMC
- Maximum parsimony
- Neighbor joining
- UPGMA
- Quartet puzzling
- Etc.

Simulation Study (Liu et al. Science 2009)

Simulation using ROSE

- Model trees with 1000 taxa
- Biologically realistic model with:
  - Varied rates of substitutions
  - Varied rates of insertions and deletions
  - Varied gap length distribution
    - Long
    - Medium
    - Short

Tree Error



True Tree Estimated Tree

- **False Negative (FN)**: an edge in the true tree that is missing from the estimated tree
- **Missing branch rate**: the percentage of edges present in the true tree but missing from the estimated tree

Alignment Error

FN

```
AACAT
A-CCG
```

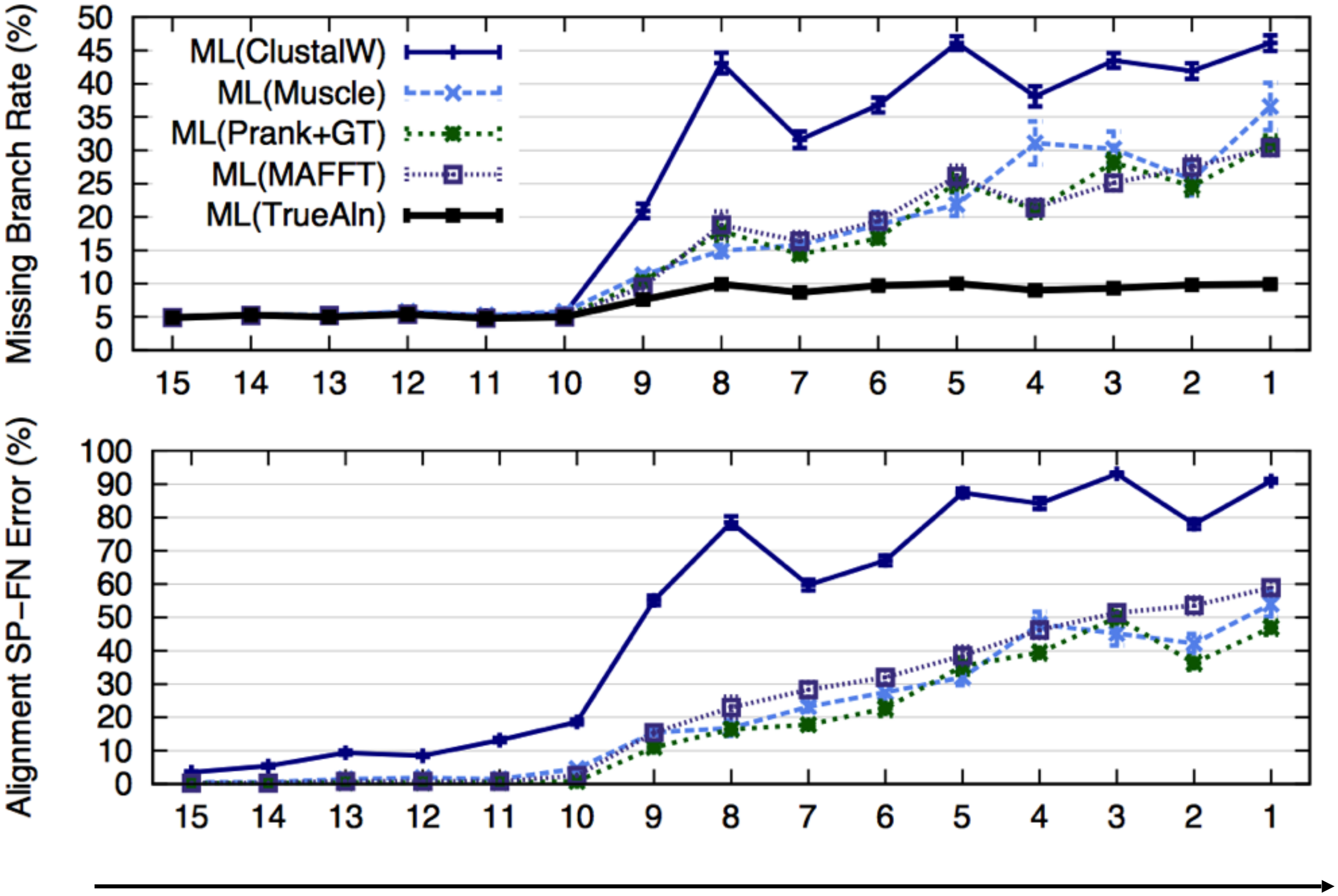
True Alignment

```
AACAT-
A-CC-G
```

Estimated Alignment

- **False Negative (FN)**: pair of nucleotides present in true alignment but missing from estimated alignment
- **Alignment SP-FN error**: percentage of paired nucleotides present in true alignment but missing from estimated alignment

Results

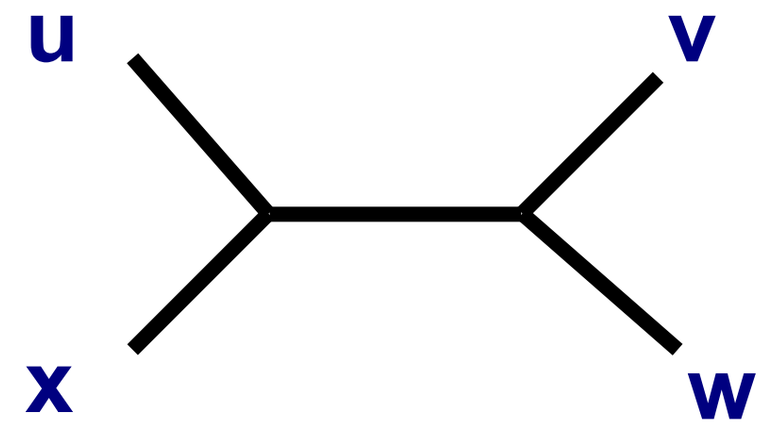

Missing Branch Rate (%)
50
45
40
35
30
25
20
15
10
5
0
ML(ClustalW)
ML(Muscle)
ML(Prank+GT)
ML(MAFFT)
ML(TrueAln)
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
Alignment SP−FN Error (%)
100
90
80
70
60
50
40
30
20
10
0
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
1000 taxon models ranked by difficulty

Problem with Two-phase Approach

- **Problem**: two-phase methods fail to return reasonable alignments and accurate trees on large and divergent datasets
  - manual alignment
  - unreliable alignments excluded from phylogenetic analysis

Simultaneous Estimation of a Tree and Alignment

```
u = CTATCACCTGACCTCCA
v = CTATCACGACCGC
w = CTGACCGC
x = CGACCGACA
```



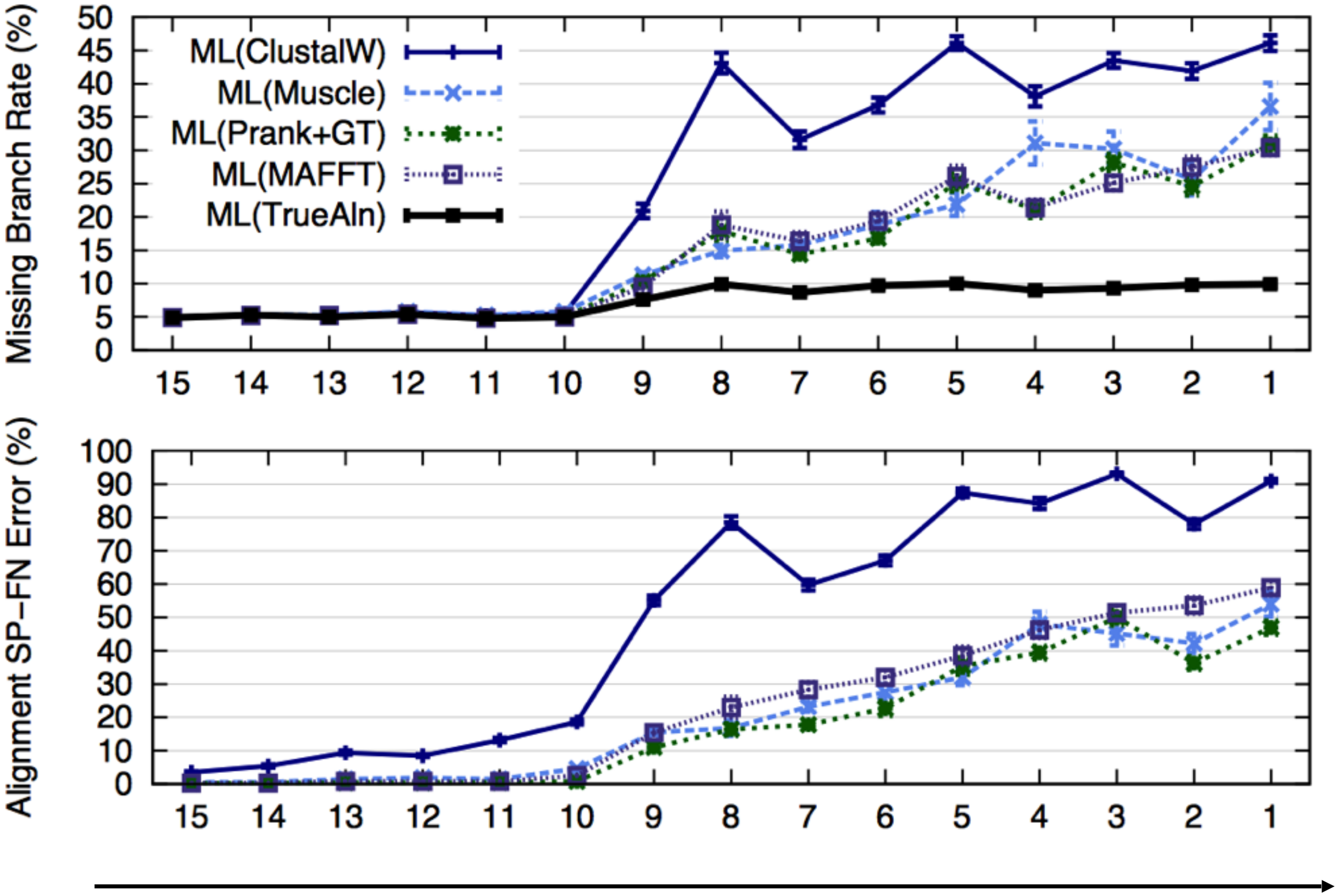
and

```
u = CTATCACCTGACCTCCA
v = CTATCAC--GACCGC--
w = CT-------GACCGC--
x = ---TCAC--GACCGACA
```

Existing Methods for Alignment and Tree Inference

- Two-phase methods
  - Infer an alignment, then use the alignment to infer a tree
  - Inaccurate on data sets with thousands of sequences
- Methods based on statistical models
  - Limited to datasets with a few hundred taxa
  - Unknown accuracy on larger datasets
- Parsimony-based methods
  - Slower than two-phase methods
  - No more accurate than two-phase methods

Results


Missing Branch Rate (%)
50
45
40
35
30
25
20
15
10
5
0
ML(ClustalW)
ML(Muscle)
ML(Prank+GT)
ML(MAFFT)
ML(TrueAln)
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
Alignment SP−FN Error (%)
100
90
80
70
60
50
40
30
20
10
0
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
1000 taxon models ranked by difficulty

Problem with Two-phase Approach

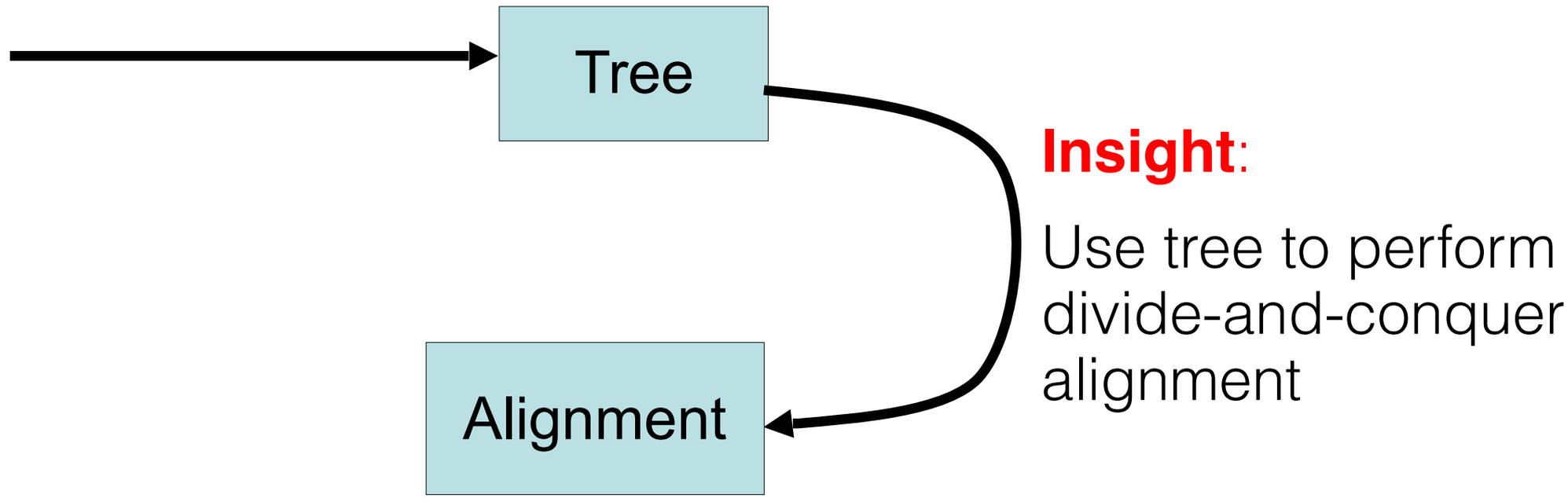
- **Problem**: two-phase methods fail to return reasonable alignments and accurate trees on large and divergent datasets
- **Insight**: divide-and-conquer to constrain dataset divergence and size

SATé Algorithm

Obtain initial alignment and
estimated ML tree

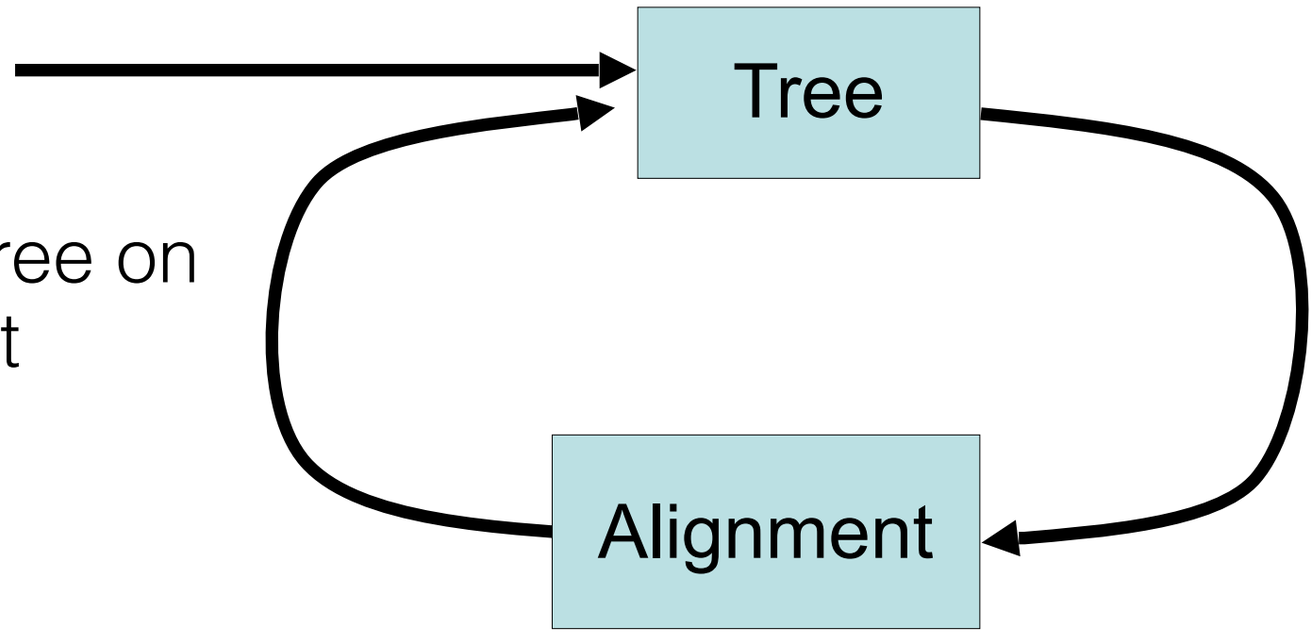
SATé Algorithm

Obtain initial alignment and estimated ML tree



SATé Algorithm

Obtain initial alignment and estimated ML tree

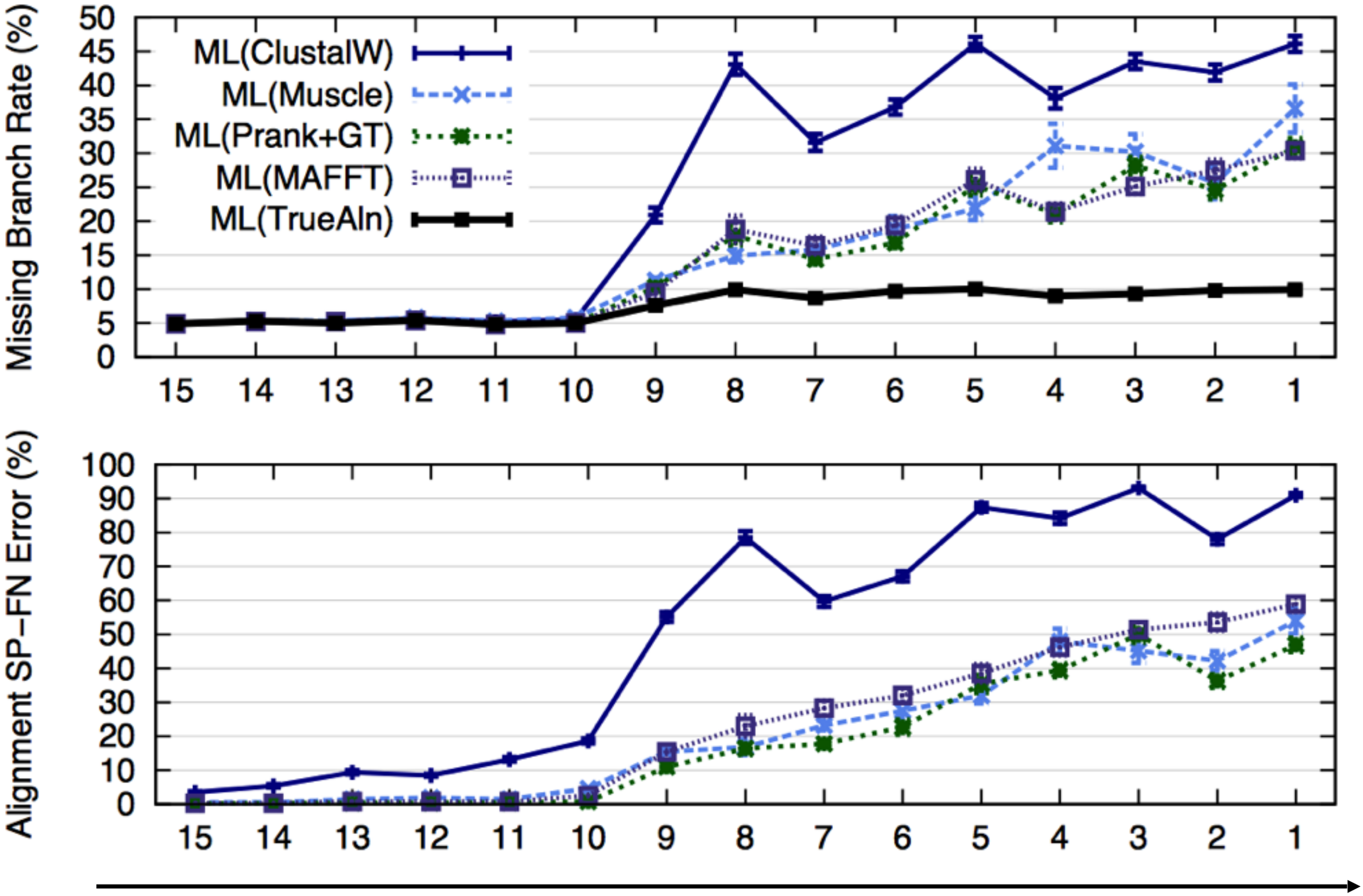


Estimate ML tree on new alignment

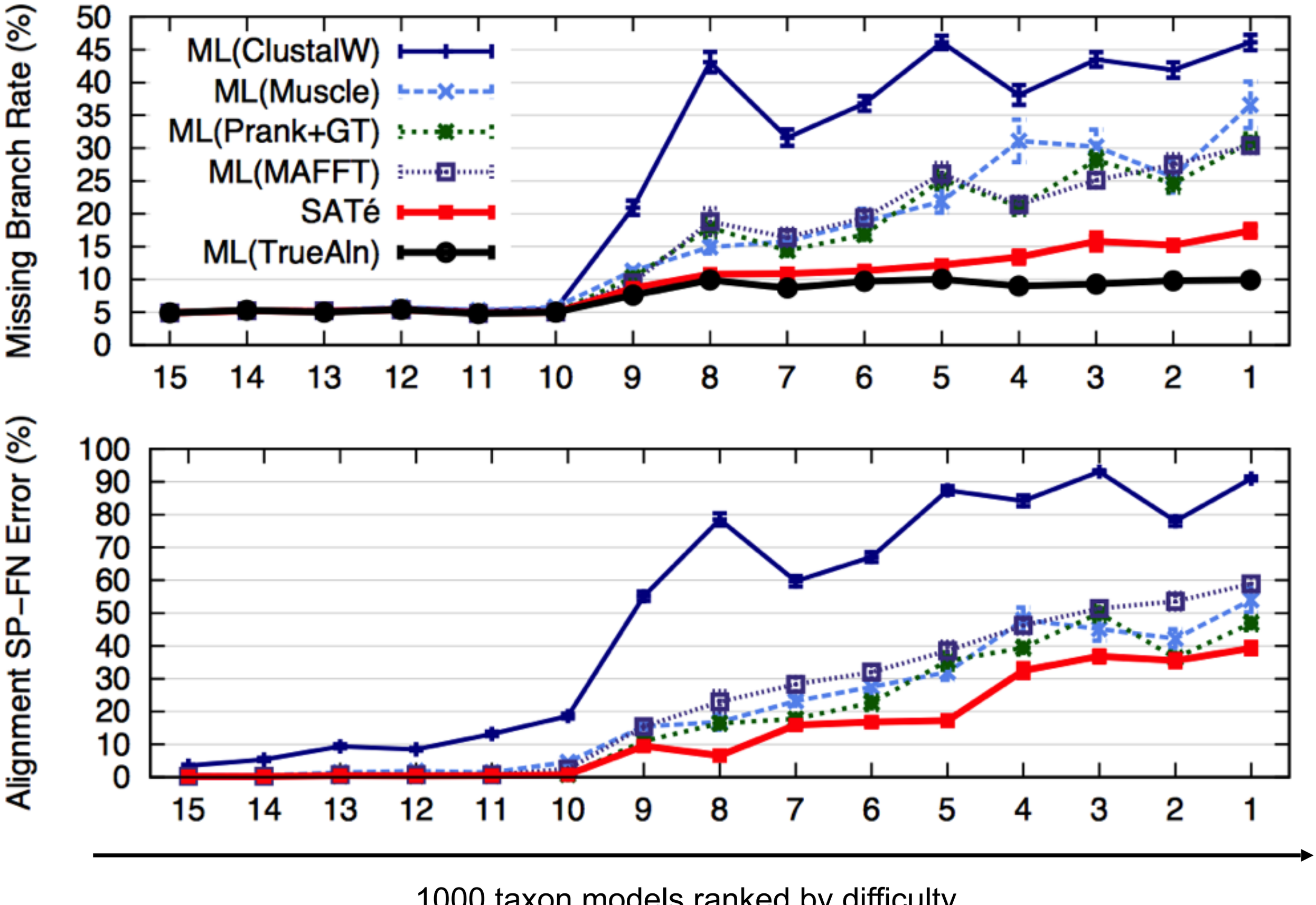
Insight:

Use tree to perform divide-and-conquer alignment

Results


Missing Branch Rate (%)
50
45
40
35
30
25
20
15
10
5
0
ML(ClustalW)
ML(Muscle)
ML(Prank+GT)
ML(MAFFT)
ML(TrueAln)
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
Alignment SP−FN Error (%)
100
90
80
70
60
50
40
30
20
10
0
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
1000 taxon models ranked by difficulty

Results


Missing Branch Rate (%)
50
45
40
35
30
25
20
15
10
5
0
ML(ClustalW)
ML(Muscle)
ML(Prank+GT)
ML(MAFFT)
SATé
ML(TrueAln)
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
Alignment SP−FN Error (%)
100
90
80
70
60
50
40
30
20
10
0
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
1000 taxon models ranked by difficulty

