Detecting Evolutionary Inter-Gene Heterogeneity in *Borrelia burgdorferi*

ELISA LOZA
Department of Mathematical Sciences
University of Bath

Contents

Contents

1. What is a *phylogenetic analysis*?

2. Conventional (homogeneous) model for likelihood-based phylogenetic inference.

Contents

1. What is a *phylogenetic analysis*?
2. Conventional (homogeneous) model for likelihood-based phylogenetic inference.
3. Downsides of the homogeneous model.

Contents

1. What is a *phylogenetic analysis*?
2. Conventional (homogeneous) model for likelihood-based phylogenetic inference.
3. Downsides of the homogeneous model.
4. An improved model that accounts for heterogeneity.

Contents

1. What is a *phylogenetic analysis*?
2. Conventional (homogeneous) model for likelihood-based phylogenetic inference.
3. Downsides of the homogeneous model.
4. An improved model that accounts for heterogeneity.
5. Applications to *Borrelia burgdorferi* data.

Phylogenetic likelihood methods

- Phylogenetics is the reconstruction and analysis of trees and other parameters to describe and understand the evolution of organisms.

Phylogenetic likelihood methods

- Phylogenetics is the reconstruction and analysis of trees and other parameters to describe and understand the evolution of organisms.

- Likelihood-based phylogenetic analyses start by observing the aligned DNA sequences of *s* organisms:

TCAAGCTATACCCGAT…

TATACCAGCTATAGCT…

CAAAGCTATACCCGAT…

CAAAGCTATACCCGAT…

⋮

The homogeneous model

T C AAGCTATACCCGAT…GC T
T A TACCAGCTATAGCT…GC A
C A AAGCTATACCCGAT…CA A
C A AAGCTATACCCGAT…CC T

- The homogeneous model for independent observations $y_1 = (T,T,C,C)'$, $y_2 = (C,A,A,A)'$, …, $y_n = (T,A,A,T)'$, is:

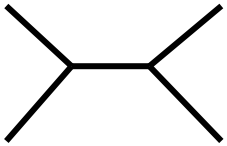
$$y_i \sim f(\cdot \mid \mathcal{X}, \mathbf{t}, Q) \quad \text{independently for } i = 1, 2, \ldots, n$$

Model parameters

$y_i \sim f(\cdot \mid \mathcal{X}, \mathbf{t}, Q)$ independently for $i = 1, 2, \ldots, n$

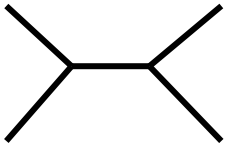
Model parameters

$y_i \sim f(\cdot \mid \mathcal{Y}, \mathbf{t}, Q)$ independently for $i = 1, 2, \ldots, n$

- A bifurcating tree with s leaves, 

Model parameters

$y_i \sim f(\cdot \mid \mathcal{Y}, \mathbf{t}, Q)$ independently for $i = 1, 2, \ldots, n$

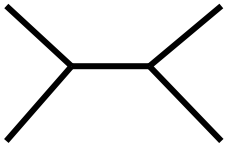
- A bifurcating tree with *s* leaves, 

- A set of positive real-valued branch lengths,

$$\mathbf{t} = (t_1, t_2, \ldots, t_5)$$

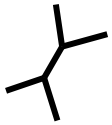
Model parameters

$y_i \sim f(\cdot \mid \mathcal{Y}, \mathbf{t}, Q)$ independently for $i = 1, 2, \ldots, n$

- A bifurcating tree with *s* leaves, 

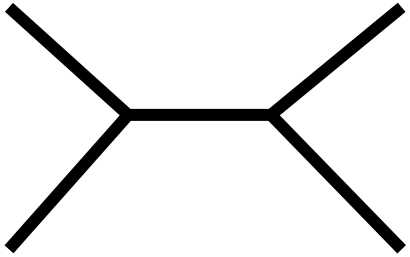
- A set of positive real-valued branch lengths,

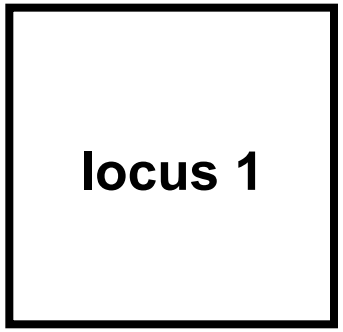
$$\mathbf{t} = (t_1, t_2, \ldots, t_5)$$

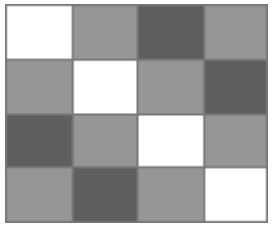
- A rate matrix Q specifying a Markov process of character substitution along 

	$r_{AC}\,\pi_C$	$r_{AG}\,\pi_G$	$r_{AT}\,\pi_T$
$r_{AC}\,\pi_A$		$r_{CG}\,\pi_G$	$r_{CT}\,\pi_T$
$r_{AG}\,\pi_A$	$r_{CG}\,\pi_C$		$r_{GT}\,\pi_T$
$r_{AT}\,\pi_A$	$r_{CT}\,\pi_C$	$r_{GT}\,\pi_G$	

