

# Maximum likelihood phylogenies inference with codon models

Maria Anisimova  
ETH Zürich

MCEB 2013, France

**ETH**

Eidgenössische Technische Hochschule Zürich  
Swiss Federal Institute of Technology Zurich



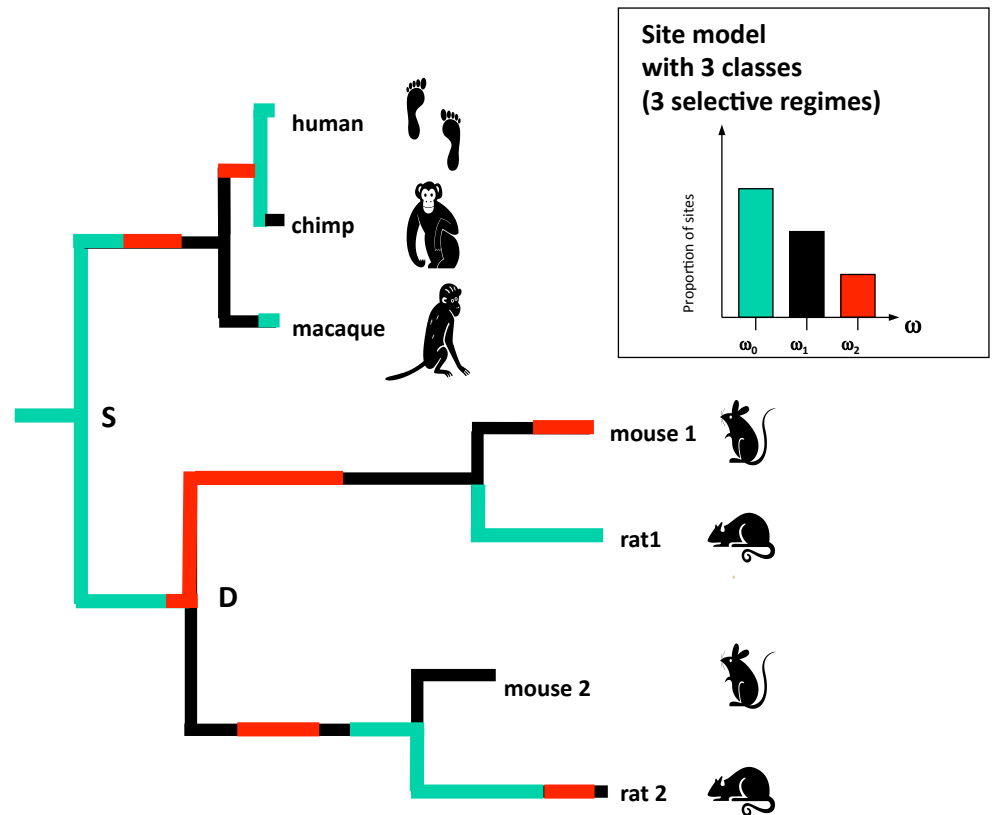
# Why codon models?

Better models for protein-coding genes:

- Incorporate structure of **genetic code** (unlike DNA models)
- Include **synonymous** substitutions (unlike AA models)
- Allow to model **selection** on protein (unlike AA & DNA models)

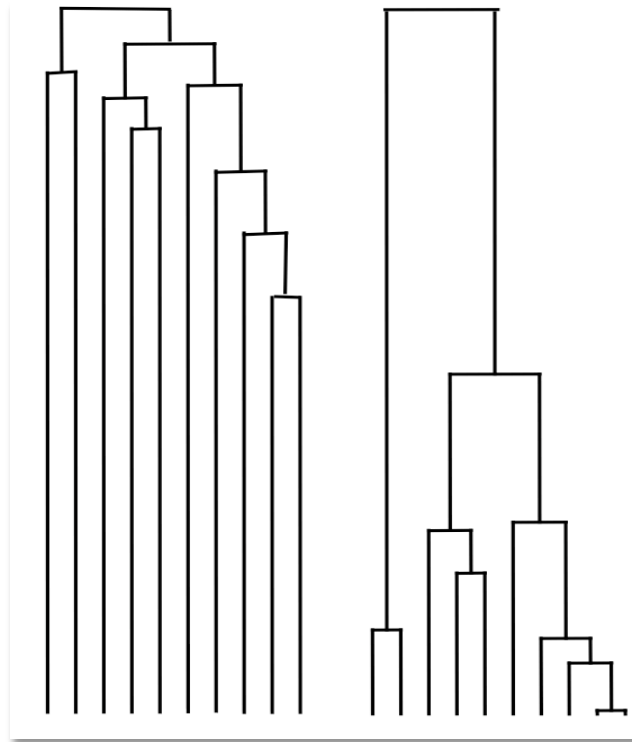
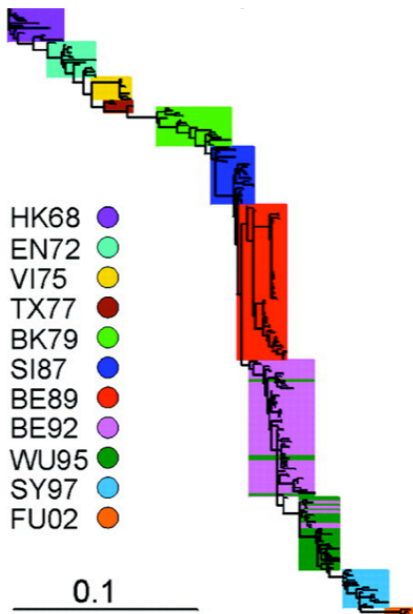
# The many faces of codon models

