Phylogenetic analysis

Laurent Abi-Rached

February 29, 2012

Phylogenetic analysis: applications

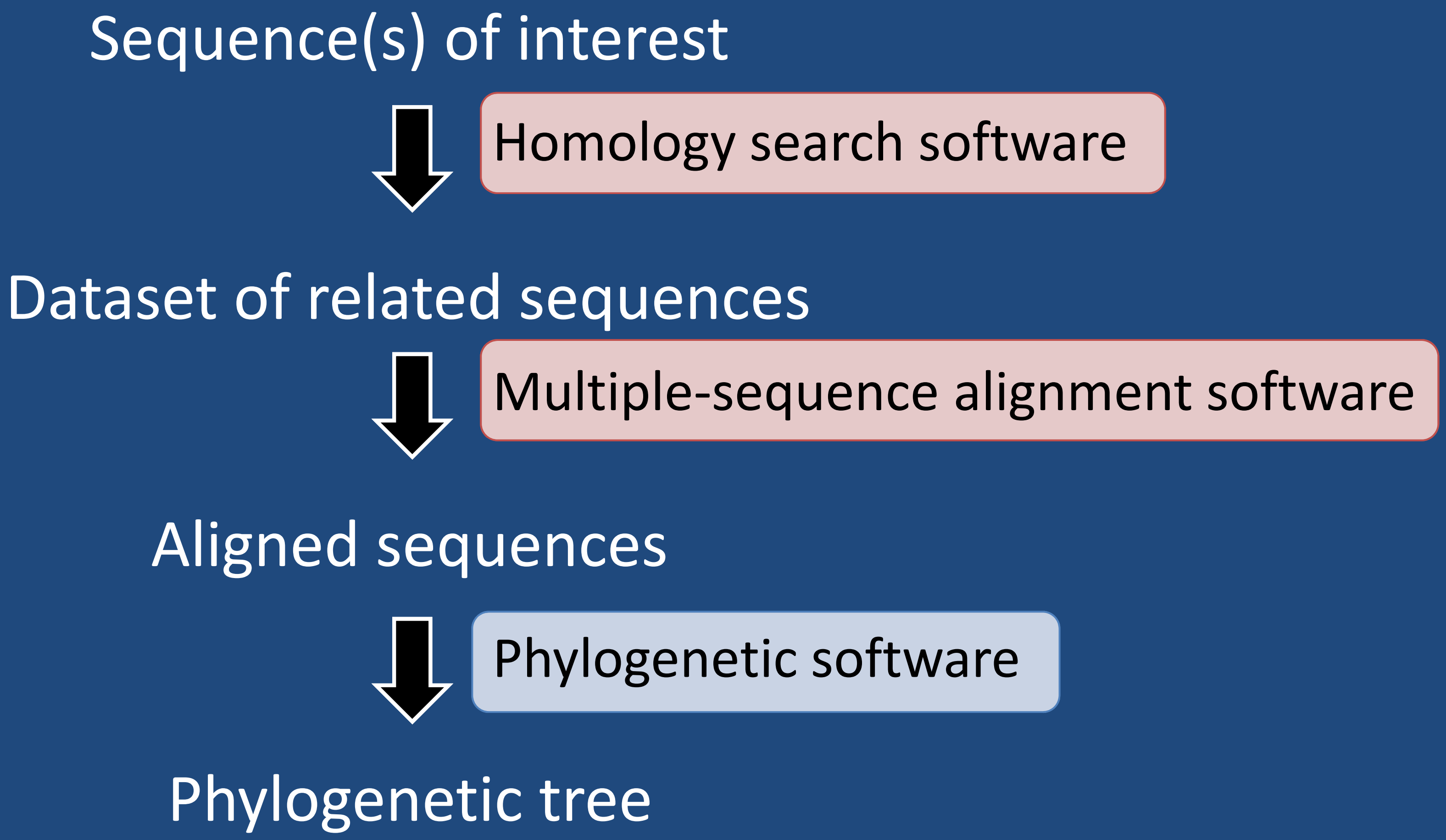
Taxonomic studies

Study of multigenic families

Basis for other types of analysis:

- Selection analysis
- Ancestral sequence reconstruction
- Divergence time analysis
- Functional divergence analysis
- Host-pathogen co-evolution

Phylogenetic analysis: principles



Sequence(s) of interest

Homology search software

Dataset of related sequences

Multiple-sequence alignment software

Aligned sequences

Phylogenetic software

Phylogenetic tree

Phylogenetic analysis: methods

Distance

MEGA: Molecular Evolutionary Genetics Analysis

(Tamura et al. Molecular Biology and Evolution 2011)

Maximum parsimony

PAUP: Phylogenetic Analysis Using Parsimony

(Swofford, D. L. 2001)

Maximum likelihood

RAxML: Randomized Axelerated Maximum Likelihood

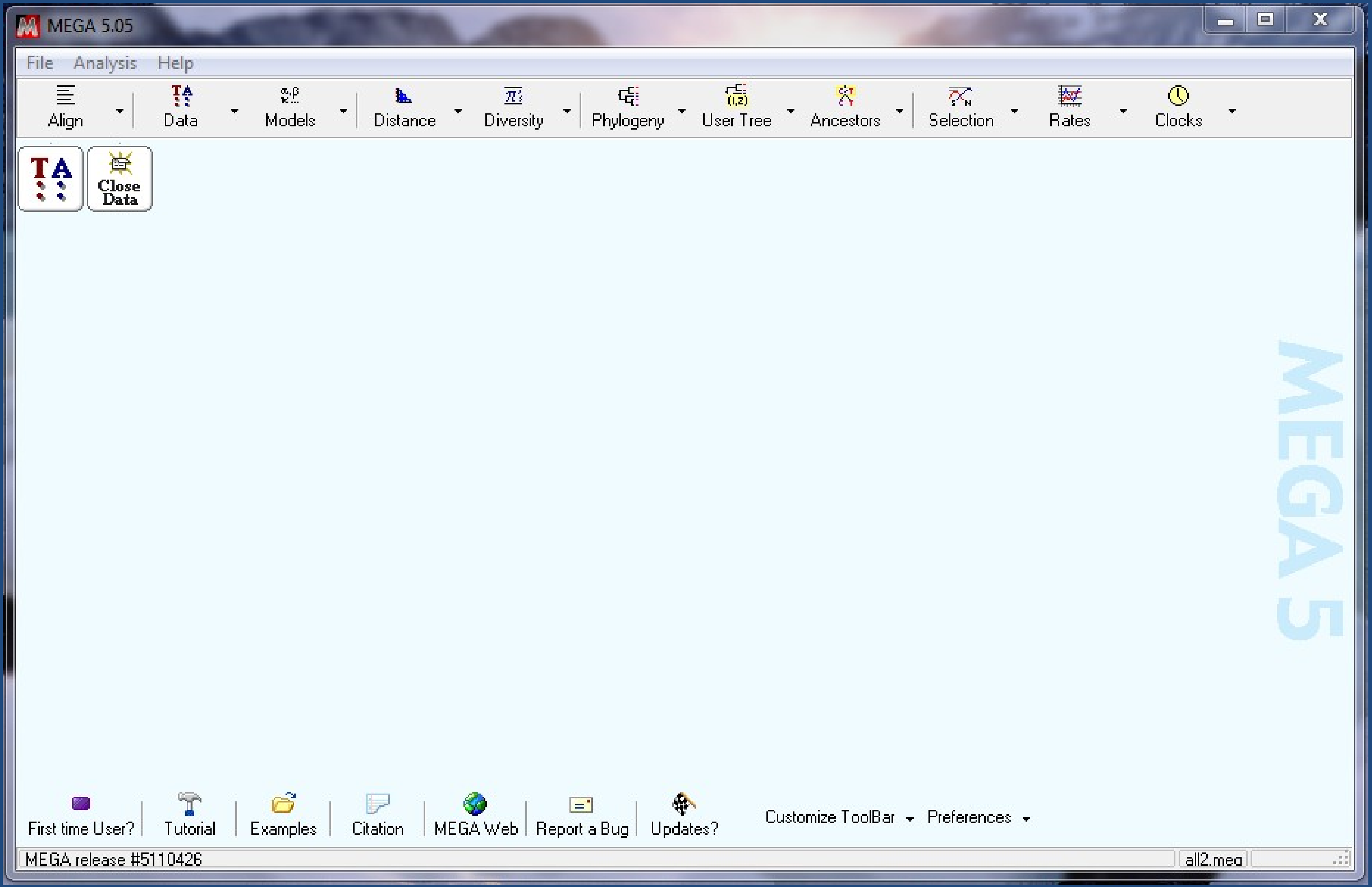
(Stamatakis, A. Bioinformatics 2006)

Bayesian

MrBayes

(Ronquist et al, Systematic Biology 2012)

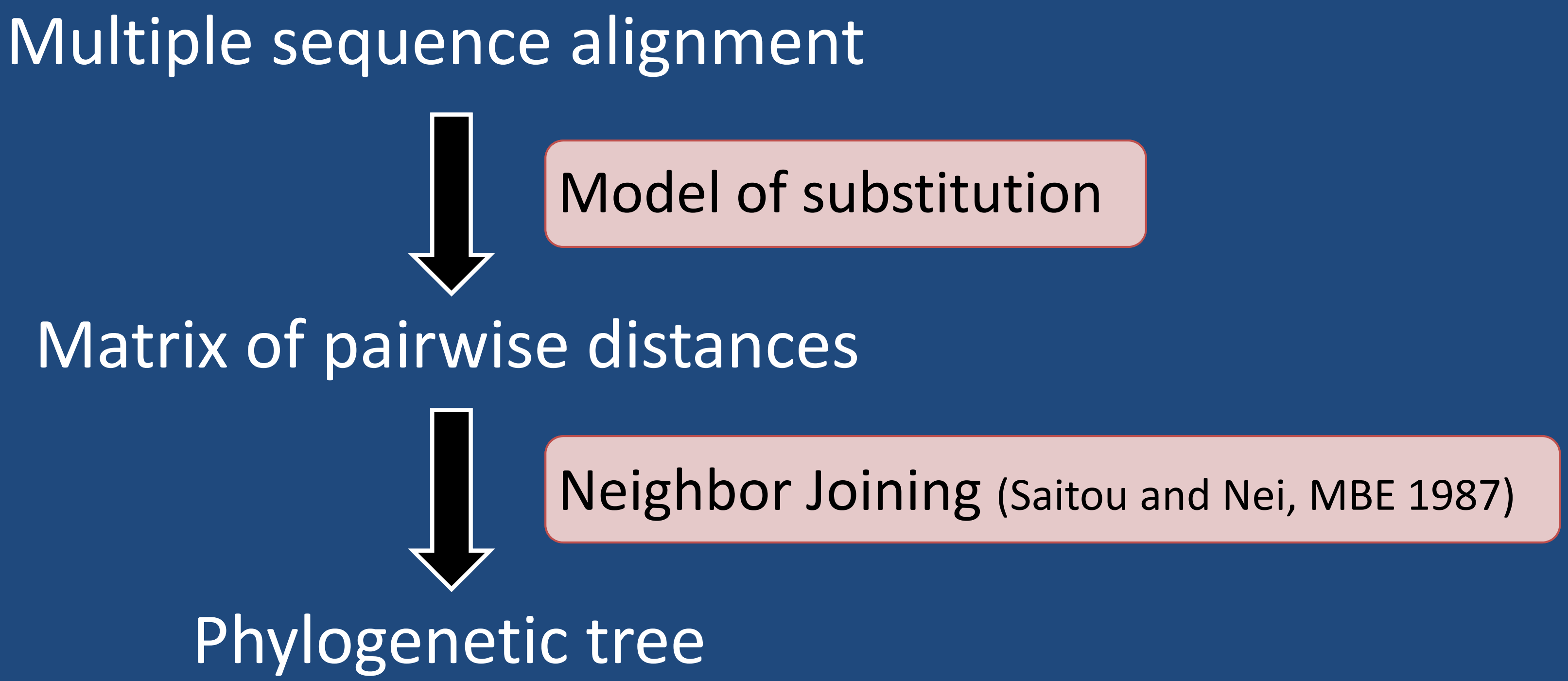
Distance methods: MEGA



MEGA: Molecular Evolutionary Genetics Analysis

(Tamura et al. Molecular Biology and Evolution 2011)

Distance methods: Neighbor-Joining (NJ)



Multiple sequence alignment

Model of substitution

Matrix of pairwise distances

Neighbor Joining (Saitou and Nei, MBE 1987)

Phylogenetic tree

Distance methods:
example of DNA models

p-distance

This distance is the proportion (*p*) of nucleotide sites at which two sequences being compared are different. It is obtained by dividing the number of nucleotide differences by the total number of nucleotides compared. It does not make any correction for multiple substitutions at the same site or substitution rate biases (for example, differences in the transitional and transversional rates).