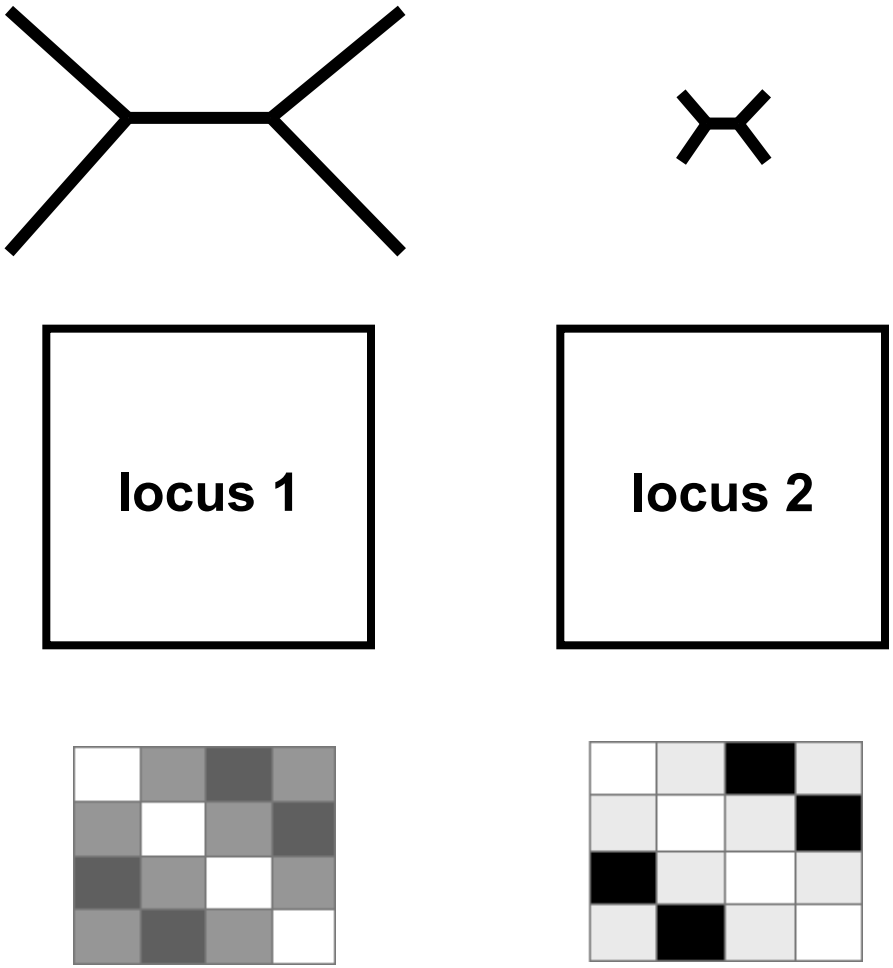
DNA data may be not homogeneous



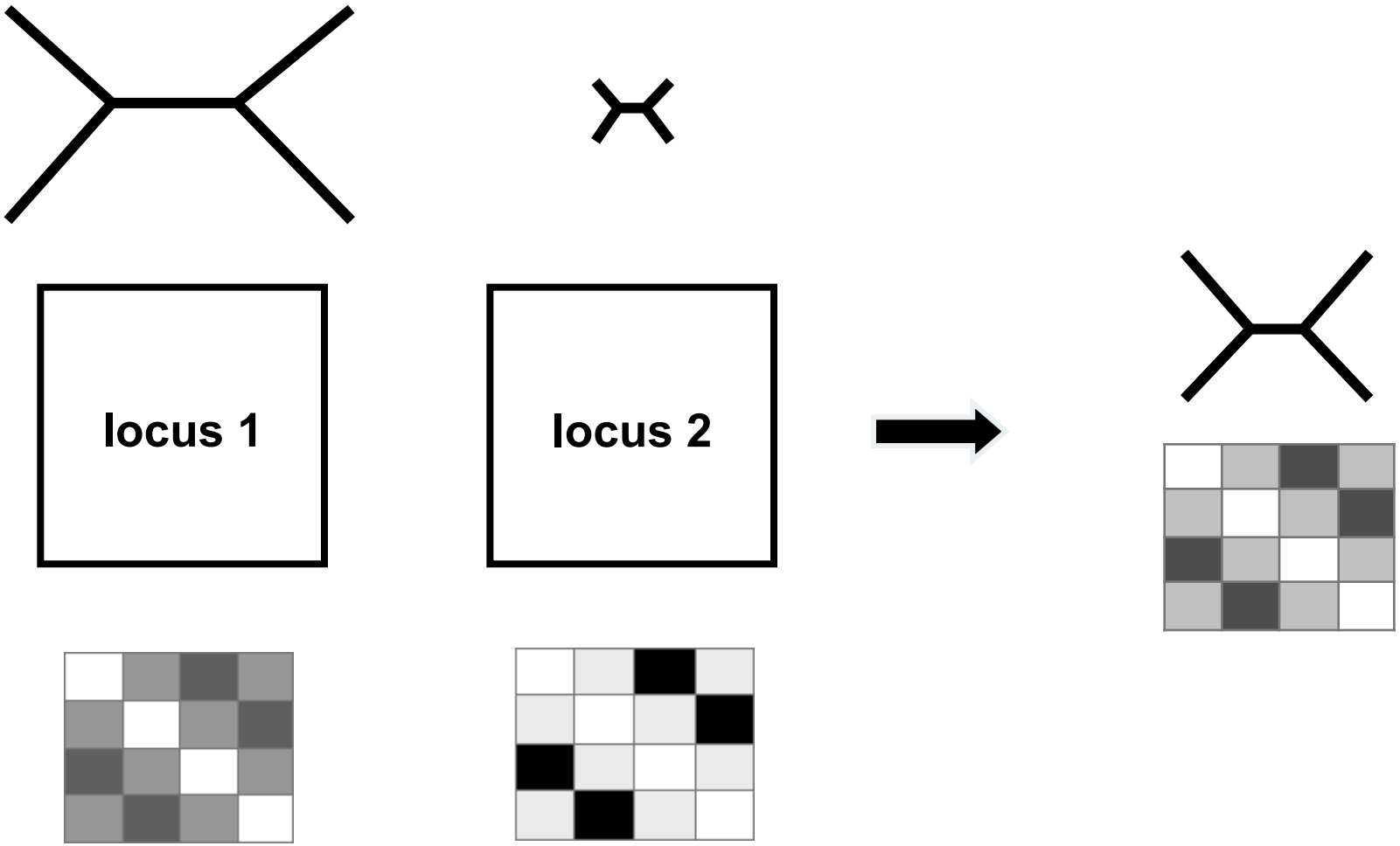




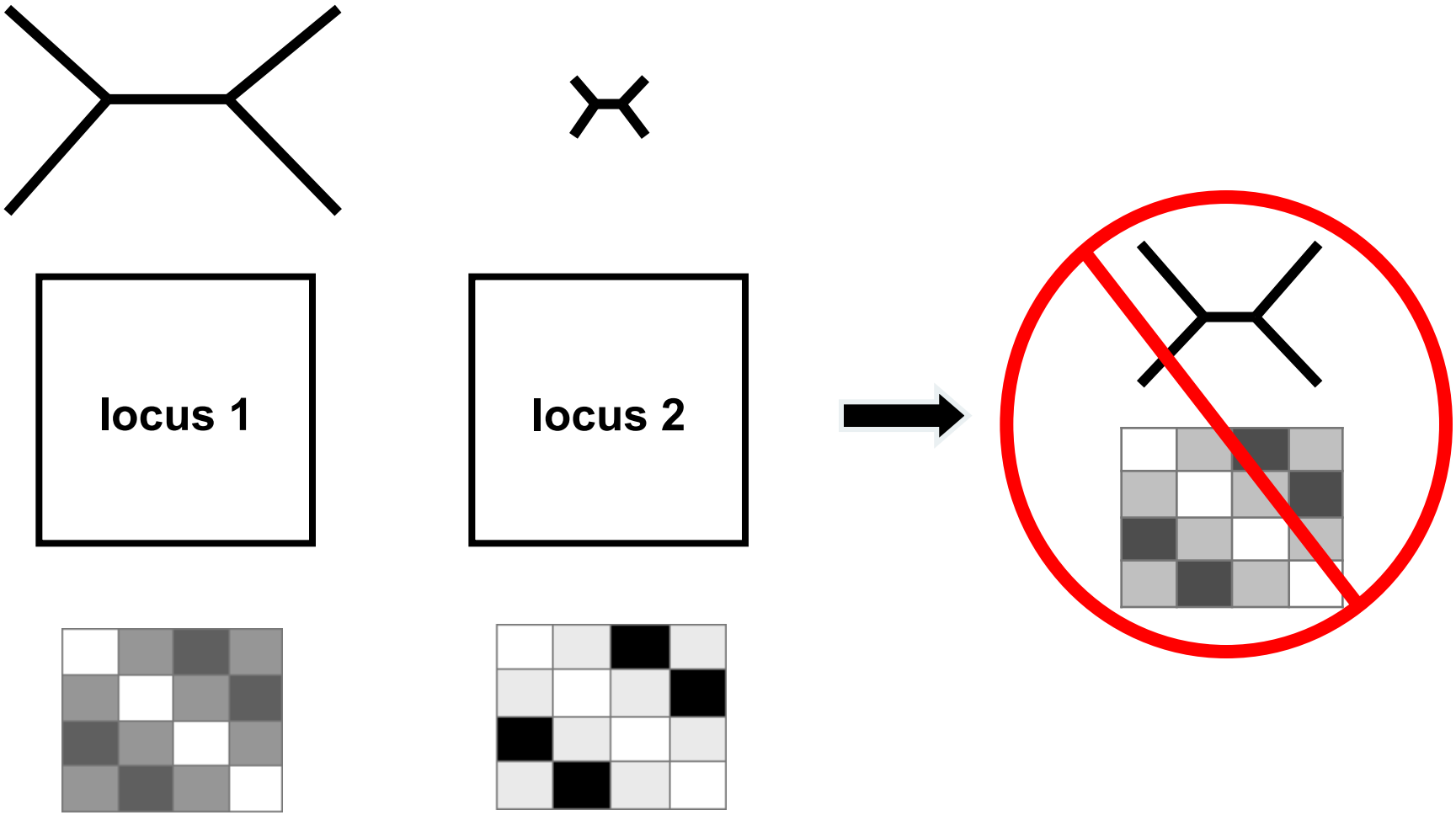
DNA data may be not homogeneous



DNA data may be not homogeneous



DNA data may be not homogeneous



Borrelia burgdorferi

- *Borrelia burgdorferi* is one of the bacterial species responsible for Lyme disease.

Borrelia burgdorferi

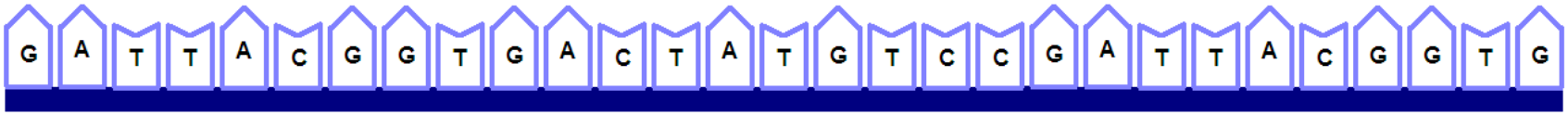
- *Borrelia burgdorferi* is one of the bacterial species responsible for Lyme disease.

- To fully understand the disease, it is crucial to unveil the evolutionary properties of its genetic variants (strains).

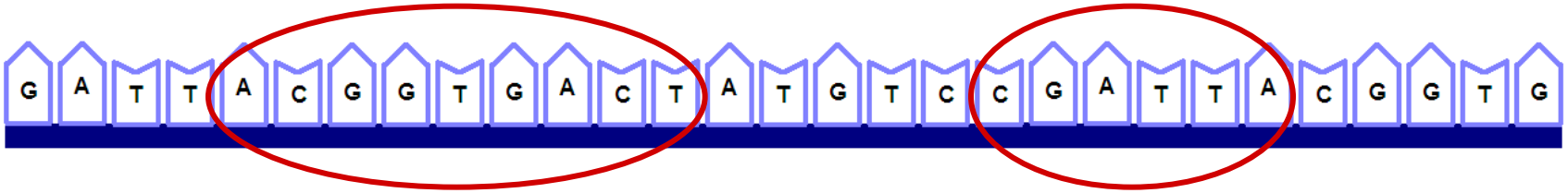
Borrelia burgdorferi

- *Borrelia burgdorferi* is one of the bacterial species responsible for Lyme disease.
- To fully understand the disease, it is crucial to unveil the evolutionary properties of its genetic variants (strains).
- Phylogenetic analysis is an essential tool.

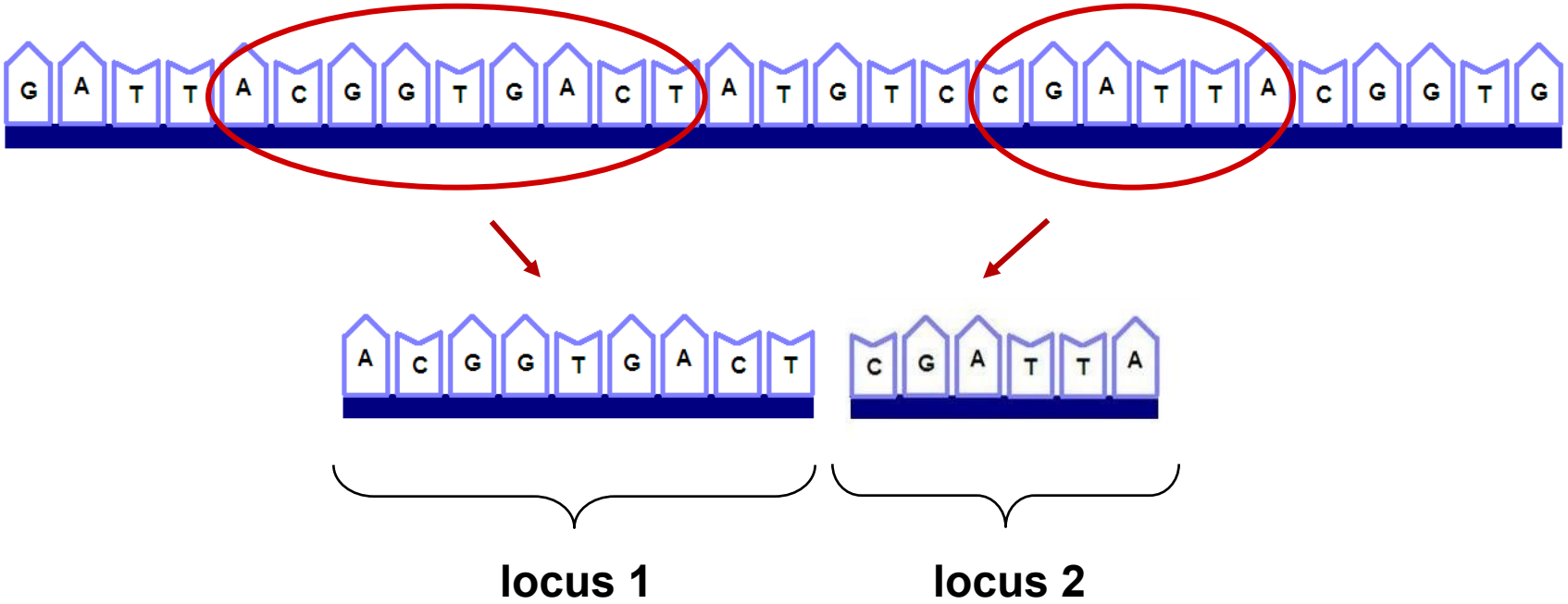
Identification of *B. burgdorferi* strains



Identification of *B. burgdorferi* strains



Identification of *B. burgdorferi* strains



Are the loci congruent in evolution, such that valid inferences can be made under a homogeneous phylogenetic model?

The Q + *t* mixture model

- Finite mixture models provide a natural way to model heterogeneous data.

$$f(\cdot \mid \chi, t, Q)$$

The Q + *t* mixture model

- Finite mixture models provide a natural way to model heterogeneous data.

$$f(\cdot \mid \chi, t, Q) \quad + \quad f(\cdot \mid \chi, t, Q)$$

The Q + *t* mixture model

- Finite mixture models provide a natural way to model heterogeneous data.

$$f(\cdot \mid \chi, t, Q) + f(\cdot \mid \chi, t, Q) + \ldots + f(\cdot \mid \chi, t, Q)$$

The Q + *t* mixture model

- Finite mixture models provide a natural way to model heterogeneous data.

$$\mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \ldots + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q})$$

for $\mathbf{w} + \mathbf{w} + \ldots + \mathbf{w} = 1$

The Q + *t* mixture model

- Finite mixture models provide a natural way to model heterogeneous data.

$$y_i \sim \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \ldots + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q})$$

for $\mathbf{w} + \mathbf{w} + \ldots + \mathbf{w} = 1$

and ind. for $i = 1, 2, \ldots, n$

A *branch-length* mixture model

$$y_i \sim \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \ldots + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q})$$

for $\mathbf{w} + \mathbf{w} + \ldots + \mathbf{w} = 1$

and ind. for $i = 1, 2, \ldots, n$

A *branch-length* mixture model

$$y_i \sim \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \ldots + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q})$$

for $\mathbf{w} + \mathbf{w} + \ldots + \mathbf{w} = 1$

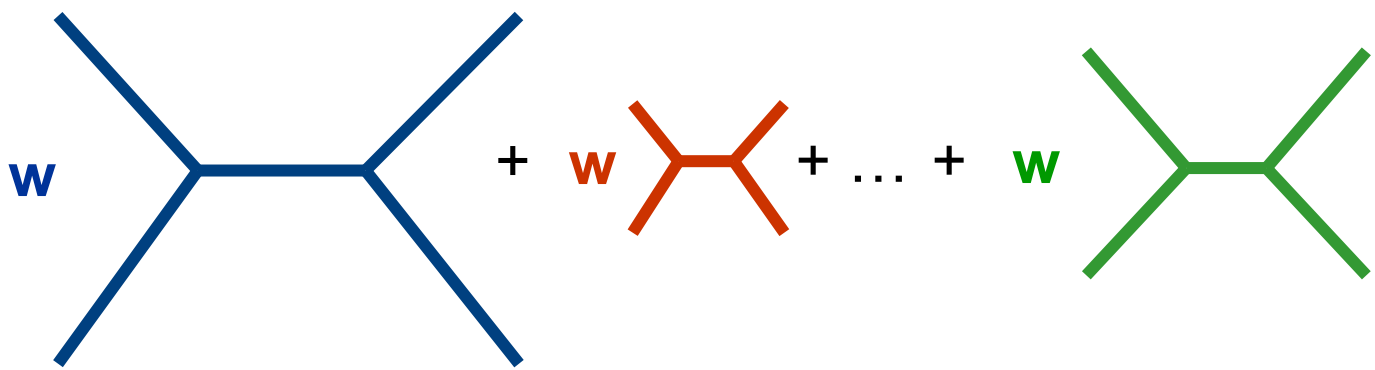
and ind. for $i = 1, 2, \ldots, n$

A *branch-length* mixture model

$$y_i \sim \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \ldots + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q})$$