- Detecting selection
- Changes in time and space
- Studies of codon bias
- Predicting codon regions
- Inferring phylogenies
- Dating speciation events
- Ancestral reconstruction
- Improved alignment
- Inferring gene features (phyloHMMs, netHMMS)
- Simulation of protein-coding genes

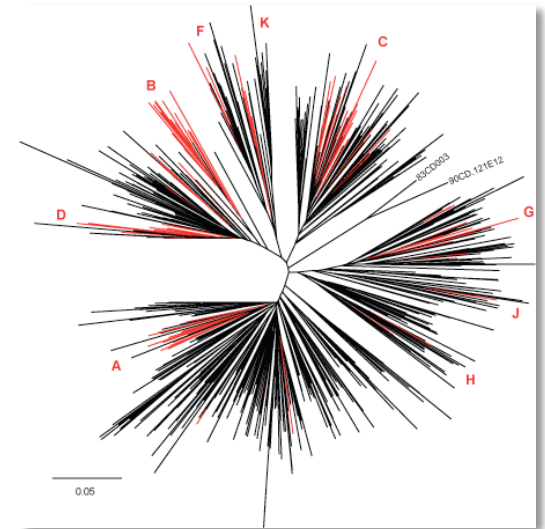


# Selection affects tree shape

Antigenic / genetic evolution of influenza A  
Smith et al. 2004 Science



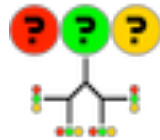
Positive selection No selection



Phylogeny of HIV-1 subtypes  
Archer and Robertson 2007, AIDS

Effect of selection on the topology of genealogical trees

Leonardo P. Maia, Alexandre Colato, José F. Fontanari\*



# CodonPhyML

## Hundreds of codon models

- Parametric, empirical, semi-parametric
- Comparable likelihoods for AA, DNA & codon models (Seo, Kishino)

## Inheritance from PhyML

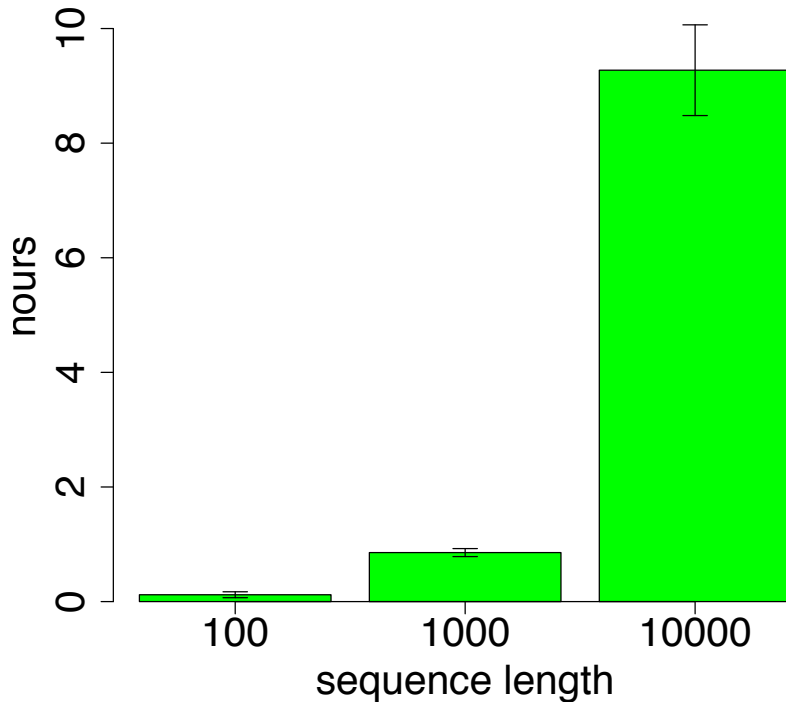
- Fast tree search heuristics (NNI + SPR)
- High accuracy, nice convergence properties
- Fast branch support (Anisimova *et al* 2011, Syst Biol)

## High performance computing

- BLAS, LAPACK, OpenMP
- Heuristic using  $\exp(Qt)$  via Taylor
- Blocking heuristic (FixQ)