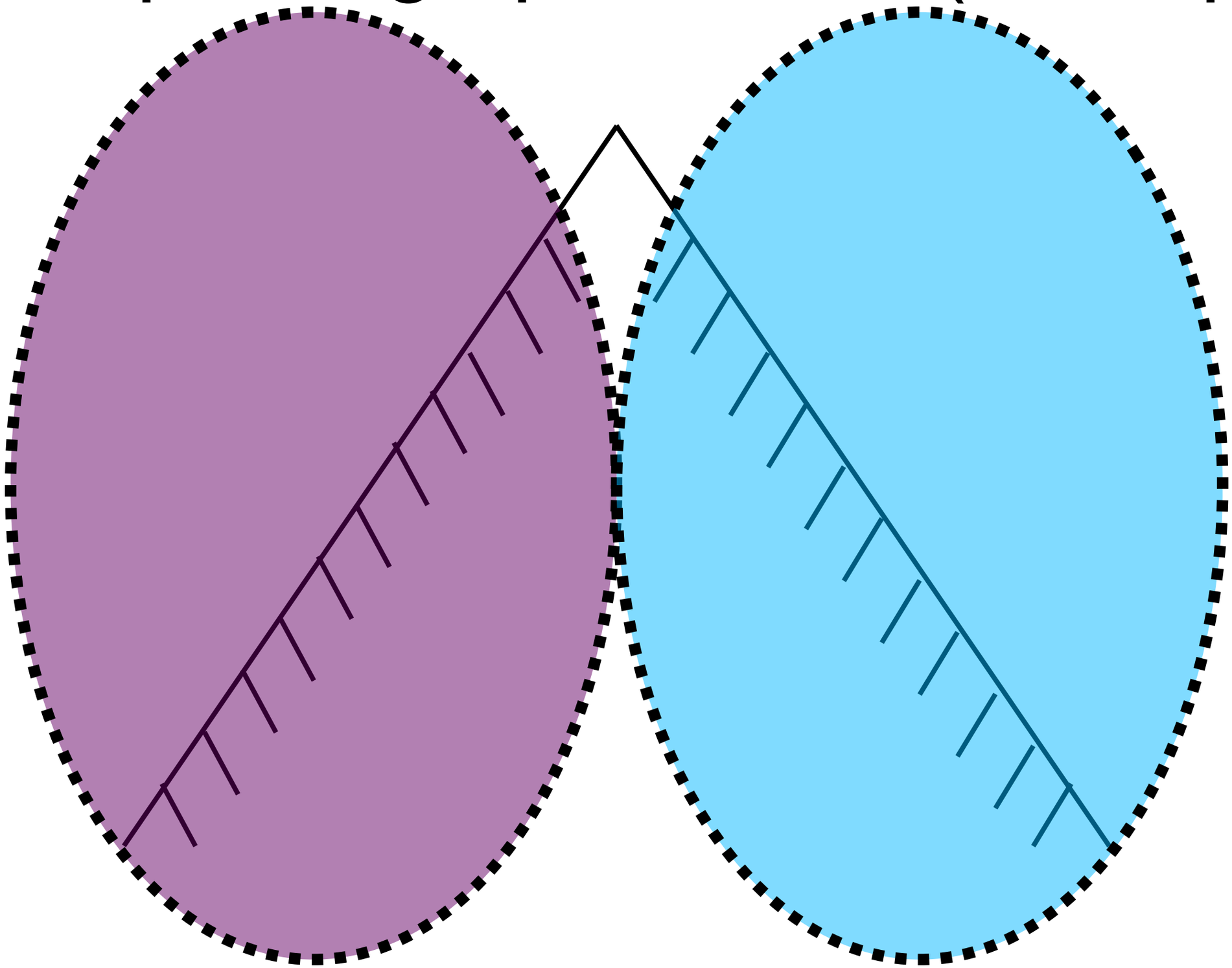
Improving Upon SATé

- **Problem**: sometimes, subproblems have too many taxa or too divergent

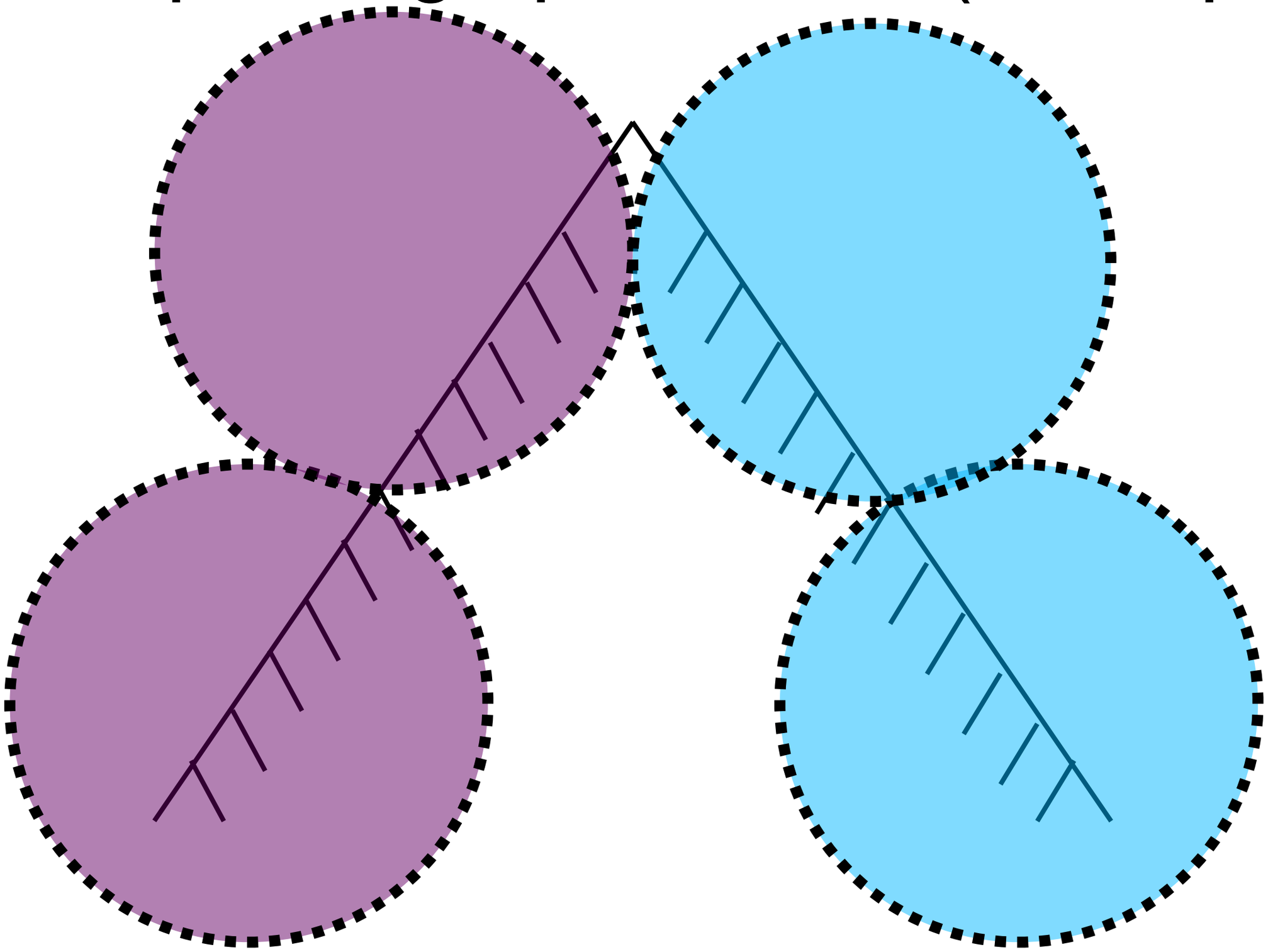
Improving Upon SATé

- **Problem**: sometimes, subproblems have too many taxa or too divergent
- **Insight:** recurse

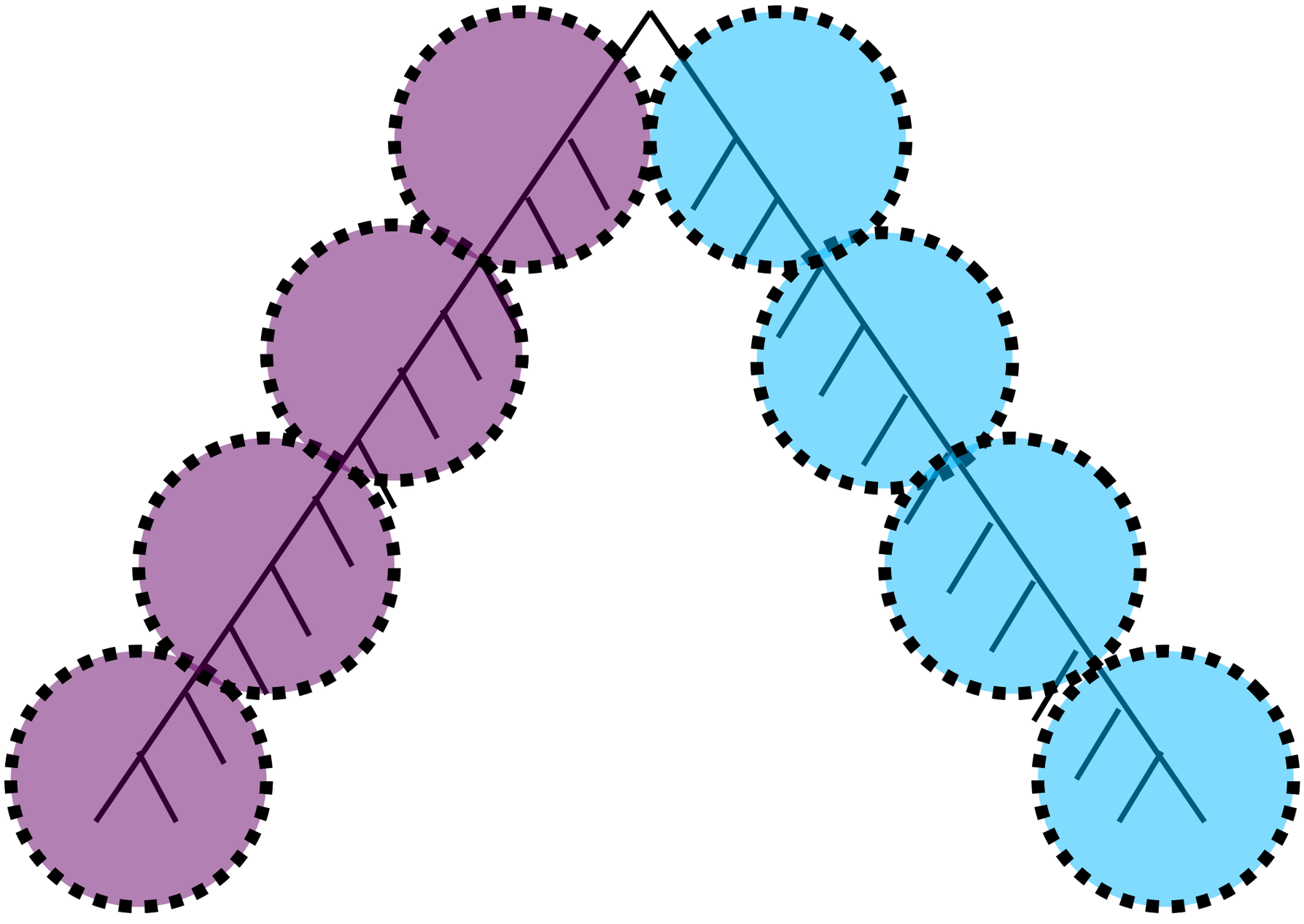
Improving upon SATé (Example)



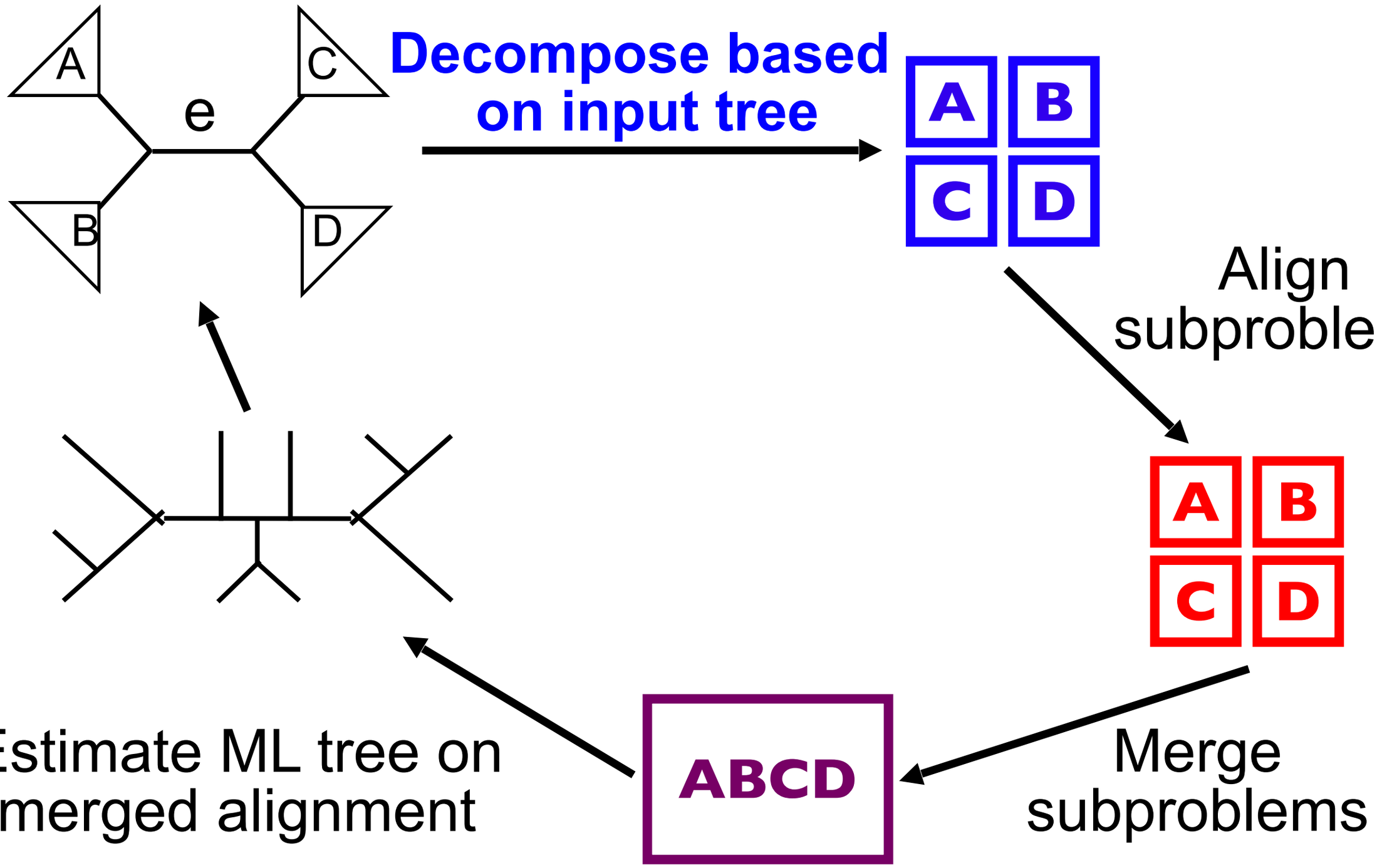
Improving upon SATé (Example)



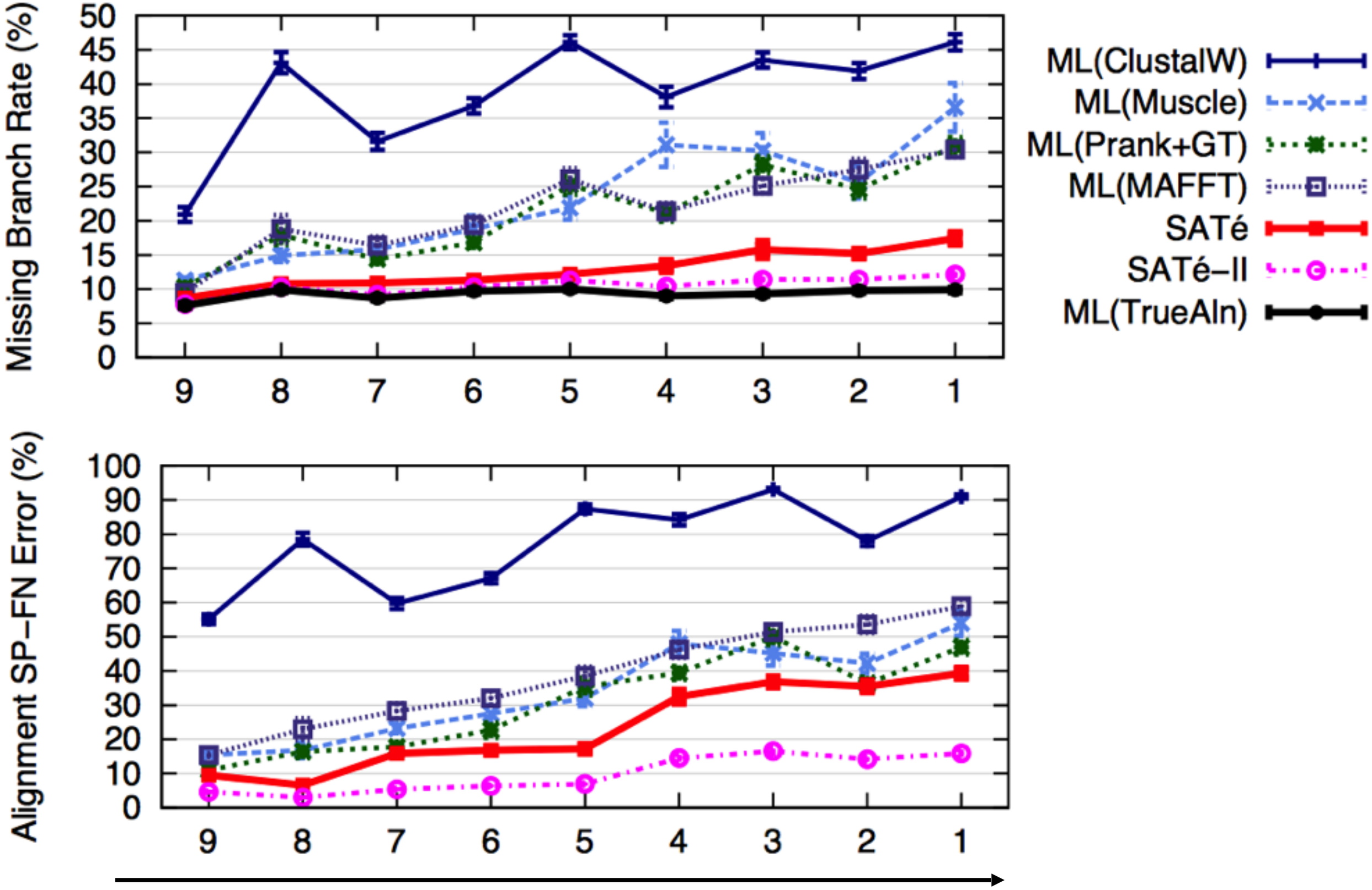
Improving upon SATé (Example)



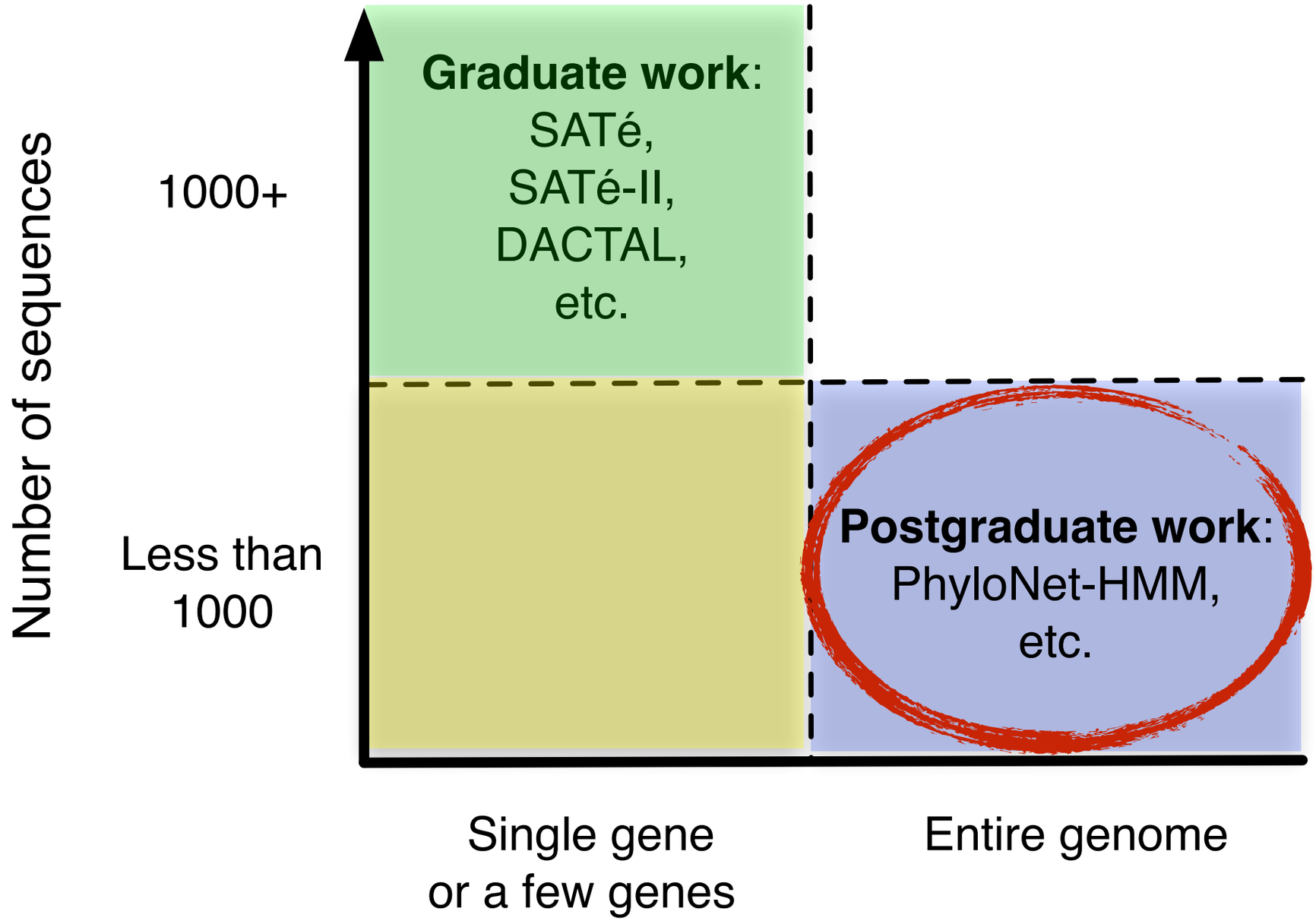
Insight: recurse



Results


Missing Branch Rate (%)
50
45
40
35
30
25
20
15
10
5
0
9
8
7
6
5
4
3
2
1
ML(ClustalW)
ML(Muscle)
ML(Prank+GT)
ML(MAFFT)
SATé
SATé–II
ML(TrueAln)
Alignment SP–FN Error (%)
100
90
80
70
60
50
40
30
20
10
0
9
8
7
6
5
4
3
2
1
1000 taxon models ranked by difficulty

Outline for Today's Talk



HPC Challenges

- Email from UTCS IT staff. I had the most computations by several orders of magnitude of any user at UTCS.

- If you let me, I'll come and take over your clusters too.

- In all seriousness, my research has some fantastic low-hanging fruit for HPC contributions, particularly regarding parallel algorithms. <point to HPC researchers in the room>

Hybridization

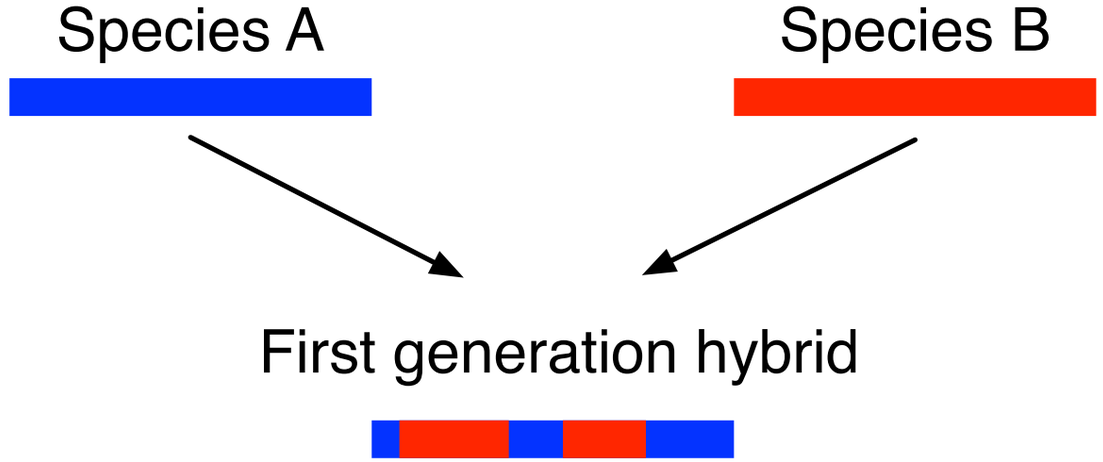
House mouse lacking warfarin resistance gene

Different mouse species carrying warfarin resistance gene

Hybrid mouse carrying warfarin resistance gene

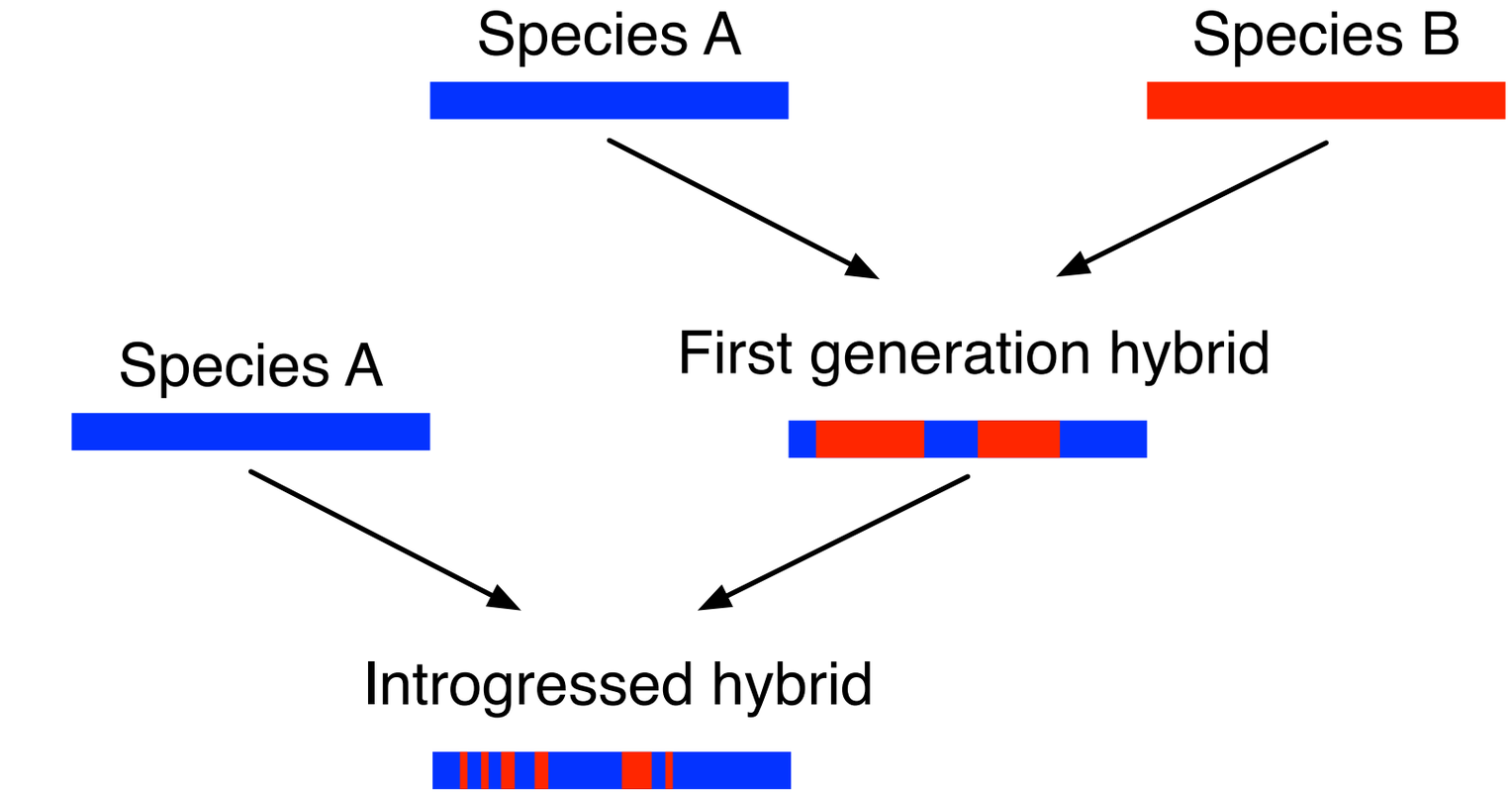
Song *et al.* 2011. Images adapted from Dejager *et al.* 2009 and the Jackson Laboratory.