for $\mathbf{w} + \mathbf{w} + \ldots + \mathbf{w} = 1$

and ind. for $i = 1, 2, \ldots, n$



A *branch-length* mixture model

$$y_i \sim \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \ldots + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q})$$

for $\mathbf{w} + \mathbf{w} + \ldots + \mathbf{w} = 1$

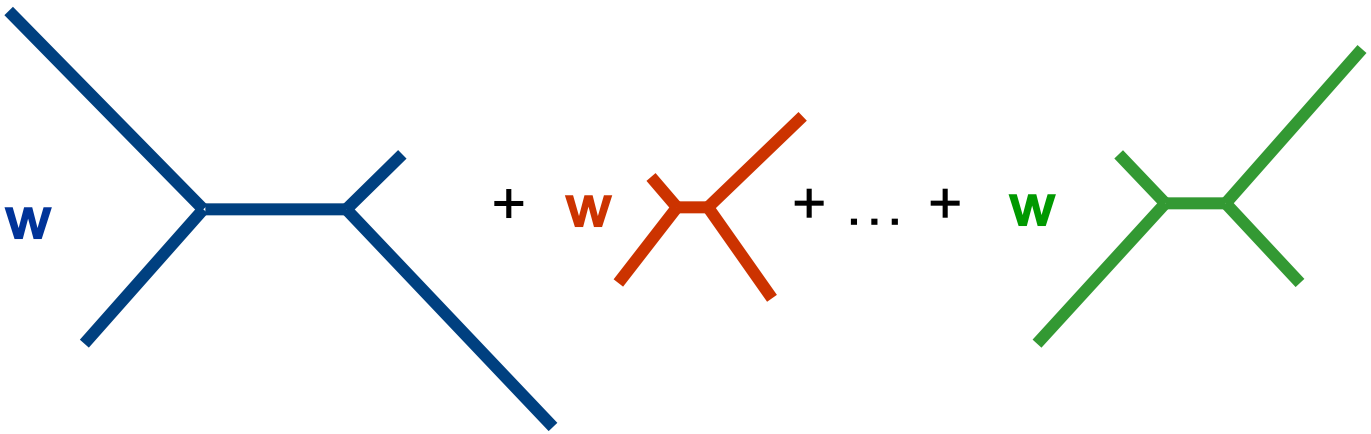
and ind. for $i = 1, 2, \ldots, n$

A *branch-length* mixture model

$$y_i \sim \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q}) + \ldots + \mathbf{w}\, f(\cdot \mid \chi, \mathbf{t}, \mathbf{Q})$$

for $\mathbf{w} + \mathbf{w} + \ldots + \mathbf{w} = 1$

and ind. for $i = 1, 2, \ldots, n$

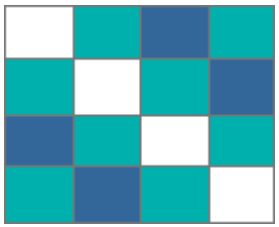
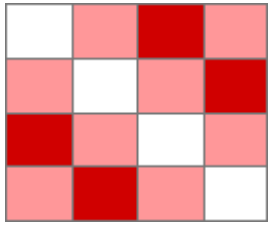
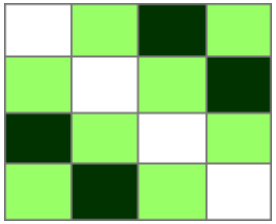


The Q + *t* mixture model

$$y_i \sim \mathbf{w}\, f(\cdot \mid \chi, t, Q) + \mathbf{w}\, f(\cdot \mid \chi, t, Q) + \ldots + \mathbf{w}\, f(\cdot \mid \chi, t, Q)$$

for $\mathbf{w} + \mathbf{w} + \ldots + \mathbf{w} = 1$

and ind. for $i = 1, 2, \ldots, n$

w  + **w**  + … + **w** 

The Q + *t* mixture model

- A $\mathbf{label}_i$ identifies the specific process from which the *i*-th site is generated.

The Q + *t* mixture model

- A $\mathbf{label}_i$ identifies the specific process from which the *i*-th site is generated.

p ($\mathbf{label}_i$ = □) = ⓪

independently for $i = 1, 2, \ldots, n$

The Q + *t* mixture model

- A **label**$_i$ identifies the specific process from which the *i*-th site is generated.

$$p\,(\,\mathbf{label}_i = \blacksquare\,) = \omega$$

independently for $i = 1, 2, \ldots, n$

The Q + *t* mixture model

- A **label**$_i$ identifies the specific process from which the *i*-th site is generated.

p (**label**$_i$ = ■) = ω

independently for $i = 1, 2, \ldots, n$

The Q + *t* mixture model

- A **label**$_i$ identifies the specific process from which the *i*-th site is generated.

p (**label**$_i$ = □) = ⓪ for □ = ■ , ■ , … , ■

independently for i = *1, 2, …, n*

The Q + *t* mixture model

- Once the **label**$_i$ for site *i* is known,

$$y_i \mid \square \sim f(\cdot \mid \lambda, t, Q)$$

independently for $i = 1, 2, \ldots, n$

The Q + *t* mixture model

- Once the **label**$_i$ for site *i* is known,

$$y_i \mid \blacksquare \sim f(\cdot \mid \chi, t \quad Q)$$

independently for $i = 1, 2, \ldots, n$

The Q + *t* mixture model

- Once the **label**$_i$ for site *i* is known,

$$y_i \mid \blacksquare \sim f(\cdot \mid \chi, t, Q)$$

independently for $i = 1, 2, \ldots, n$

The Q + *t* mixture model: an example

- Consider a DNA alignment:

The Q + *t* mixture model: an example

- Consider a DNA alignment:

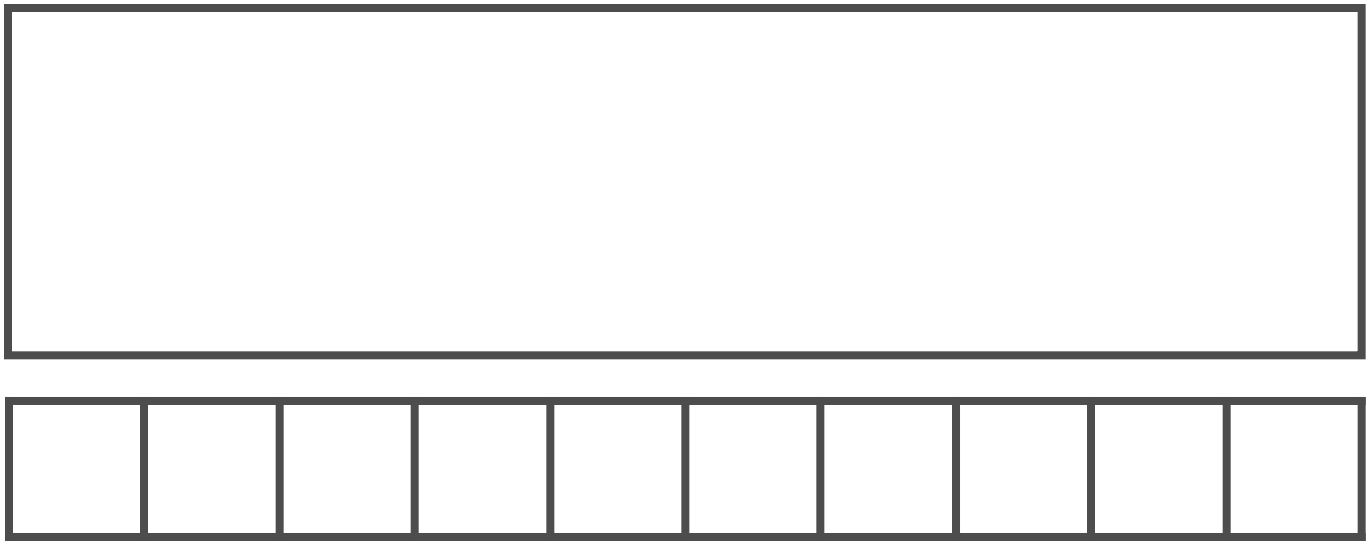
- Sites are modelled by:

$$y_i \sim \mathbf{w\, f(\cdot \mid \chi, t, Q)} + \mathbf{w\, f(\cdot \mid \chi, t, Q)}$$

independently for $i = 1, 2, \ldots, n$

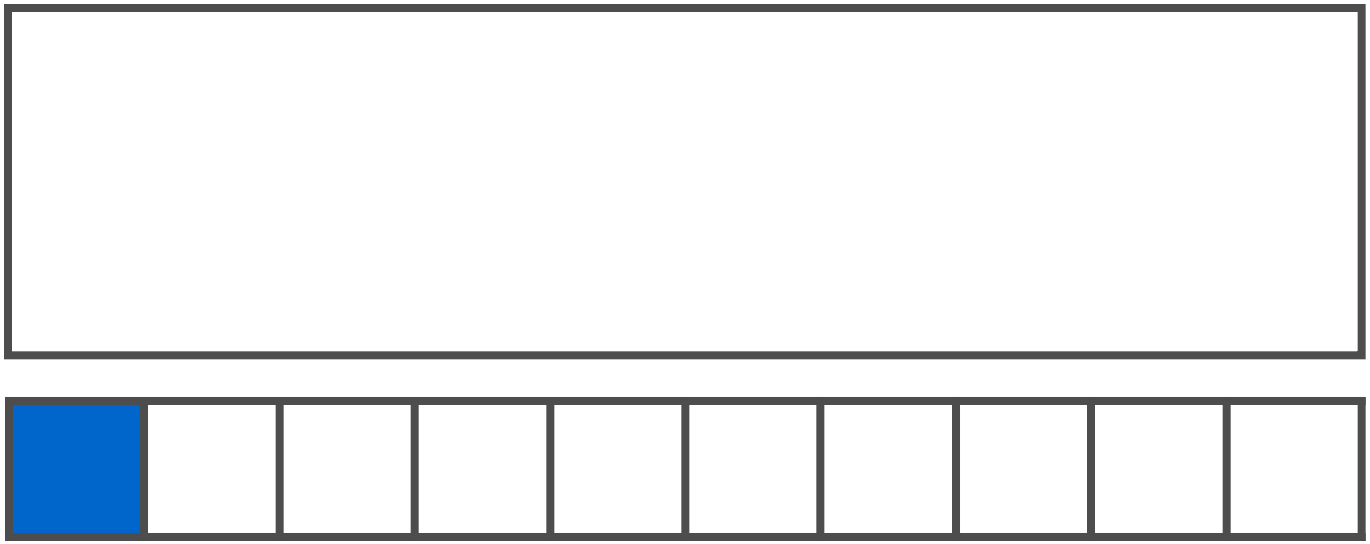
The Q + *t* mixture model: an example

- Consider a DNA alignment:



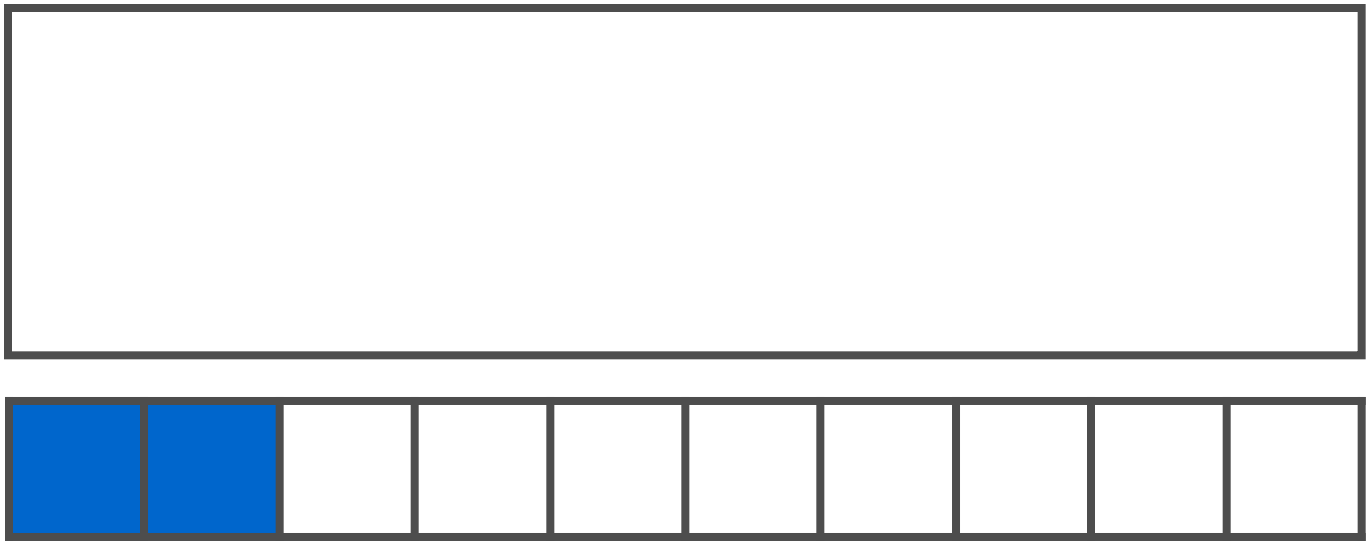
The Q + *t* mixture model: an example

- Consider a DNA alignment:



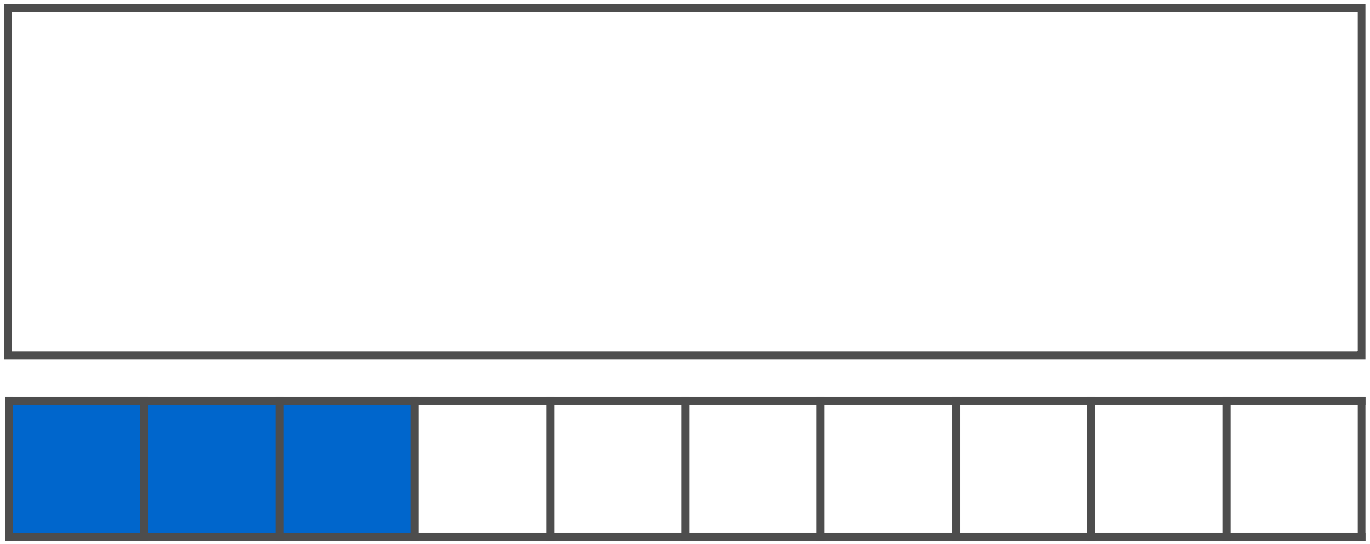
The Q + *t* mixture model: an example

- Consider a DNA alignment:



The Q + *t* mixture model: an example

- Consider a DNA alignment:

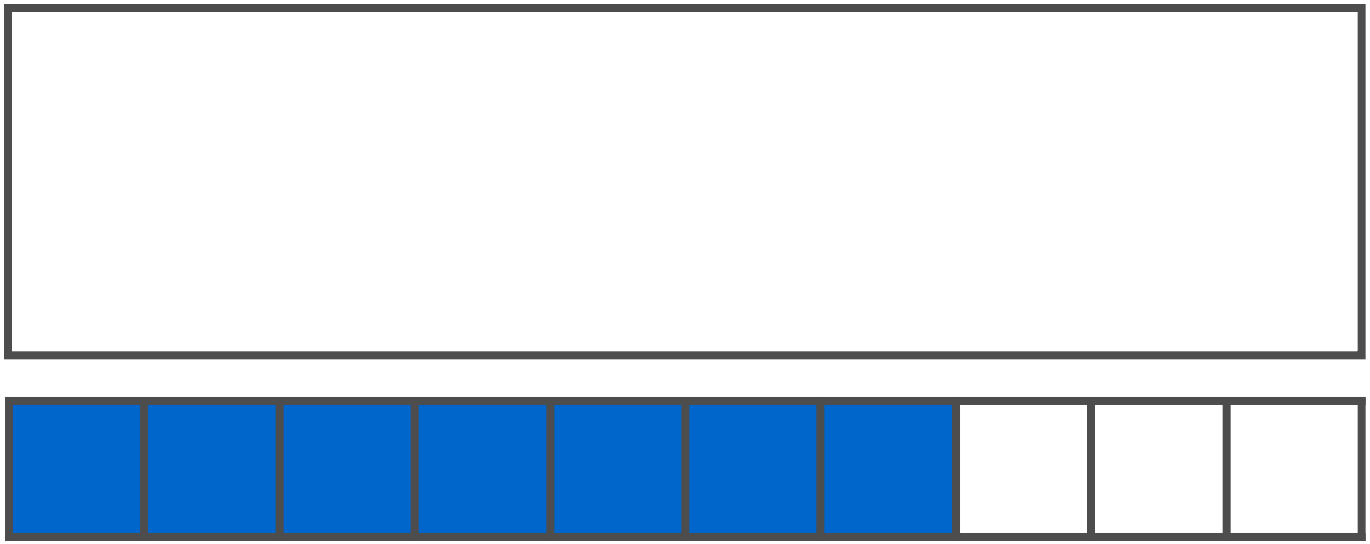


The Q + *t* mixture model: an example

- Consider a DNA alignment:

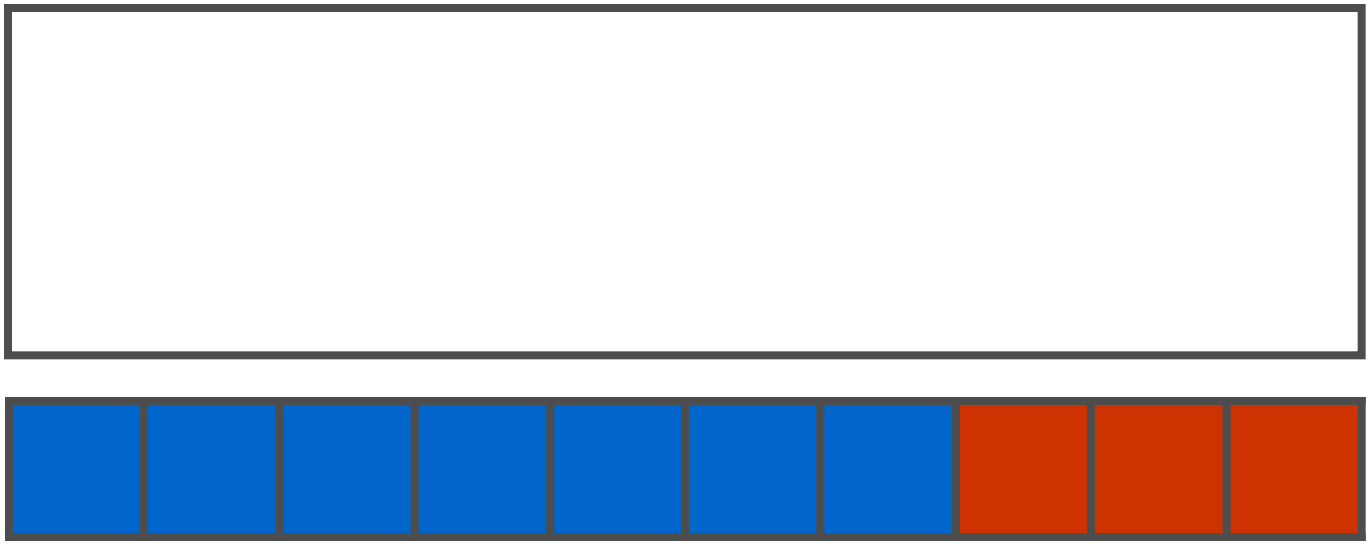
The Q + *t* mixture model: an example

- Consider a DNA alignment:



The Q + *t* mixture model: an example

- Consider a DNA alignment:



The Q + *t* mixture model: an example

$y_i \mid \blacksquare \sim f(\cdot \mid \lambda, t, Q)$

independently for $i = 1, 2, \ldots, m$

The Q + *t* mixture model: an example

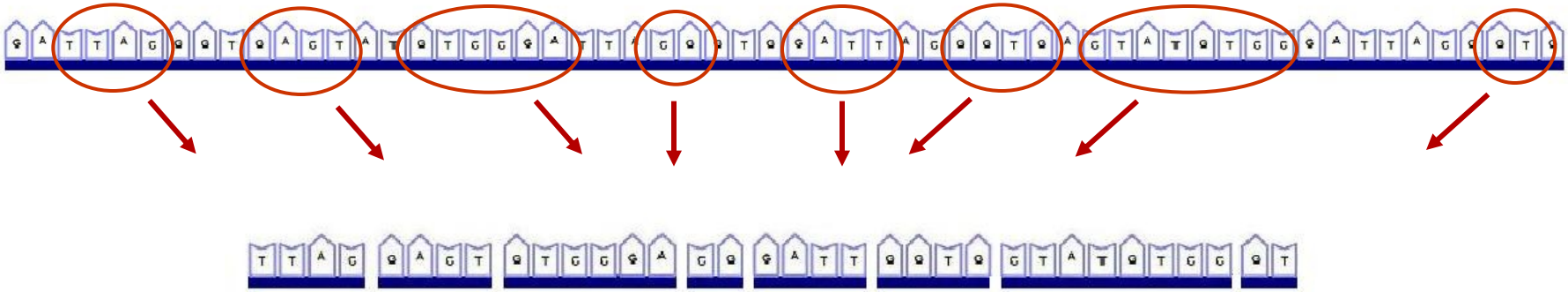
$y_i \mid \blacksquare \sim f(\cdot \mid \lambda, t, Q)$

independently for $i = m+1, \ldots, n$

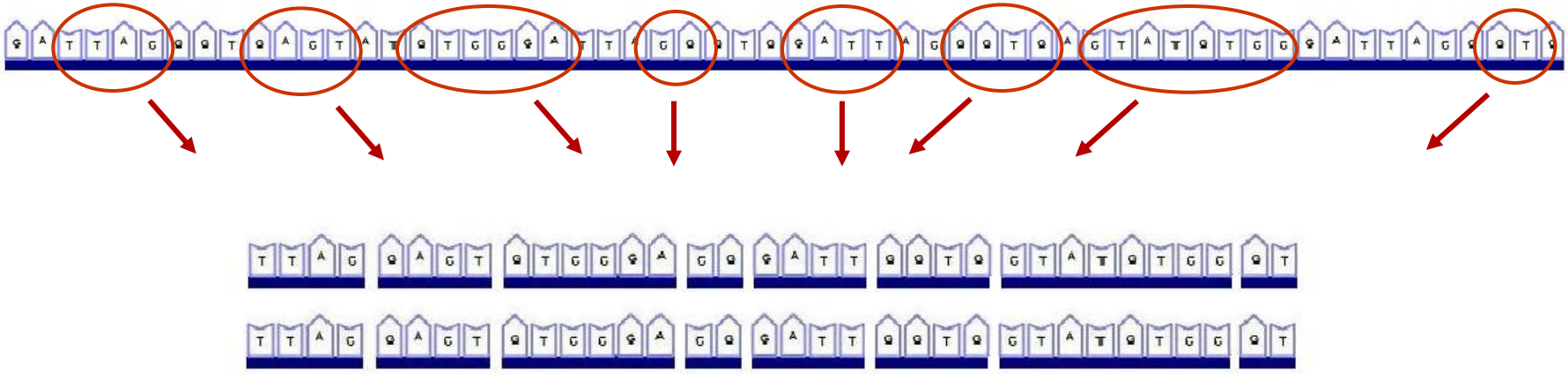
Analysis of *B. burgdorferi*: the 'housekeeping genes' alignment