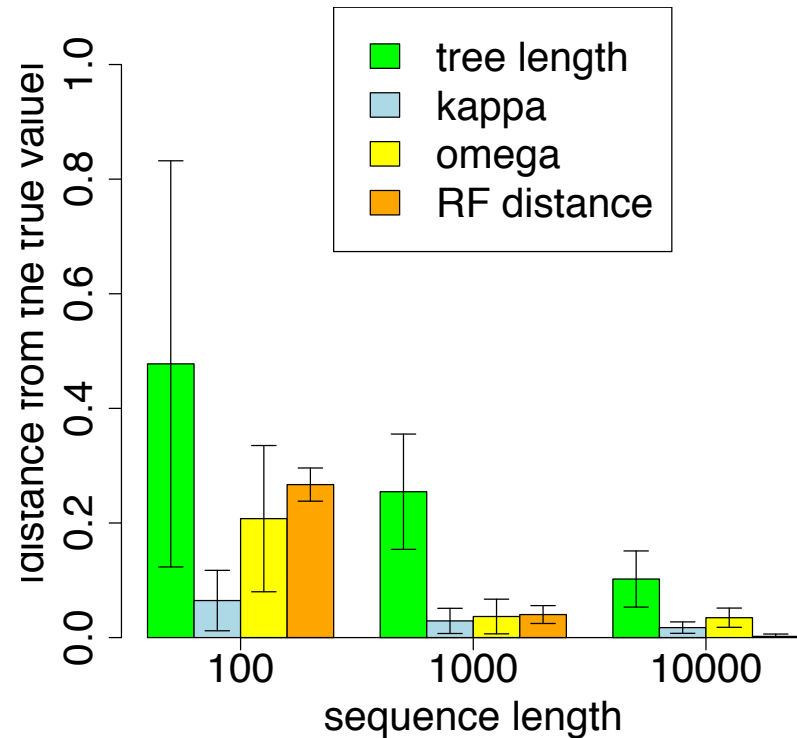
# CodonPhyML: Simulations

100 Taxa – NNI



100 taxa 1000 codons  
in < 1 hour

100 Taxa – NNI



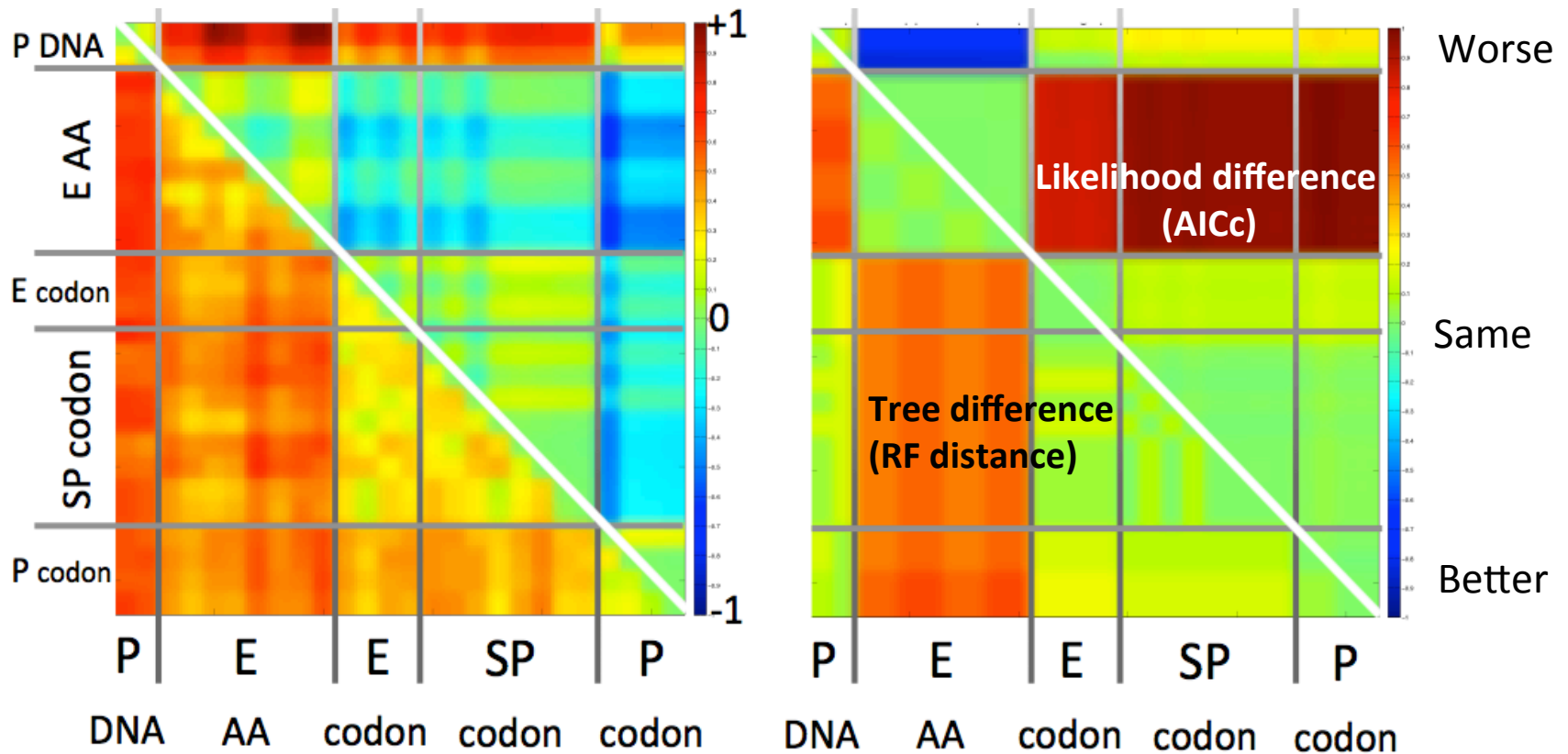
Nice convergence  
properties

# CodonPhyML:

## Model comparison on real data

Model types: DNA, AA, codon

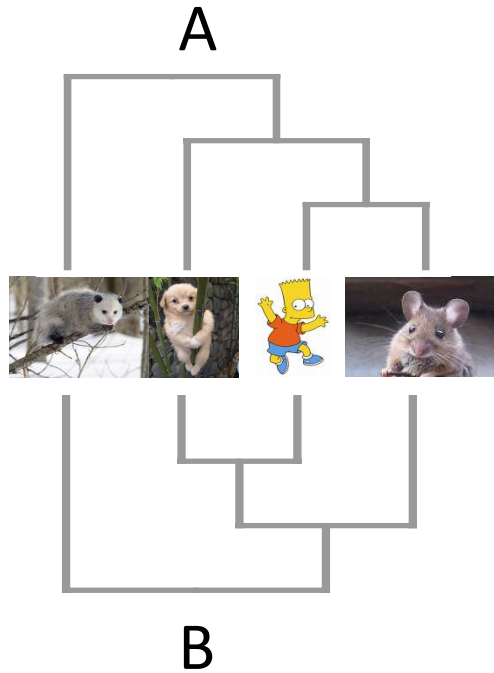
E = empirical, SP = semi-parametric, P = parametric