Tamura-Nei distance

The Tamura-Nei model (1993) corrects for multiple hits, taking into account the differences in substitution rate between nucleotides and the inequality of nucleotide frequencies. It distinguishes between transitional substitution rates between purines and transversional substitution rates between pyrimidines.

Distance methods:
example of amino acid models

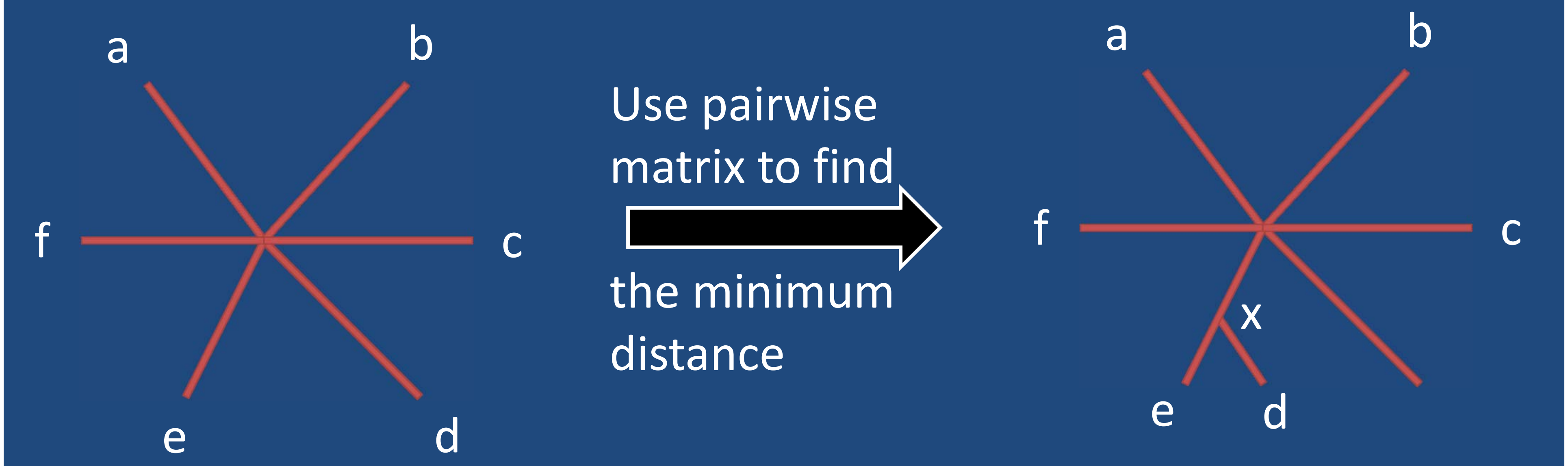
p-distance (Amino acids)

This distance is the proportion (*p*) of amino acid sites at which the two sequences to be compared are different. It is obtained by dividing the number of amino acid differences by the total number of sites compared. It does not make any correction for multiple substitutions at the same site.

Poisson Correction (PC) distance

The Poisson correction distance assumes equal amino acid frequencies while correcting for multiple substitutions at the same site.

Distance methods:
the NJ reconstruction



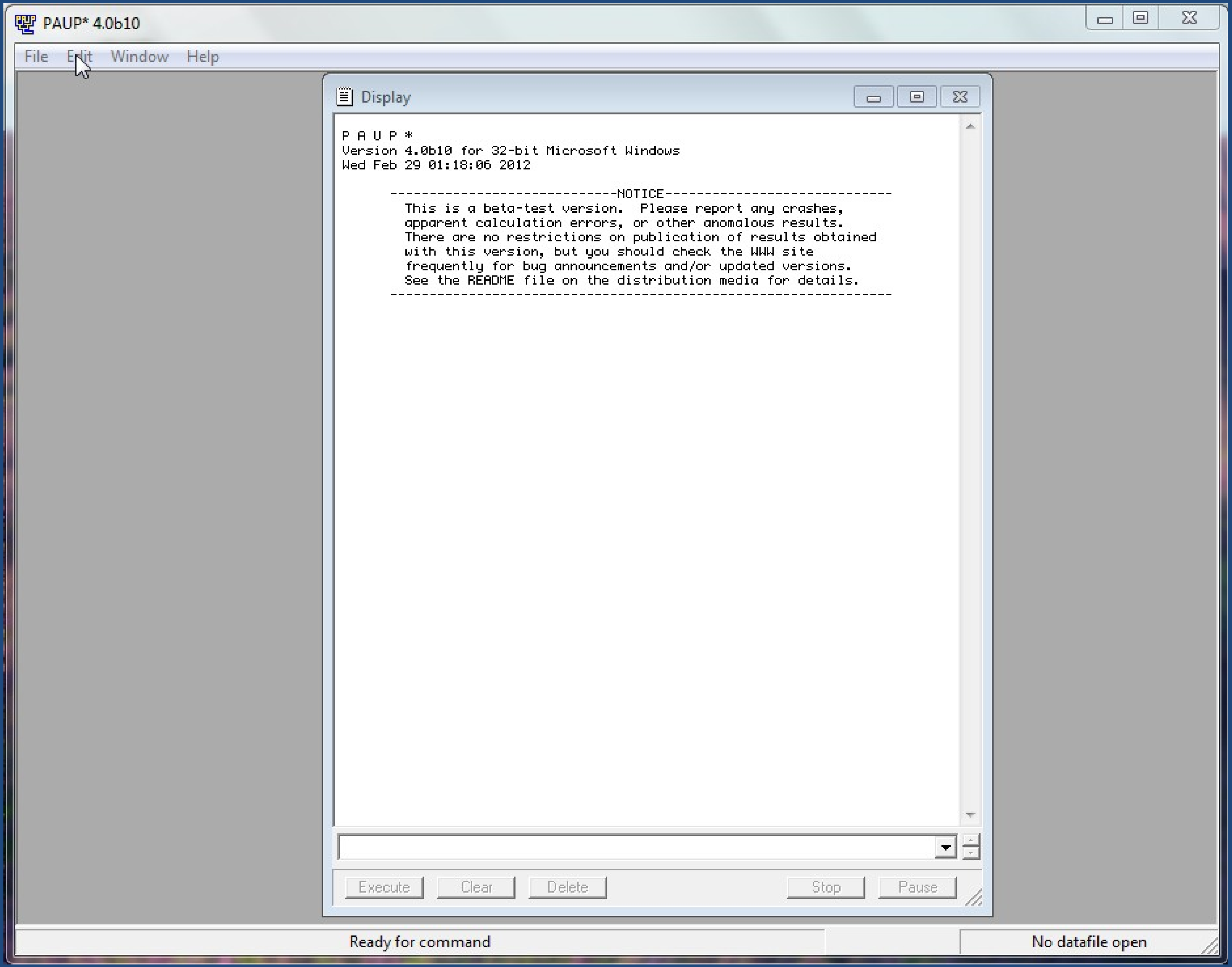
Use pairwise matrix to find the minimum distance

Method: - fast

- very useful for ‘screening’ datasets (recombination analysis)

(- performs well with distantly related sequences)

Maximum parsimony



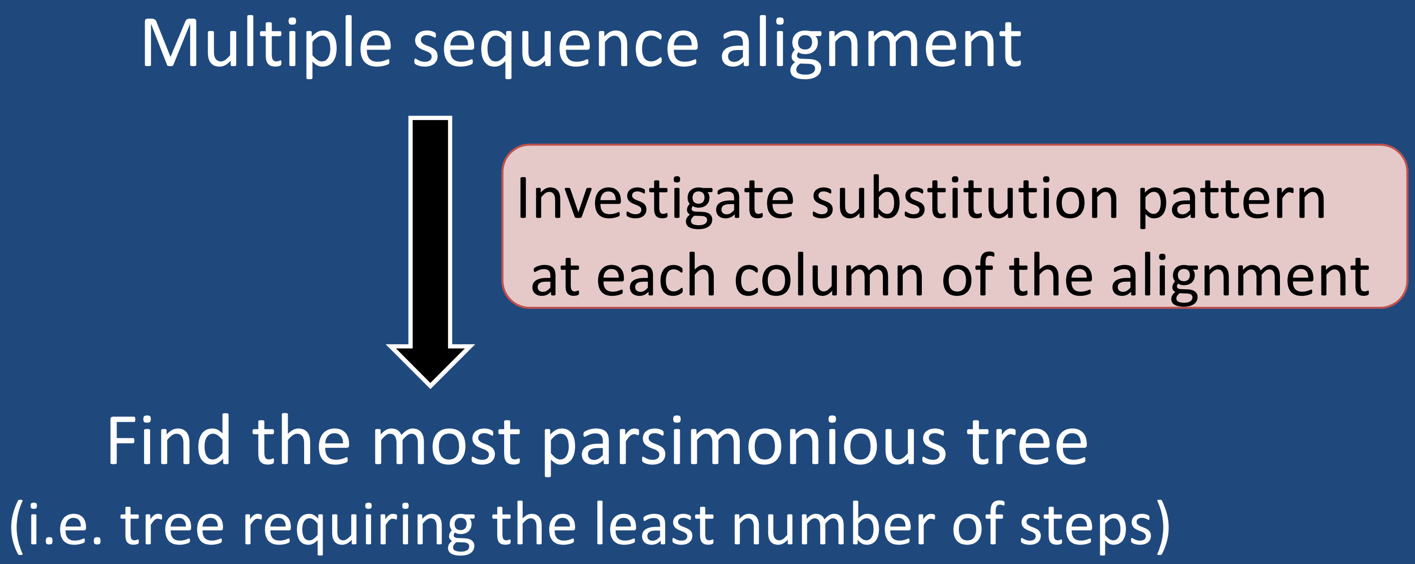
```
P A U P *
Version 4.0b10 for 32-bit Microsoft Windows
Wed Feb 29 01:18:06 2012

      -----------------------------NOTICE-----------------------------
        This is a beta-test version.  Please report any crashes,
        apparent calculation errors, or other anomalous results.
        There are no restrictions on publication of results obtained
        with this version, but you should check the WWW site
        frequently for bug announcements and/or updated versions.
        See the README file on the distribution media for details.
      ----------------------------------------------------------------
```

PAUP: Phylogenetic Analysis Using Parsimony

(Swofford, D. L. 2001)

Maximum parsimony

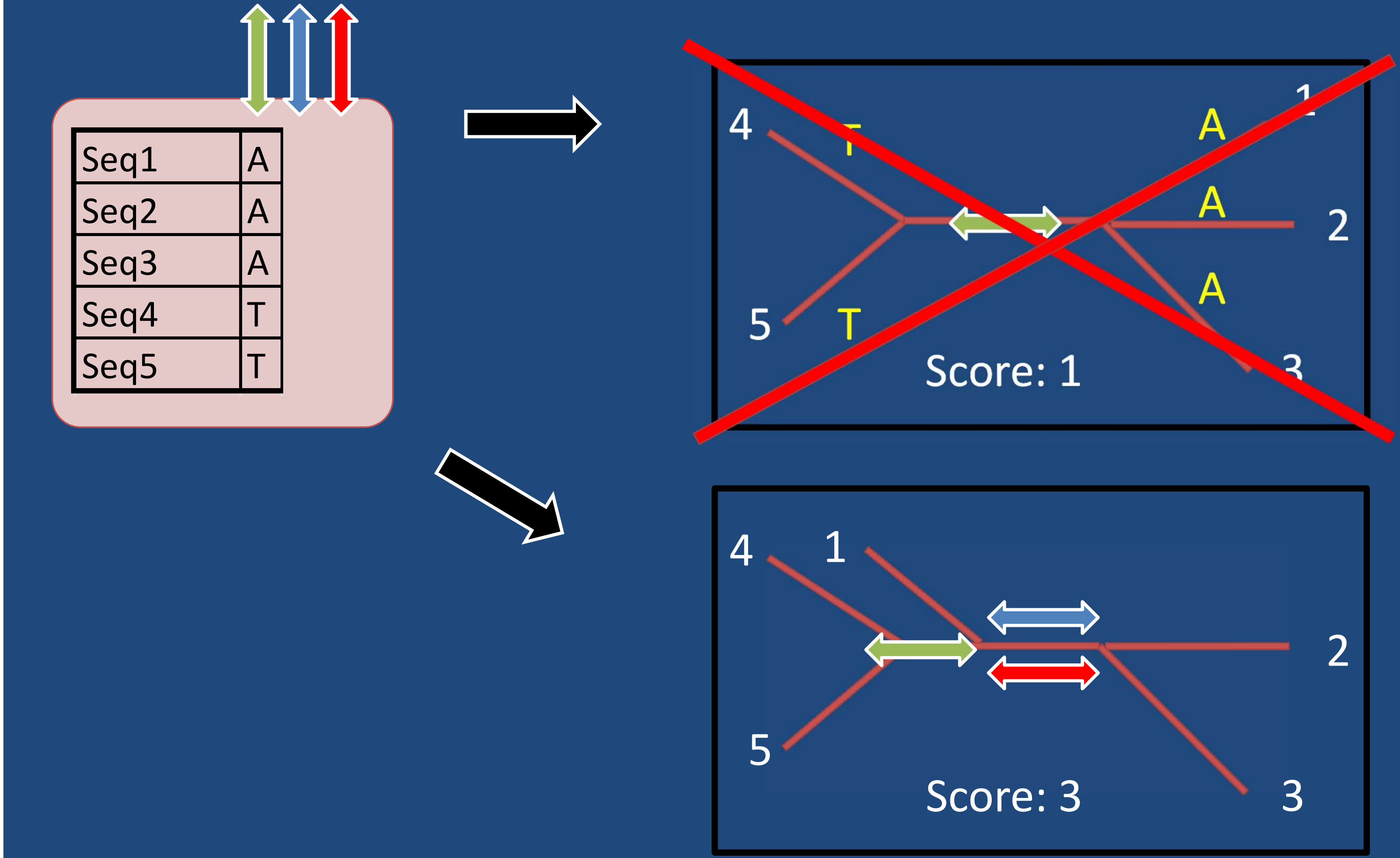


Multiple sequence alignment

Investigate substitution pattern at each column of the alignment

Find the most parsimonious tree
(i.e. tree requiring the least number of steps)