Analysis of *B. burgdorferi*: the 'housekeeping genes' alignment

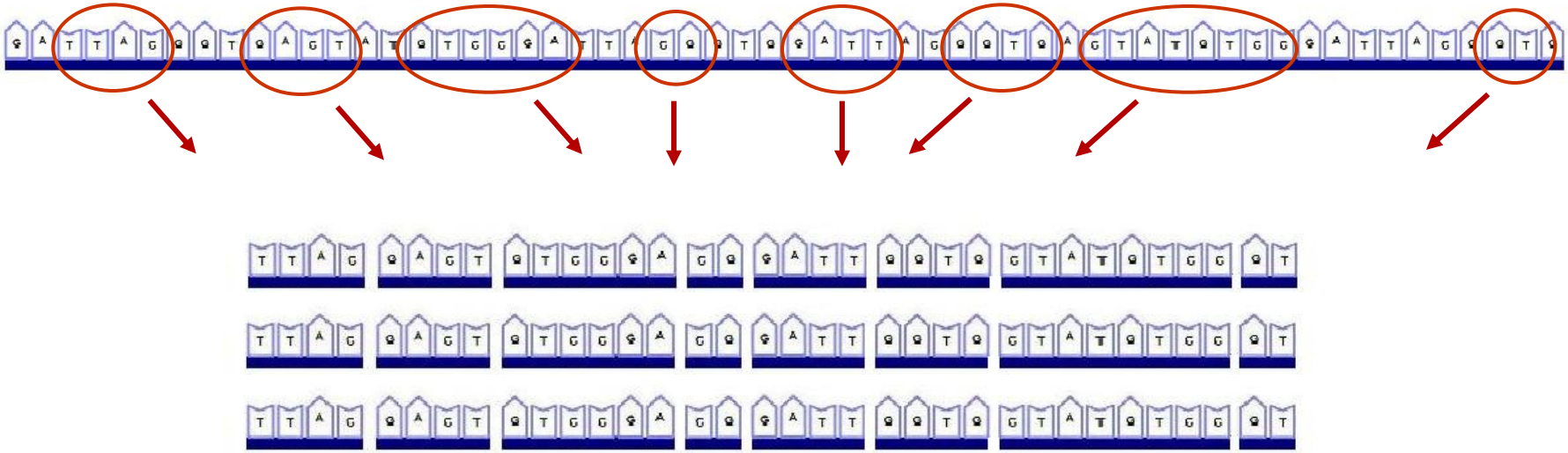
Analysis of *B. burgdorferi*: the 'housekeeping genes' alignment



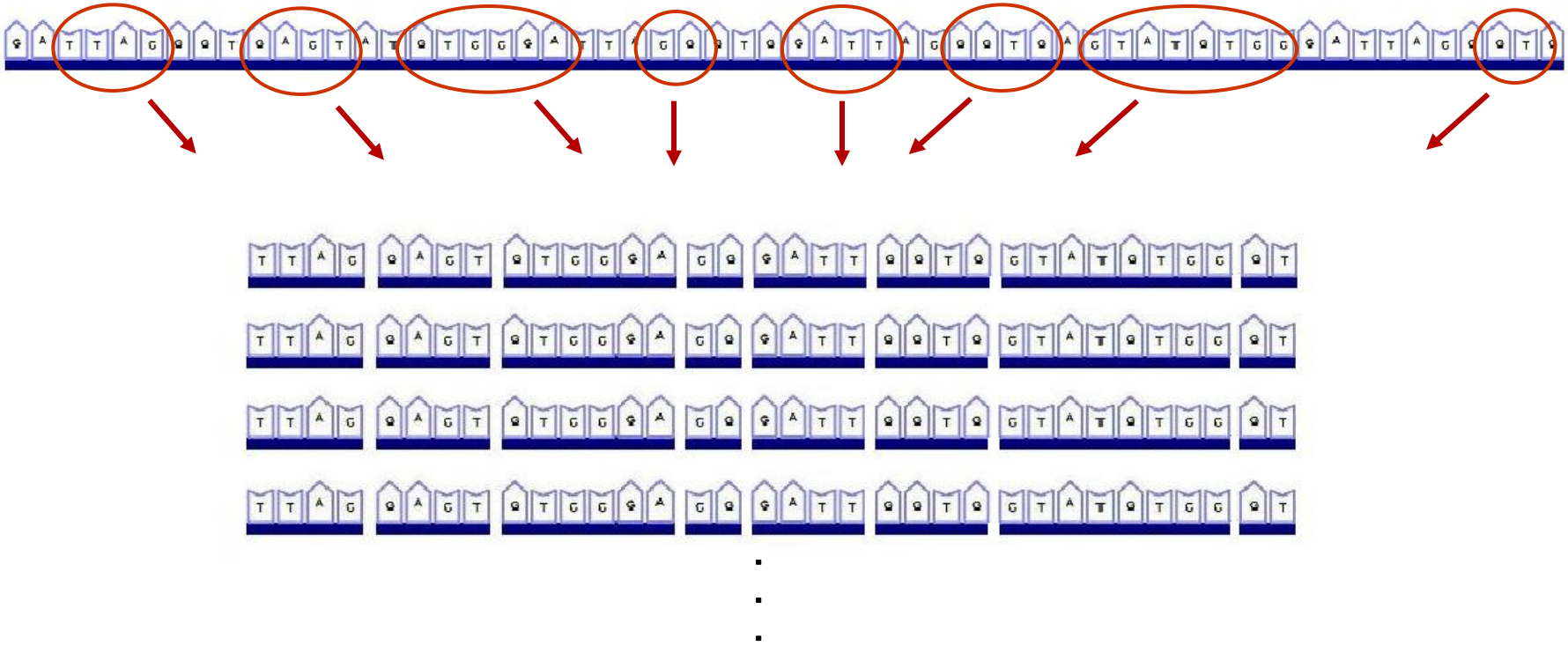
Analysis of *B. burgdorferi*: the 'housekeeping genes' alignment



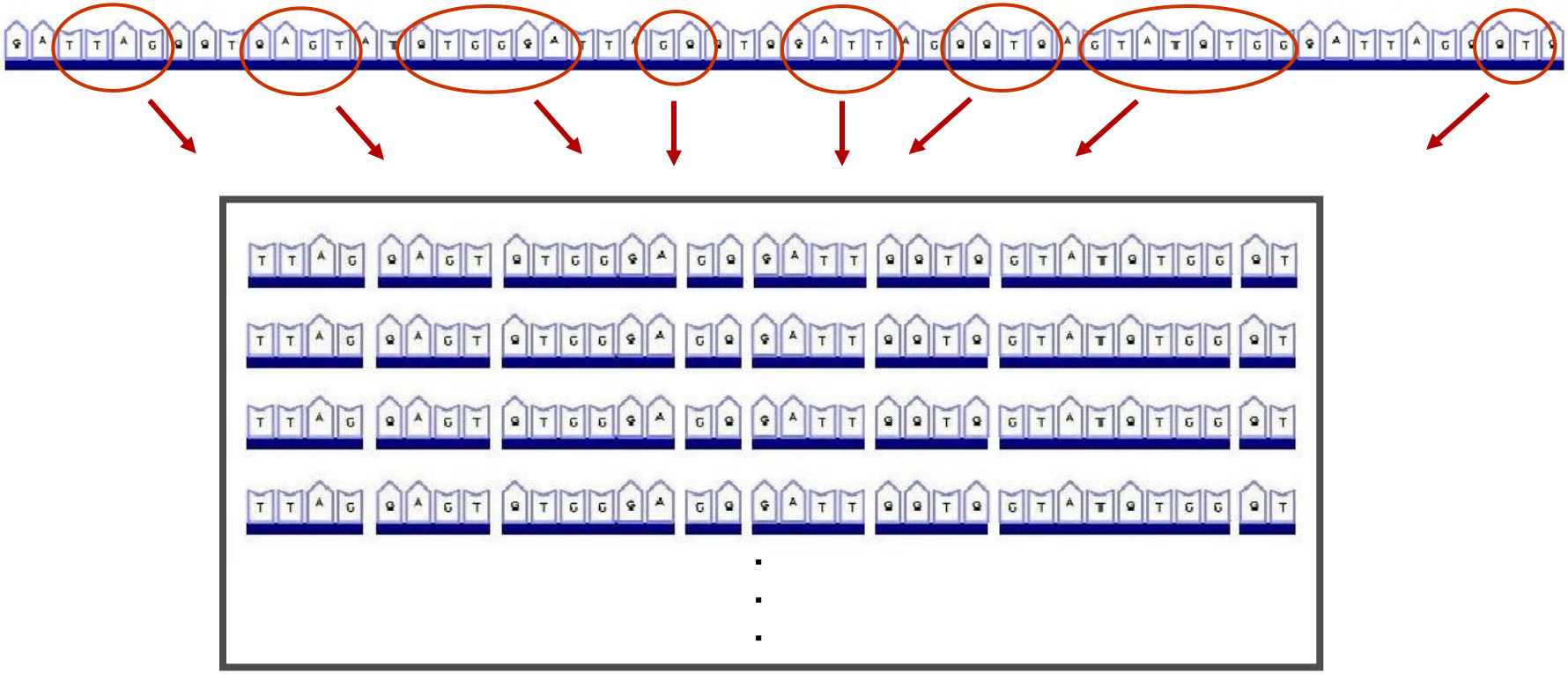
Analysis of *B. burgdorferi*: the 'housekeeping genes' alignment