# Human-mouse or human-dog?

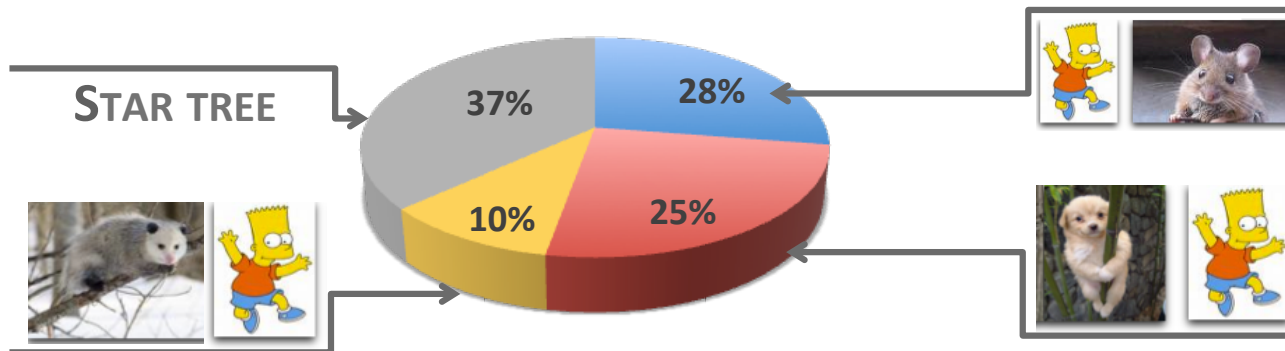
## Tree B: Primates with Carnivores

- Janke et al (1994)
- Reyes et al (2000)
- Arnason et al (2002)
- Misawa & Janke (2003)
- Jorgesen et al (2005)
- Kullberg et al (2006)
- Cannarozzi et al (2007)
- Huttley et al (2007)

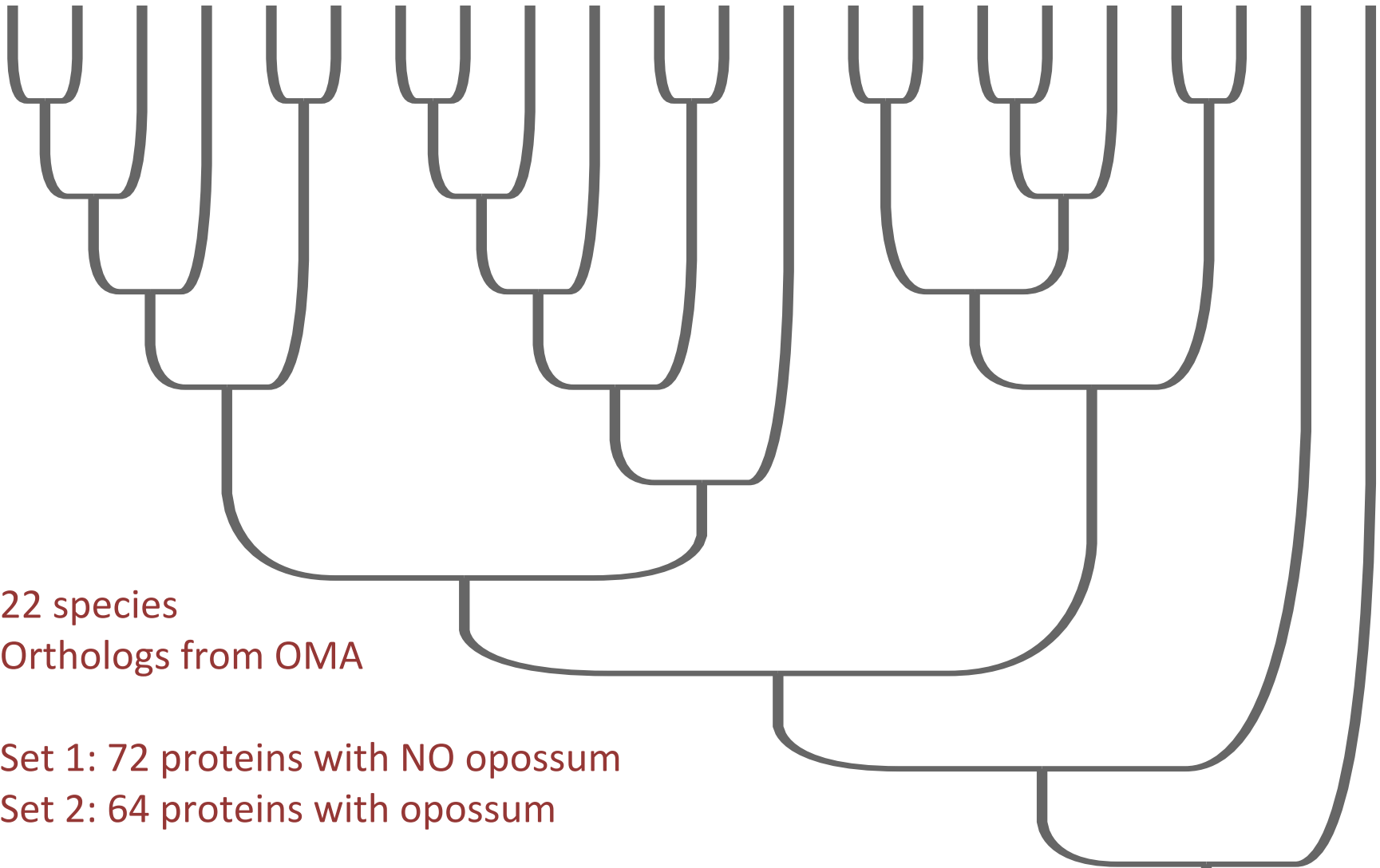
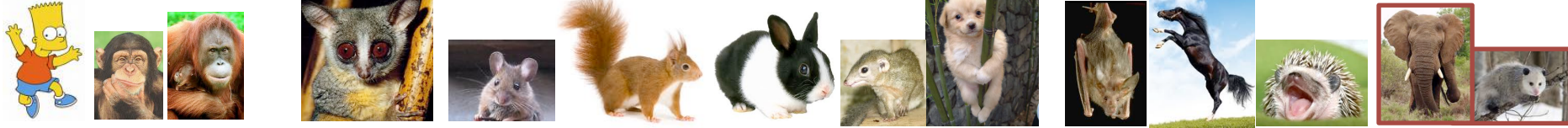