Maximum parsimony



Seq1	A
Seq2	A
Seq3	A
Seq4	T
Seq5	T

Maximum parsimony

Exact searches are often too slow

↓

Have to use heuristic approaches

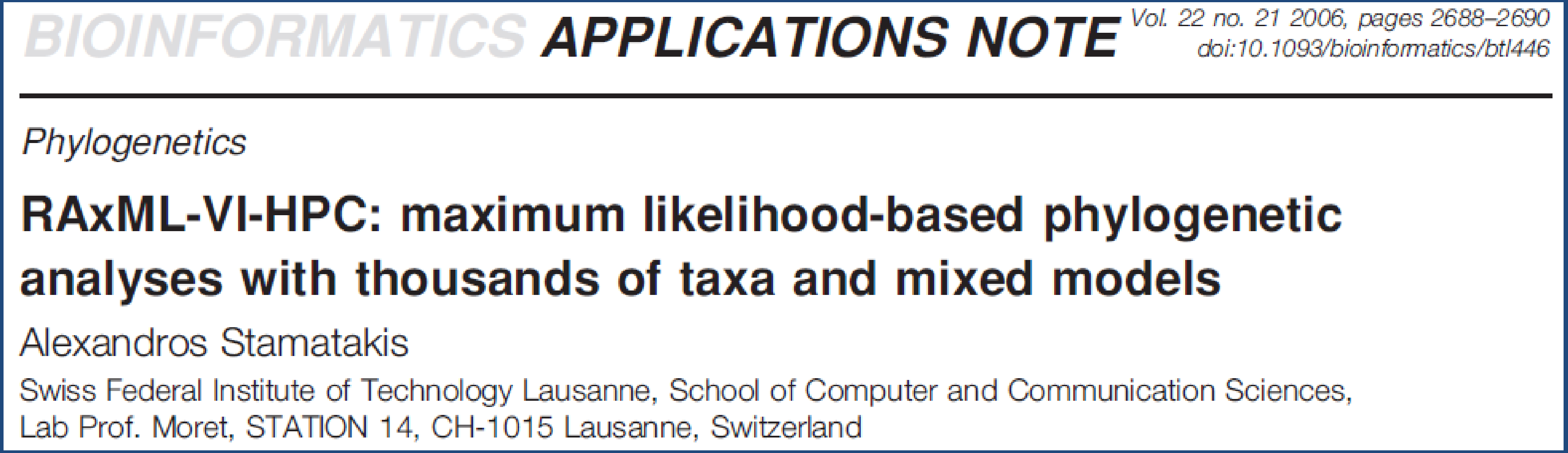
(i.e. Tree bisection and reconnection (TBR) branch swapping)

Method: - simple

- relatively fast with heuristic approaches

(- performs well with closely related sequences)

Maximum likelihood approach



BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 22 no. 21 2006, pages 2688–2690*
doi:10.1093/bioinformatics/btl446

Phylogenetics

RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models

Alexandros Stamatakis

Swiss Federal Institute of Technology Lausanne, School of Computer and Communication Sciences,
Lab Prof. Moret, STATION 14, CH-1015 Lausanne, Switzerland

RAxML: Randomized Axelerated Maximum Likelihood

(Stamatakis, A. Bioinformatics 2006)

"New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0."

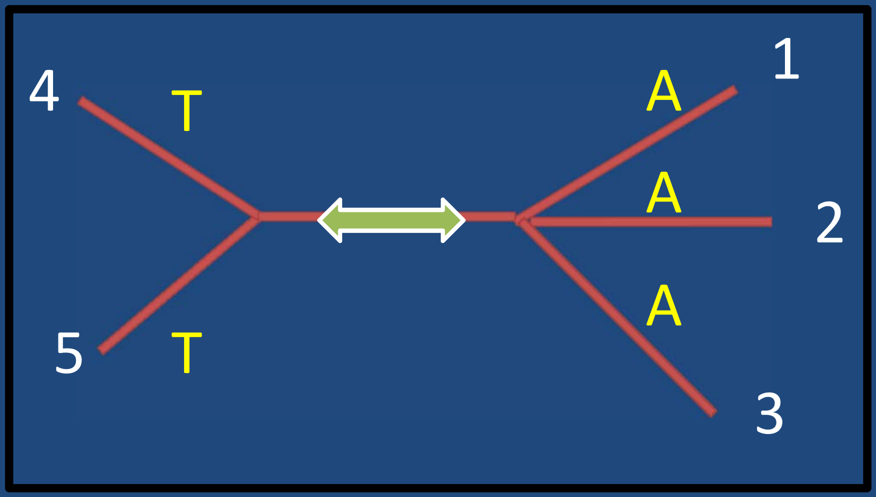
Guindon S. et al. Systematic Biology, 59(3):307-21, 2010.

Maximum likelihood approach

Likelihood function

Seq1	A	A	T
Seq2	A	T	A
Seq3	A	T	A
Seq4	T	A	T
Seq5	T	A	T

Column #1



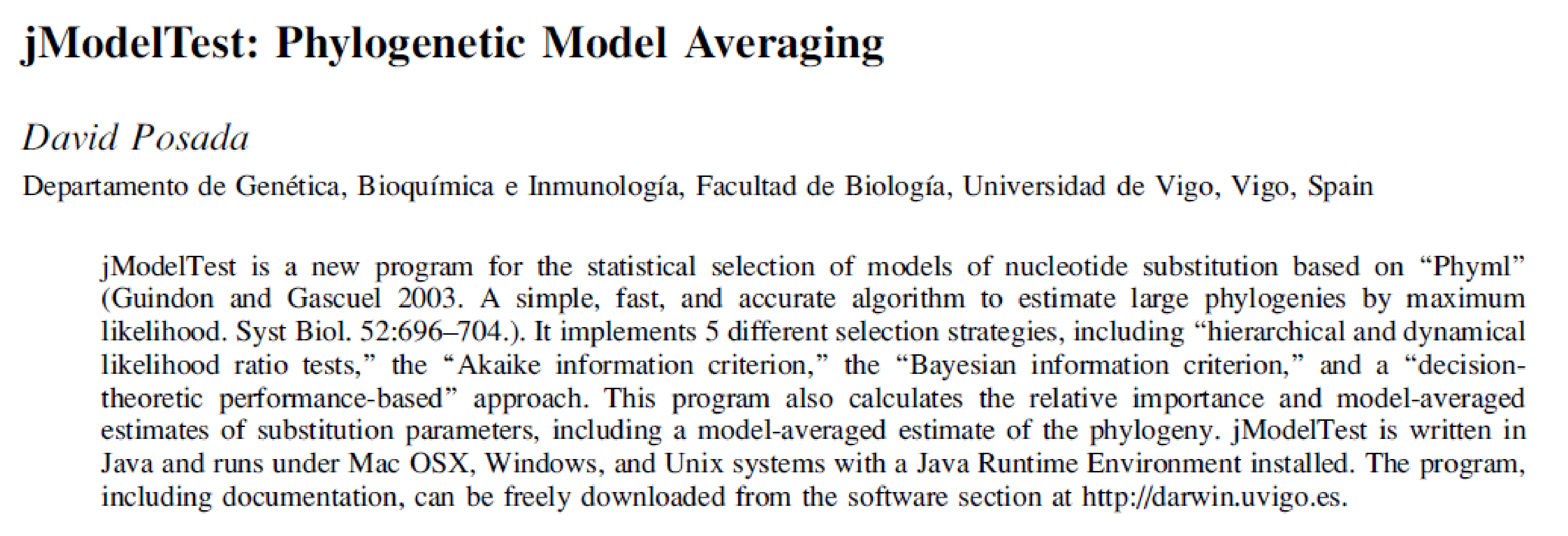
<u>Likelihood calculation:</u>

```
Given a model of substitution:
  for each possible tree
    for each column of the alignment
      calculate the likelihood of the column, given the tree
```

<u>Method:</u> - slow

- most accurate

Maximum likelihood approach: selecting a model of substitution



jModelTest: Phylogenetic Model Averaging

David Posada

Departamento de Genética, Bioquímica e Inmunología, Facultad de Biología, Universidad de Vigo, Vigo, Spain

jModelTest is a new program for the statistical selection of models of nucleotide substitution based on "Phyml" (Guindon and Gascuel 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.). It implements 5 different selection strategies, including "hierarchical and dynamical likelihood ratio tests," the "Akaike information criterion," the "Bayesian information criterion," and a "decision-theoretic performance-based" approach. This program also calculates the relative importance and model-averaged estimates of substitution parameters, including a model-averaged estimate of the phylogeny. jModelTest is written in Java and runs under Mac OSX, Windows, and Unix systems with a Java Runtime Environment installed. The program, including documentation, can be freely downloaded from the software section at http://darwin.uvigo.es.

Posada D. 2008. Molecular Biology and Evolution 25: 1253-1256.