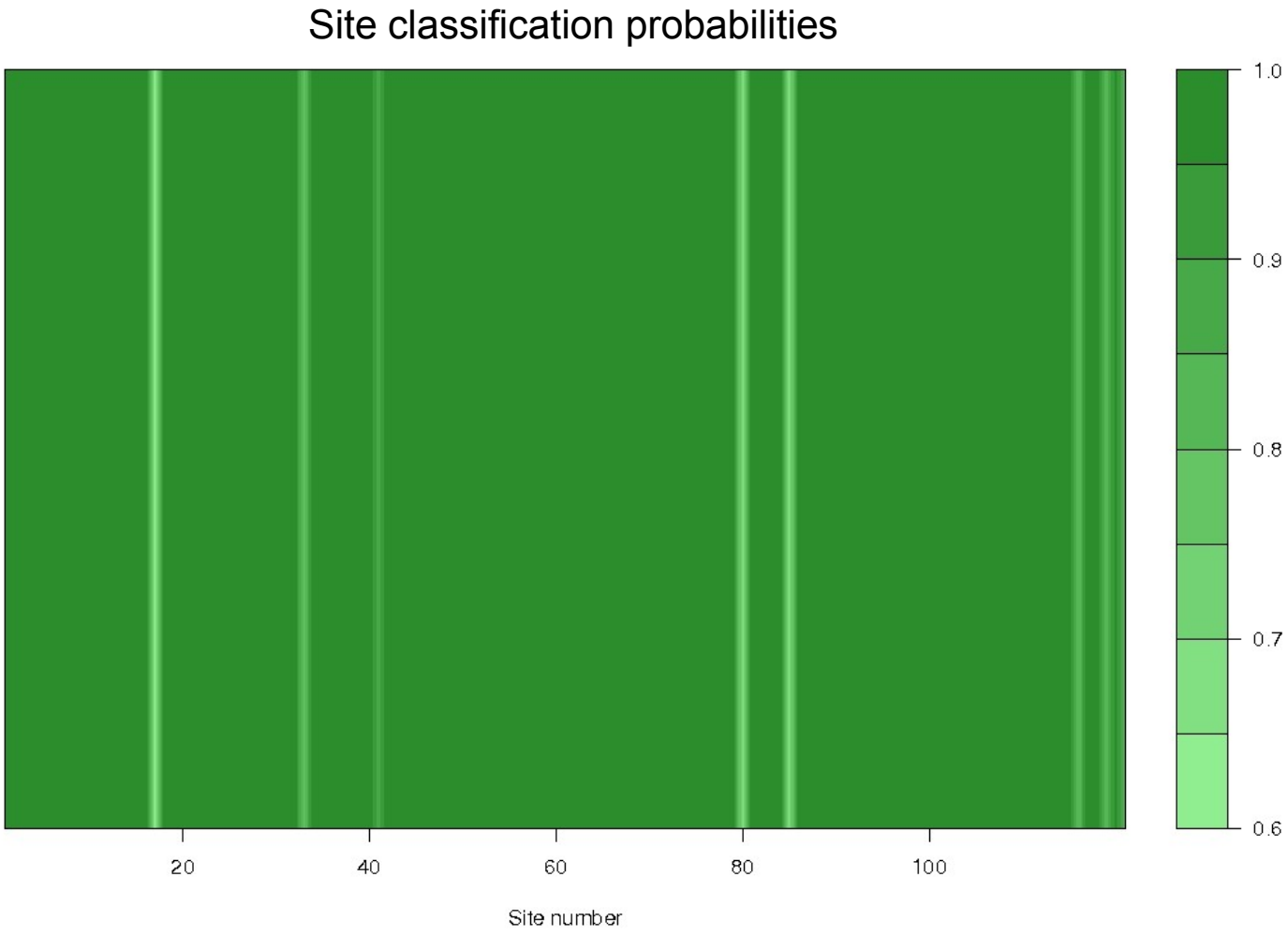
Analysis of *B. burgdorferi*: the 'housekeeping genes' alignment



Analysis of *B. burgdorferi*: the 'housekeeping genes' alignment

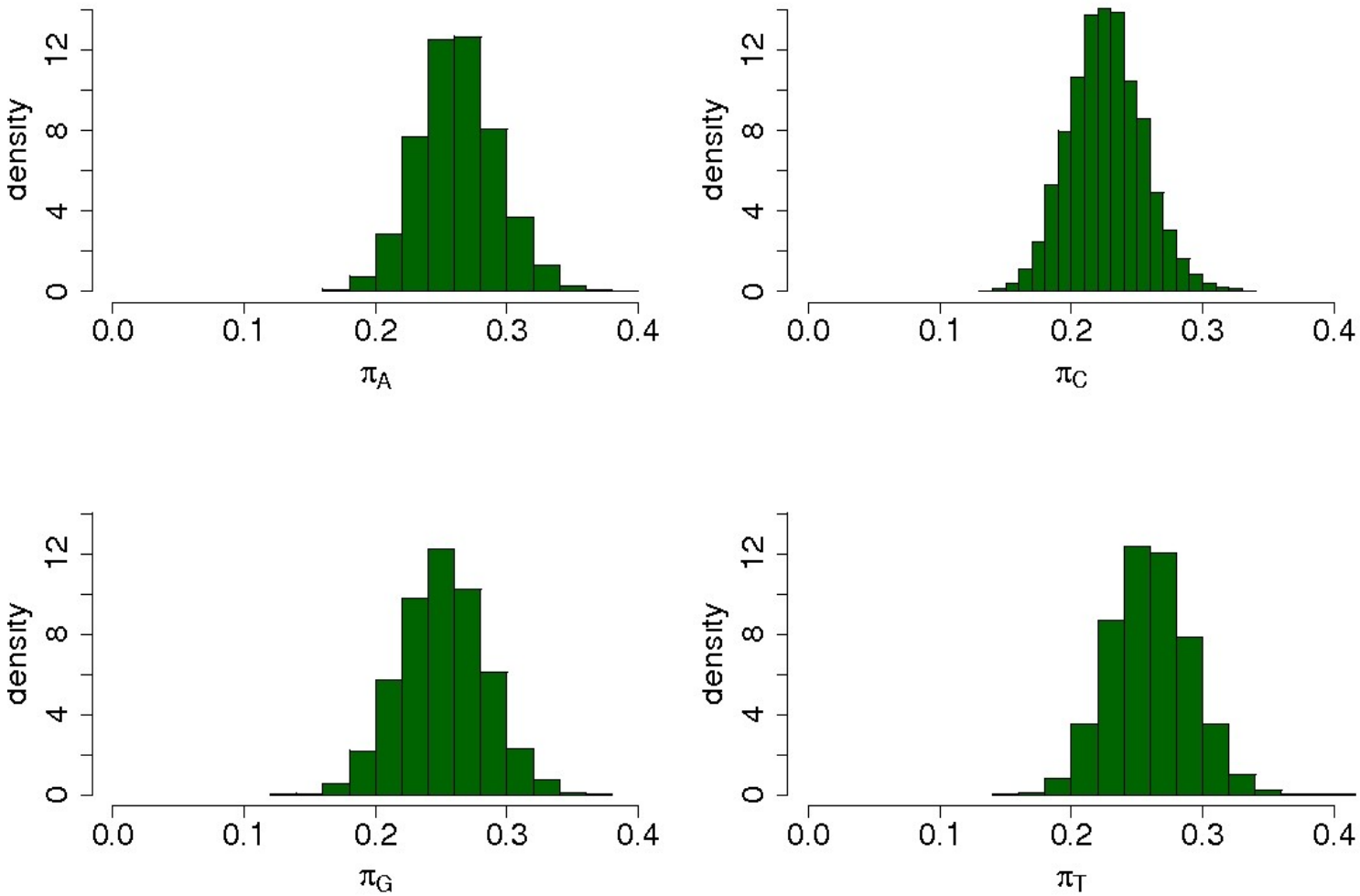


Analysis of *B. burgdorferi*: the 'housekeeping genes' alignment



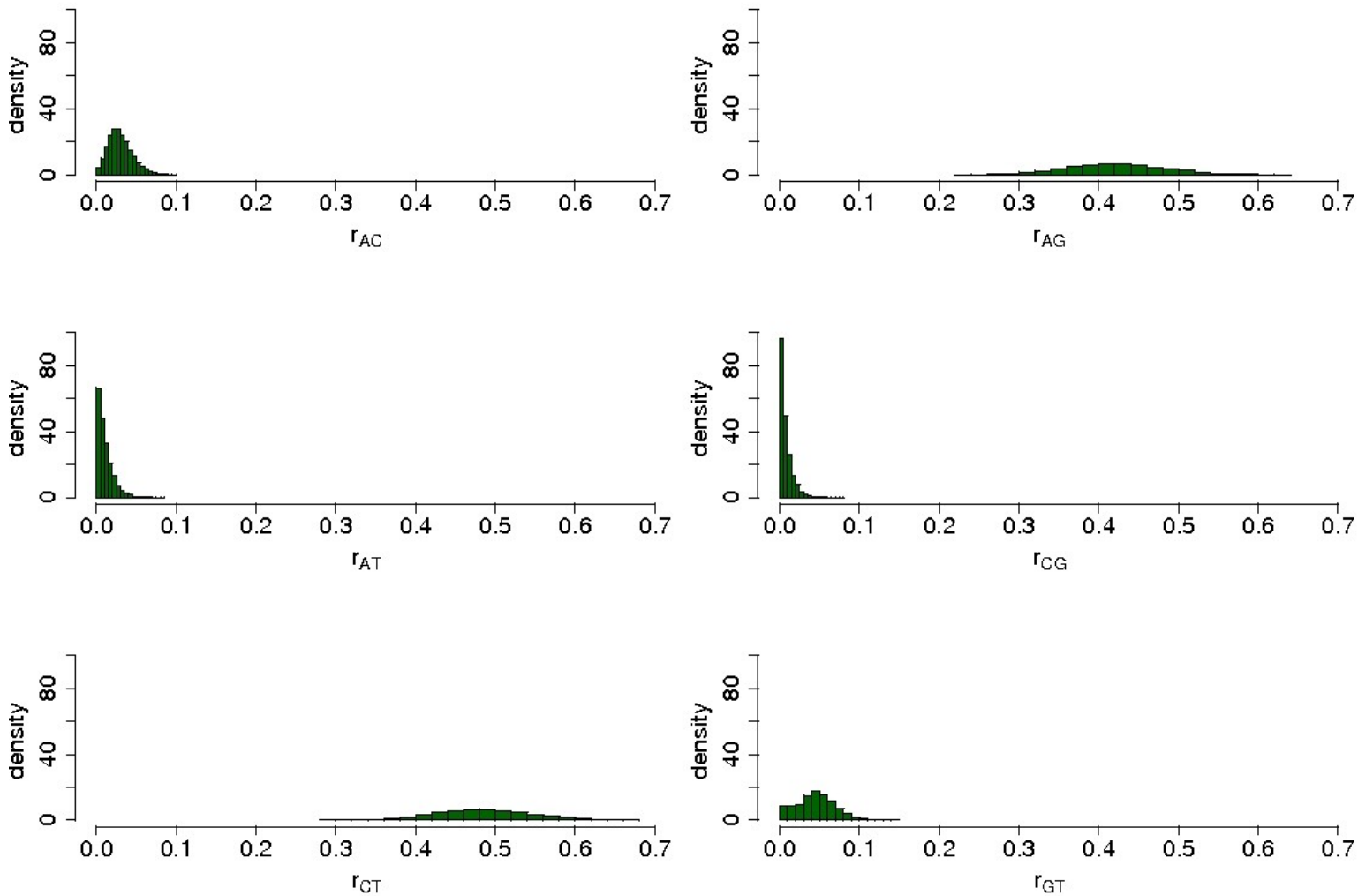
Analysis of *B. burgdorferi*: the 'housekeeping genes' alignment

Posterior densities of stationary frequencies



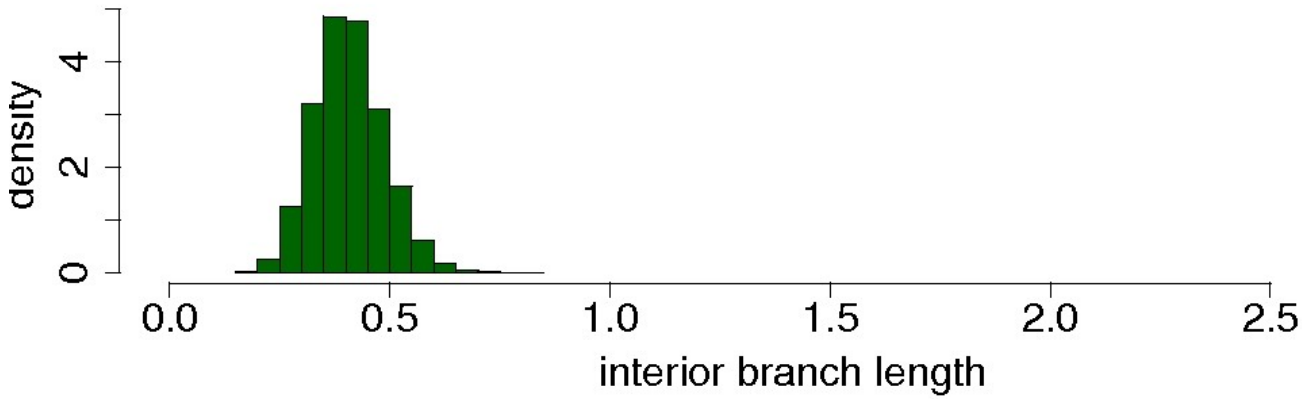
Analysis of *B. burgdorferi*: the 'housekeeping genes' alignment