Introgression



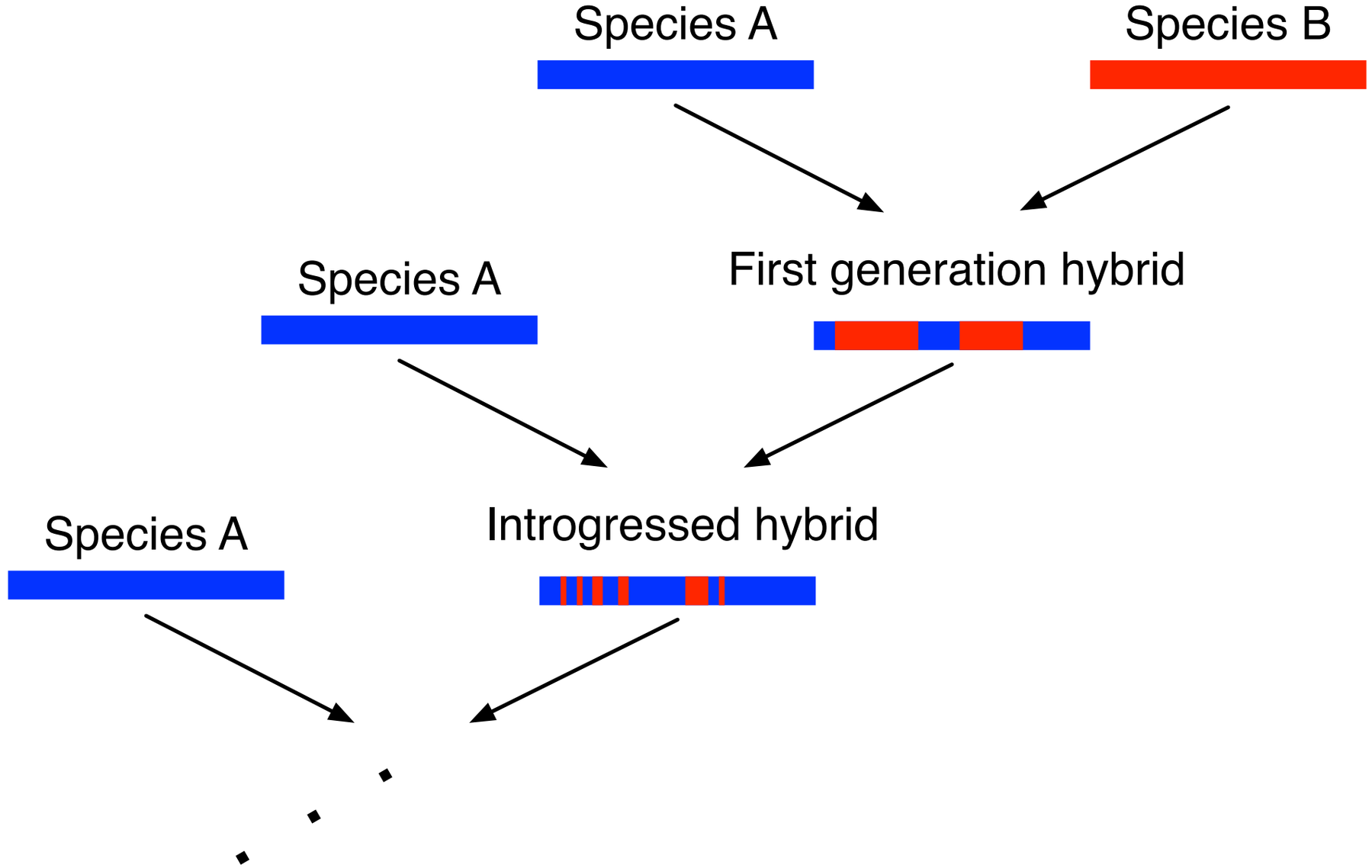
Approximately half of genome
from A and half of genome from B

Introgression



More than half of genome from A
and less than half of genome from B

Introgression

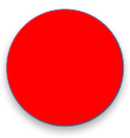


Naïve Sliding Windows

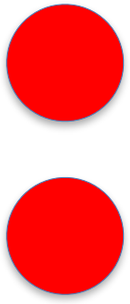
1. Break the genome into segments using a sliding-window (or other approaches)

2. Estimate a local tree in between every pair of breakpoints

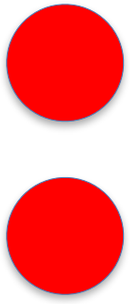
Sliding Windows (Example)



Sliding Windows (Example)

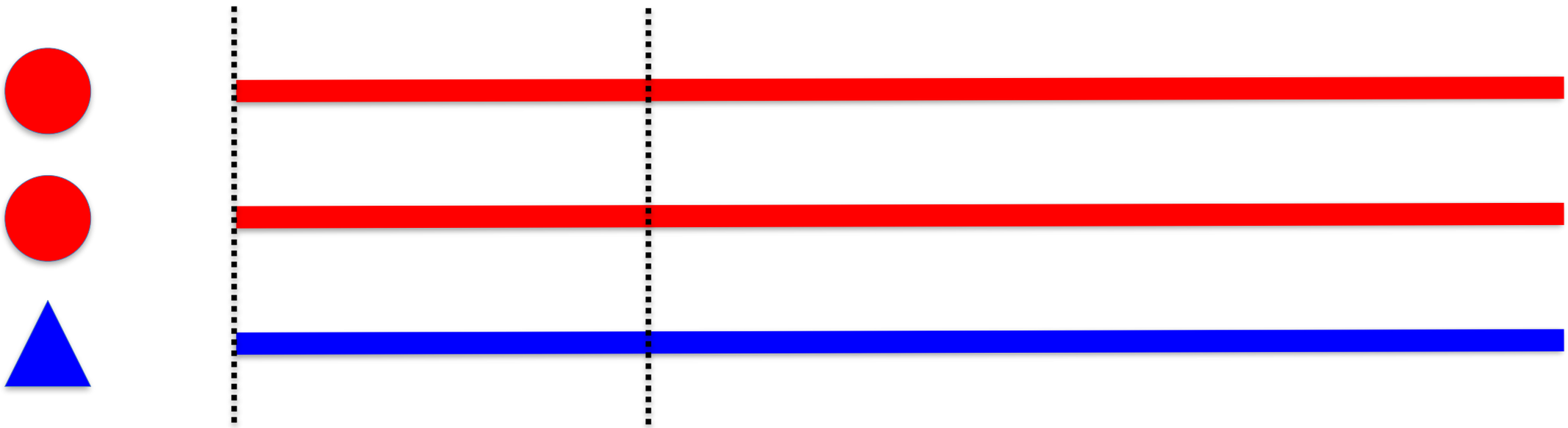


Sliding Windows (Example)

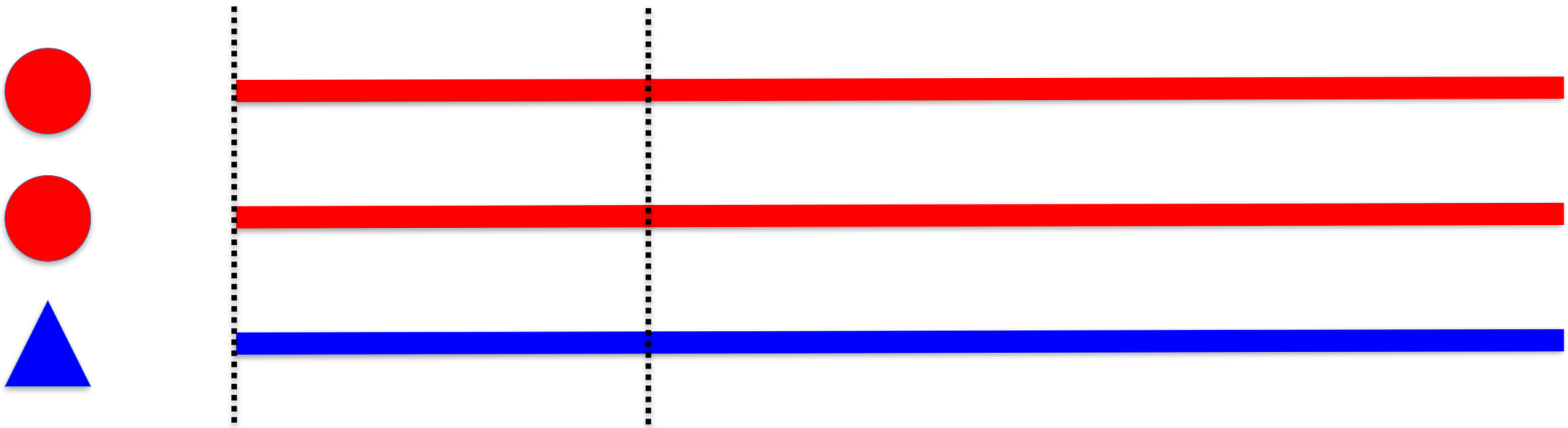


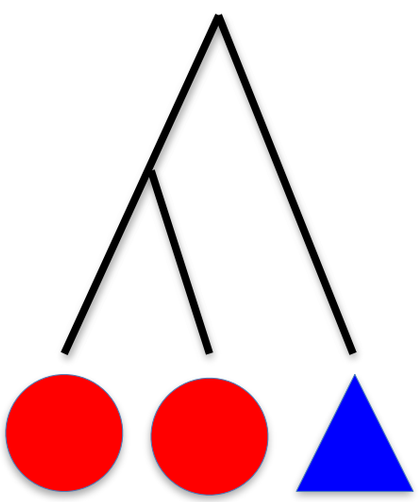
Sliding Windows (Example)

Sliding Windows (Example)

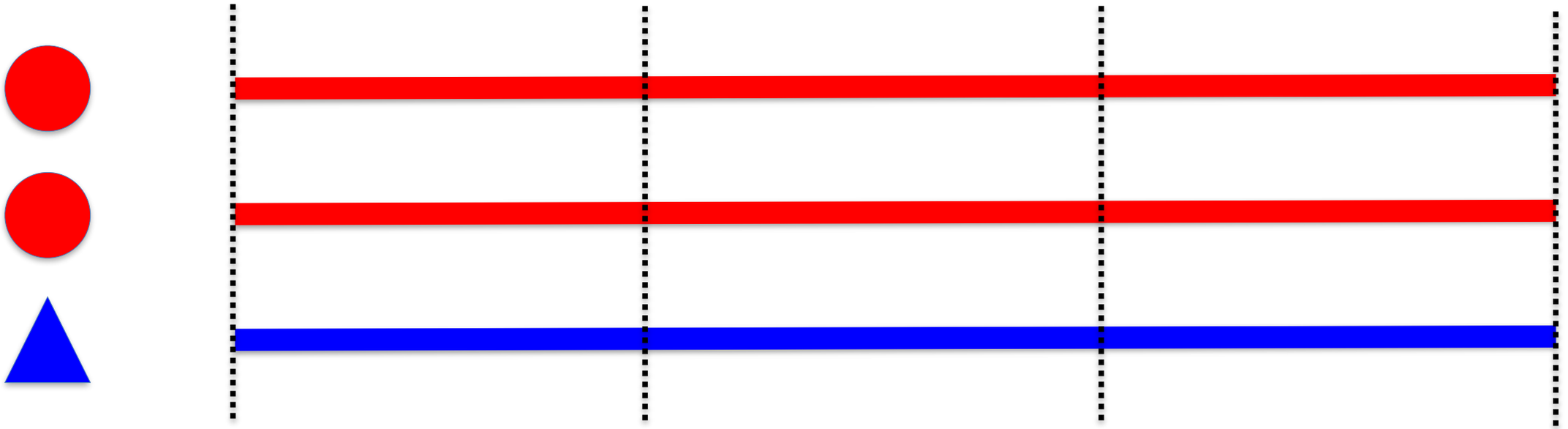


Sliding Windows (Example)

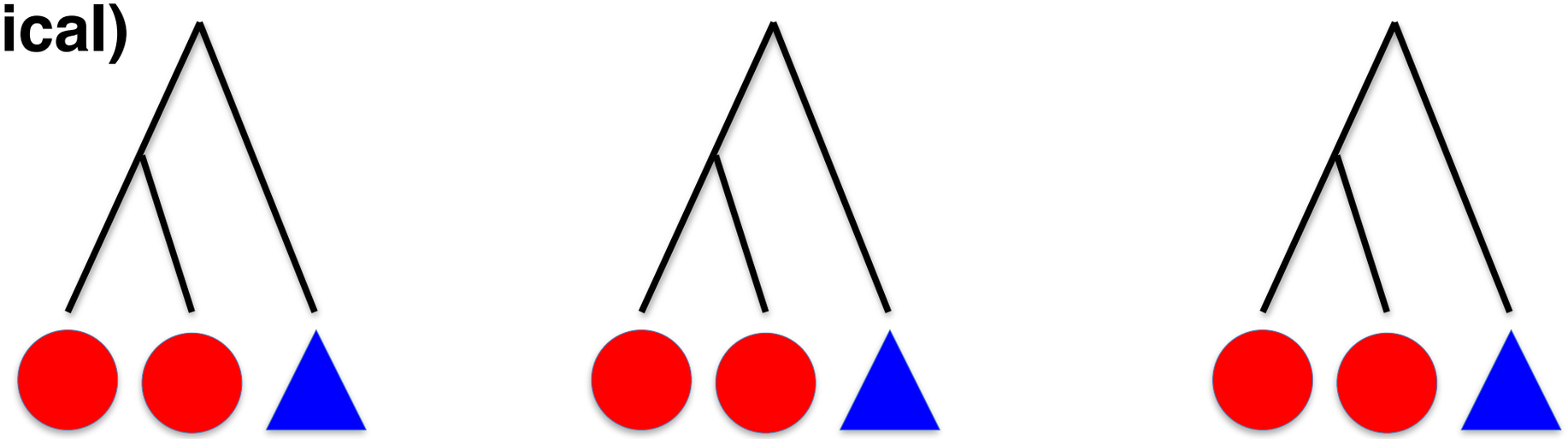




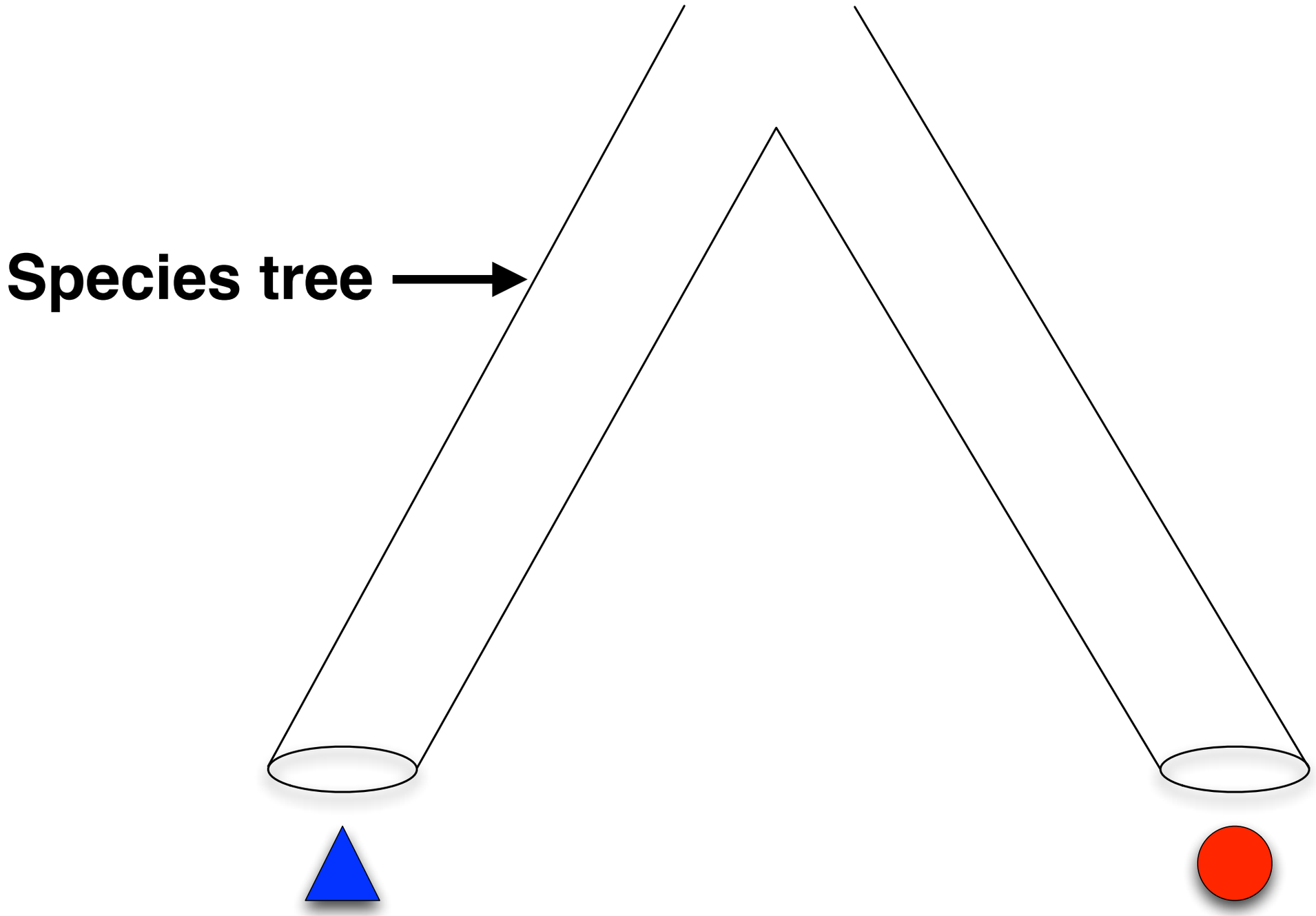
Sliding Windows (Example)



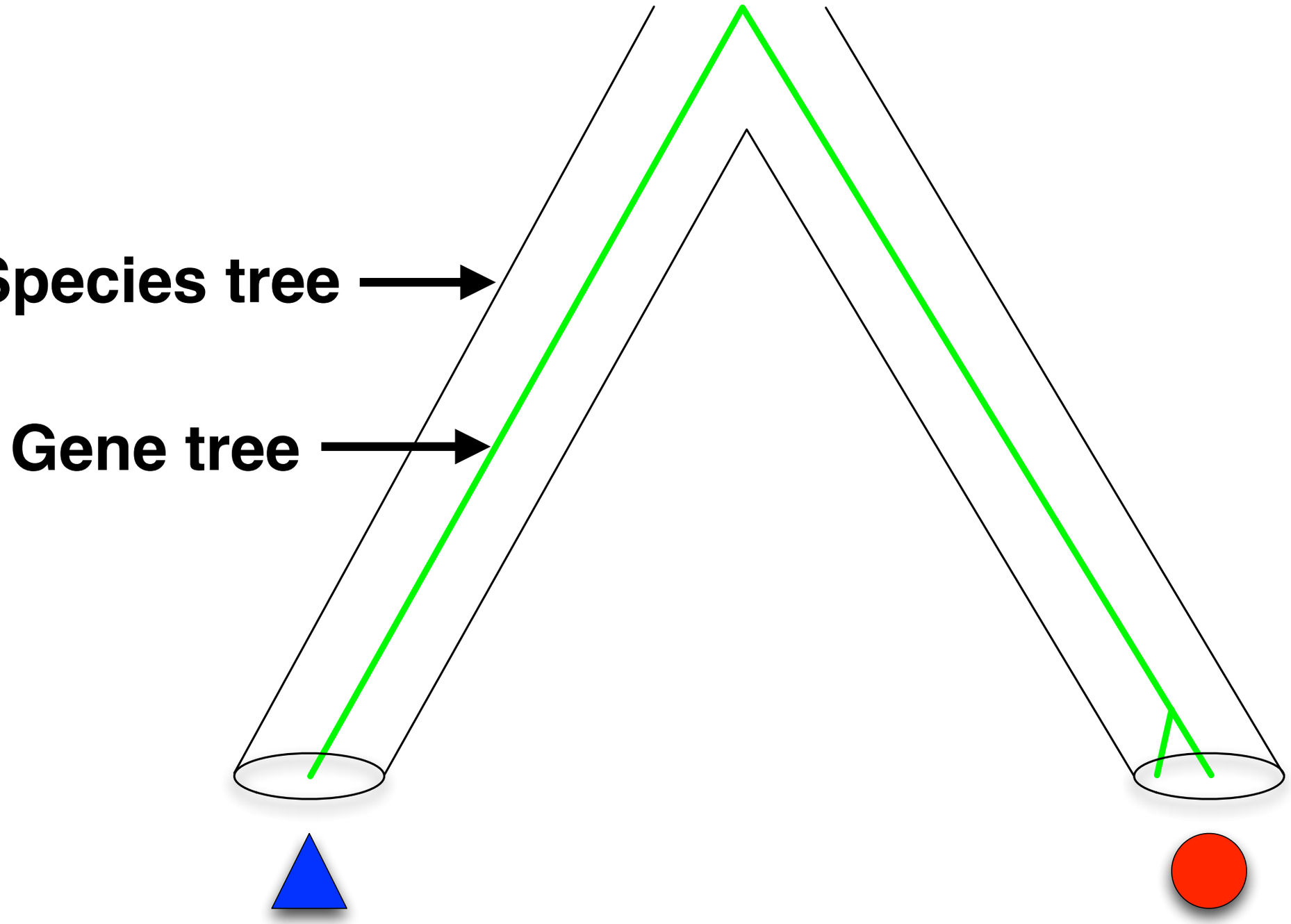
Gene trees (all identical)



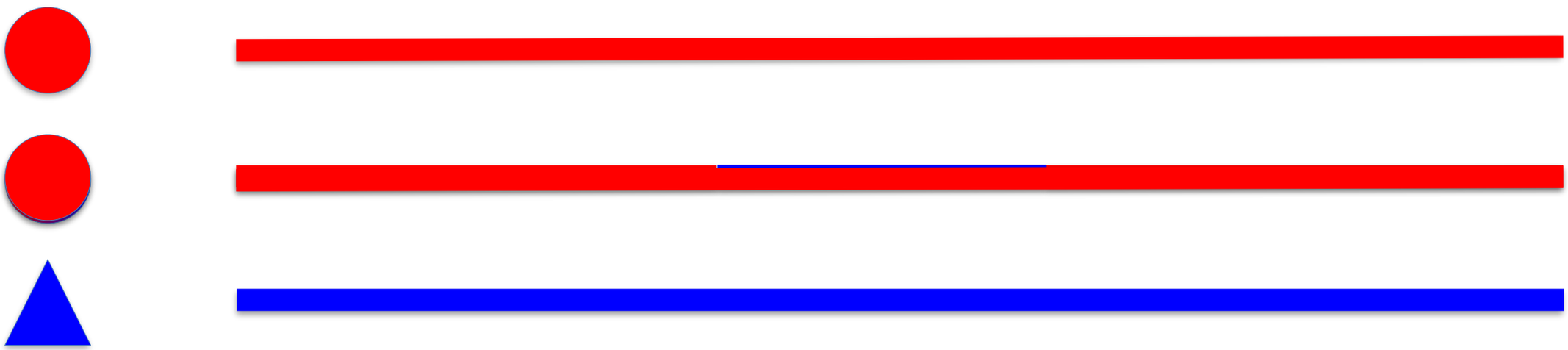
A Gene Tree in a Species Tree (Example)



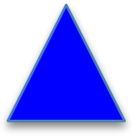
A Gene Tree in a Species Tree (Example)

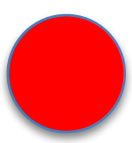


Sliding Windows (Example)

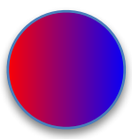


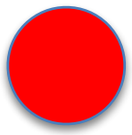
Sliding Windows (Example)



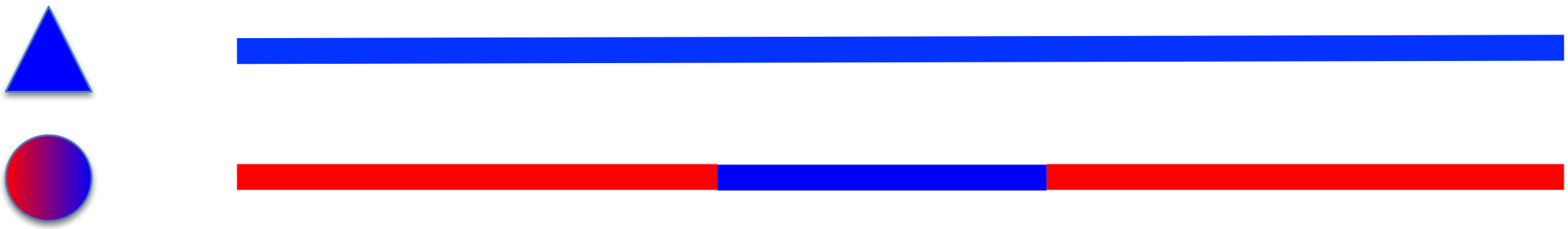


Sliding Windows (Example)

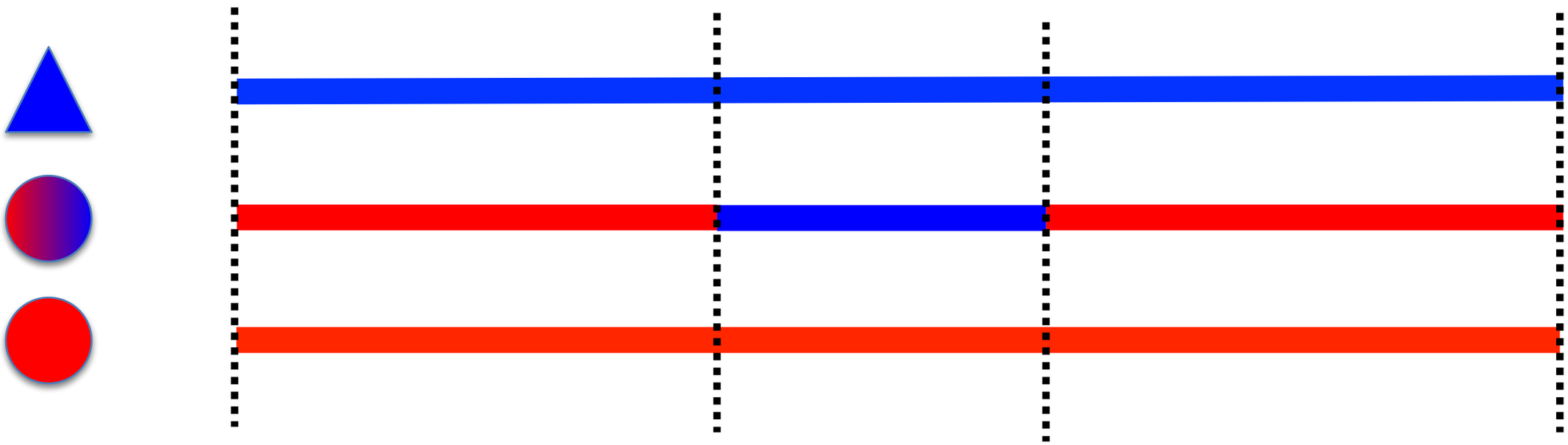




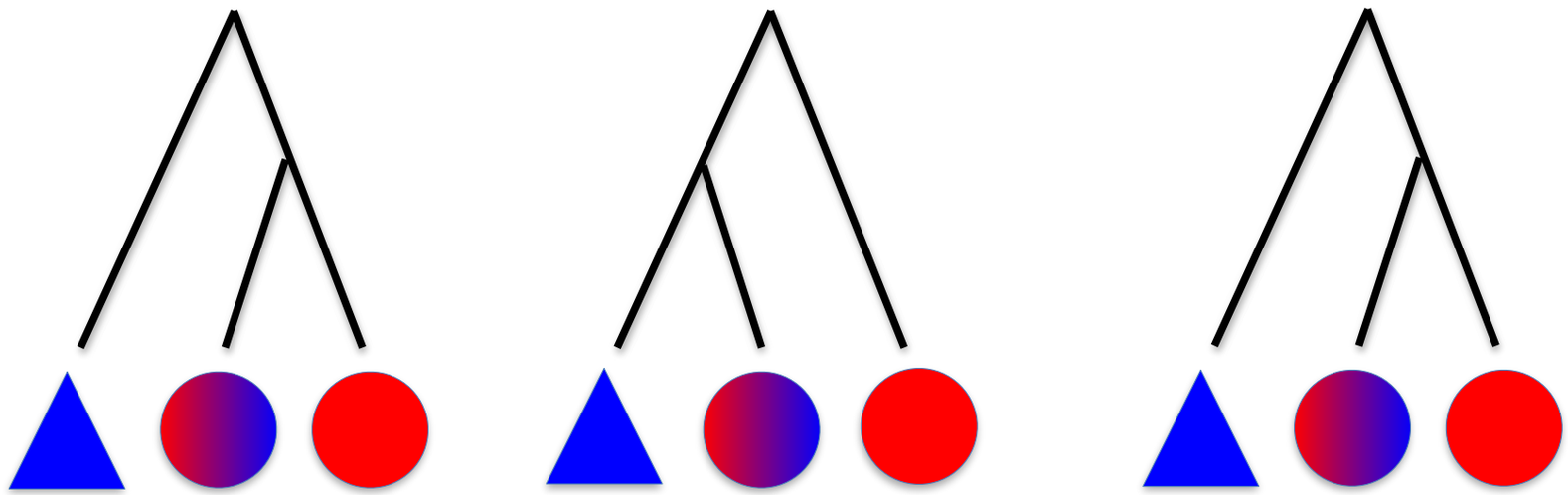
Sliding Windows (Example)