Maximum likelihood approach: selecting a model of substitution

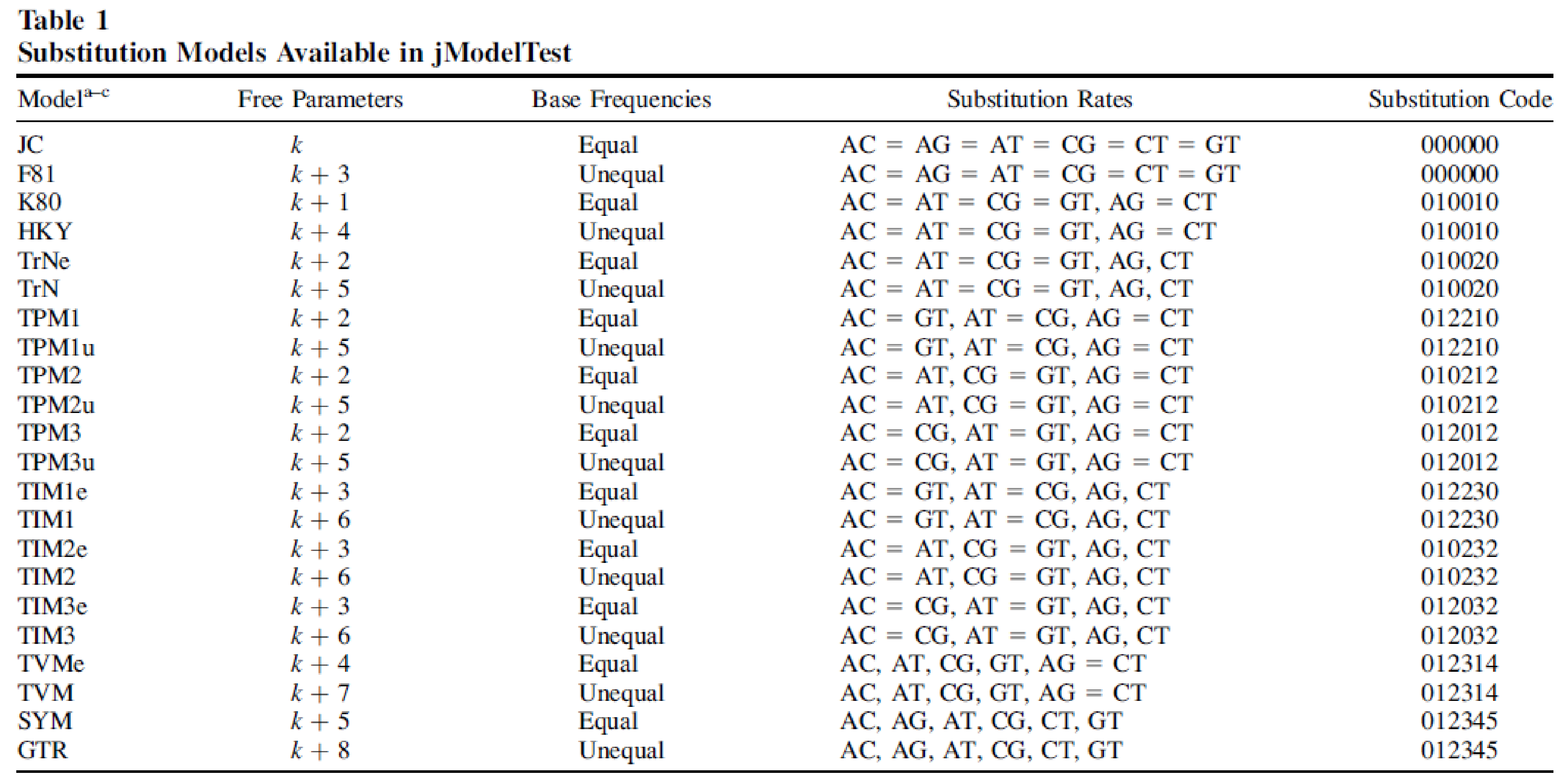


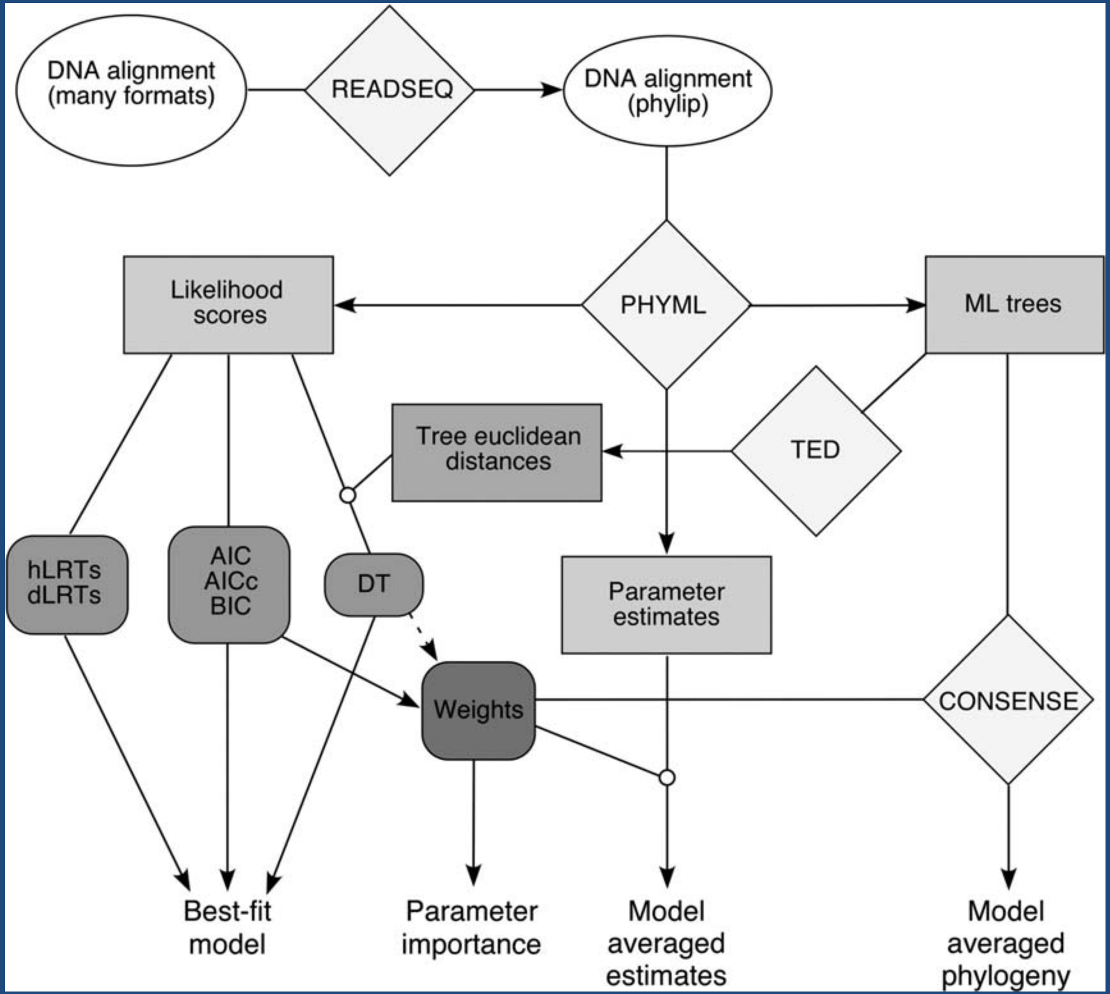
Table 1
Substitution Models Available in jModelTest

Model[a–c]	Free Parameters	Base Frequencies	Substitution Rates	Substitution Code
JC	k	Equal	AC = AG = AT = CG = CT = GT	000000
F81	$k + 3$	Unequal	AC = AG = AT = CG = CT = GT	000000
K80	$k + 1$	Equal	AC = AT = CG = GT, AG = CT	010010
HKY	$k + 4$	Unequal	AC = AT = CG = GT, AG = CT	010010
TrNe	$k + 2$	Equal	AC = AT = CG = GT, AG, CT	010020
TrN	$k + 5$	Unequal	AC = AT = CG = GT, AG, CT	010020
TPM1	$k + 2$	Equal	AC = GT, AT = CG, AG = CT	012210
TPM1u	$k + 5$	Unequal	AC = GT, AT = CG, AG = CT	012210
TPM2	$k + 2$	Equal	AC = AT, CG = GT, AG = CT	010212
TPM2u	$k + 5$	Unequal	AC = AT, CG = GT, AG = CT	010212
TPM3	$k + 2$	Equal	AC = CG, AT = GT, AG = CT	012012
TPM3u	$k + 5$	Unequal	AC = CG, AT = GT, AG = CT	012012
TIM1e	$k + 3$	Equal	AC = GT, AT = CG, AG, CT	012230
TIM1	$k + 6$	Unequal	AC = GT, AT = CG, AG, CT	012230
TIM2e	$k + 3$	Equal	AC = AT, CG = GT, AG, CT	010232
TIM2	$k + 6$	Unequal	AC = AT, CG = GT, AG, CT	010232
TIM3e	$k + 3$	Equal	AC = CG, AT = GT, AG, CT	012032
TIM3	$k + 6$	Unequal	AC = CG, AT = GT, AG, CT	012032
TVMe	$k + 4$	Equal	AC, AT, CG, GT, AG = CT	012314
TVM	$k + 7$	Unequal	AC, AT, CG, GT, AG = CT	012314
SYM	$k + 5$	Equal	AC, AG, AT, CG, CT, GT	012345
GTR	$k + 8$	Unequal	AC, AG, AT, CG, CT, GT	012345

Posada D. 2008. Molecular Biology and Evolution 25: 1253-1256.

Maximum likelihood approach: selecting a model of substitution

jModelTest pipeline.



Posada D. 2008. Molecular Biology and Evolution 25: 1253-1256.

Phylogenetic analysis: methods

Distance

MEGA: Molecular Evolutionary Genetics Analysis

(Tamura et al. Molecular Biology and Evolution 2011)

Maximum parsimony

PAUP: Phylogenetic Analysis Using Parsimony

(Swofford, D. L. 2001)

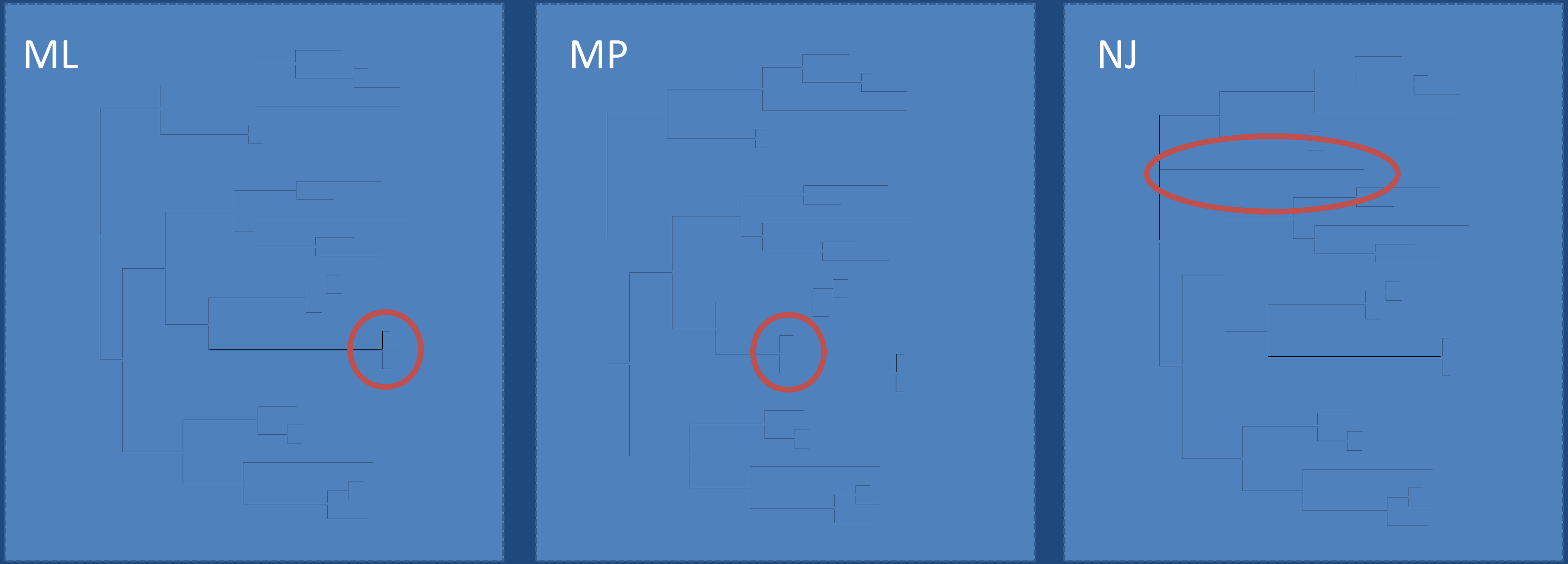
Maximum likelihood

RAxML: Randomized Axelerated Maximum Likelihood

(Stamatakis, A. Bioinformatics 2006)

$\Rightarrow$Consistency !

Comparing tree topologies



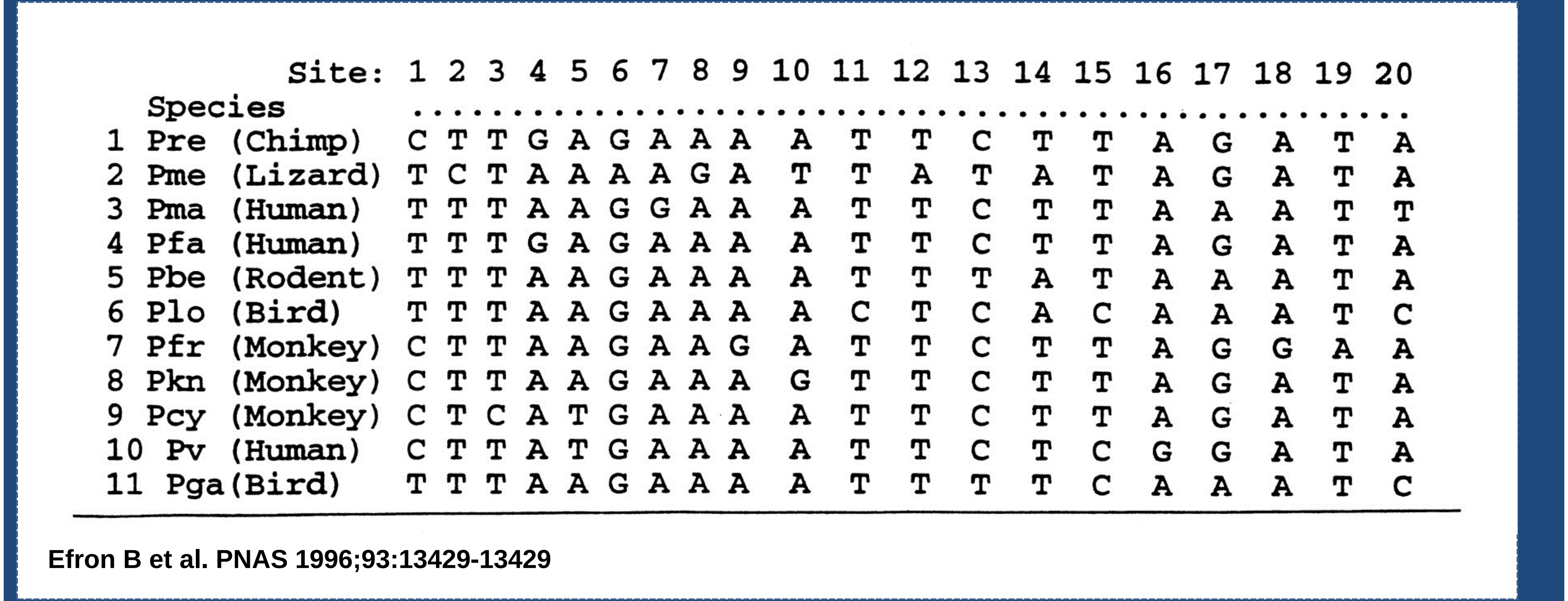
Shimodaira-Hasegawa test of alternative phylogenetic hypotheses (SH test)
(Shimodaira and Hasegawa, Molecular Biology and Evolution, 1999)