## Tree A: Primates with Rodents

- Shoshani & McKenna (1998)
- Murphy et al (2001)
- Lin et.al (2002)
- Amrine-Madsen et al (2003)
- Thomas et.al (2003)
- Bashir et.al (2006)
- Lunter (2007)







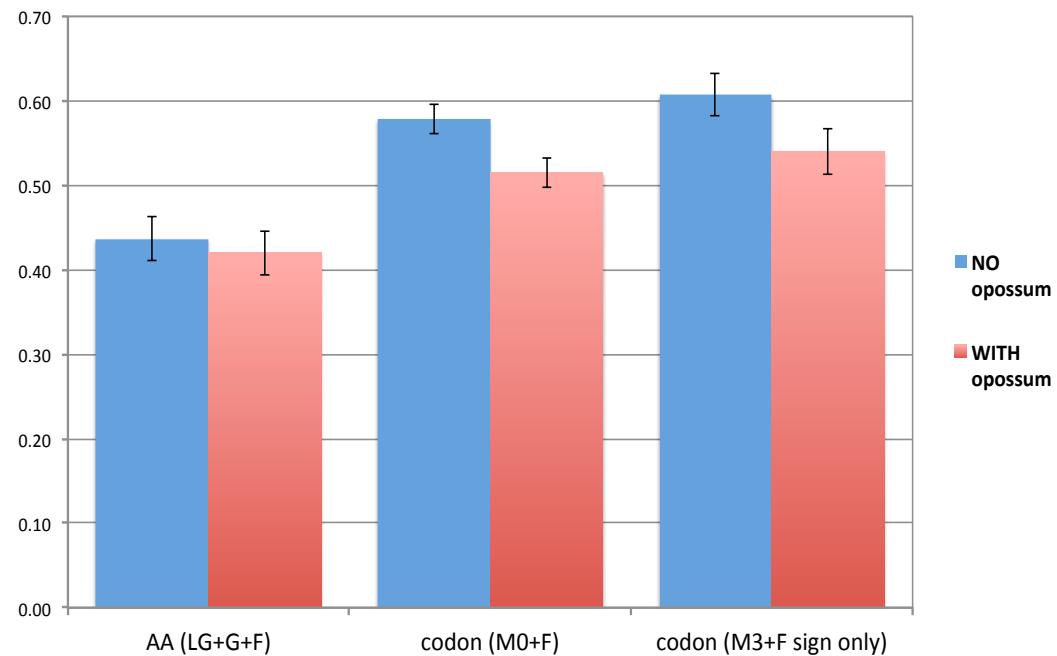
22 species  
Orthologs from OMA

Set 1: 72 proteins with NO opossum  
Set 2: 64 proteins with opossum

# Comparison by distance to reference tree

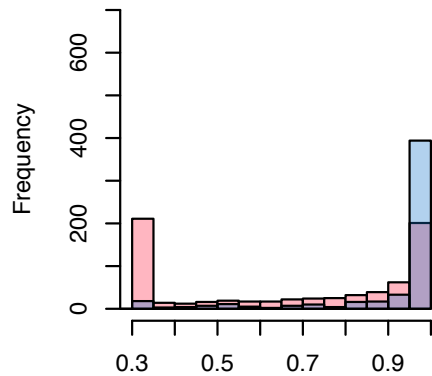
	Best fit (NO opossum)	Best fit (With oposs)
Codon M0	<b>70/72</b>	<b>62/64</b>
AA LG+ $\Gamma$	0/72	0/64
DNA GTR+ $\Gamma$	2/72	2/64