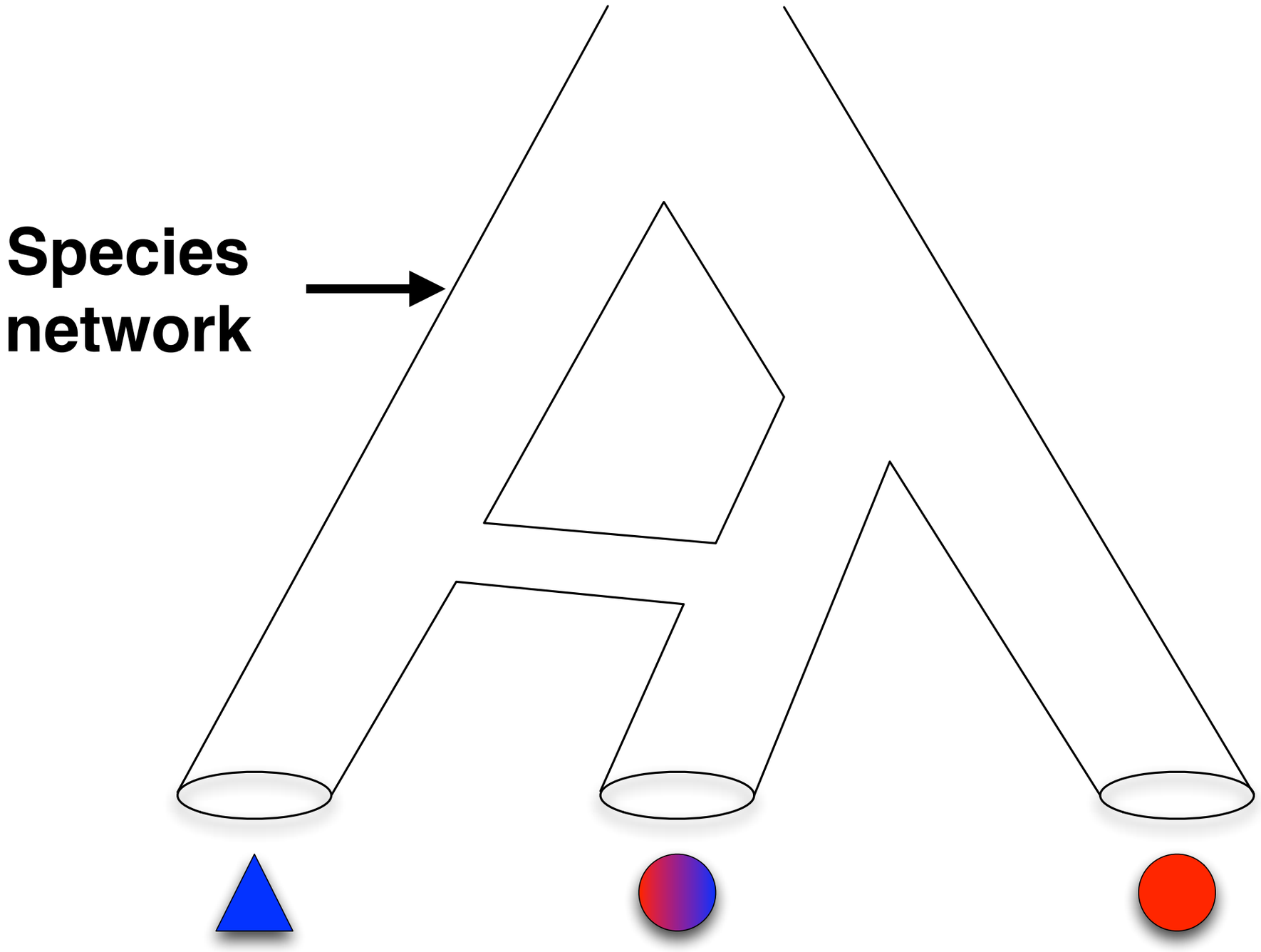
Sliding Windows (Example)



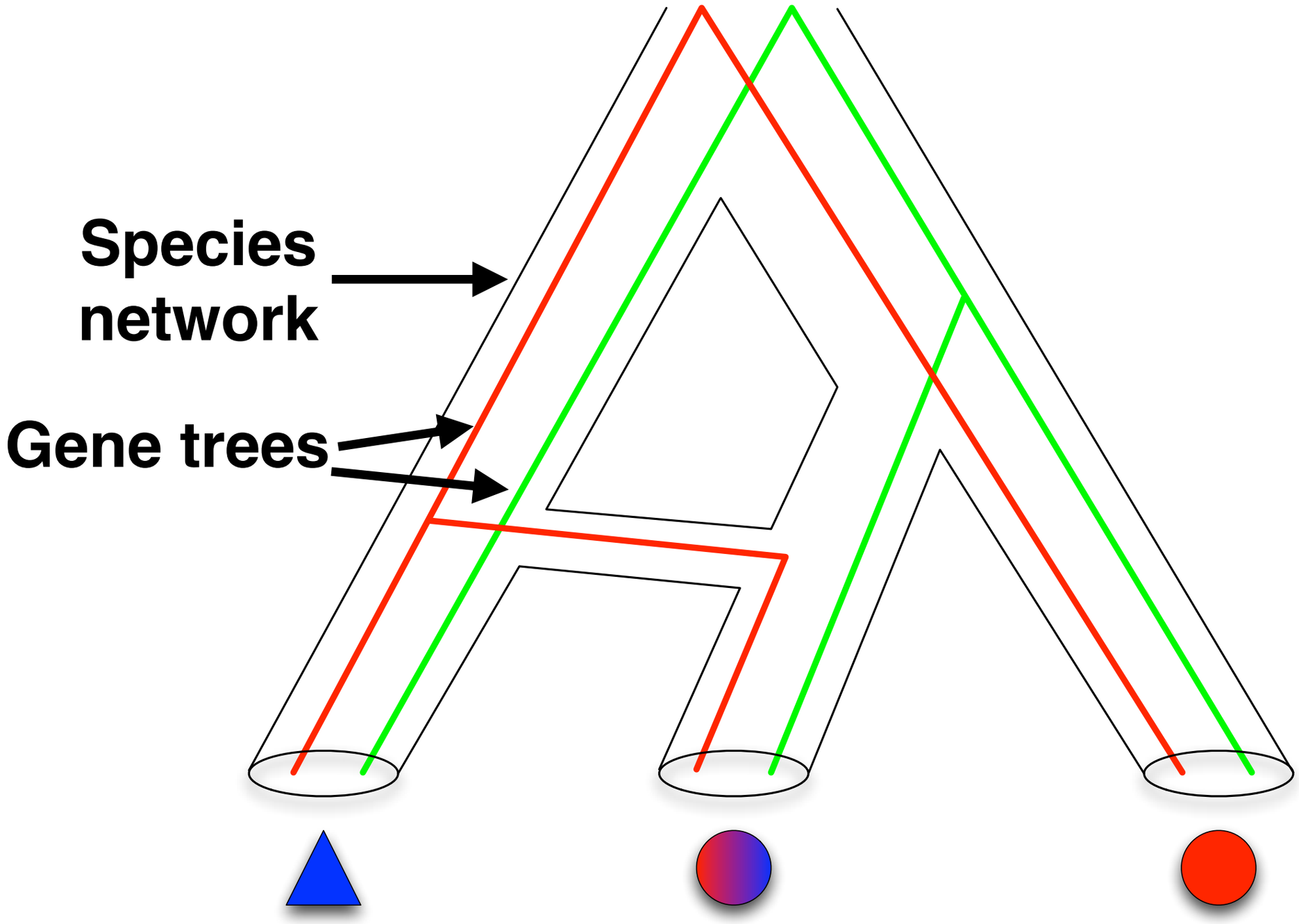
Gene tree incongruence!



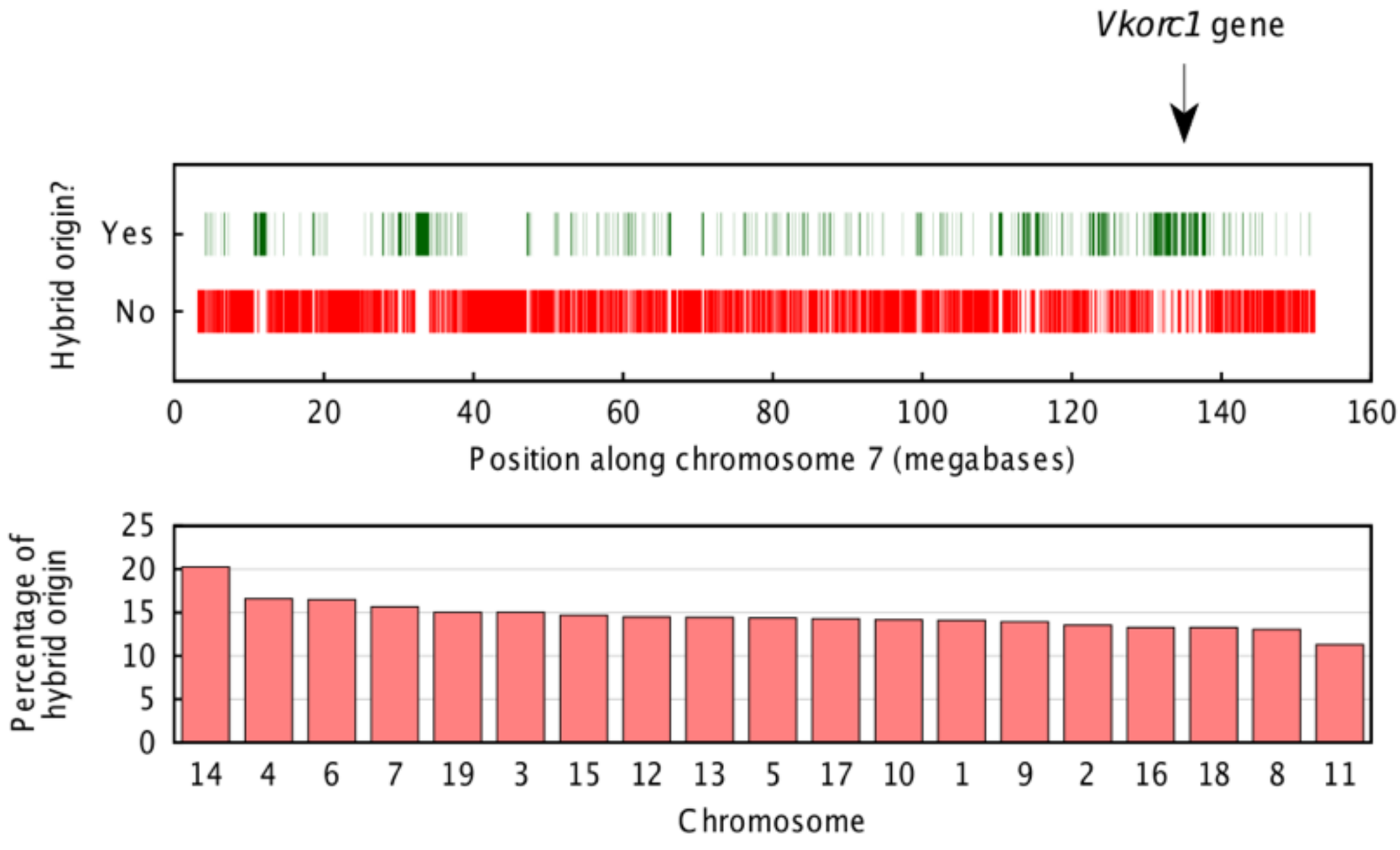
“Horizontal” Gene Tree Incongruence (Example)



“Horizontal” Gene Tree Incongruence (Example)



Sliding Windows: Results



Liu et al. in prep.

Sliding Windows Approach Is Too Simplistic

- Gene tree incongruence can occur for reasons other than introgression
- The organisms in our study included “vertical” gene tree incongruence due to:
  - Incomplete lineage sorting
  - Recombination

Approach #2: Reconciliation

- For a gene tree *g* and a species network *N*, Yu *et al.* 2012 proposed an algorithm to calculate $P[g|N]$, accounting for introgression and incomplete lineage sorting
- Motivates the following optimization problem:

1. Estimate a set of gene trees *G* using Sliding Windows Approach

2. Under the model of Yu *et al.* 2012, choose:

$$\arg\max_{N} \prod_{g \in G} P[g|N]$$

Approach #2: Reconcile Gene Trees with Species Network

- Relevant prior theoretical work:
  - Degnan and Salter 2005
    - Probability $P[g|T]$ of observing a gene tree g given a species tree T
    - Accounts for incomplete lineage sorting only
  - Yu *et al.* 2012
    - Probability $P[g|N]$ of observing a gene tree g given a species network N
    - Accounts for introgression and incomplete lineage sorting

Issues with Reconciliation-based Approaches

- Assumes that gene trees are correct
  - Estimated gene trees typically contain some error
- Assumes that each genome position is identically and independently distributed
  - Biologically unrealistic since adjacent nucleotides tend to be inherited together
- Doesn’t capture recombination

Problem: Computational Introgression Detection

Input:

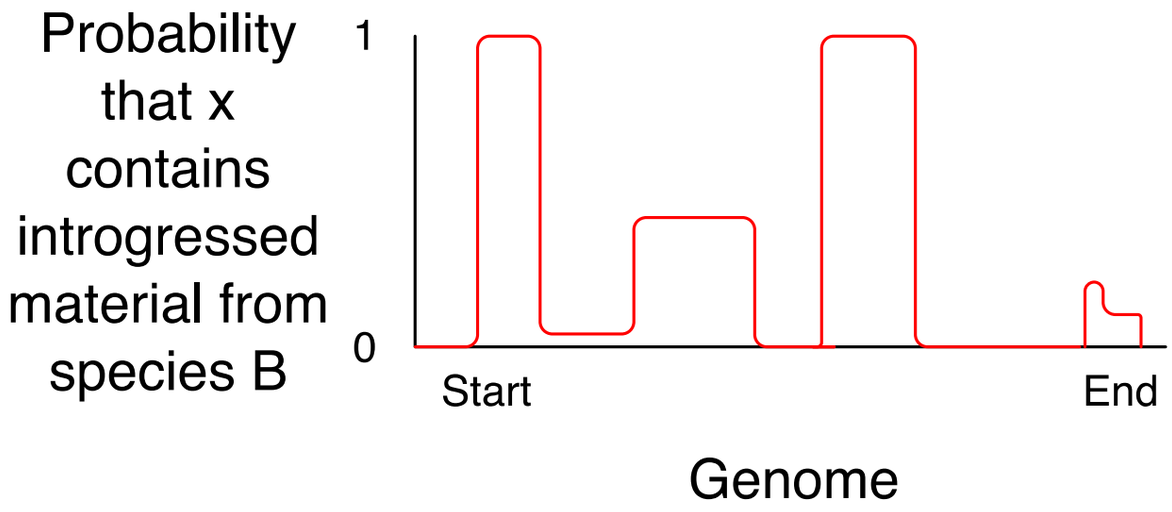
Species	Genome ID	Introgressed?
A	x	Unknown
A	a	No
…	…	…
A	a	No
B	b	No
…	…	…
B	b	No

Problem: Computational Introgression Detection

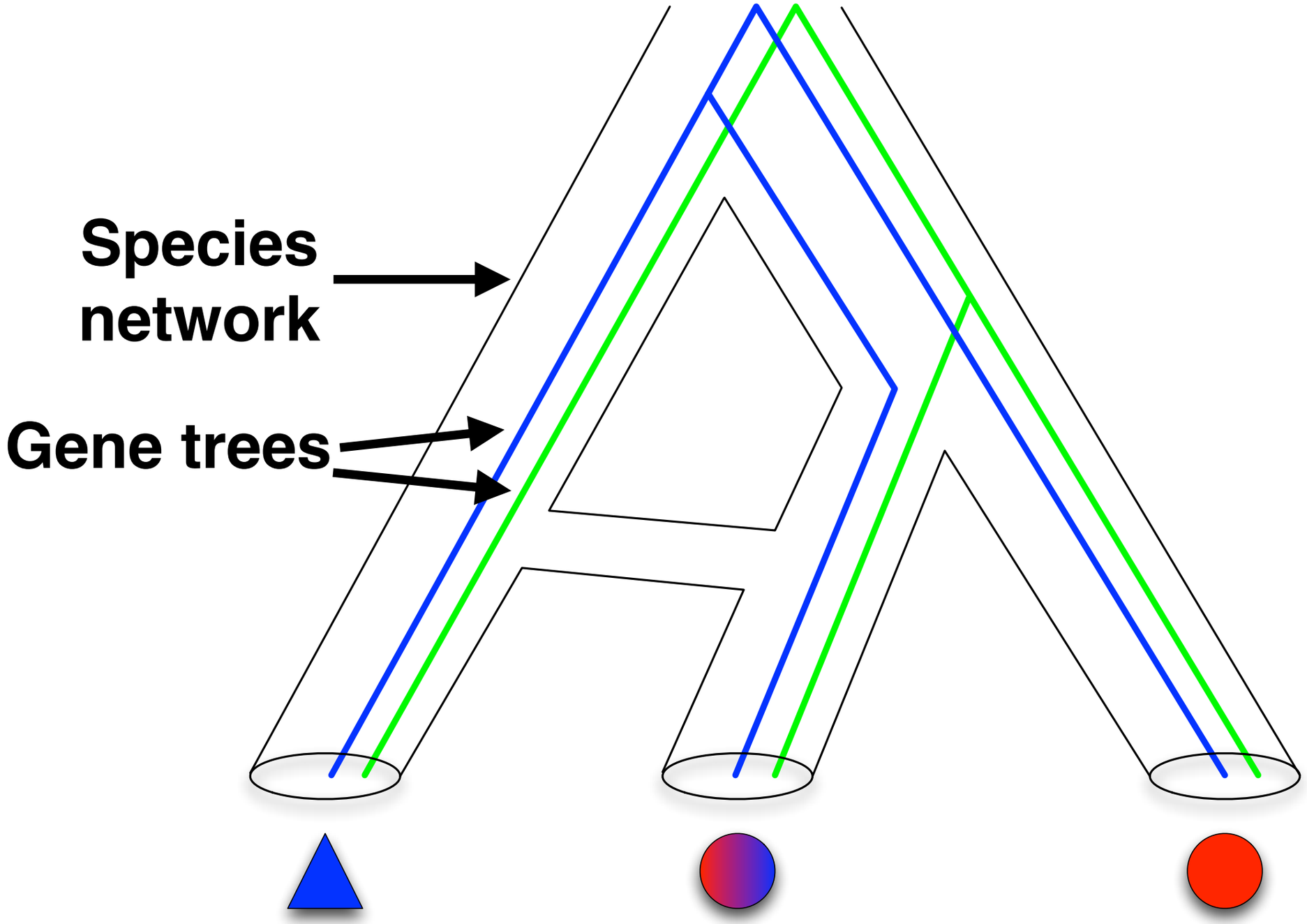
Input:

Species	Genome ID	Introgressed?
A	x	Unknown
A	a	No
…	…	…
A	a	No
B	b	No
…	…	…
B	b	No

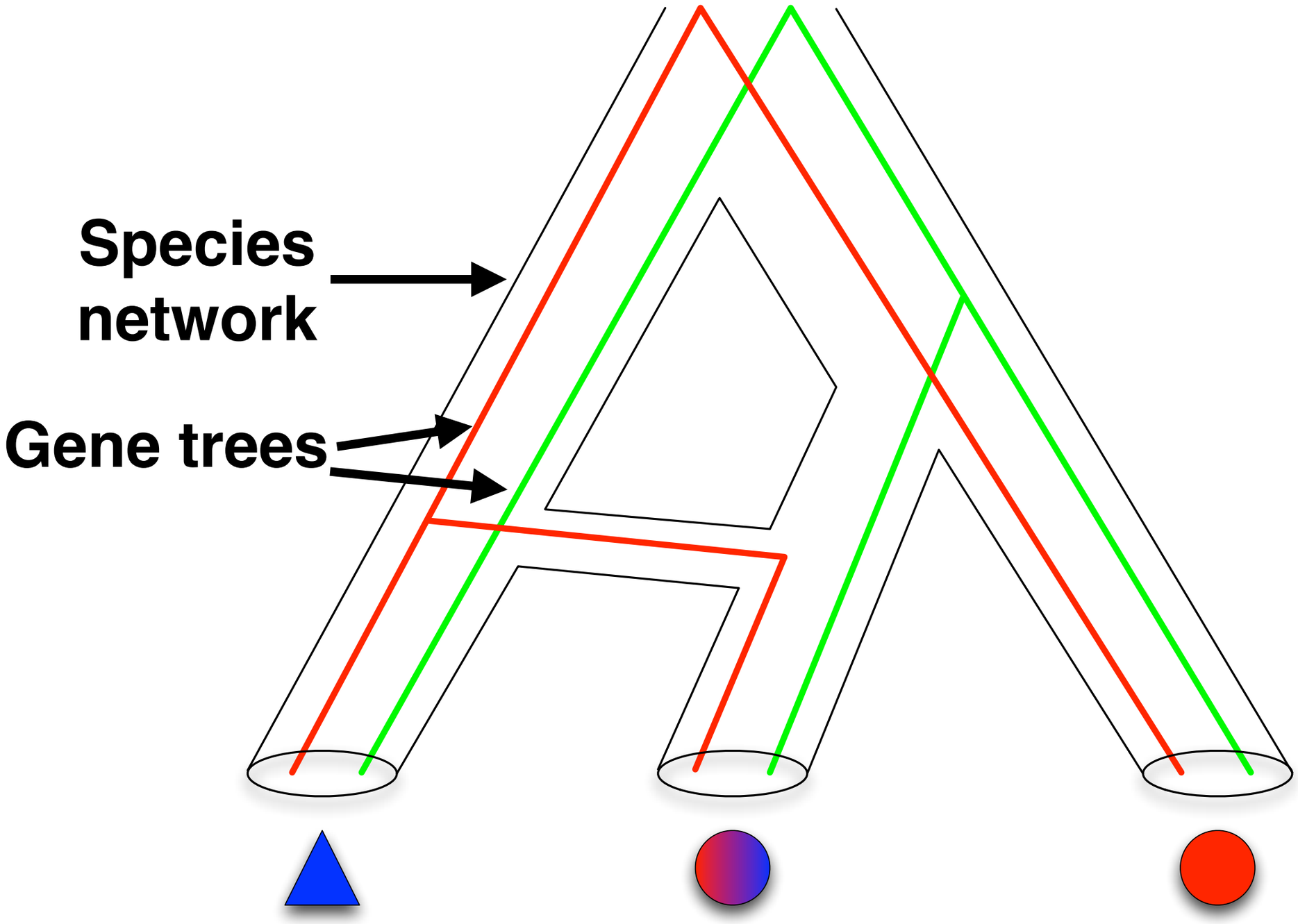
Output:



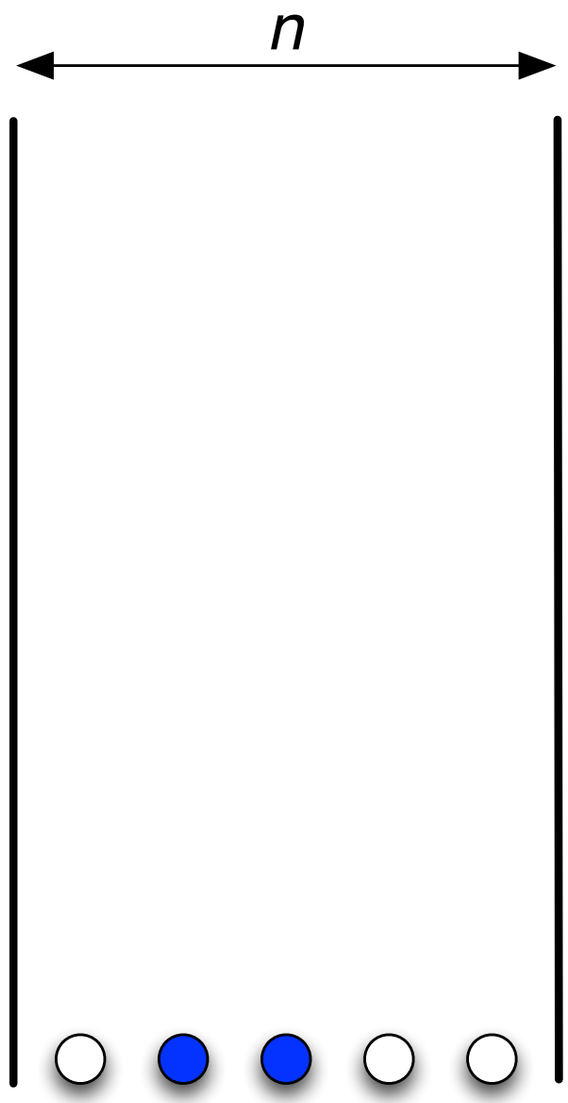
“Vertical” Gene Tree Incongruence (Example)