Null hypothesis: all trees are equally good explanation of the data

-> Resampling approach

Reliability of the phylogenetic trees: non-parametric bootstrap

Felsenstein, J. 1985 Evolution



	Site:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Species																				
1	Pre (Chimp)	C	T	T	G	A	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
2	Pme (Lizard)	T	C	T	A	A	A	A	G	A	T	T	A	T	A	T	A	G	A	T	A
3	Pma (Human)	T	T	T	A	A	G	G	A	A	A	T	T	C	T	T	A	A	A	T	T
4	Pfa (Human)	T	T	T	G	A	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
5	Pbe (Rodent)	T	T	T	A	A	G	A	A	A	A	T	T	T	A	T	A	A	A	T	A
6	Plo (Bird)	T	T	T	A	A	G	A	A	A	A	C	T	C	A	C	A	A	A	T	C
7	Pfr (Monkey)	C	T	T	A	A	G	A	A	G	A	T	T	C	T	T	A	G	G	A	A
8	Pkn (Monkey)	C	T	T	A	A	G	A	A	A	G	T	T	C	T	T	A	G	A	T	A
9	Pcy (Monkey)	C	T	C	A	T	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
10	Pv (Human)	C	T	T	A	T	G	A	A	A	A	T	T	C	T	C	G	G	A	T	A
11	Pga (Bird)	T	T	T	A	A	G	A	A	A	A	T	T	T	T	C	A	A	A	T	C

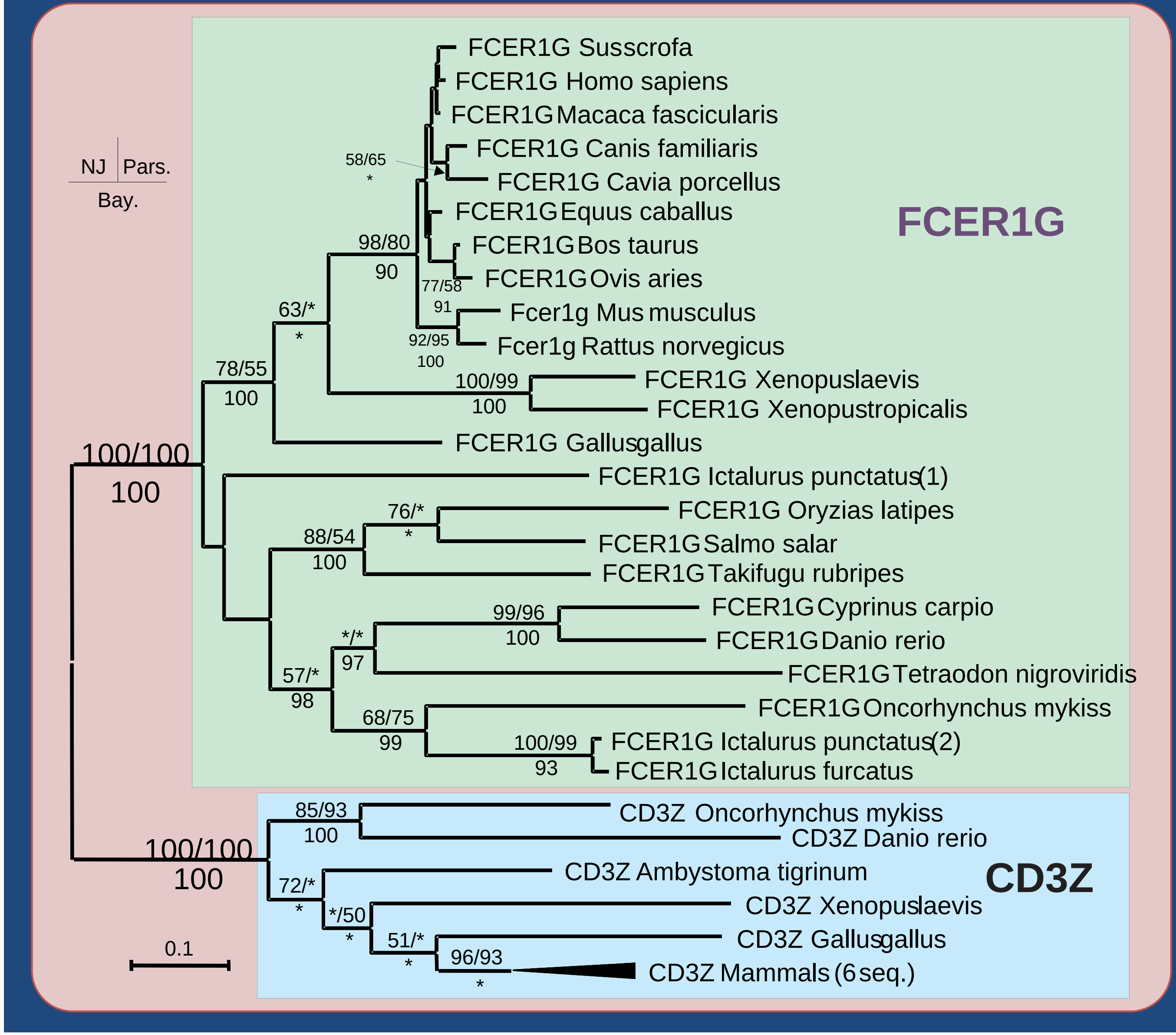
Efron B et al. PNAS 1996;93:13429-13429

⇒ Make *n* pseudo replicates of the original dataset

⇒ Generate a phylogenetic tree for each of the *n* pseudo replicates

⇒ Make a consensus of the *n* phylogenetic trees

Reliability of the phylogenetic trees: non-parametric bootstrap



Bootstrap support:

<50

50-70

70-90

>90

Consistency!

Consistency

Three phylogenetic methods lead to:

- the same well-supported clades

- the same clades but some are poorly-supported with one (or more) method

- very poor support with all methods

- supported divergence between methods

⬇

Investigate dataset

Common problems in phylogenetic analysis: dataset

- Alignment problems

⇒ Improve alignment or restrict data to well-aligned segments
⇒ Analyze domains/exons separately

- Recombination

⇒ Isolate the recombinant sequences and/or segments
⇒ Analyze domains/exons separately

- Functional divergence between paralogs

⇒ Identify and discard the positions

- Sequences with bias in sequence composition

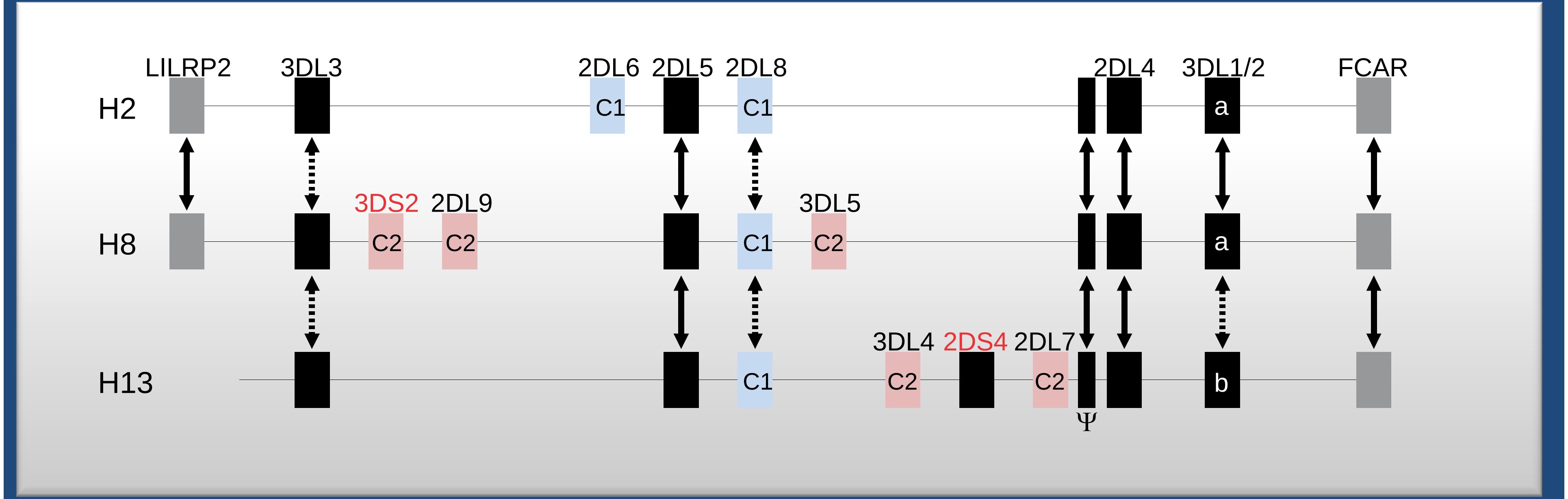
⇒ Discard them or use appropriate methods

- Sequences with long branches / impact on the root of the tree

- Lack of sequences in key taxonomic groups

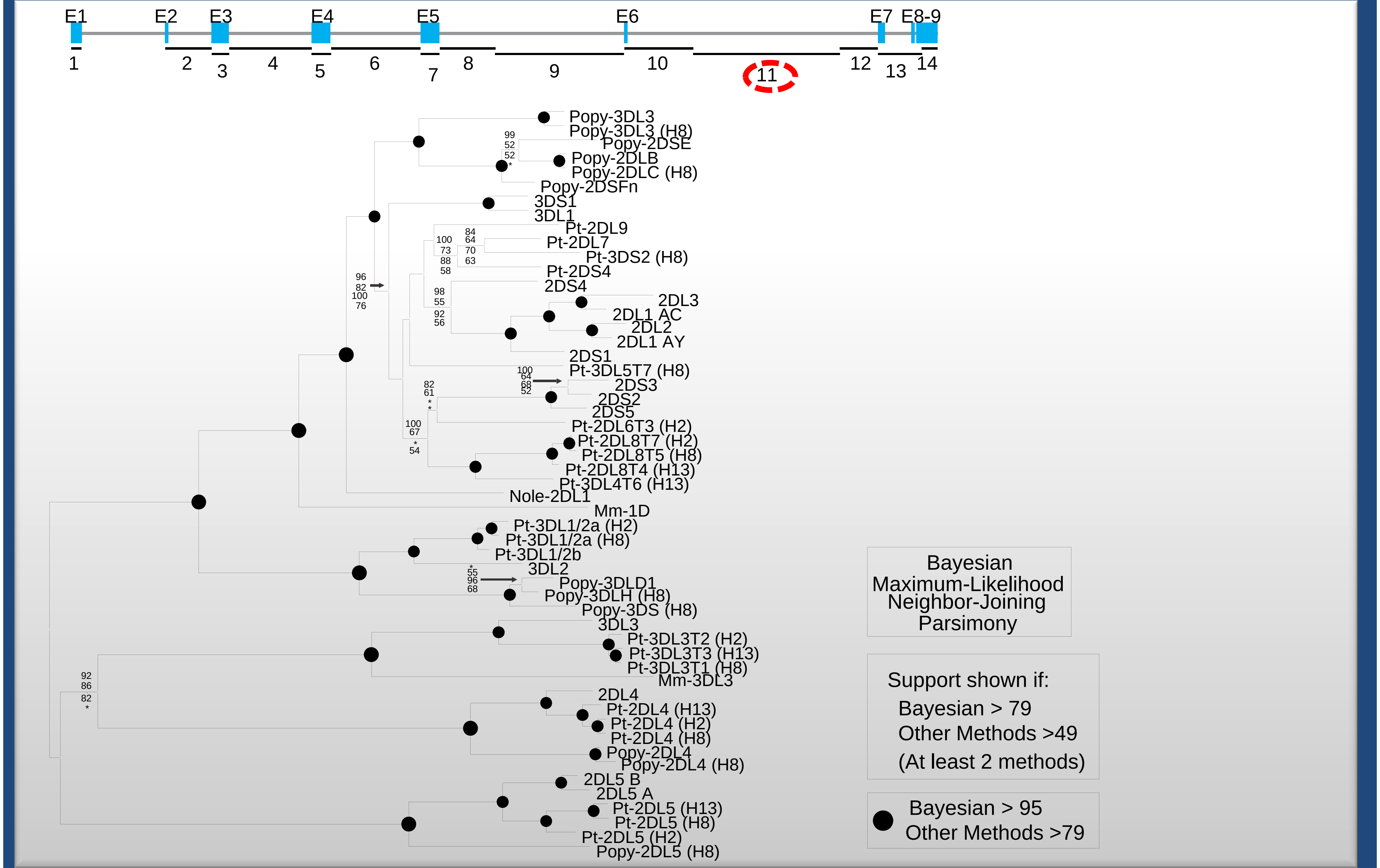
⇒ Obtain more sequences (data mining, exp. approaches)