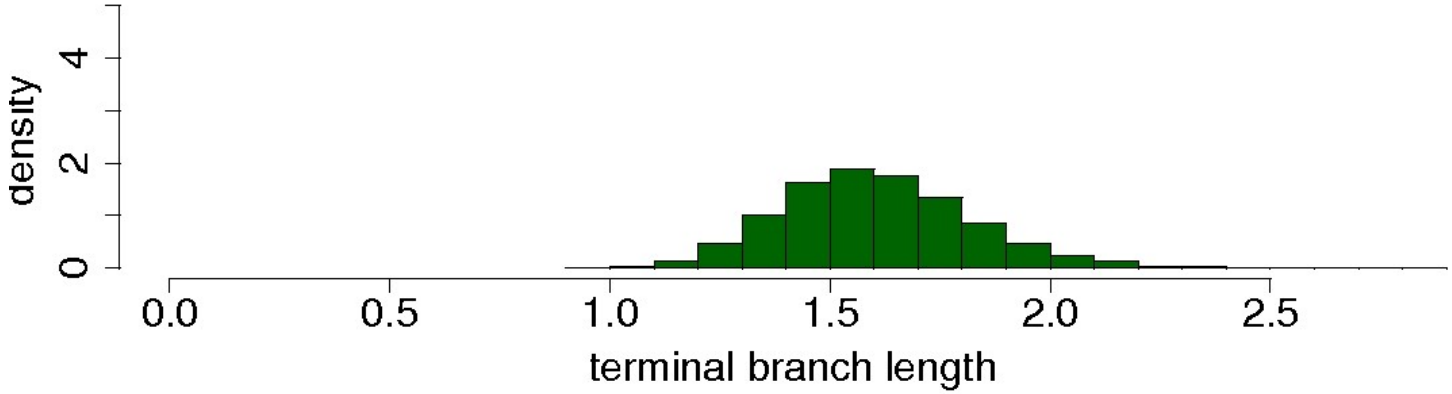
Posterior densities of substitution rates



Analysis of *B. burgdorferi*: the 'housekeeping genes' alignment

Posterior densities of branch lengths

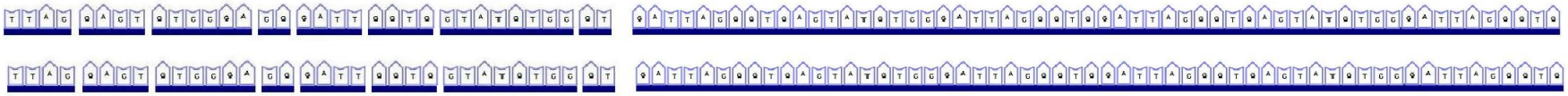




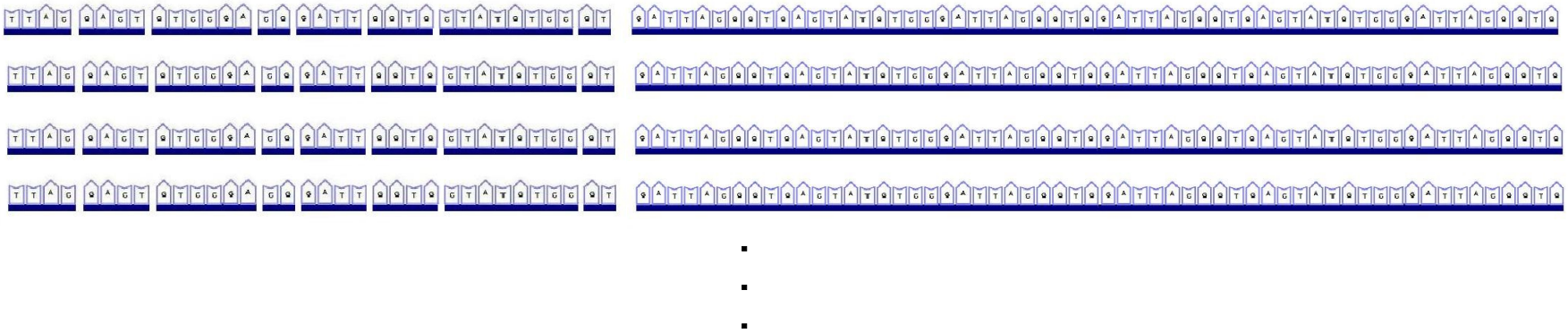
Analysis of *B. burgdorferi*: the 'housekeeping g.|ospC' alignment

Analysis of *B. burgdorferi*: the 'housekeeping g.|ospC' alignment

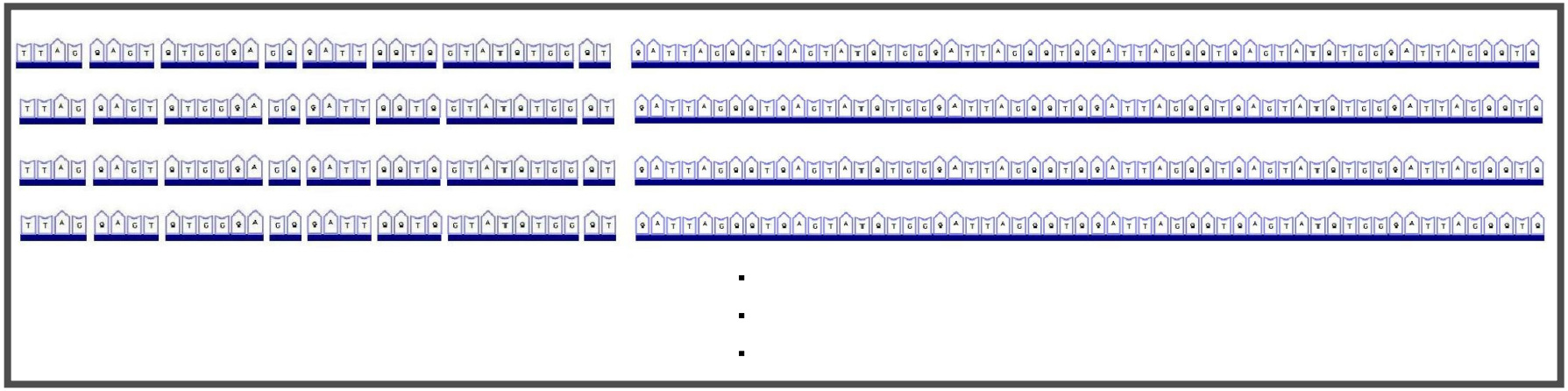
Analysis of *B. burgdorferi*: the 'housekeeping g.|ospC' alignment



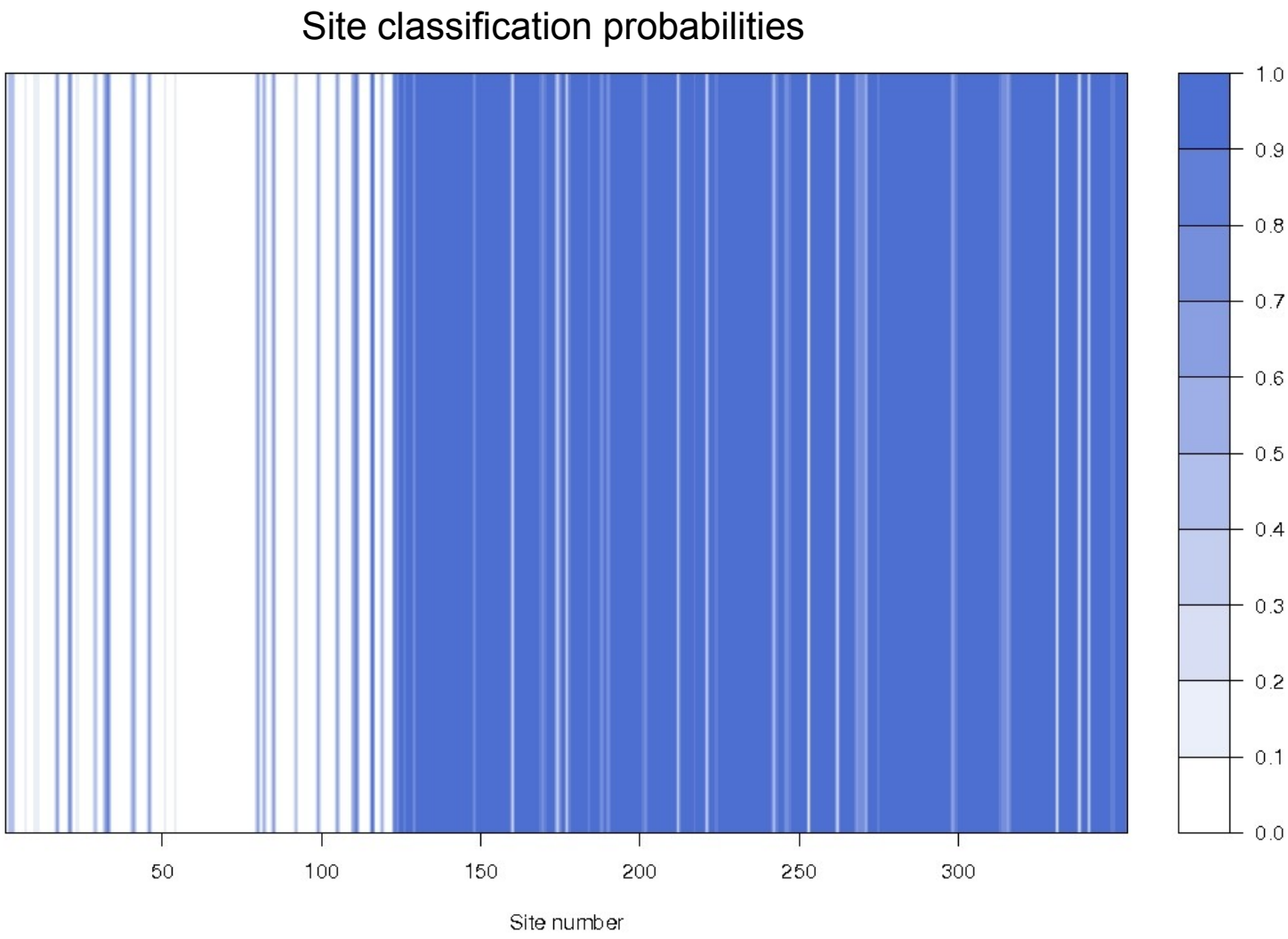
Analysis of *B. burgdorferi*: the 'housekeeping g.|ospC' alignment