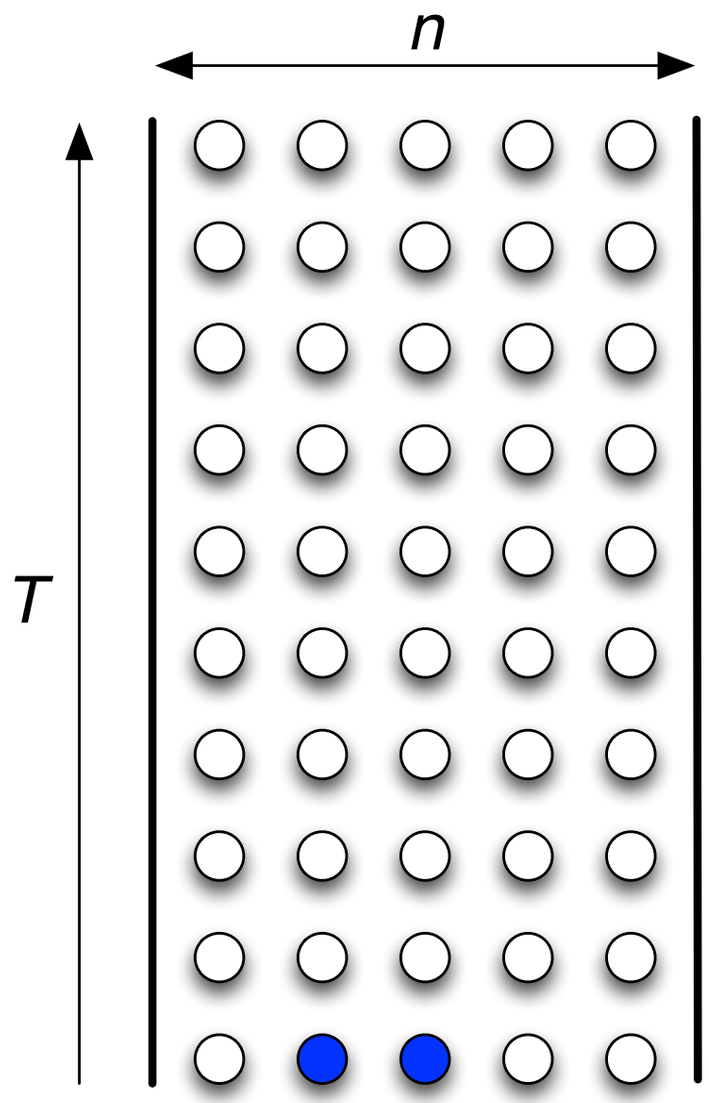
“Horizontal” Gene Tree Incongruence (Example)



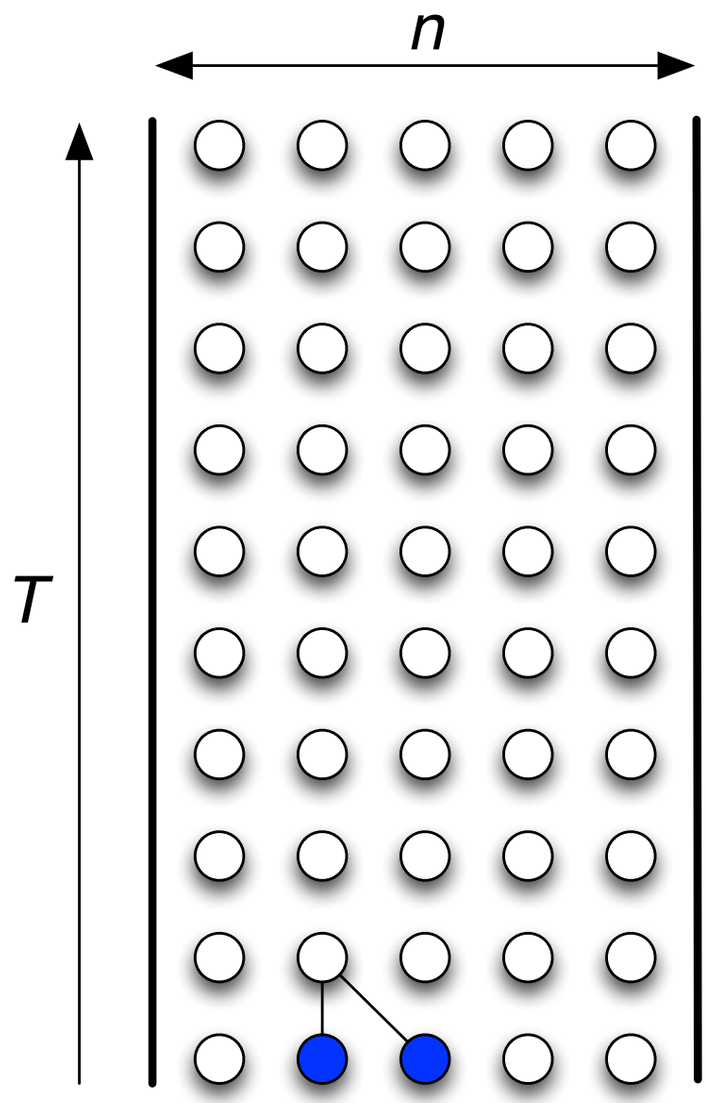
The Probability of A Coalescence Event: Discrete Generations



The Probability of A Coalescence Event: Discrete Generations

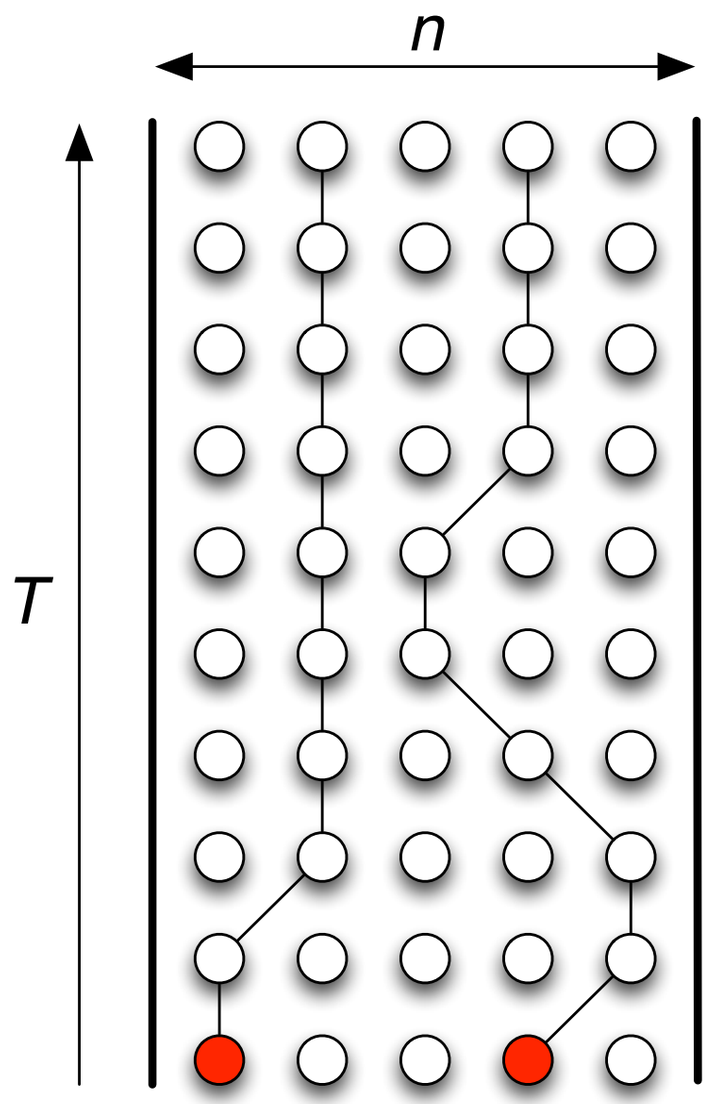


The Probability of A Coalescence Event: Discrete Generations

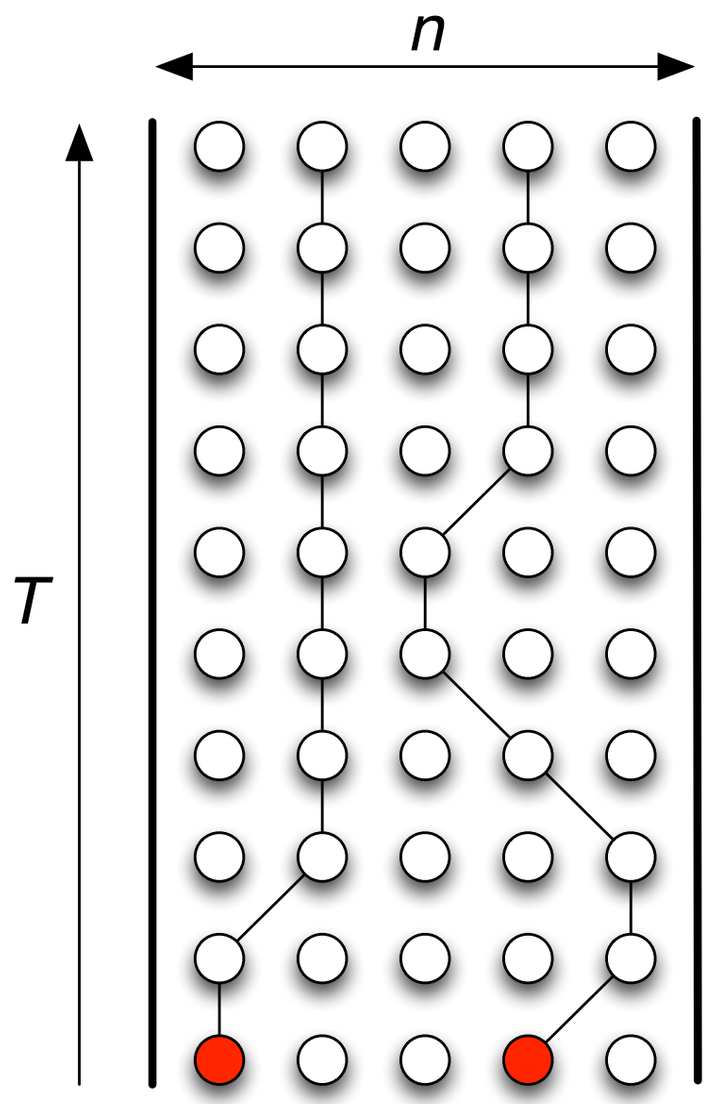


$$\Pr(\text{two lineages coalesce in 1 generation}) = \frac{1}{n}$$

The Probability of A Coalescence Event: Discrete Generations

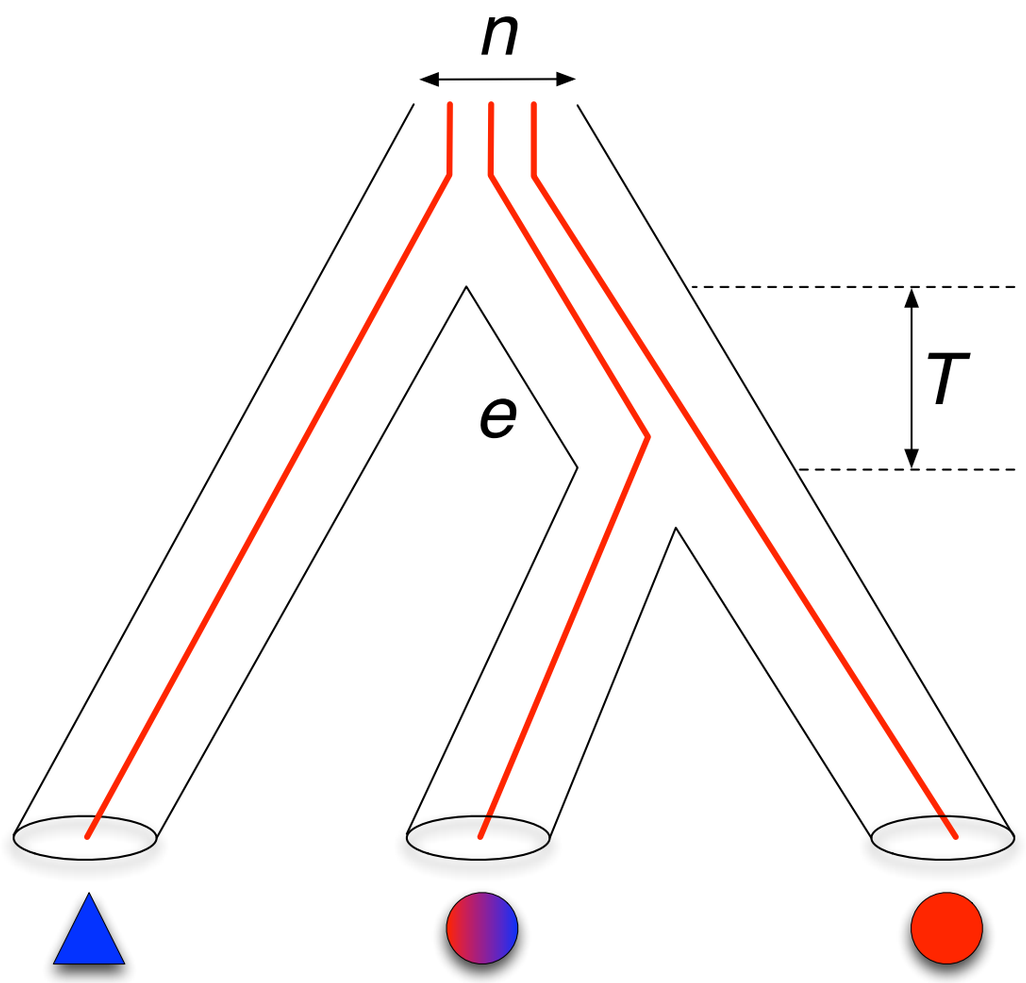


The Probability of A Coalescence Event: Discrete Generations



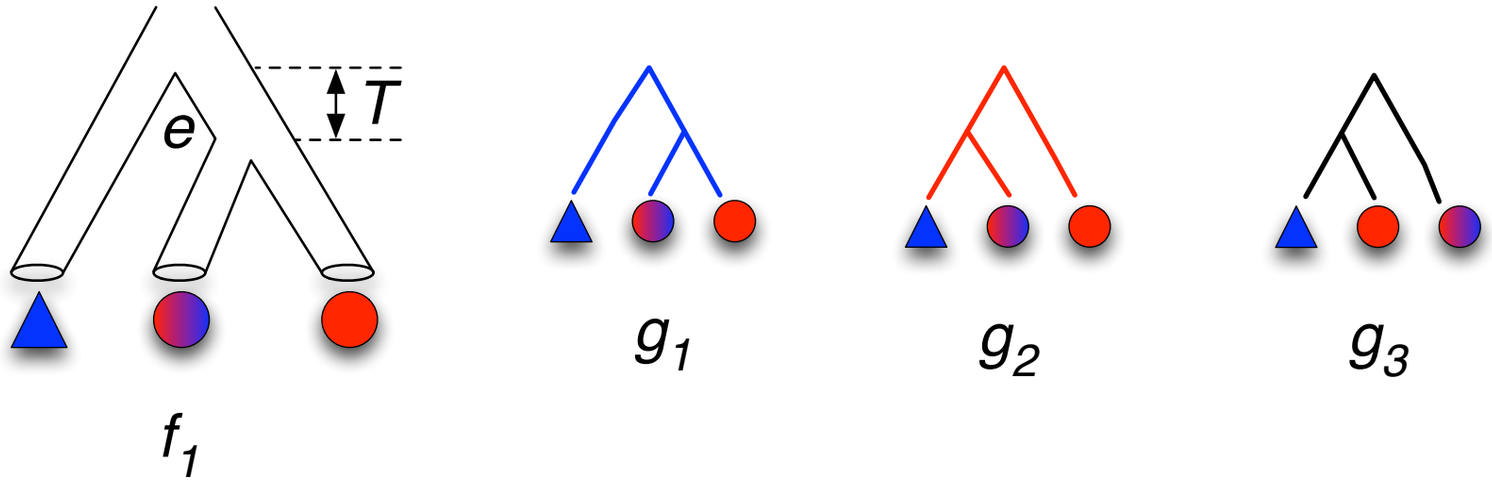
$$\Pr(\text{two lineages don't coalesce in } T \text{ generations}) = (1 - \frac{1}{n})^{T-1}$$

The Probability of a Gene Tree in a Species Tree: Discrete Generations



$$\Pr(\text{two lineages don't coalesce in } T \text{ generations}) = (1 - \frac{1}{n})^{T-1}$$

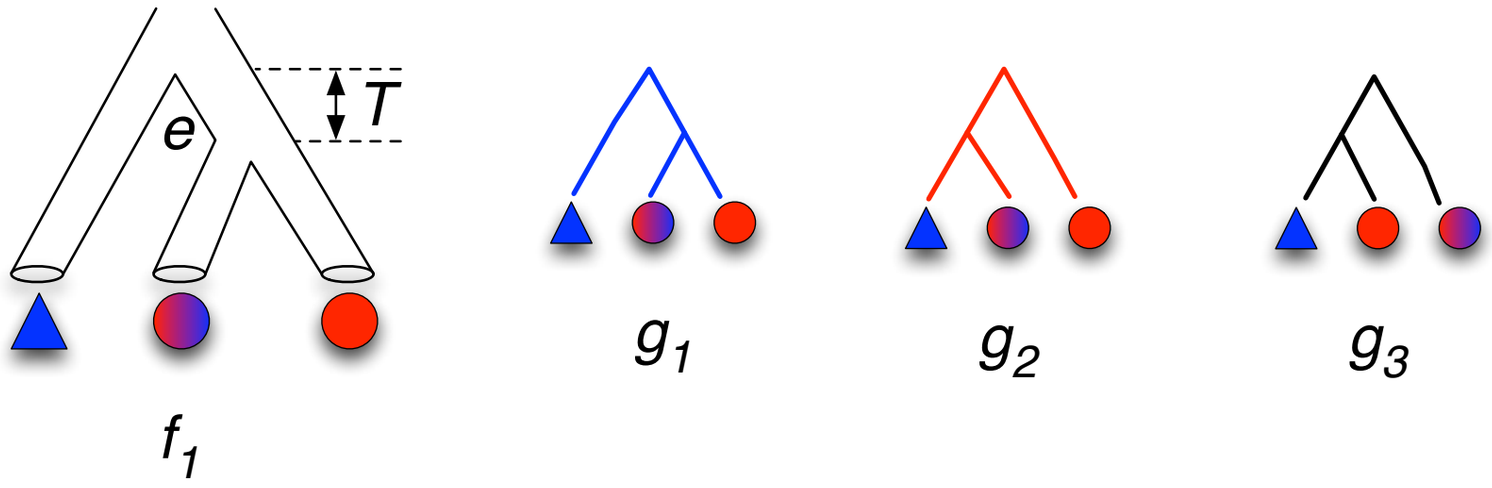
The Probability of a Gene Tree in a Species Tree: Discrete Generations



$$\Pr(\text{two lineages don't coalesce in } T \text{ generations}) = (1 - \frac{1}{n})^{T-1}$$

$$\Pr(g_2|f_1, T) = \Pr(g_3|f_1, T) = \frac{1}{3}(1 - \frac{1}{n})^{T-1}$$

The Probability of a Gene Tree in a Species Tree: Discrete Generations



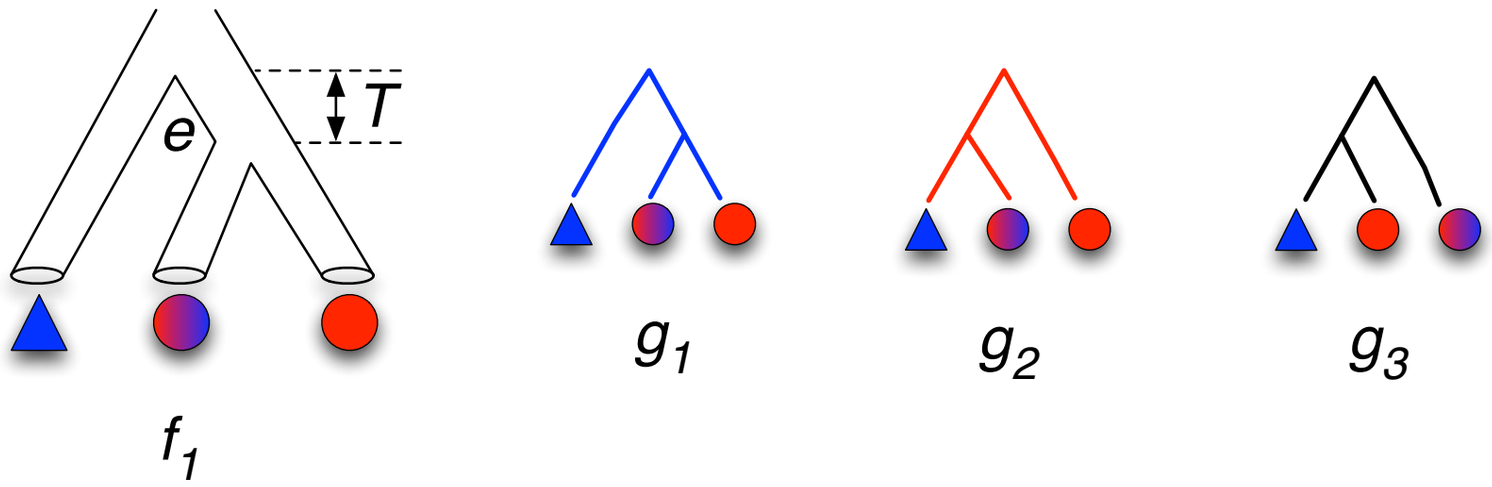
$$\Pr(\text{two lineages don't coalesce in } T \text{ generations}) = (1 - \frac{1}{n})^{T-1}$$

$$\Pr(g_2|f_1, T) = \Pr(g_3|f_1, T) = \frac{1}{3}(1 - \frac{1}{n})^{T-1}$$

$$\Pr(\text{two lineages coalesce in } T \text{ generations}) = 1 - (1 - \frac{1}{n})^{T-1}$$

$$\Pr(g_1|f_1, T) = 1 - \frac{2}{3}(1 - \frac{1}{n})^{T-1}$$

The Probability of a Gene Tree in a Species Tree



$$\Pr(\text{two lineages don't coalesce in time } T) = \frac{T^0 e^{-T}}{0!}$$
$$= e^{-T}$$
$$\Pr(g_2|f_1, T) = \Pr(g_3|f_1, T) = \frac{1}{3}e^{-T}$$
$$\Pr(\text{two lineages coalesce in time } T) = 1 - e^{-T}$$
$$\Pr(g_1|f_1, T) = 1 - \frac{2}{3}e^{-T}$$

The Probability of a Gene Tree in a Species Tree

- The probability of *u* lineages coalescing into *v* lineages in time *T* (Rosenberg 2002 and others):

$$p_{uv}(T) = \sum_{k=v}^{u} e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-v}}{v!(k-v)!(v+k-1)} \times \prod_{y=0}^{k-1} \frac{(v+y)(u-y)}{(u+y)}.$$

- The probability of a gene tree topology *g* given a containing species tree (Ψ,λ) (Degan and Salter 2005):

$$P_{\psi,\boldsymbol{\lambda}}(G = g) = \sum_{\mathbf{h} \in H_{\psi}(g)} \frac{w(\mathbf{h})}{d(\mathbf{h})} \prod_{b=1}^{n-2} \frac{w_b(\mathbf{h})}{d_b(\mathbf{h})} p_{u_b(\mathbf{h})v_b(\mathbf{h})}(\lambda_b).$$

PhyloNet-HMM

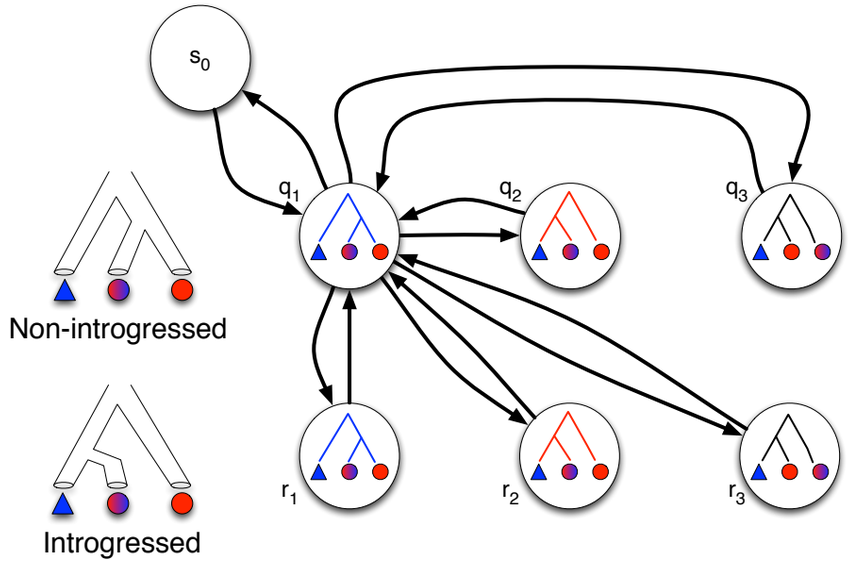
- Each hidden state s_i is associated with a gene tree $g(s_i)$ contained within a "parental" tree $f(s_i)$
- The set of HMM parameters λ consists of
  - The initial state distribution π
  - Transition probabilities

$$a_{ij} = \begin{cases} P(g(s_i)|f(s_i)) \cdot \gamma & \text{if } s_i \text{ and } s_j \text{ in different rows} \\ P(g(s_i)|f(s_i)) \cdot (1-\gamma) & \text{if } s_i \text{ and } s_j \text{ in same row} \end{cases}$$

where γ is the "horizontal" parental tree switching frequency.

  - The emission probabilities $b_i = P(O_t|g(s_i))$

PhyloNet-HMM



- Each hidden state s_i is associated with a gene tree $g(s_i)$ contained within a "parental" tree $f(s_i)$.
- Let q_t be PhyloNet-HMM's hidden state at time t, where $1 \leq t \leq k$ and k is the length of the input observation sequence O.
- The set of HMM parameters λ consists of:
  - Transition probabilities $A=\{a_{ij}\}$, where

$$a_{ij} = \begin{cases} \gamma \Pr[g(s_i)|f(s_i)] & \text{if } s_i \text{ and } s_j \text{ are in different rows} \\ (1-\gamma) \Pr[g(s_i)|f(s_i)] & \text{if } s_i \text{ and } s_j \text{ are in same row} \end{cases}$$

    and γ is the "vertical" parental tree switching frequency and Pr[g(s
  - The emission probabilities $b_i = \Pr[O_t|g(s_i)]$ under a model of nucleotide substitution (e.g., Jukes-Cantor (1969))
  - The initial state distribution $\pi_i = \Pr[q_1 = s_i]$

First HMM-related Problem

- Let q_t be PhyloNet-HMM's hidden state at time t, where $1 \leq t \leq k$ and k is the length of the input observation sequence O.

- What is the likelihood of the model given the observed DNA sequences O?

  - Forward algorithm calculates "prefix" probability $\alpha_t(i)$

  - Backward algorithm calculates "suffix" probability $\beta_t(i)$

  - Model likelihood is $P[O|\lambda] = \sum_{i=1}^{N} \alpha_k(i).$

First HMM-related Problem

- Let q_t be PhyloNet-HMM's hidden state at time t, where $1 \leq t \leq k$ and k is the length of the input observation sequence O.
- What is the likelihood of the model given the observed DNA sequences O?

$$\alpha_t(i) = P[O_1, O_2, \ldots, O_t, q_t = S_i | \lambda].$$

  - Forward algorithm calculates the "prefix" probability

$$\beta_t(i) = P[O_{t+1}, O_{t+2}, \ldots, O_k | q_t = S_i, \lambda].$$

  - Backward algorithm calculates the "suffix" probability
  - Model likelihood is $P[O|\lambda] = \sum_{i=1}^{N} \alpha_k(i).$

Second HMM-related Problem

- Which sequence of states best explains the observation sequence?

- Posterior decoding probability $\gamma_t(i)$ is the probability that the HMM is in state s_i at time t, which can be computed as:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P[O|\lambda]}.$$

Third HMM-related Problem

- How do we choose parameter values that maximize the model likelihood?