Example #1: analysis of the *KIR* locus in chimpanzee



Abi-Rached *et al* (PLoS Genetics, 2010)

Example #1: analysis of the *KIR* locus in chimpanzee



Bayesian
Maximum-Likelihood
Neighbor-Joining
Parsimony

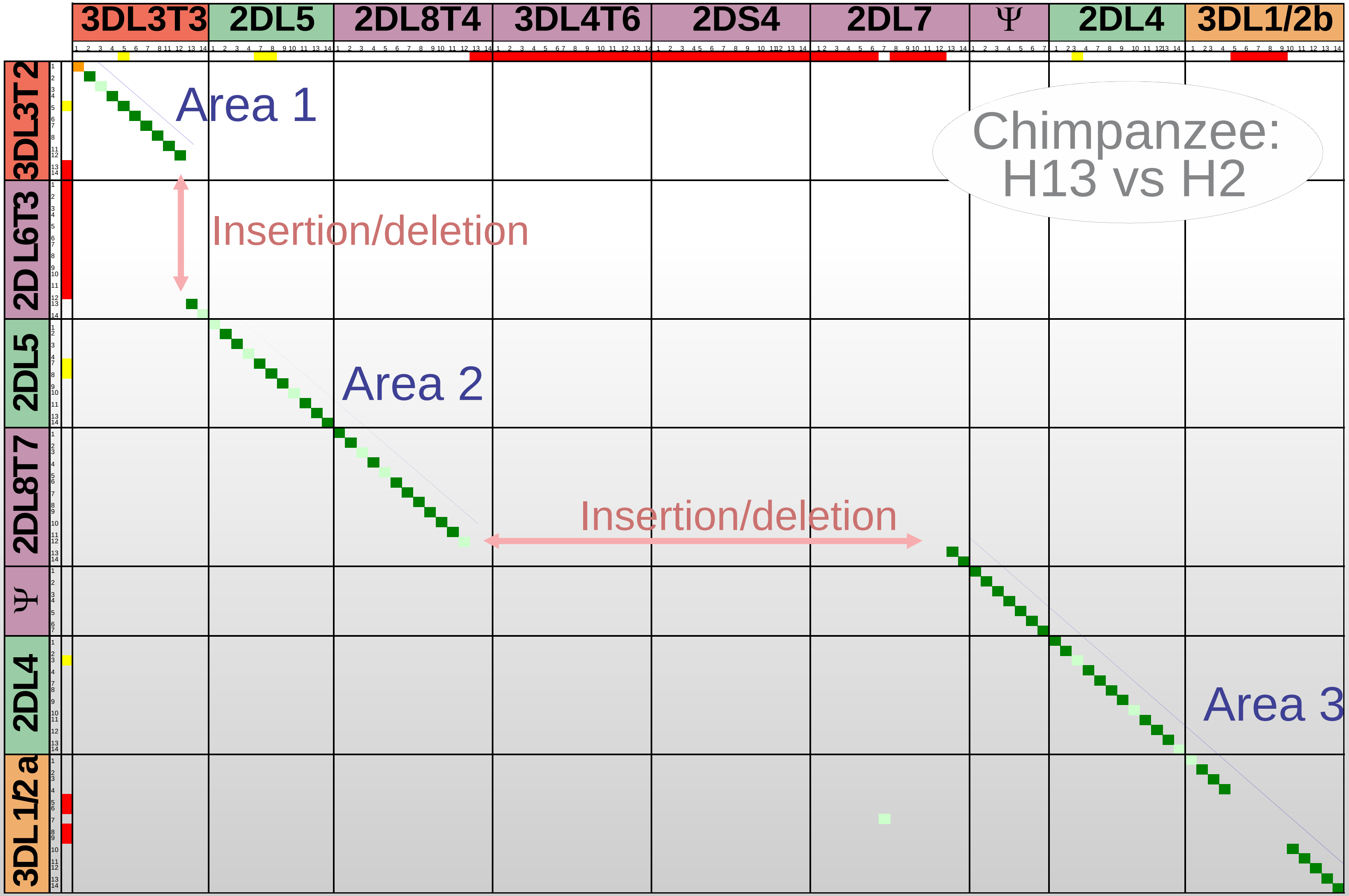
Support shown if:

Bayesian > 79
Other Methods >49
(At least 2 methods)

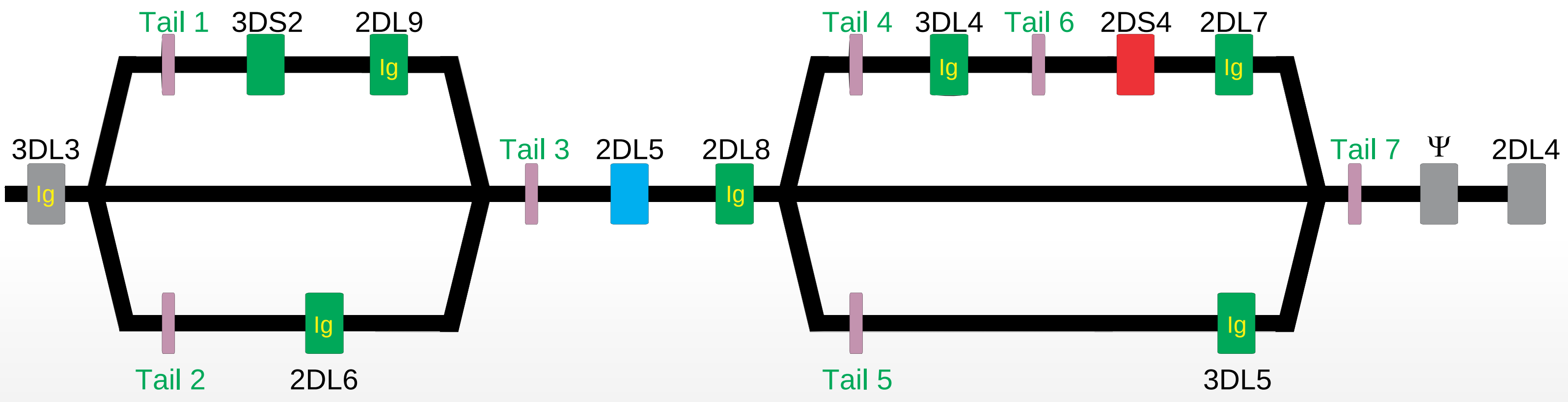
● Bayesian > 95
Other Methods >79

Abi-Rached *et al* (PLoS Genetics, 2010)

Example #1: analysis of the *KIR* locus in chimpanzee

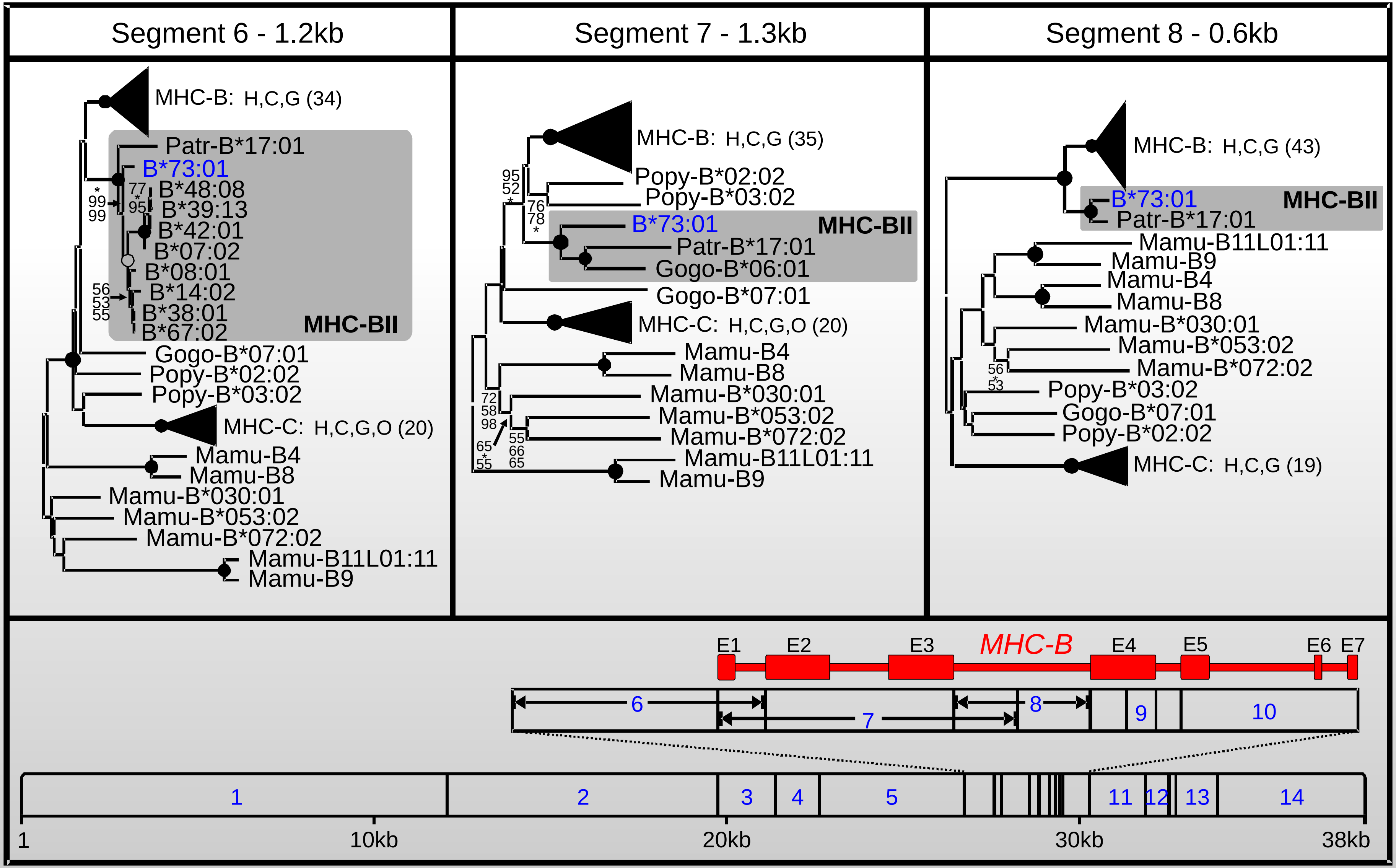


Example #1: analysis of the *KIR* locus in chimpanzee



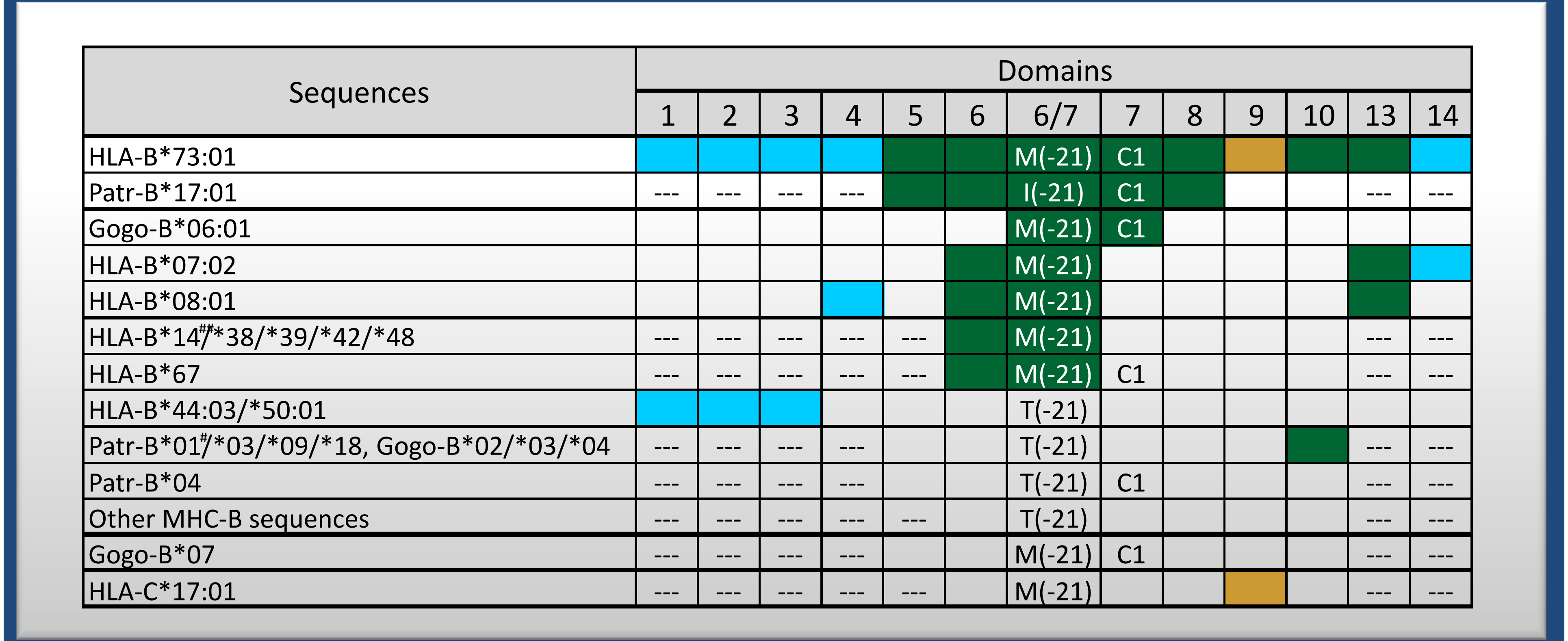
```
               250       260       270       280       290       300       310       320
         .|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|..
2DL1     HRWCSNKKNAAVMDQESAGNRTANSEDSDEQDPQEVTYTQLNHCVFTQRKITRPSQRPKTPPTDIIVYTELPNAESRSKVVSCP
T1(3DL3) ..............
T2(3DL3) ................P..K..VKR.............A..D.R....KNHSPF.EAQDN.NRYQR.HRTSKC
T3(3DL3) ................P.....VSR.............A..............L..........TS.........P........
T4(2DL8) ................P....IVKR.............AH.D.........NP..R........T..........P........
T5(2DL8) .....K..........P....IVKR.............AH.D.........NP..R........T..........P........
T6(3DL4) .C...K..........P....IVKR.............AH.D.........KP...........T..........P....F...
T7(2DL8) .....K..........P.....V.R.............A............SP..E........TS.........P........
                                 CK        ITIM 1        PKC             ITIM 2
```