Frequency of splits matching Ref. Tree

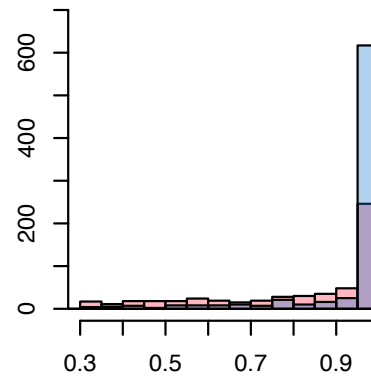


# “Correct” vs “Incorrect” partitions

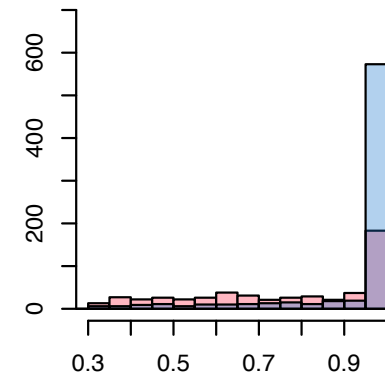
LG+G+F



M0



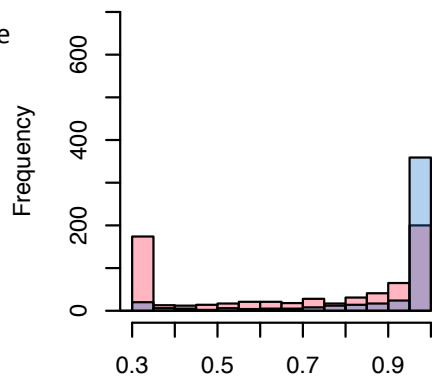
M3



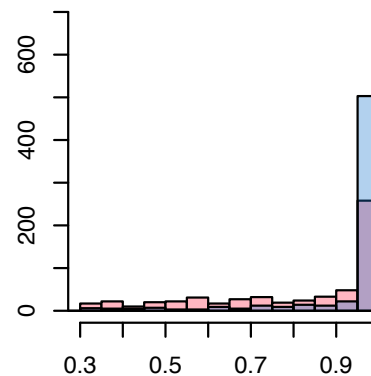
■ Splits  
matching  
Ref Tree

■ Splits  
not found  
in Ref Tree

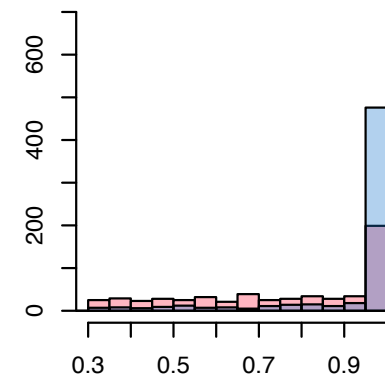
LG+G+F with opposum



M0 with opposum



M3 with opposum

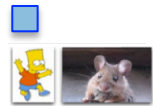


aBayes support

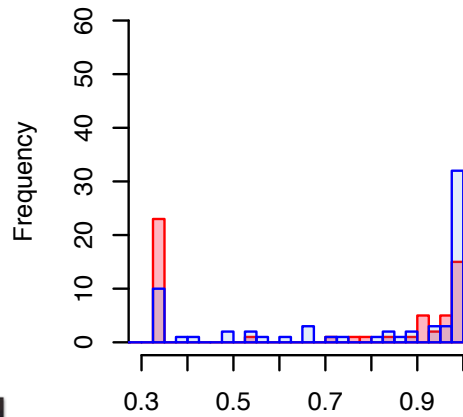
aBayes support

aBayes support

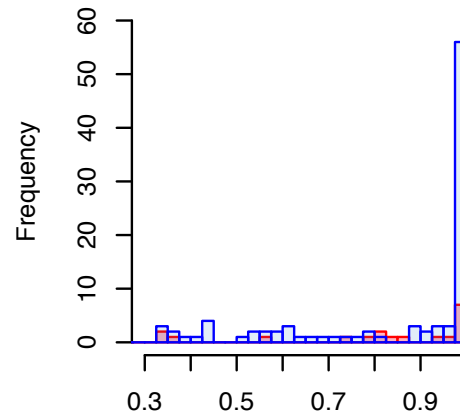
# Support for human-mouse



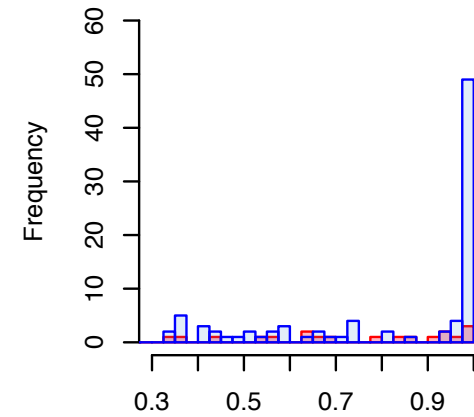
**LG+G+F**



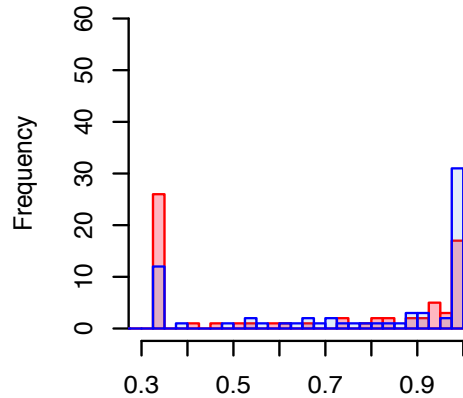