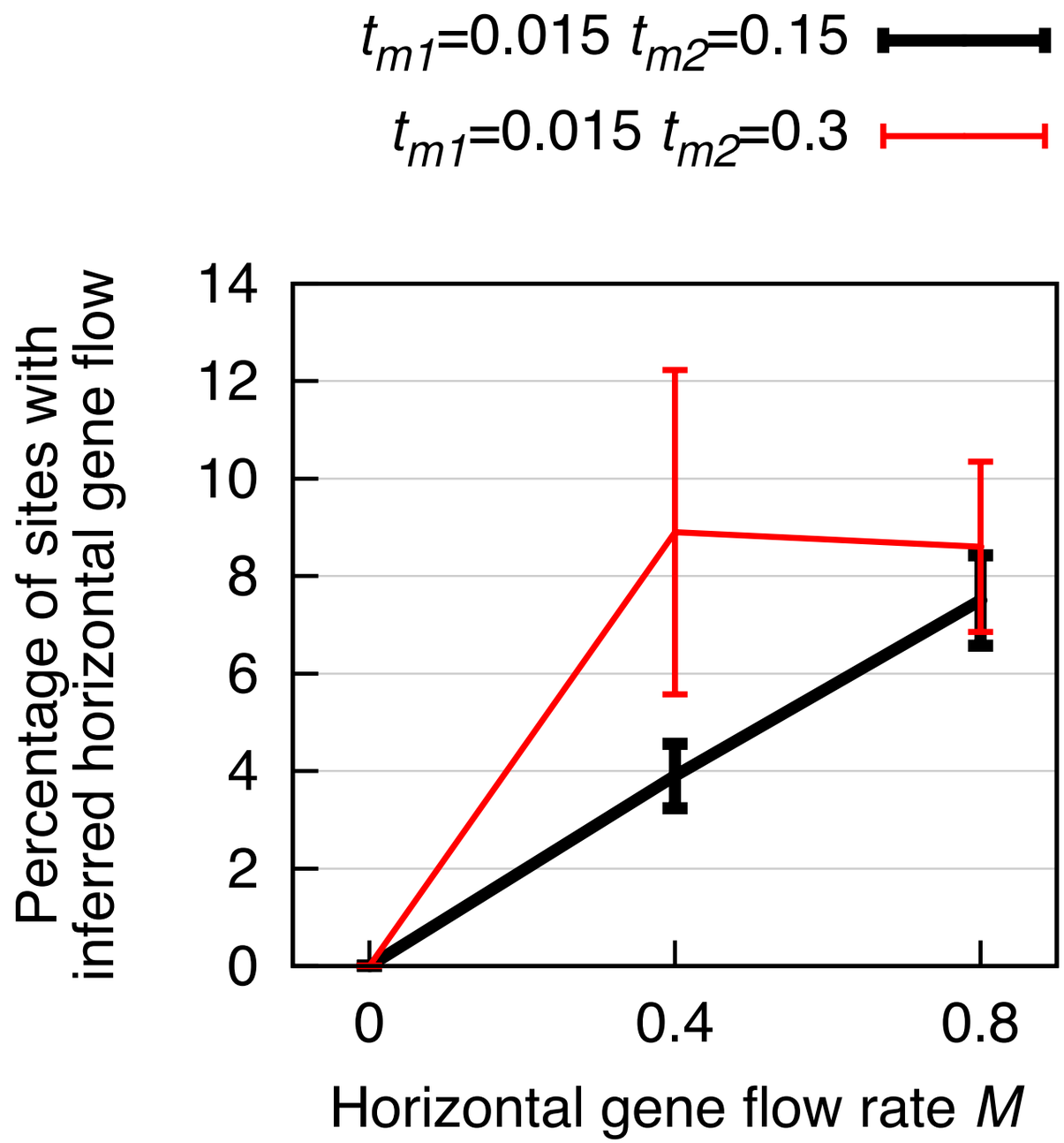
  - Perform local search to optimize the criterion

$$\arg\max_{\lambda} P[O|\lambda]$$

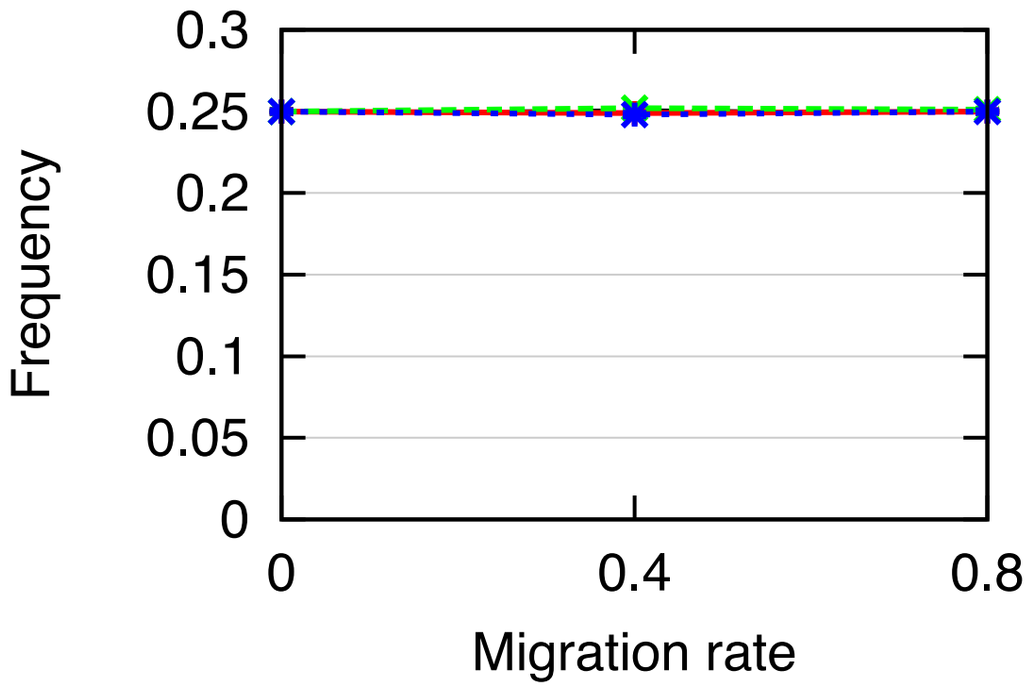
Related HMM-based Approaches

- CoalHMM (Mailund *et al.* 2012)
  - Models introgression + incomplete lineage sorting + recombination (with a simplifying assumption)
  - Currently supports two sequences only
  - Assumes that time is discretized
- Other approaches that don't account for introgression (*e.g.*, Hobolth *et al.* 2007)

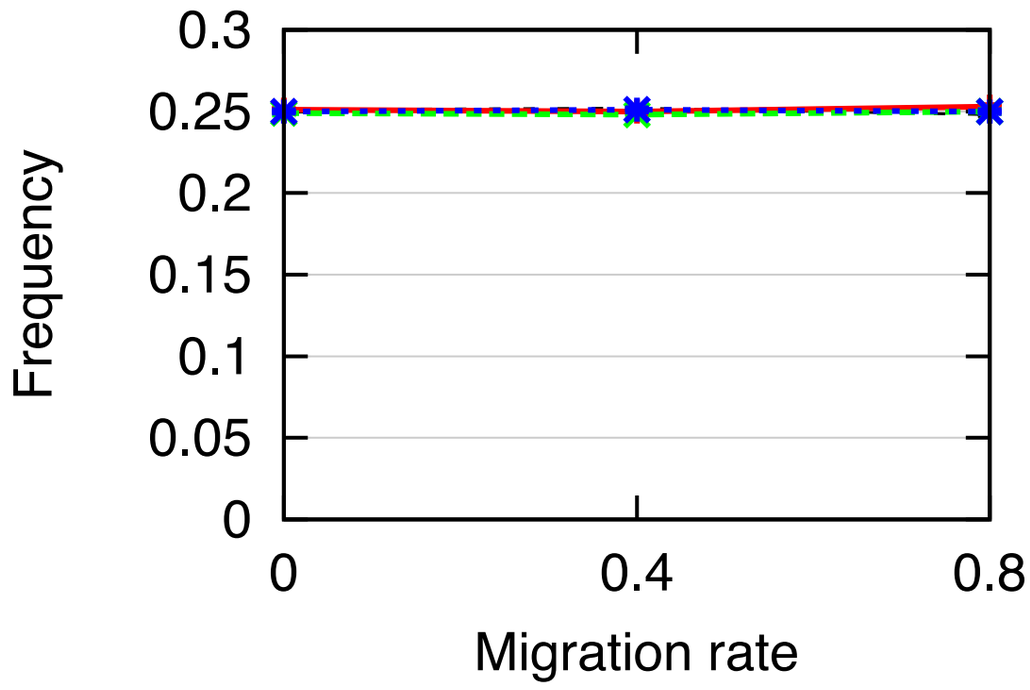
Simulation Study Results



Simulation Study Results

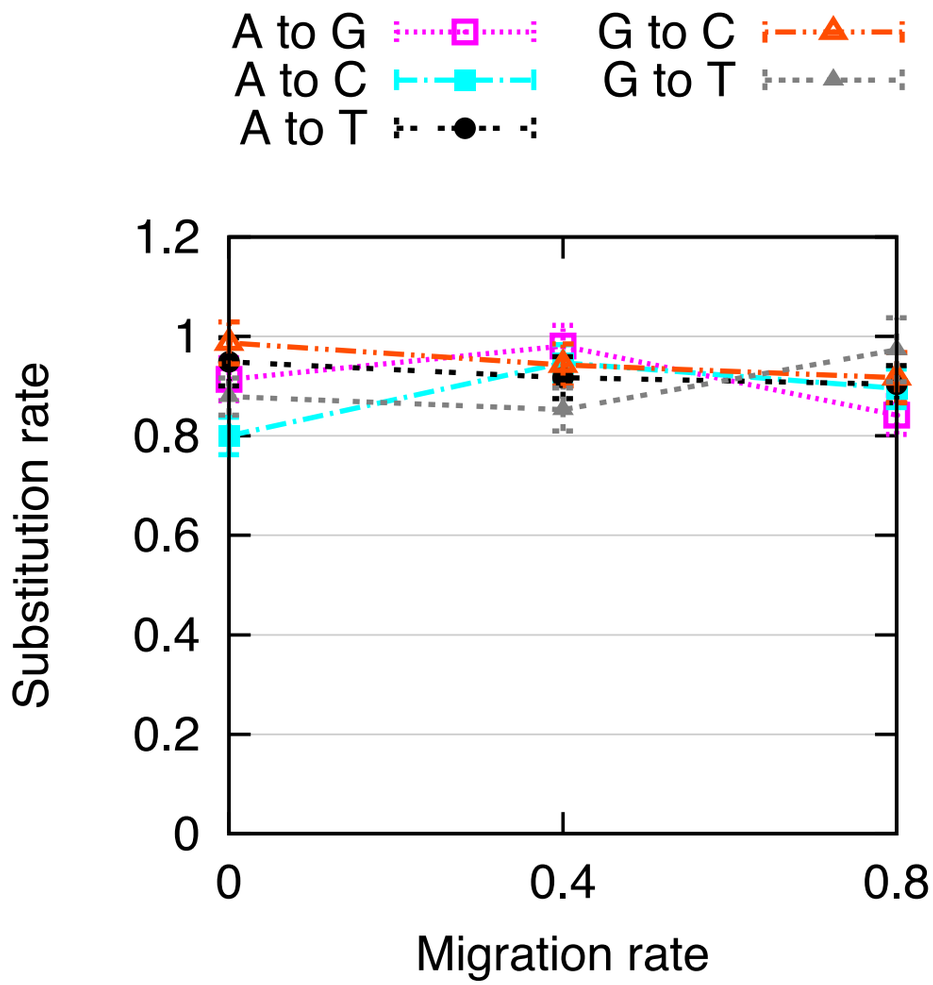


(a) $t_{m2} = 0.015, t_{m1} = 0.15$

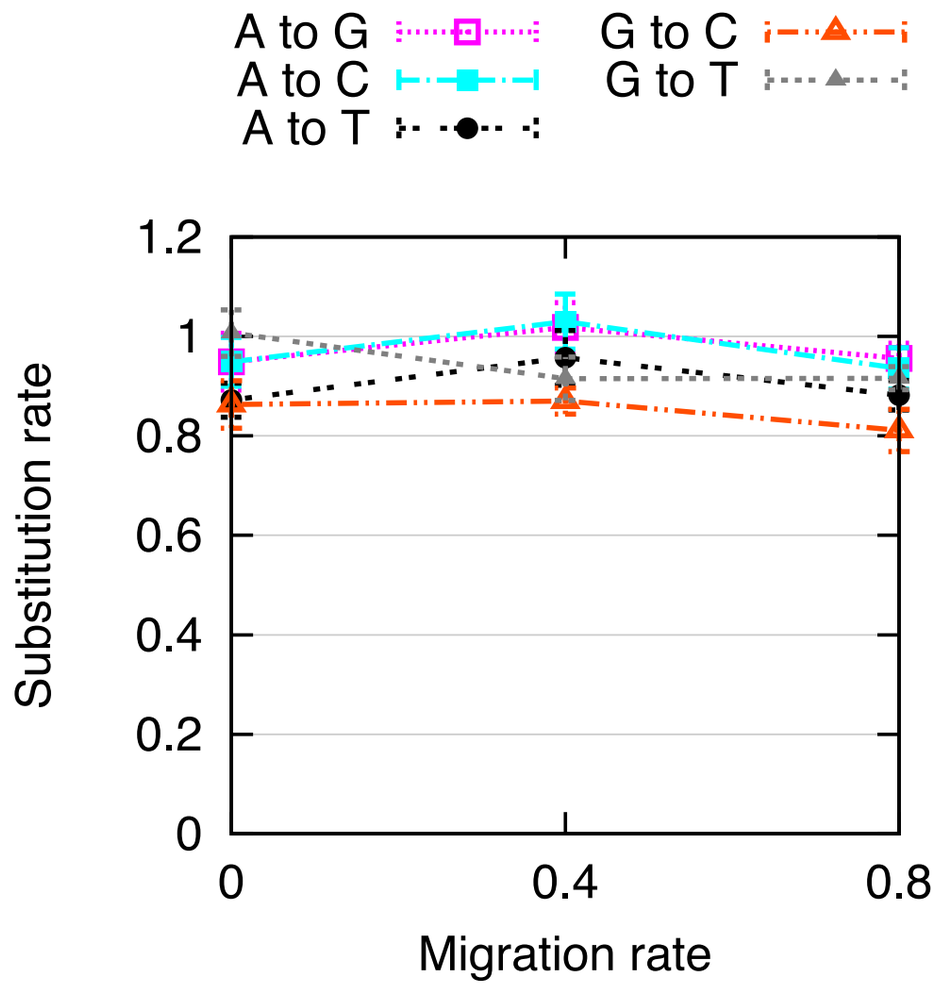


(b) $t_{m2} = 0.015, t_{m1} = 0.3$

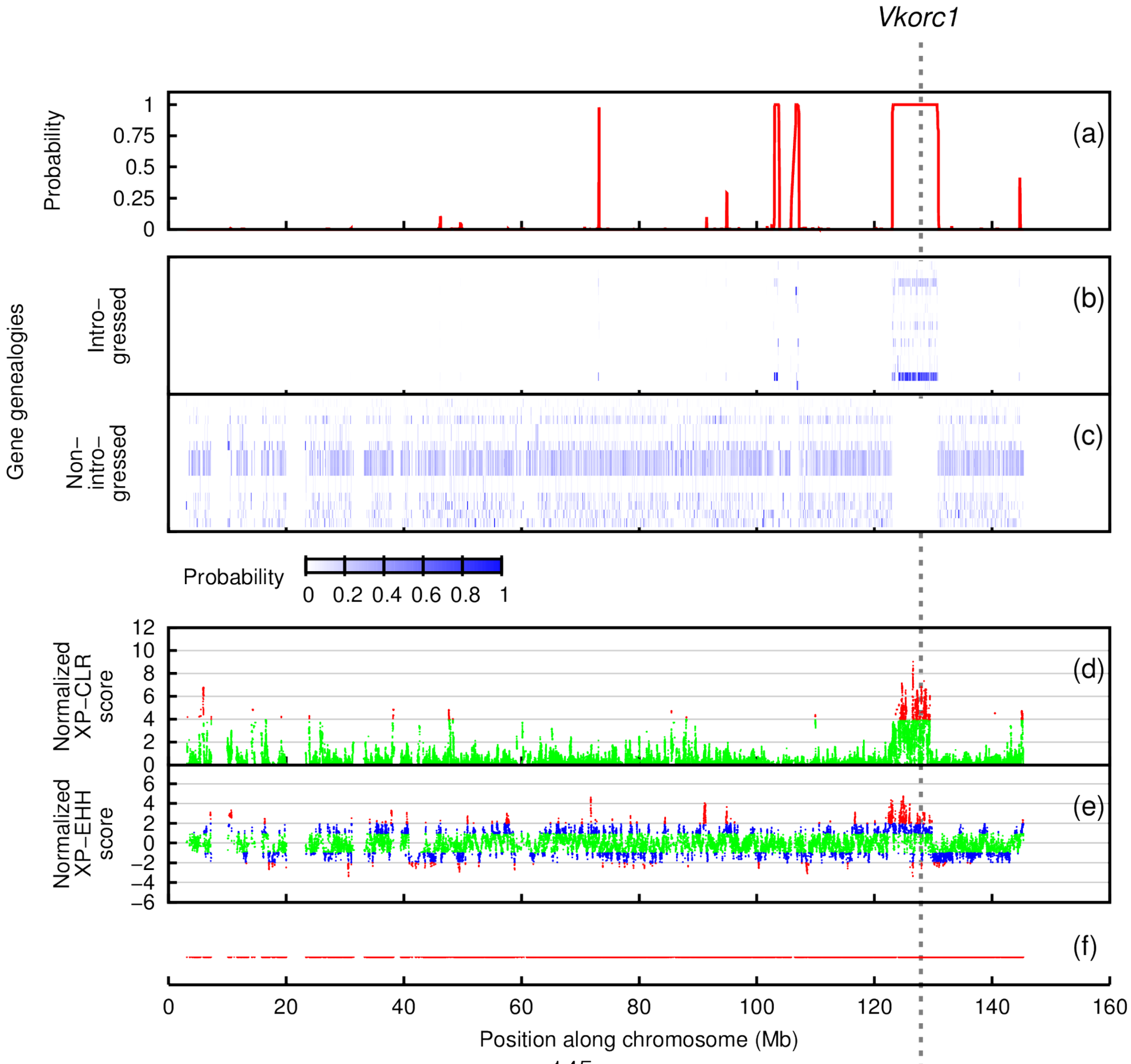
Simulation Study Results

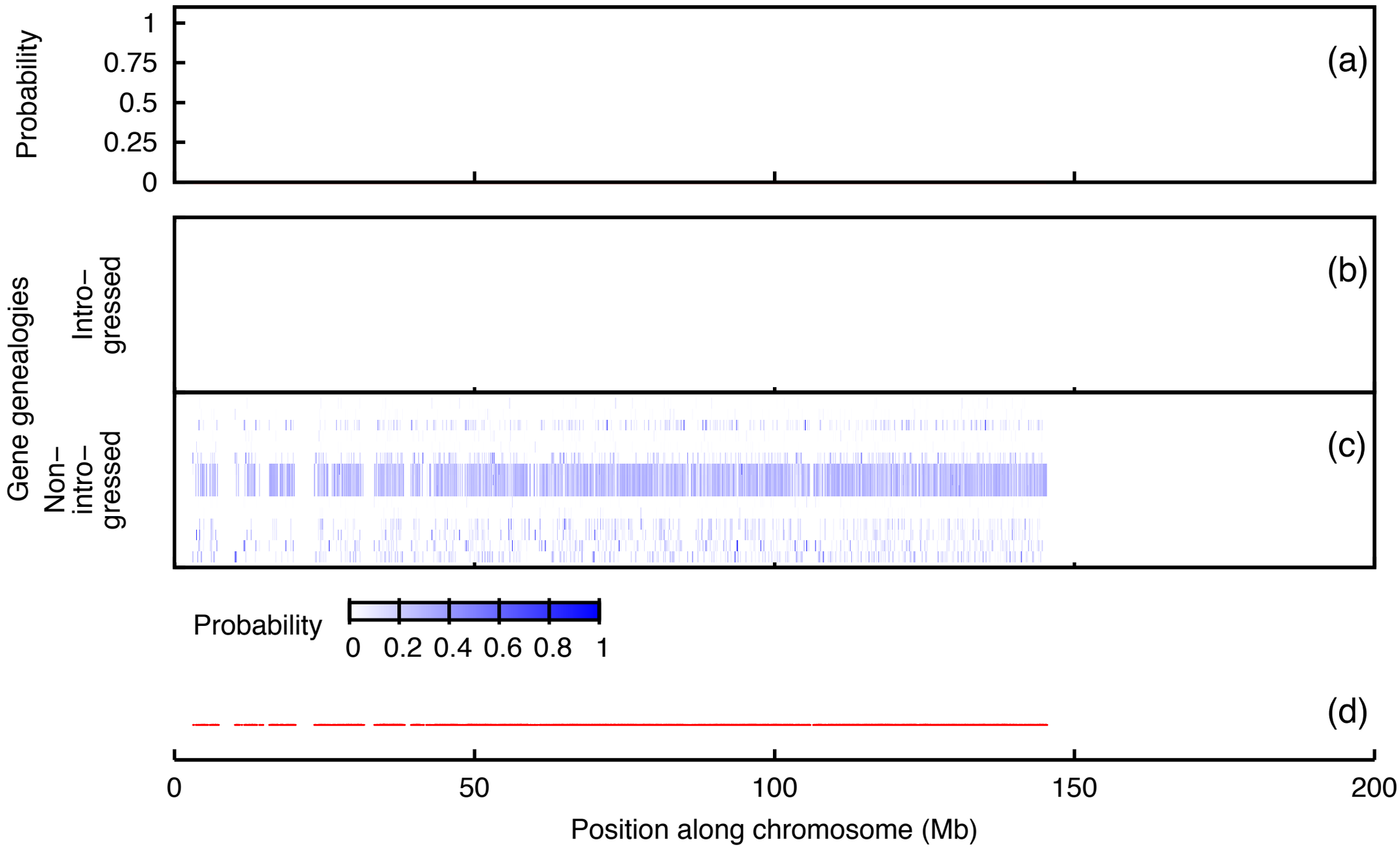


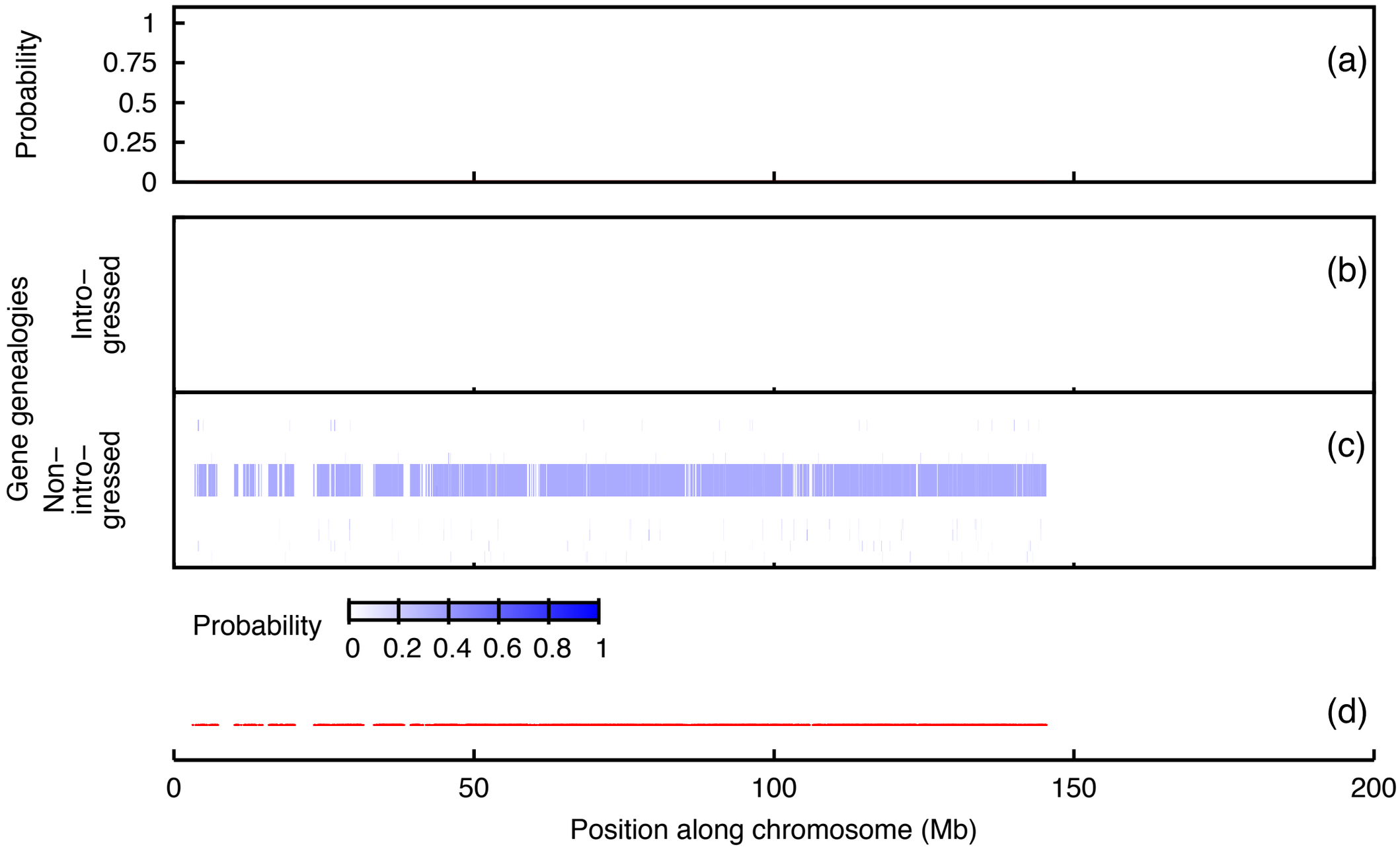
(a) $t_{m2} = 0.015, t_{m1} = 0.15$



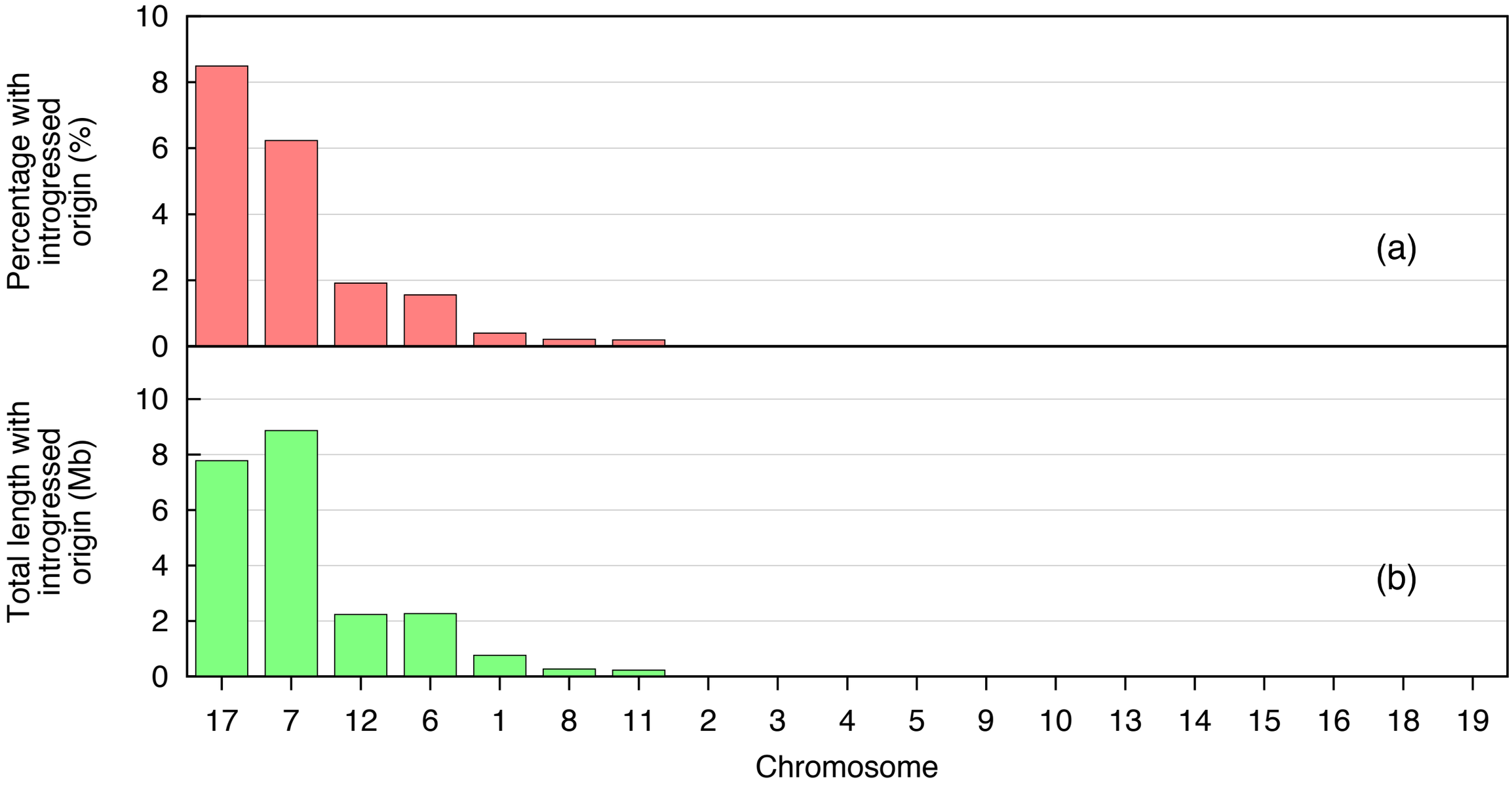
(b) $t_{m2} = 0.015, t_{m1} = 0.3$


Vkorc1
Probability
1
0.75
0.5
0.25
0
(a)
Gene genealogies
Intro–gressed
(b)
Non–intro–gressed
(c)
Probability
0 0.2 0.4 0.6 0.8 1
Normalized XP–CLR score
12
10
8
6
4
2
0
(d)
Normalized XP–EHH score
6
4
2
0
–2
–4
–6
(e)
(f)
0
20
40
60
80
100
120
140
160
Position along chromosome (Mb)


Probability
1
0.75
0.5
0.25
0
(a)
Gene genealogies
Intro–
gressed
Non–
intro–
gressed
(b)
(c)
Probability
0 0.2 0.4 0.6 0.8 1
(d)
0
50
100
150
200
Position along chromosome (Mb)


Probability
1
0.75
0.5
0.25
0
(a)
Gene genealogies
Intro–
gressed
(b)
Non–
intro–
gressed
(c)
Probability
0 0.2 0.4 0.6 0.8 1
(d)
0
50
100
150
200
Position along chromosome (Mb)

PhyloNet-HMM Scan of Whole Mouse Genomes



Liu *et al.* in prep. for PNAS.