Analysis of *B. burgdorferi*: the 'housekeeping g.|ospC' alignment

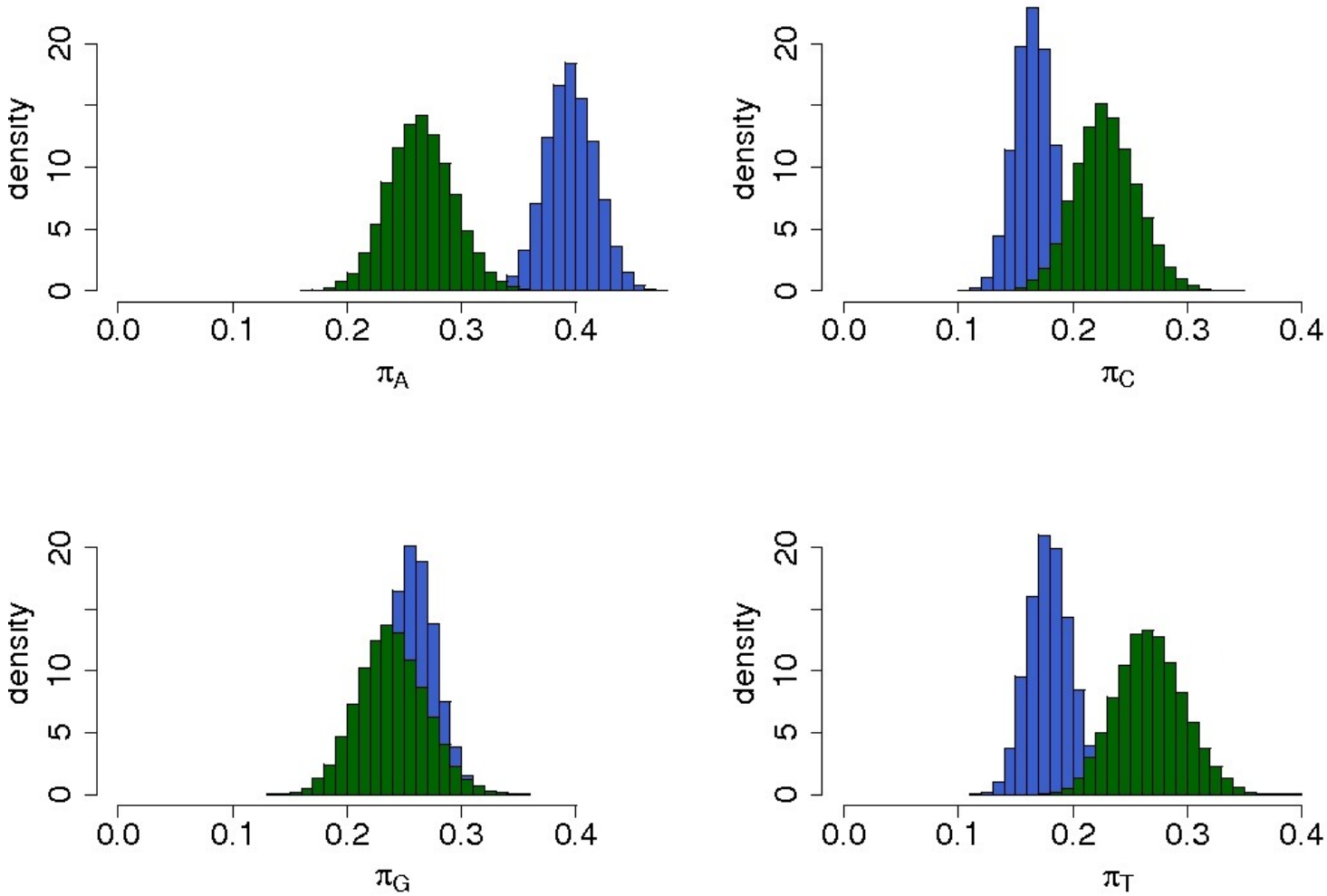


Analysis of *B. burgdorferi*: the 'housekeeping g.|ospC' alignment



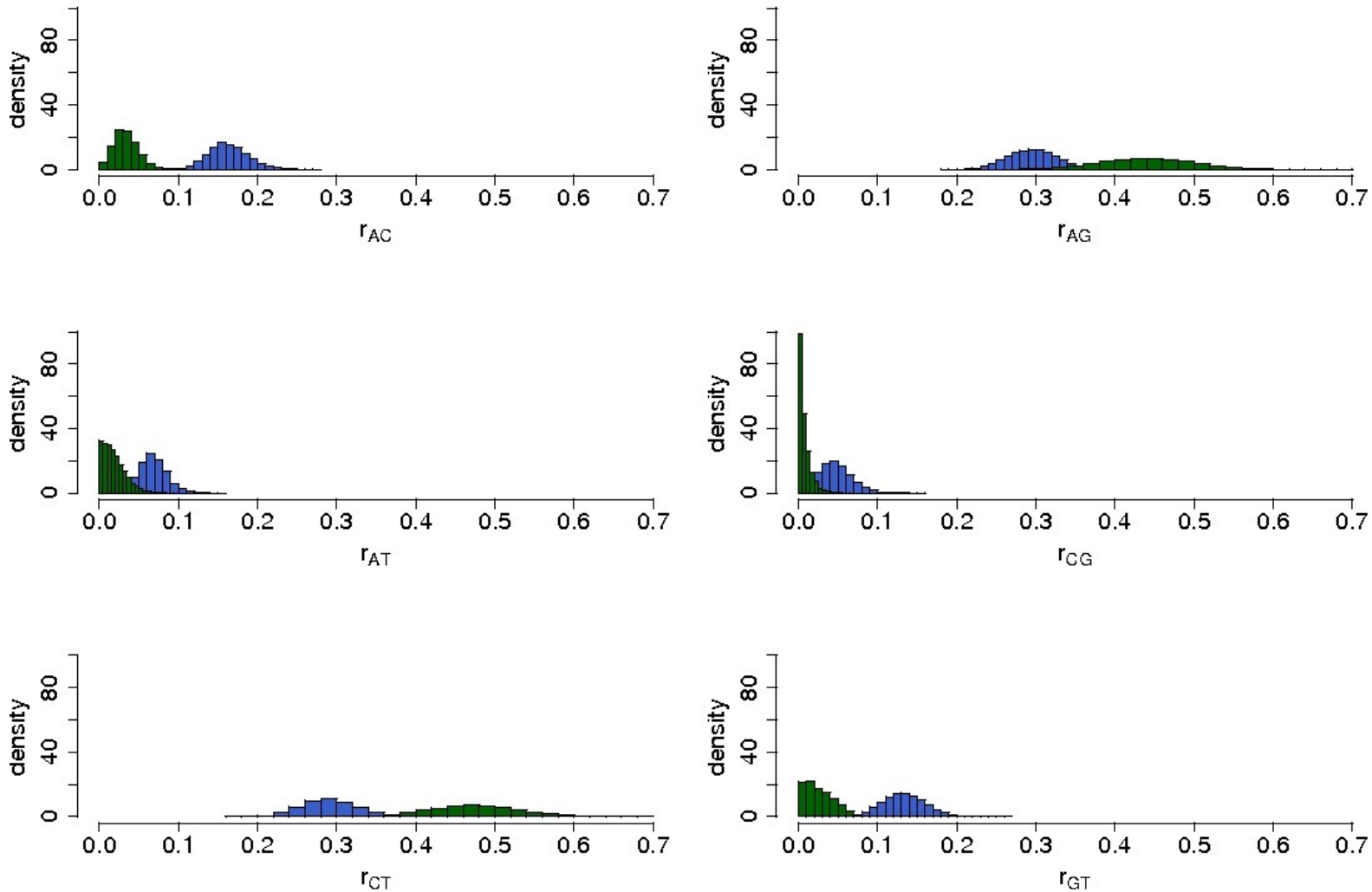
Analysis of *B. burgdorferi*: the 'housekeeping g.|ospC' alignment

Posterior densities of stationary frequencies



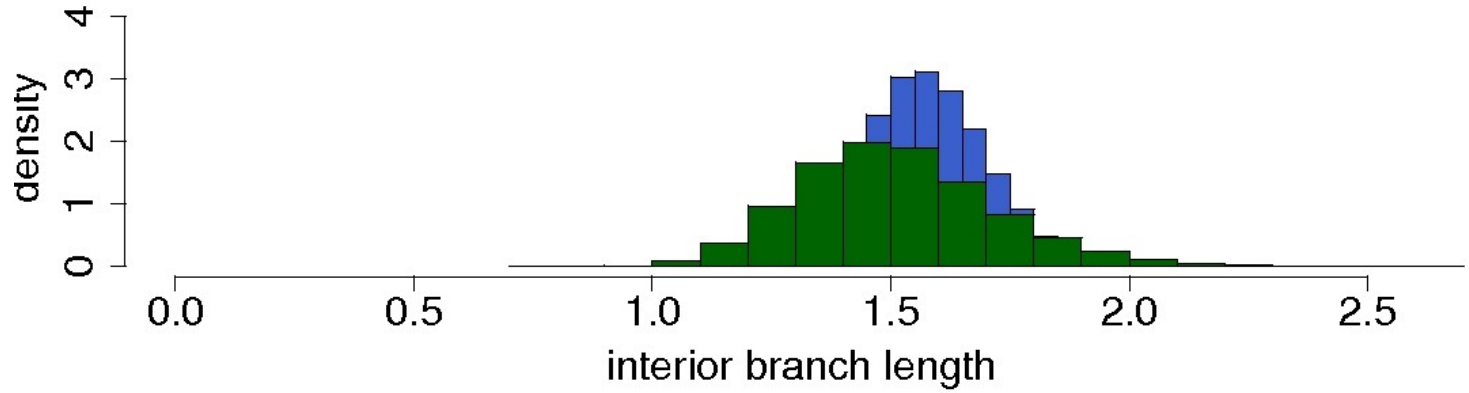
Analysis of *B. burgdorferi*: the 'housekeeping g.|ospC' alignment

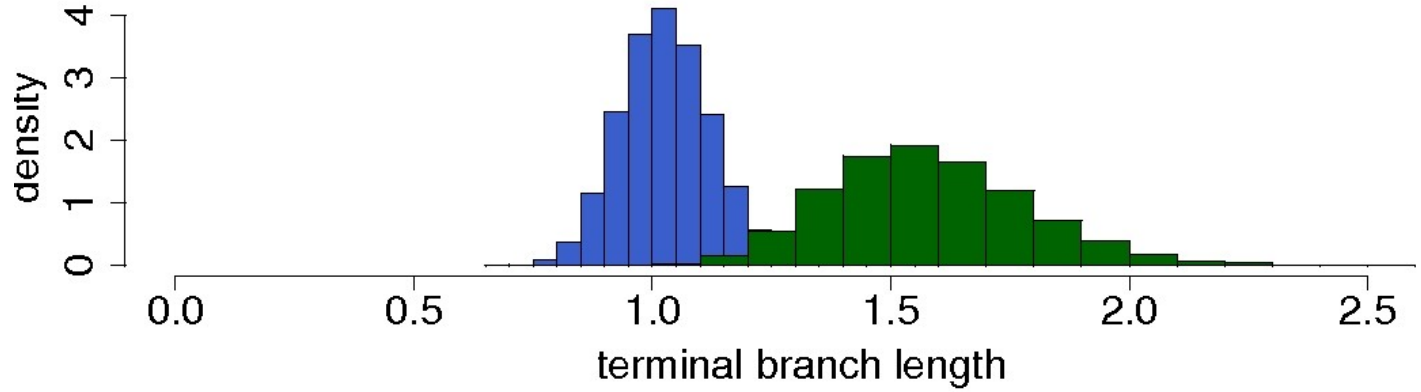
Posterior densities of substitution rates



Analysis of *B. burgdorferi*: the 'housekeeping g.|ospC' alignment

Posterior densities of branch lengths





Conclusions

- A more realistic phylogenetic model that accommodates heterogeneity.

Conclusions

- A more realistic phylogenetic model that accommodates heterogeneity.

- The *Q+t mixture model* automatically recovers the evolutionary identity of a site.

Conclusions

- A more realistic phylogenetic model that accommodates heterogeneity.
- The *Q+t mixture model* automatically recovers the evolutionary identity of a site.
- It is a suitable indicator of evolutionary homogeneity or heterogeneity among large-scale concatenations of genes.

Conclusions

- It is relevant testing for homogeneity as a concatenation of genes will produce valid inferences only when there is evolutionary congruence.

Conclusions

- It is relevant testing for homogeneity as a concatenation of genes will produce valid inferences only when there is evolutionary congruence.
- *B. burgdorferi* data is just one application of many other possibilities.

Acknowledgements

- Merrilee Hurn, Mathematical Sciences
- Tony Robinson, Mathematical Sciences
- Gabi Margos, Biology and Biochemistry
- Klaus Kurtenbach, Biology and Biochemistry

Research supported by

CONACYT Consejo Nacional de Ciencia y Tecnología