**M0**



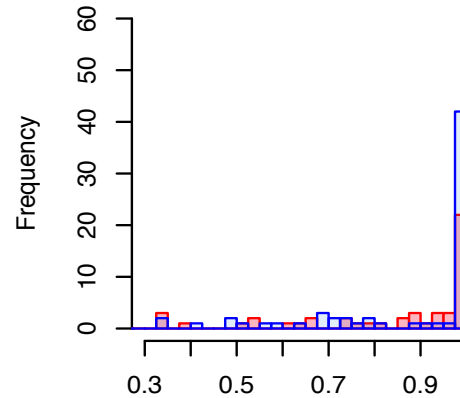
**M3**



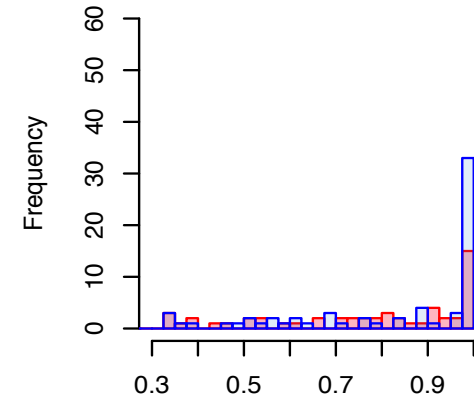
**LG+G+F with opossum**



**M0 with opossum**



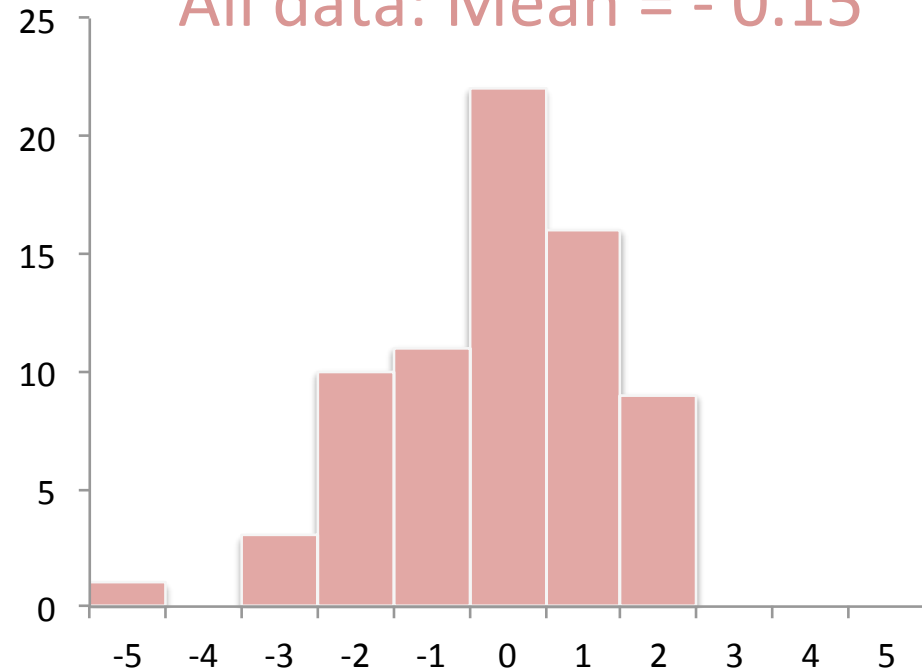
**M3 with opossum**



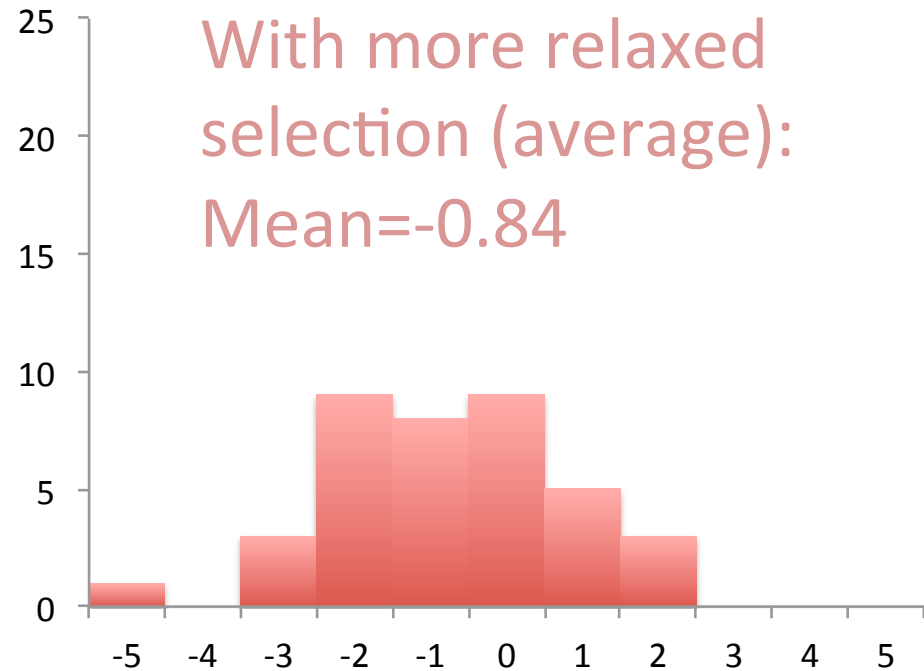
# Tree shape statistics

Shapes of inferred codon trees are less variable  
Mean #cherry under codon model is closer to “true” value,  
but not significantly different: 6.0 (aa) vs 6.3 (codon)

All data: Mean = -0.15



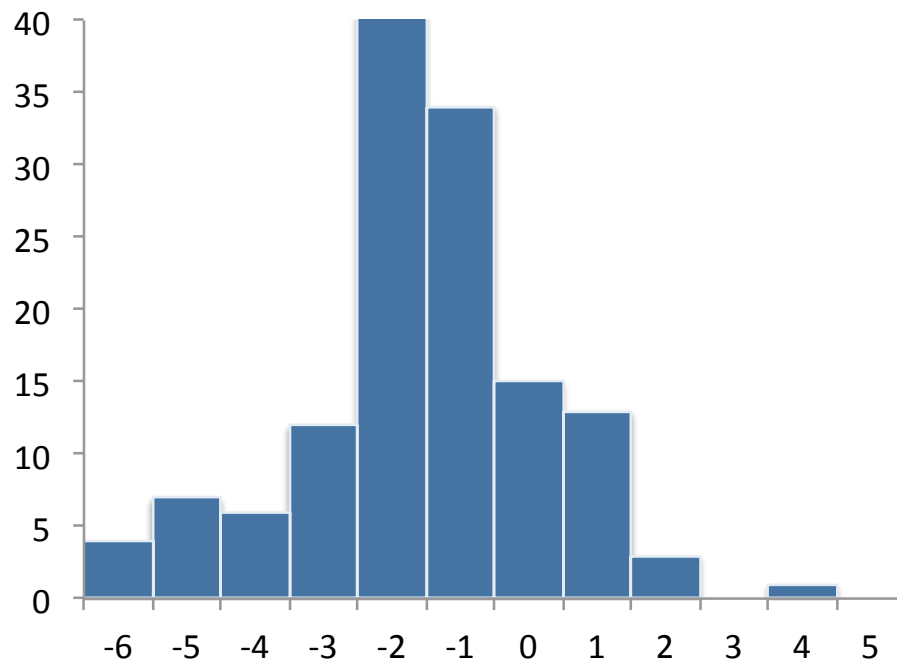
With more relaxed selection (average):  
Mean=-0.84



Difference in #cherries between AA & codon models

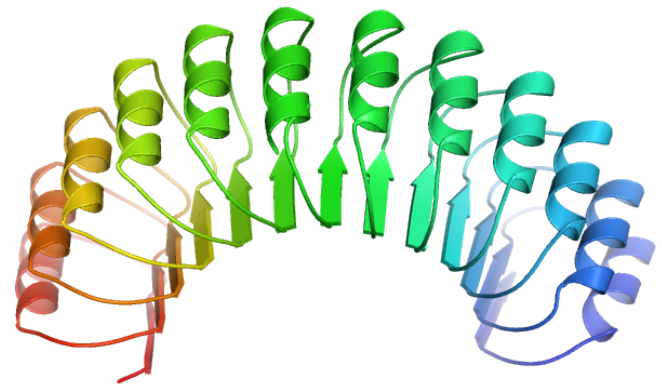
# Tree shape statistics (other data!)

All data:  
Mean = - 1.51



Difference in #cherries between codon & aa models

Bacterial GALA (type III effectors) acquired from host plants by LGT: residues under positive selection are found on the convex side of horse-shoe & involved in binding





# Acknowledgements



## Enthusiasm & togetherness

Computational Biochemistry Research Group

## Contributions

← Marcelo Zanetti

Stefan Zoller

Manuel Gil

## Discussions



Ziheng Yang

Olivier Gascuel



# Acknowledgements

See posters # 28 (Elke) and #29 Adam



## Enthusiasm & togetherness

Computational Biochemistry Research Group

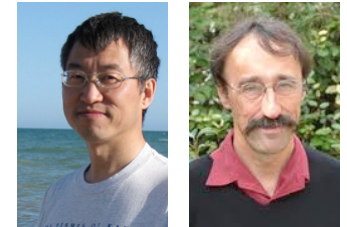
## Contributions

Marcelo Zanetti

Stefan Zoller

Manuel Gil

## Discussions



Ziheng Yang

Olivier Gascuel



Questions?