Future Direction #4

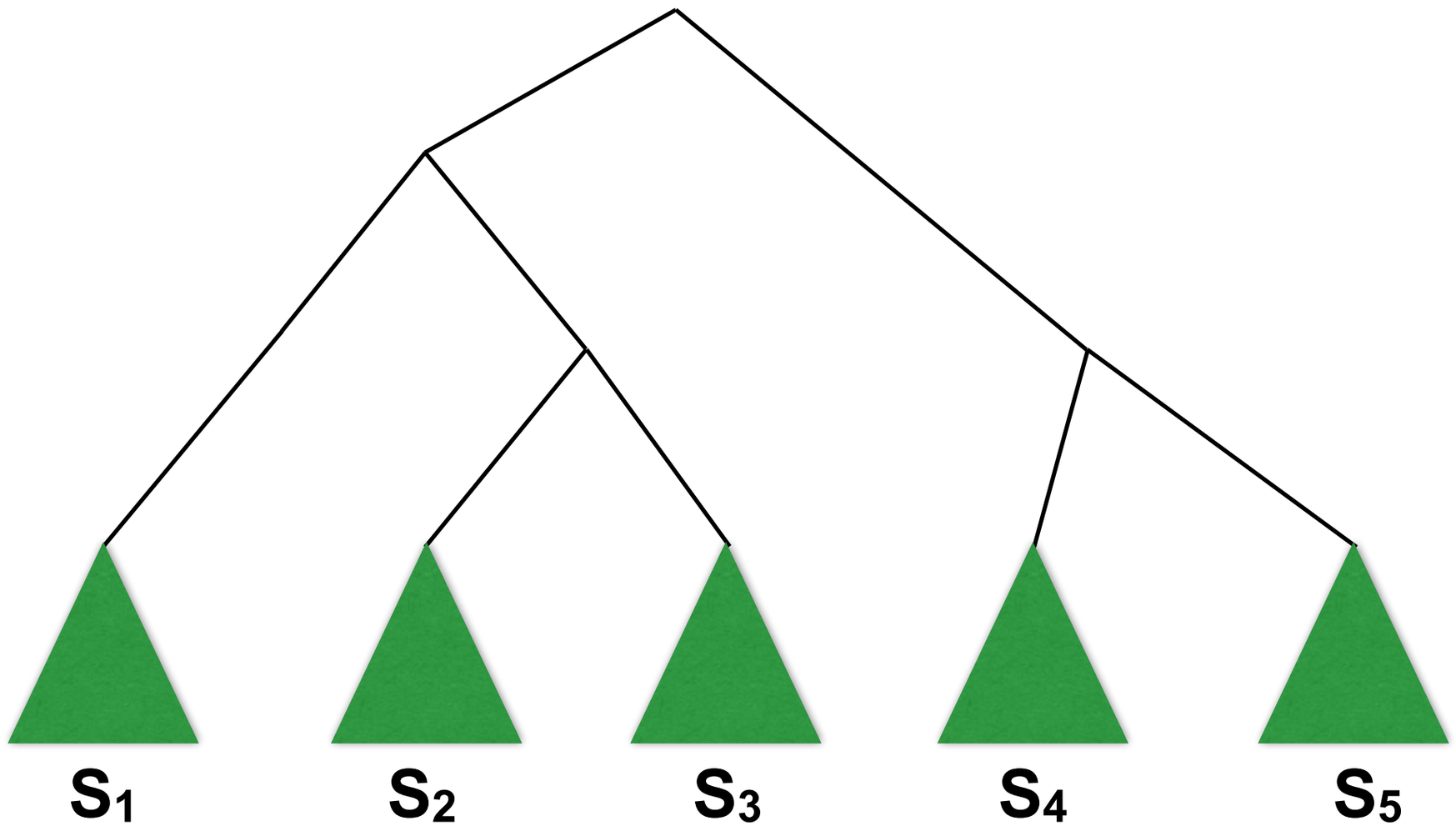
- Measures of selection under complex evolutionary scenarios.

DNA Sequence Evolution

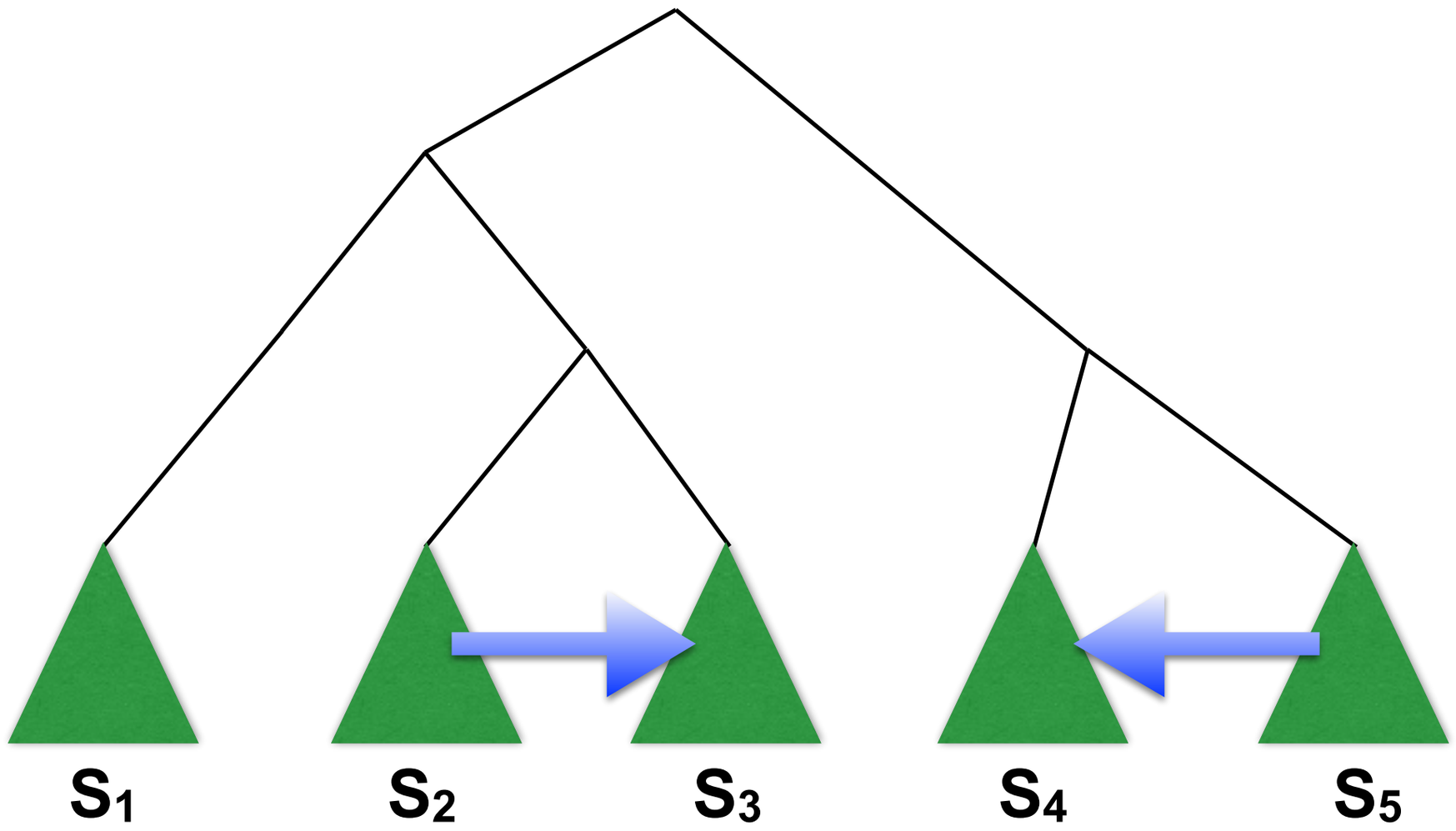
- Walk through calculation on a single edge.

- Then for a three taxon tree.

Future Direction #1

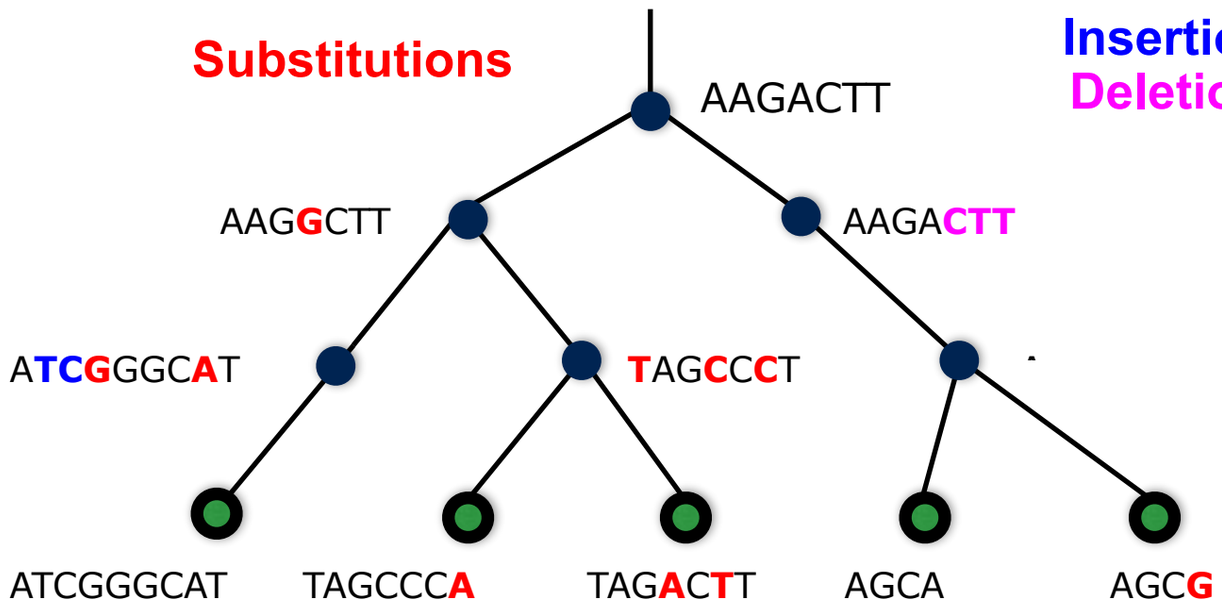


Future Direction #1



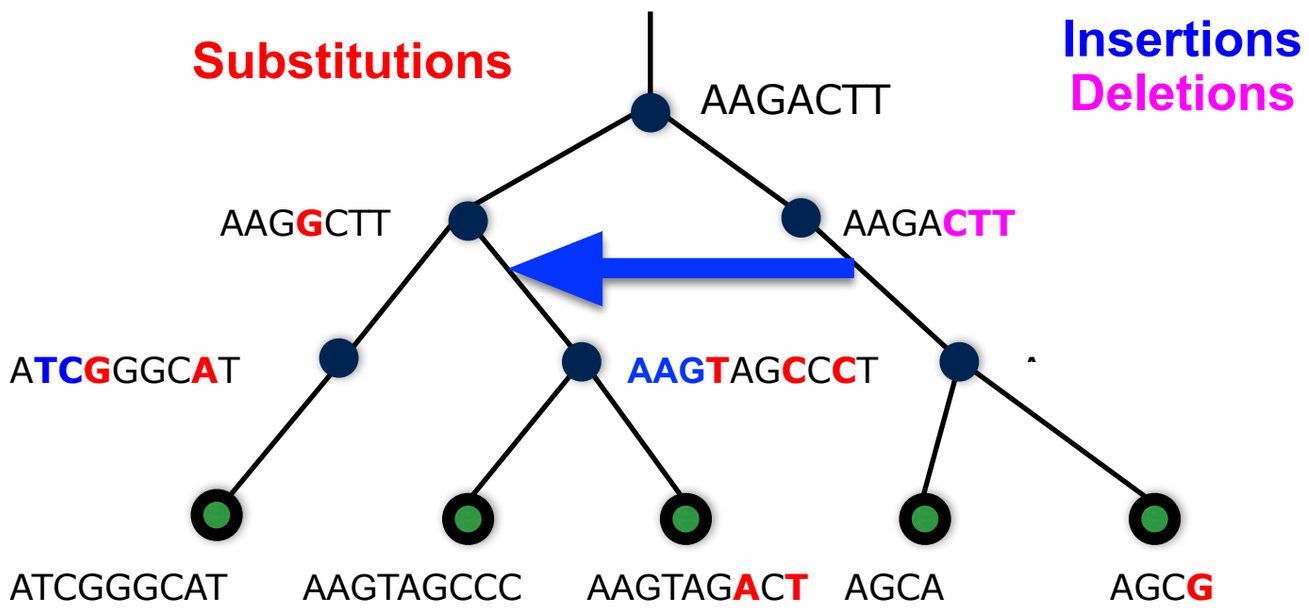
Future Direction #2

- The most widely used multiple sequence alignment methods assume that evolution is tree-like.



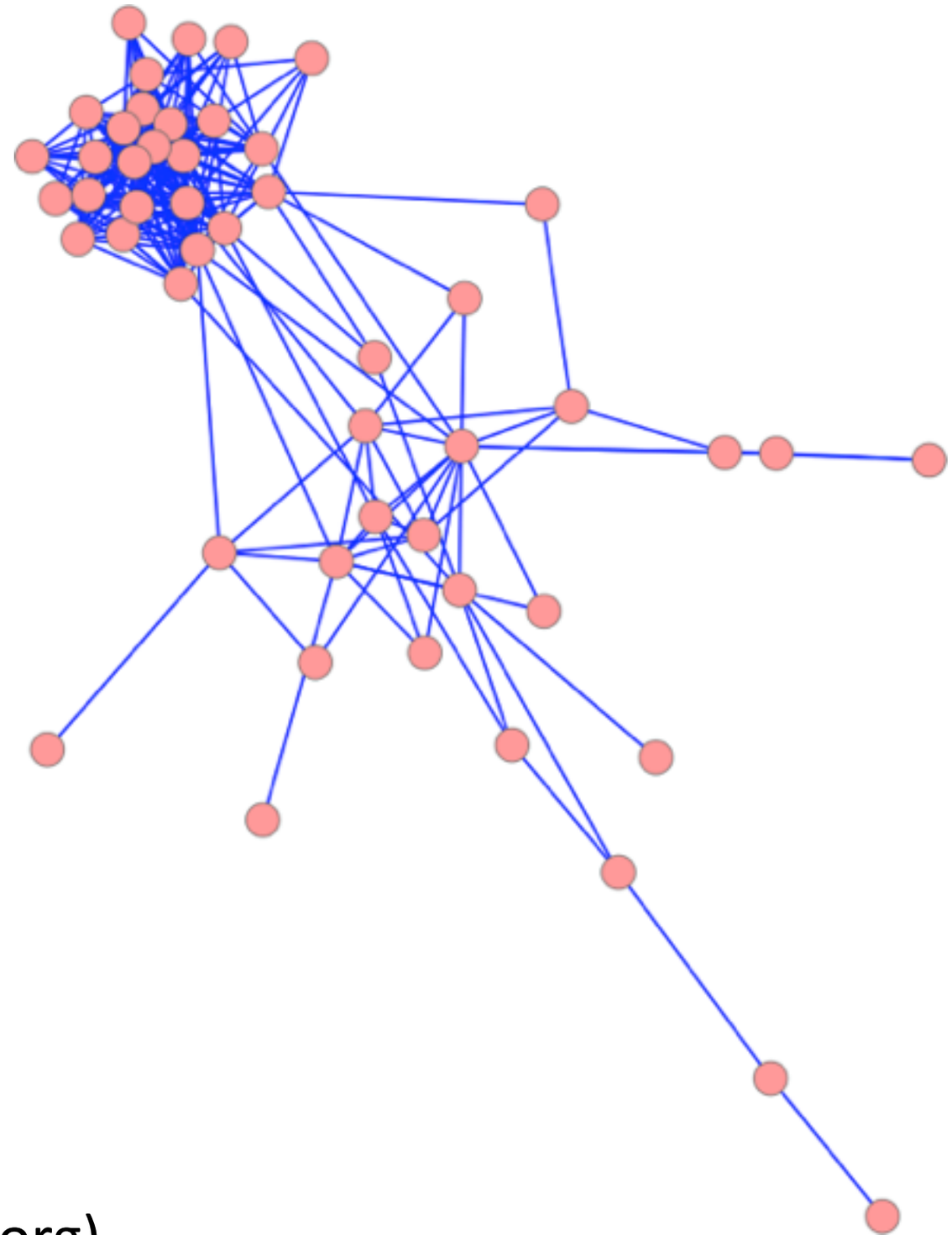
Future Direction #2

- The most widely used multiple sequence alignment methods assume that evolution is tree-like.
- I propose to extend alignment approaches to the case where evolution is not tree-like.



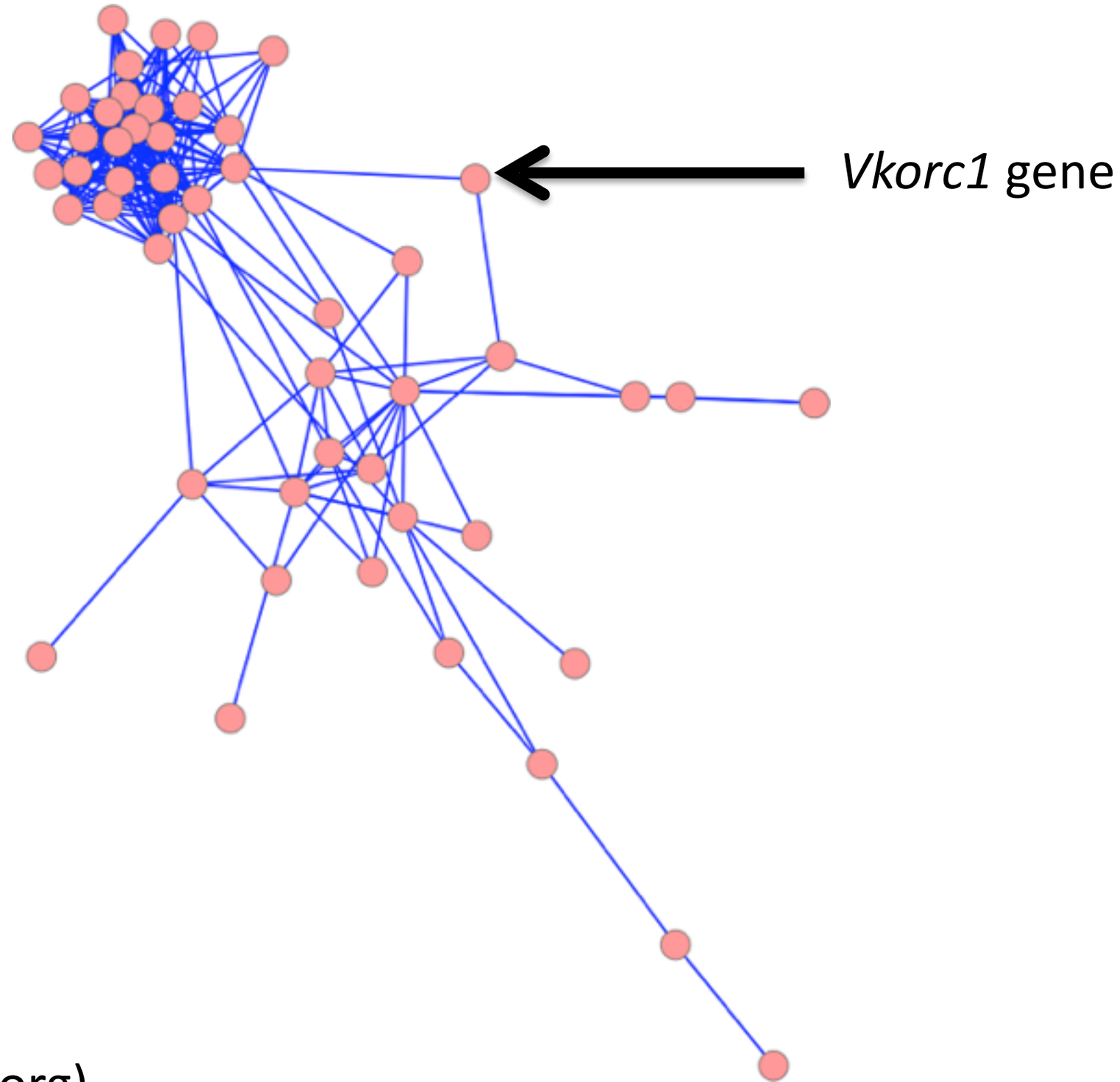
Warfarin-associated Genes with Introgressed Origin

- Each **pink node** is a **gene**.
- Each **blue link** is an **interaction** between a pair of genes.



Visualized using Cytoscape (www.cytoscape.org).

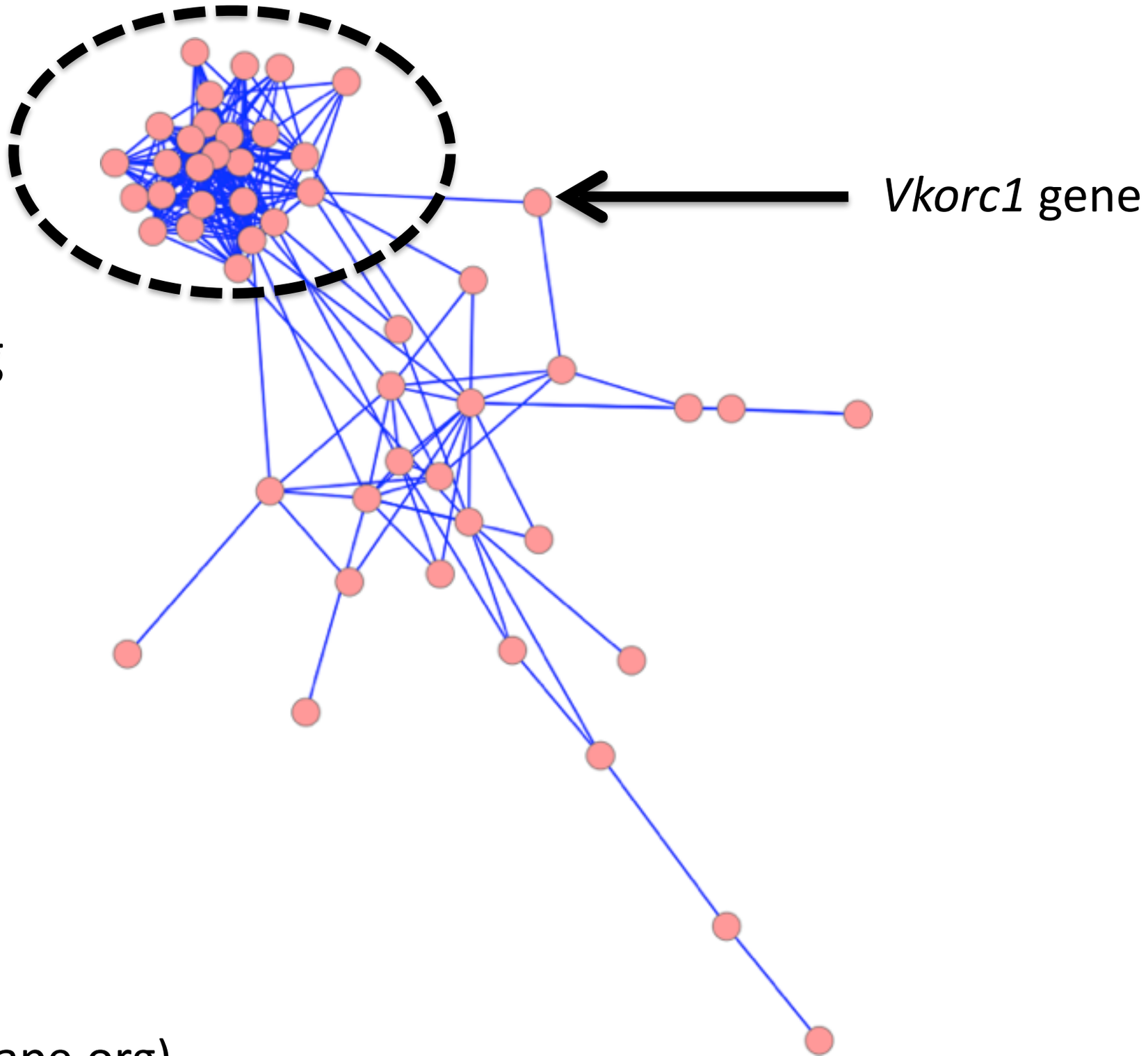
Warfarin-associated Genes with Introgressed Origin



Visualized using
Cytoscape (www.cytoscape.org).