Example #2: analysis of the *HLA-B*73* haplotype



Abi-Rached *et al* (Science, 2011)

Example #2: analysis of the *HLA-B*73* haplotype

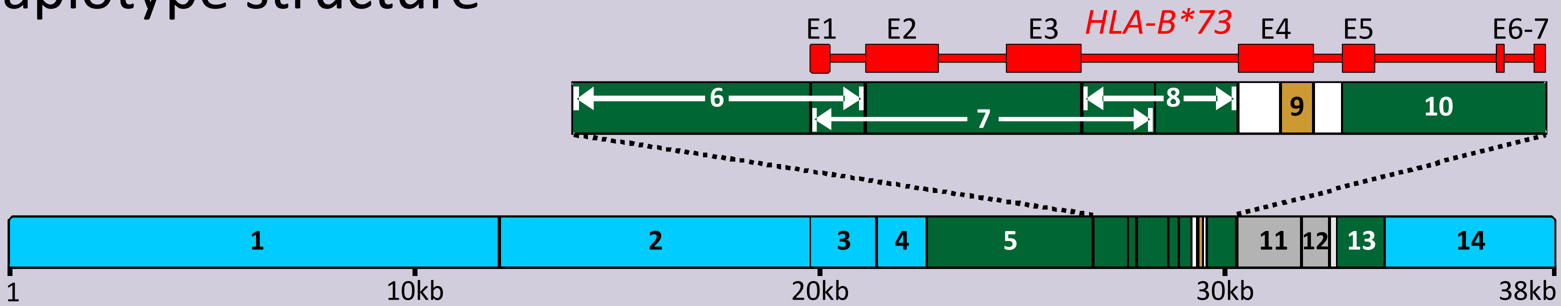


Sequences	Domains												
	1	2	3	4	5	6	6/7	7	8	9	10	13	14
HLA-B*73:01	🟦	🟦	🟦	🟦	🟩	🟩	🟩 M(-21)	🟩 C1	🟩	🟧	🟩	🟩	🟦
Patr-B*17:01	---	---	---	---	🟩	🟩	🟩 I(-21)	🟩 C1	🟩			---	---
Gogo-B*06:01							🟩 M(-21)	🟩 C1					
HLA-B*07:02						🟩	🟩 M(-21)					🟩	🟦
HLA-B*08:01				🟦		🟩	🟩 M(-21)					🟩	
HLA-B*14##/*38/*39/*42/*48	---	---	---	---	---	🟩	🟩 M(-21)					---	---
HLA-B*67	---	---	---	---	---	🟩	🟩 M(-21)	C1				---	---
HLA-B*44:03/*50:01	🟦	🟦	🟦				T(-21)						
Patr-B*01#/*03/*09/*18, Gogo-B*02/*03/*04	---	---	---	---			T(-21)				🟩	---	---
Patr-B*04	---	---	---	---			T(-21)	C1				---	---
Other MHC-B sequences	---	---	---	---	---		T(-21)					---	---
Gogo-B*07	---	---	---	---			M(-21)	C1				---	---
HLA-C*17:01	---	---	---	---	---		M(-21)			🟧		---	---

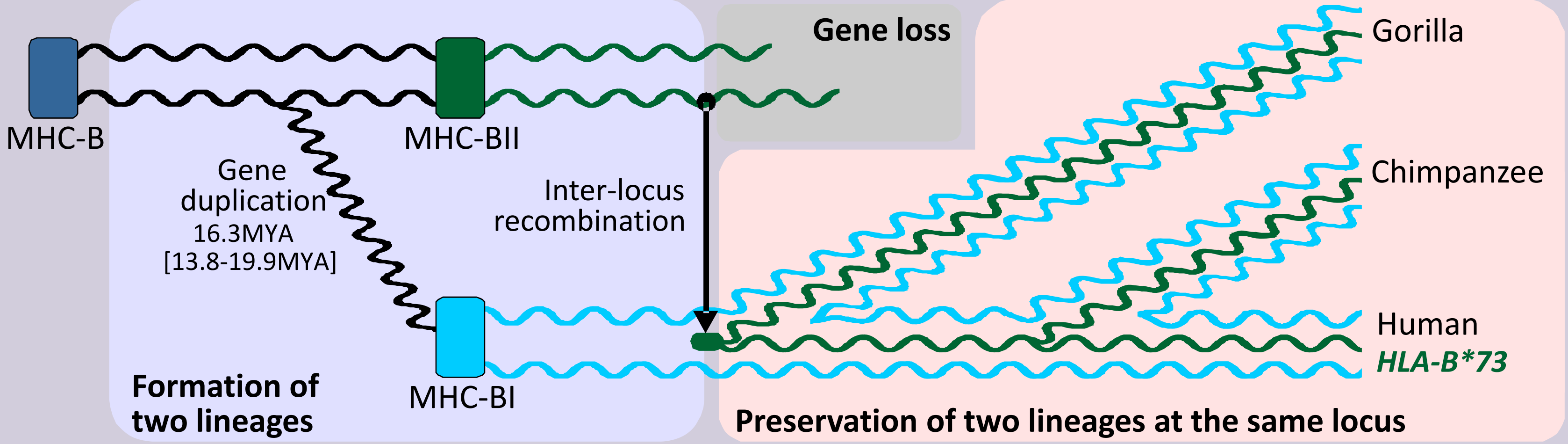
Abi-Rached *et al* (Science, 2011)

Example #2: analysis of the *HLA-B*73* haplotype

I. Haplotype structure



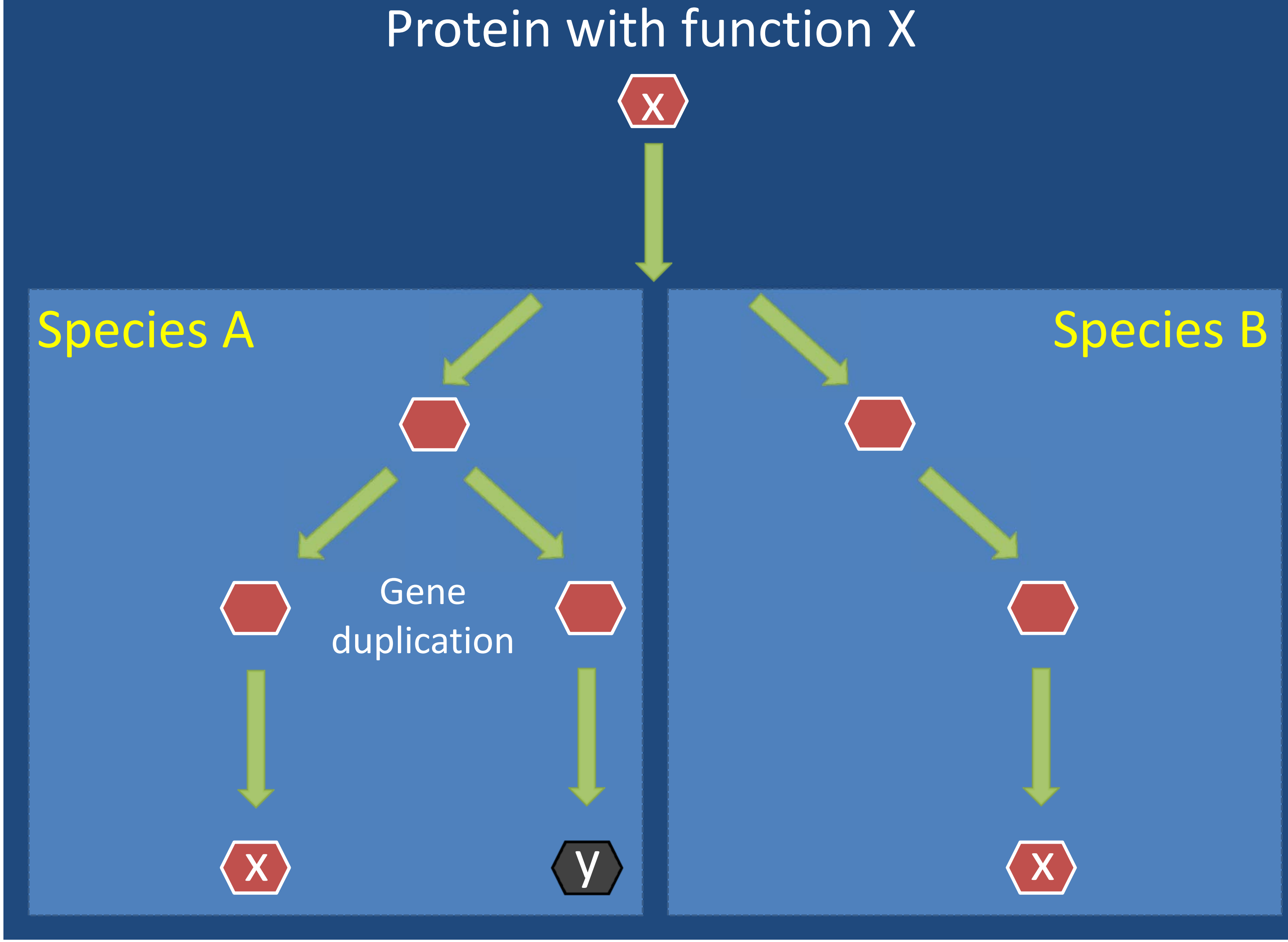
II. Phylogeny



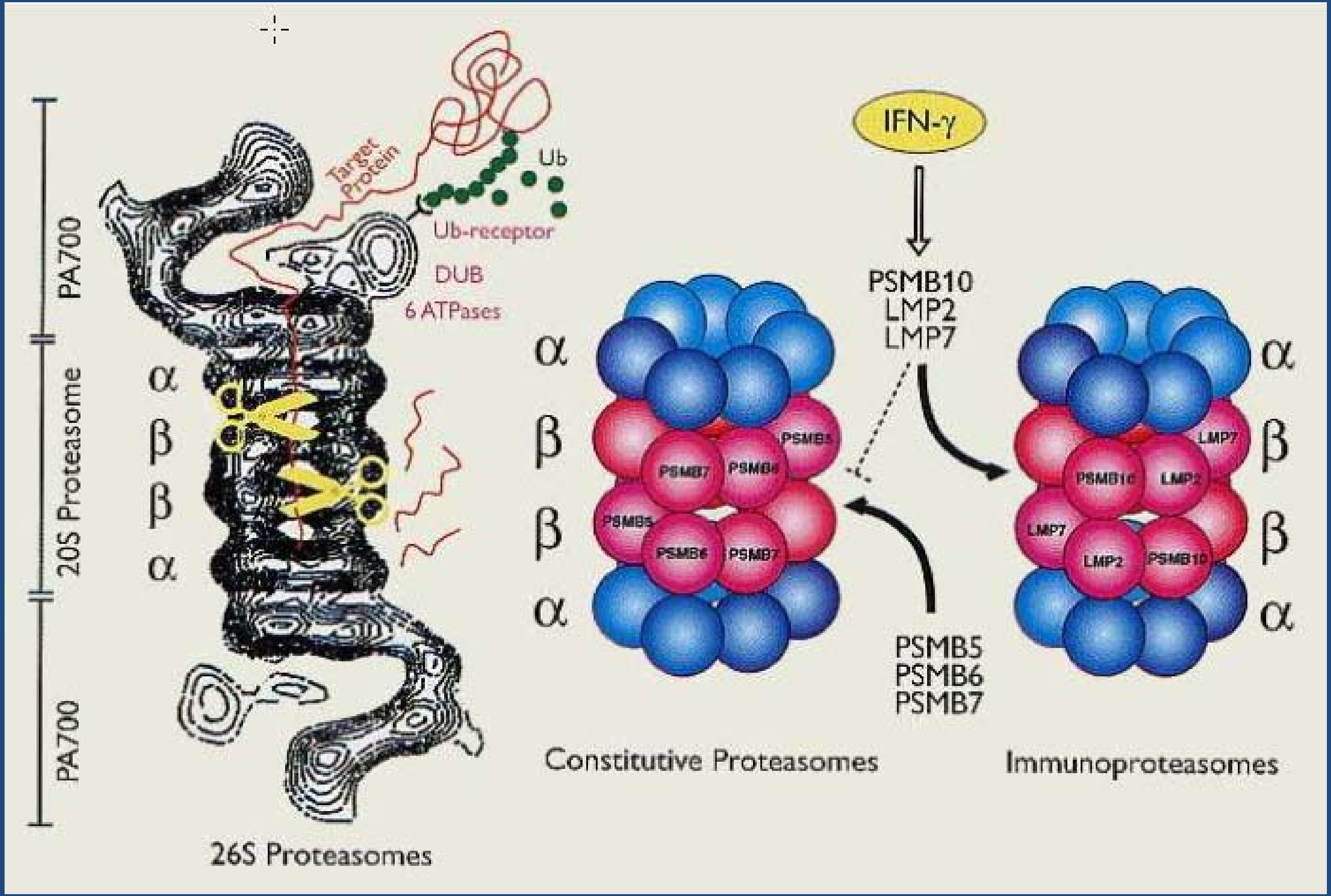
Gene loss
Gorilla
MHC-B
MHC-BII
Chimpanzee
Gene duplication
16.3MYA
[13.8-19.9MYA]
Inter-locus recombination
Human
*HLA-B*73*
MHC-BI
Formation of two lineages
Preservation of two lineages at the same locus