Warfarin-associated Genes with Introgressed Origin = New Potential Targets for Personalized Warfarin Therapy

Cytochrome P450 genes (highly studied, responsible for drug metabolism)

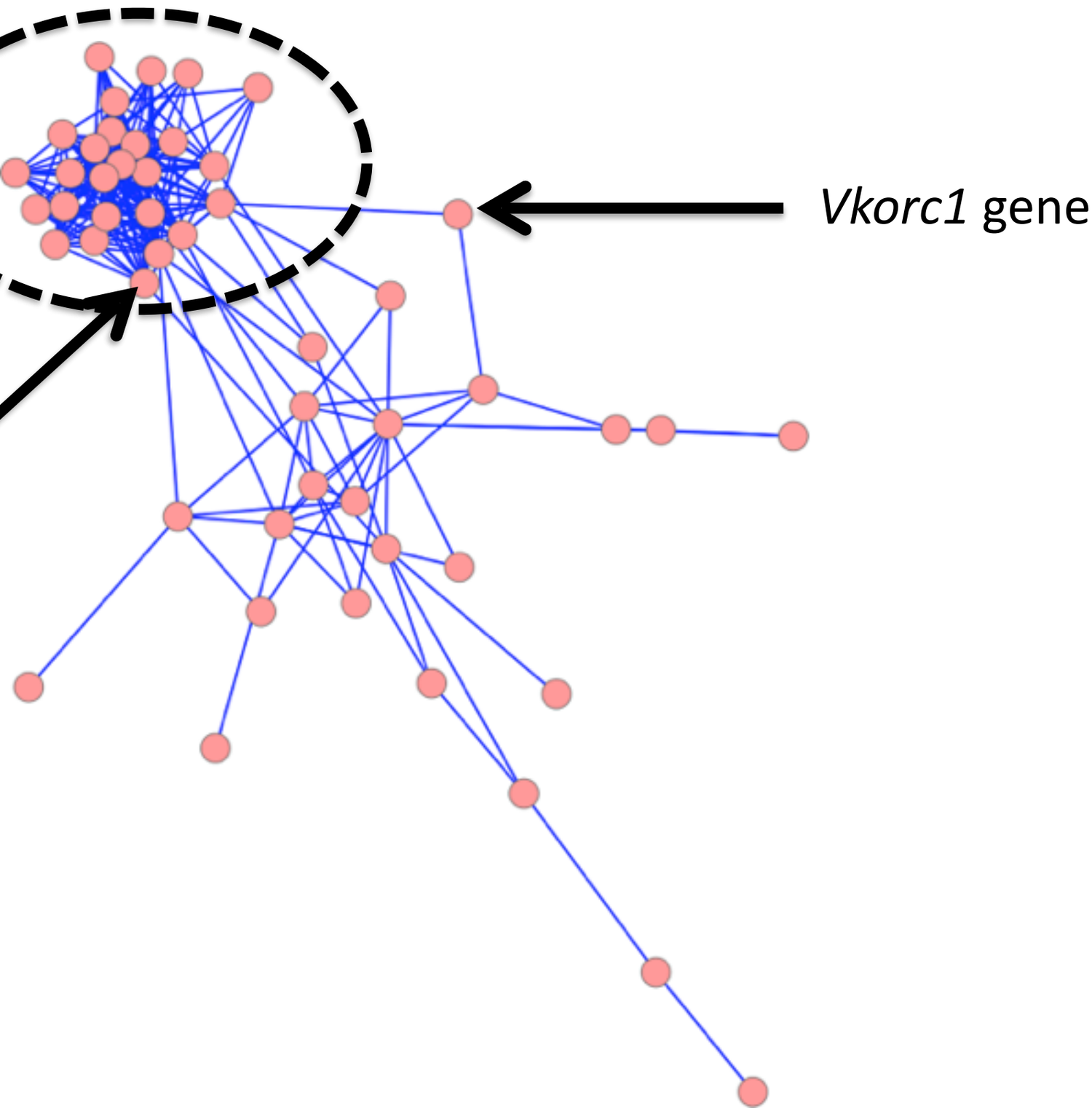


Visualized using Cytoscape (www.cytoscape.org).

Warfarin-associated Genes with Introgressed Origin = New Potential Targets for Personalized Warfarin Therapy

Cytochrome P450 genes (highly studied, responsible for drug metabolism)

Poorly understood "orphan" cytochrome P450 gene



Visualized using Cytoscape (www.cytoscape.org).

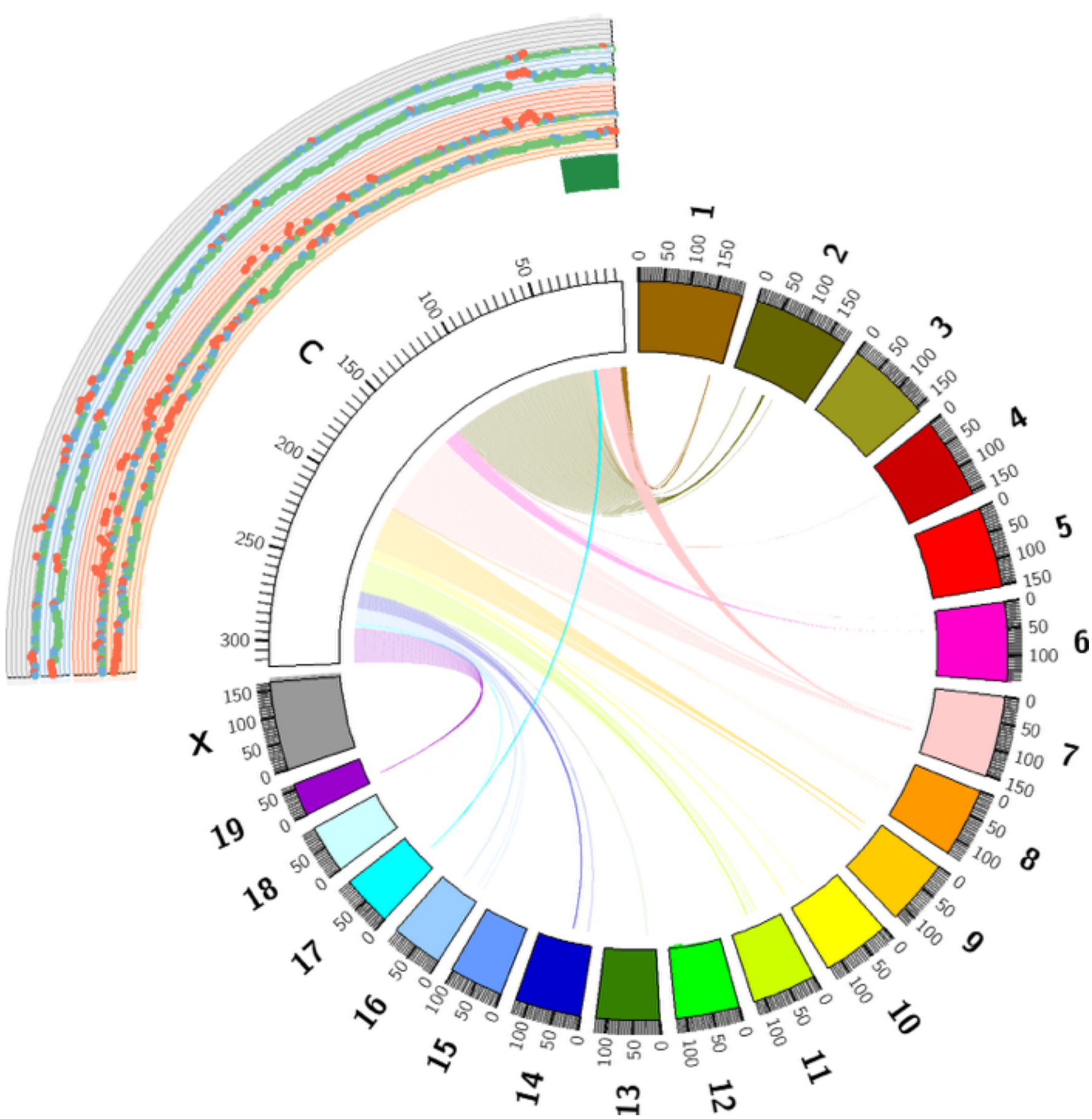
Acknowledgments

- Thanks to my postdoctoral mentors (Luay Nakhleh and Michael H. Kohn), my graduate adviser and co-adviser (Tandy Warnow and C. Randal Linder), and their labs.
- Supported in part by:
  - A training fellowship from the Keck Center of the Gulf Coast Consortia, on Rice University's NLM Training Program in Biomedical Informatics (Grant No. T15LM007093).
  - NLM (Grant No. R01LM00949405 to Luay Nakhleh)
  - NHLBI (Grant No. R01HL09100704 to Michael Kohn)

Introgression of Functional Gene Modules

Liu et al. in prep.

Introgression of a Functional Cluster of Olfactory Receptor-Related Genes



Other

- Computational approaches constitute basic research of interest to NSF (IIS, ABI)
- Wide range of applications of interest to different funding agencies, including:
  - The role of introgression in the spread of pesticide resistance in wild mice, with applications to personalized warfarin therapy (NIH)
  - The role of horizontal gene transfer in the spread of antibiotic resistance in bacteria (NIH)
  - Bacterial genomics (DOE)
  - Hybridization in plants (USDA)

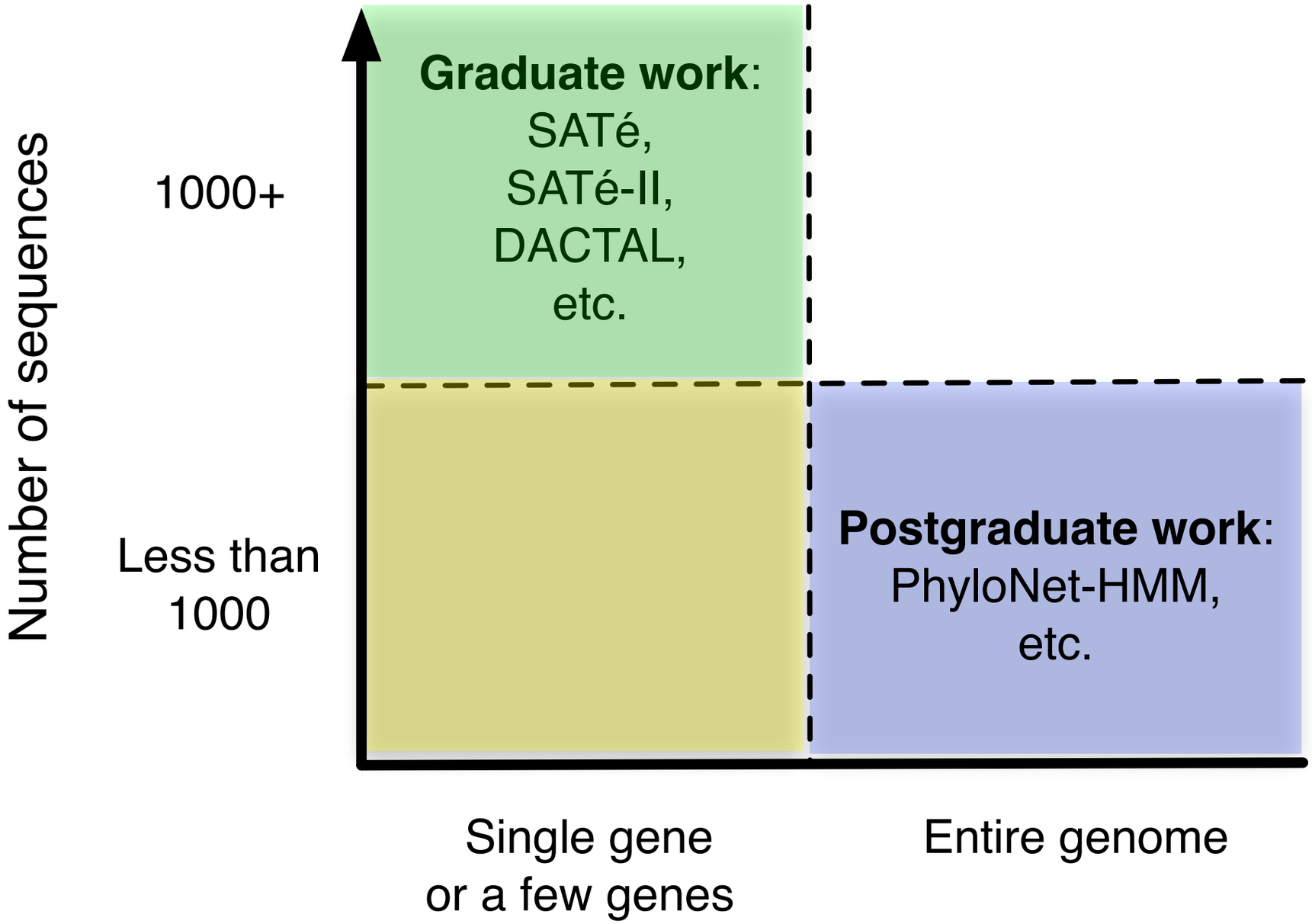
Future Direction #1

- Previous analyses (at most five genomes and a single network edge) required more than a CPU-month on a large cluster

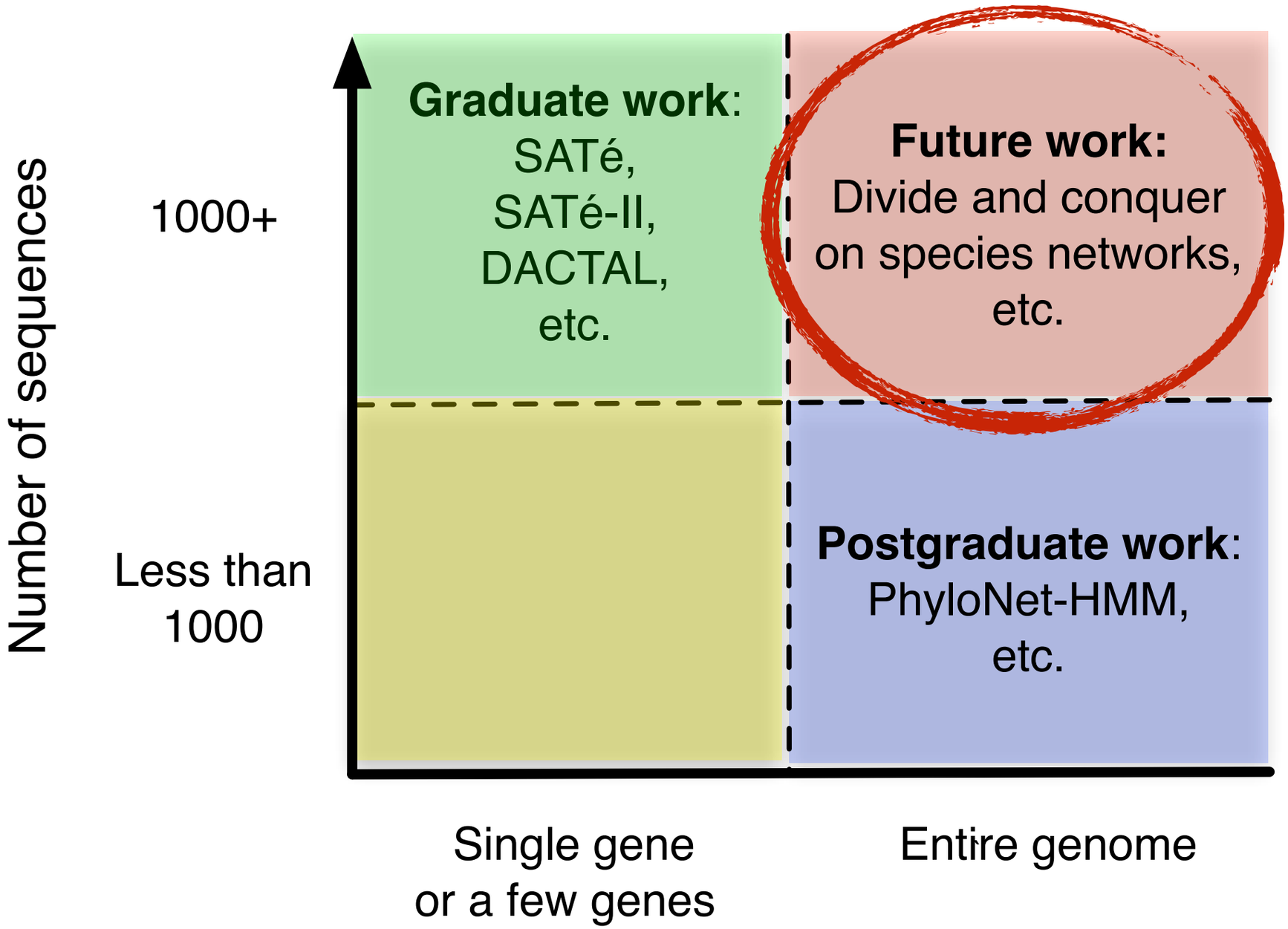
- Problem is combinatorial in both the number of genomes and the number of network edges

- Challenge: efficient and accurate network-based inference from hundreds of genomes or more

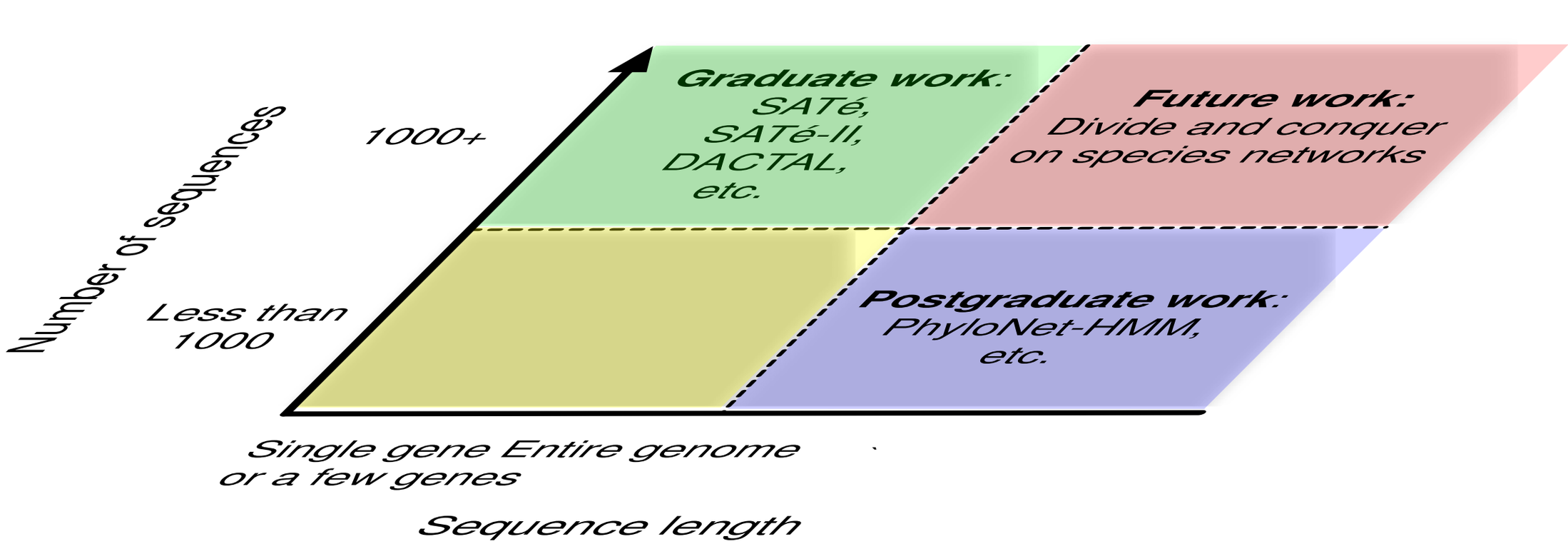
My Contributions



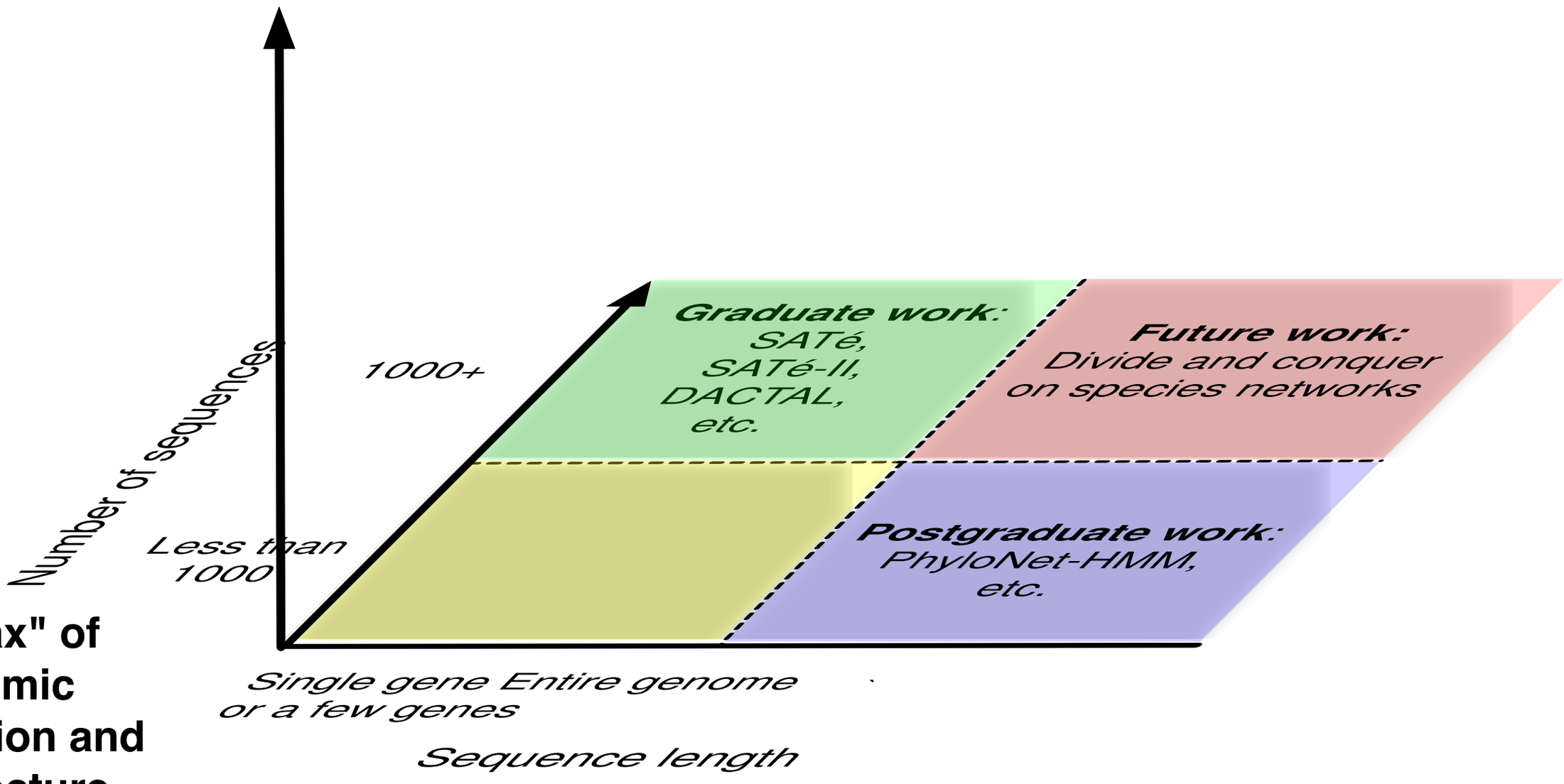
Future Direction #1



Future Direction #2

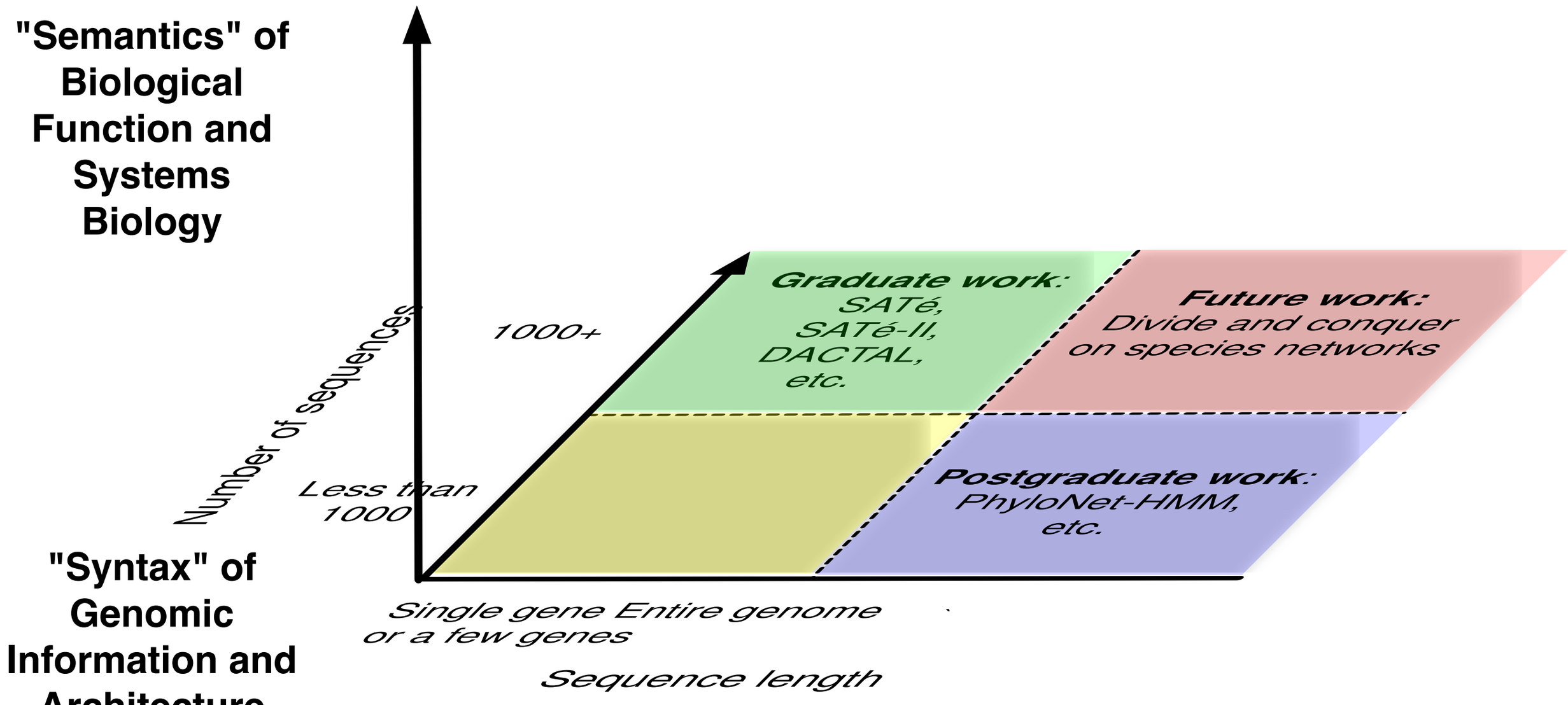


Future Direction #2



"Syntax" of Genomic Information and Architecture

Future Direction #2



Future Direction #2