Abi-Rached *et al* (Science, 2011)

Example #3: functional divergence

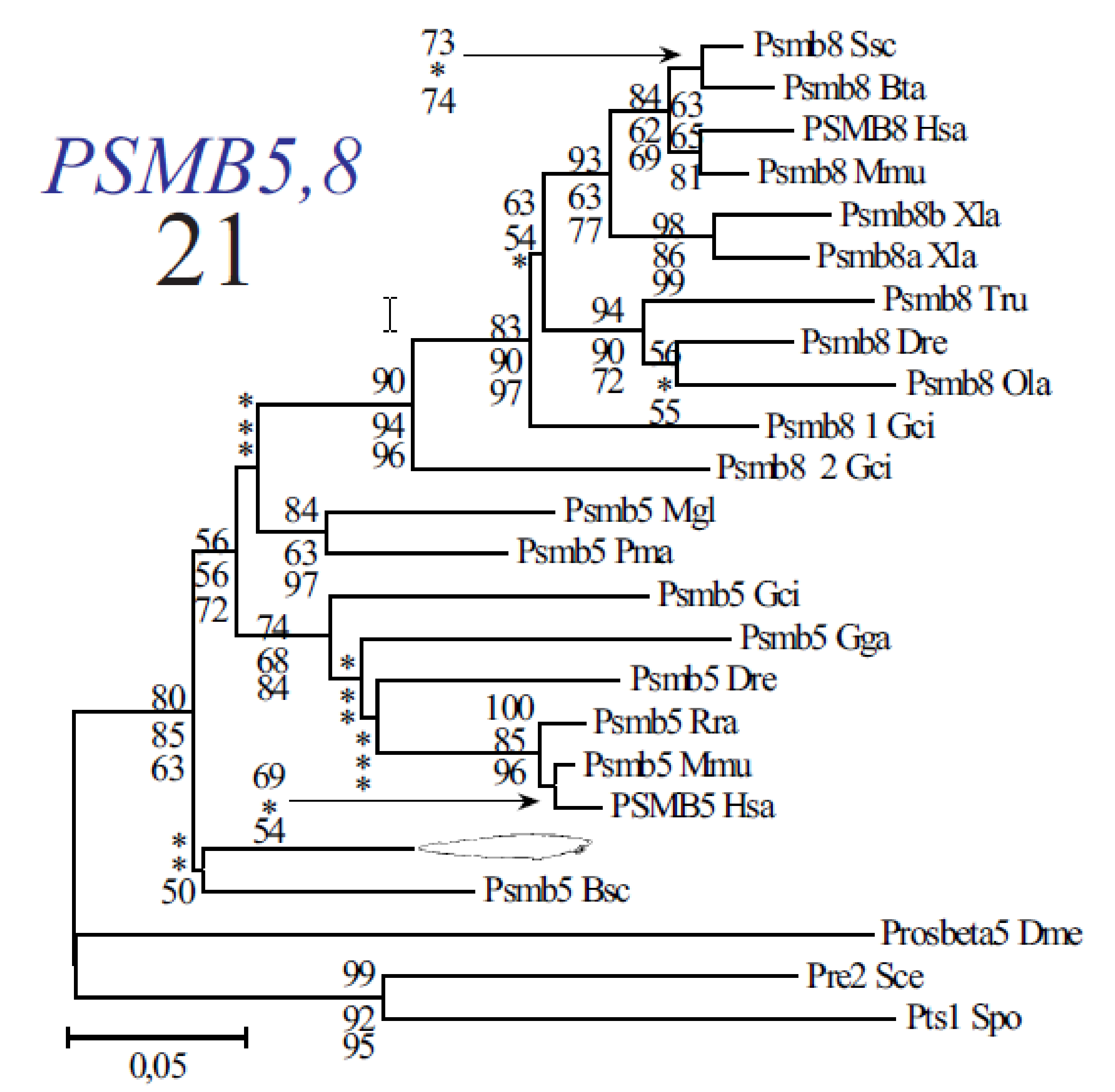


Example #3: functional divergence



Tanaka et Kasahara, 1998

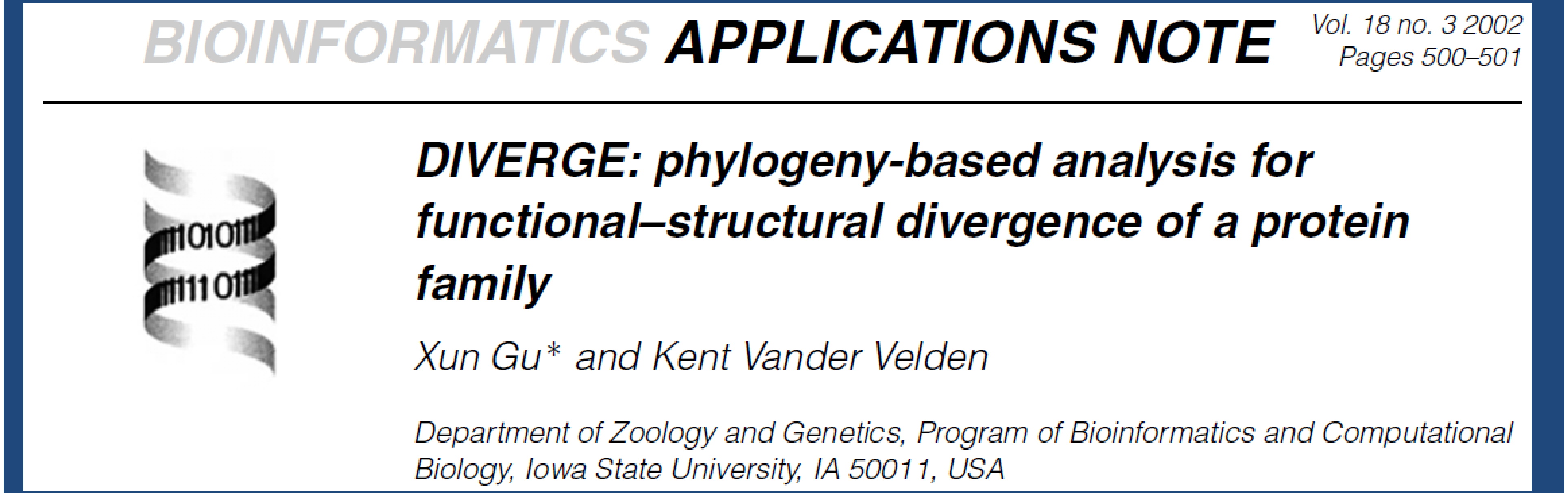
Example #3: functional divergence



Amphioxus
Cephalochordate

Abi-Rached *et al* (Nature Genetics, 2002)

Functional divergence



BIOINFORMATICS **APPLICATIONS NOTE**

Vol. 18 no. 3 2002
Pages 500–501

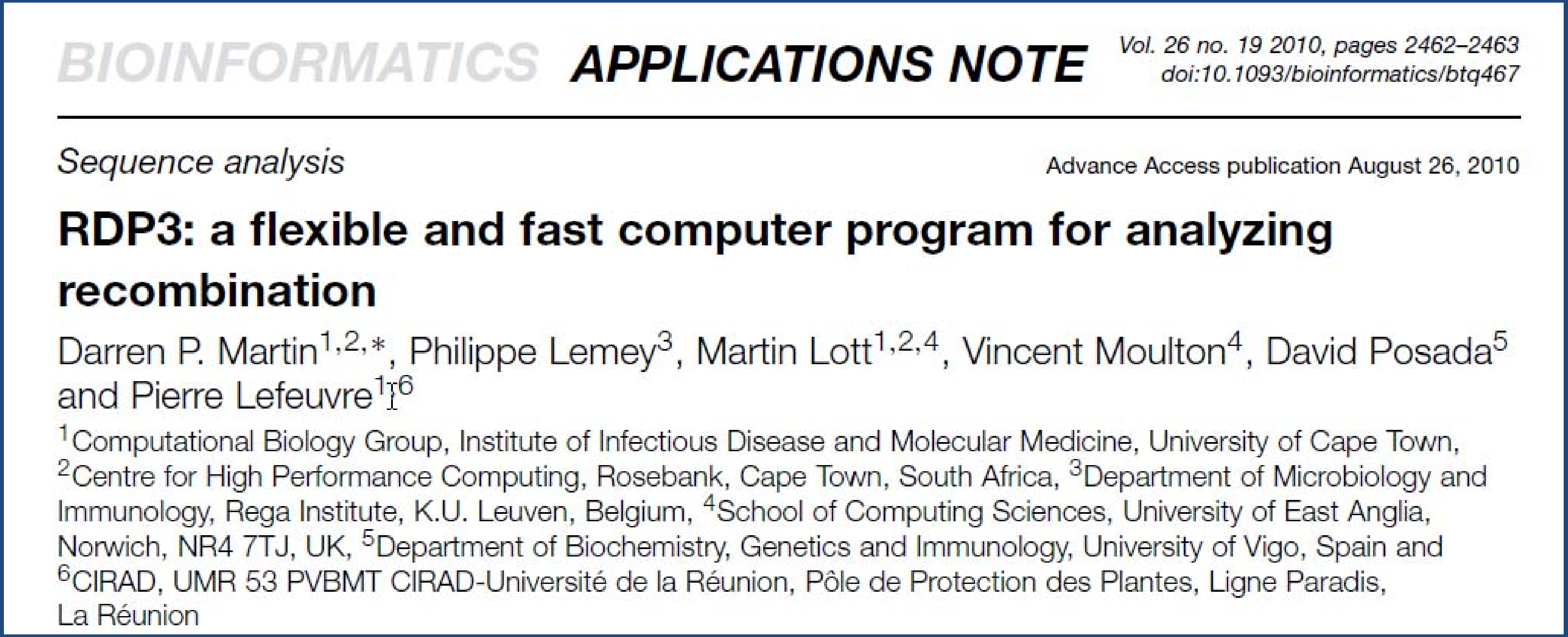
DIVERGE: phylogeny-based analysis for functional–structural divergence of a protein family

Xun Gu and Kent Vander Velden*

Department of Zoology and Genetics, Program of Bioinformatics and Computational Biology, Iowa State University, IA 50011, USA

Gu and Vander Velden, Bioinformatics. 2002 Mar;18(3):500-1.

Recombination



BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 19 2010, pages 2462–2463
doi:10.1093/bioinformatics/btq467

Sequence analysis

Advance Access publication August 26, 2010

RDP3: a flexible and fast computer program for analyzing recombination

Darren P. Martin[1,2,*], Philippe Lemey[3], Martin Lott[1,2,4], Vincent Moulton[4], David Posada[5] and Pierre Lefeuvre[1,6]

[1] Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, [2] Centre for High Performance Computing, Rosebank, Cape Town, South Africa, [3] Department of Microbiology and Immunology, Rega Institute, K.U. Leuven, Belgium, [4] School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK, [5] Department of Biochemistry, Genetics and Immunology, University of Vigo, Spain and [6] CIRAD, UMR 53 PVBMT CIRAD-Université de la Réunion, Pôle de Protection des Plantes, Ligne Paradis, La Réunion

Martin DP et al (2010). Bioinformatics 26, 2462-2463.