# GPU computing and the tree of life

Michael P. Cummings
Center for Bioinformatics and Computational Biology
University of Maryland Institute of Advanced Computer Studies

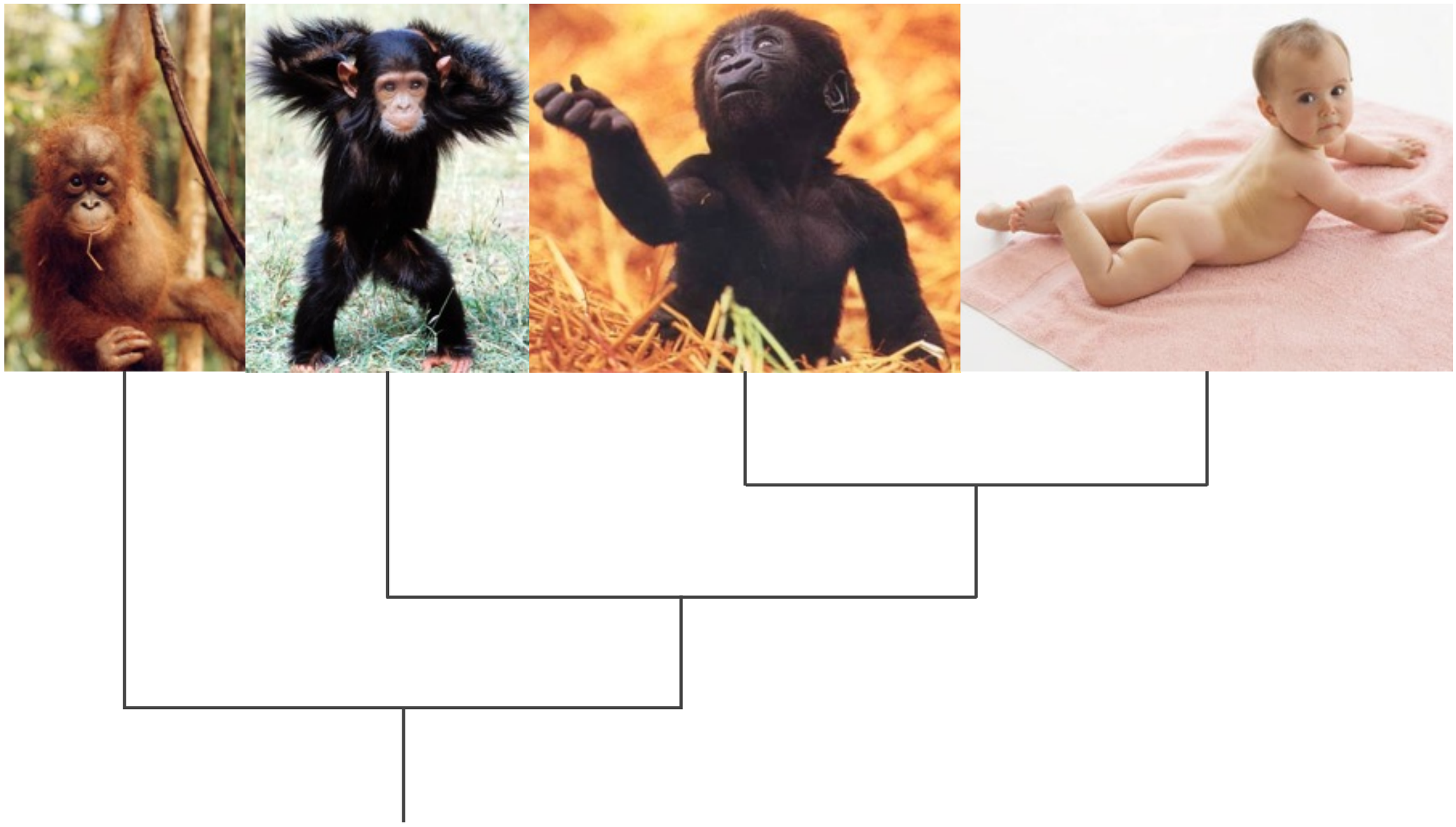
GPU summit
27 October 2014

some domain science context

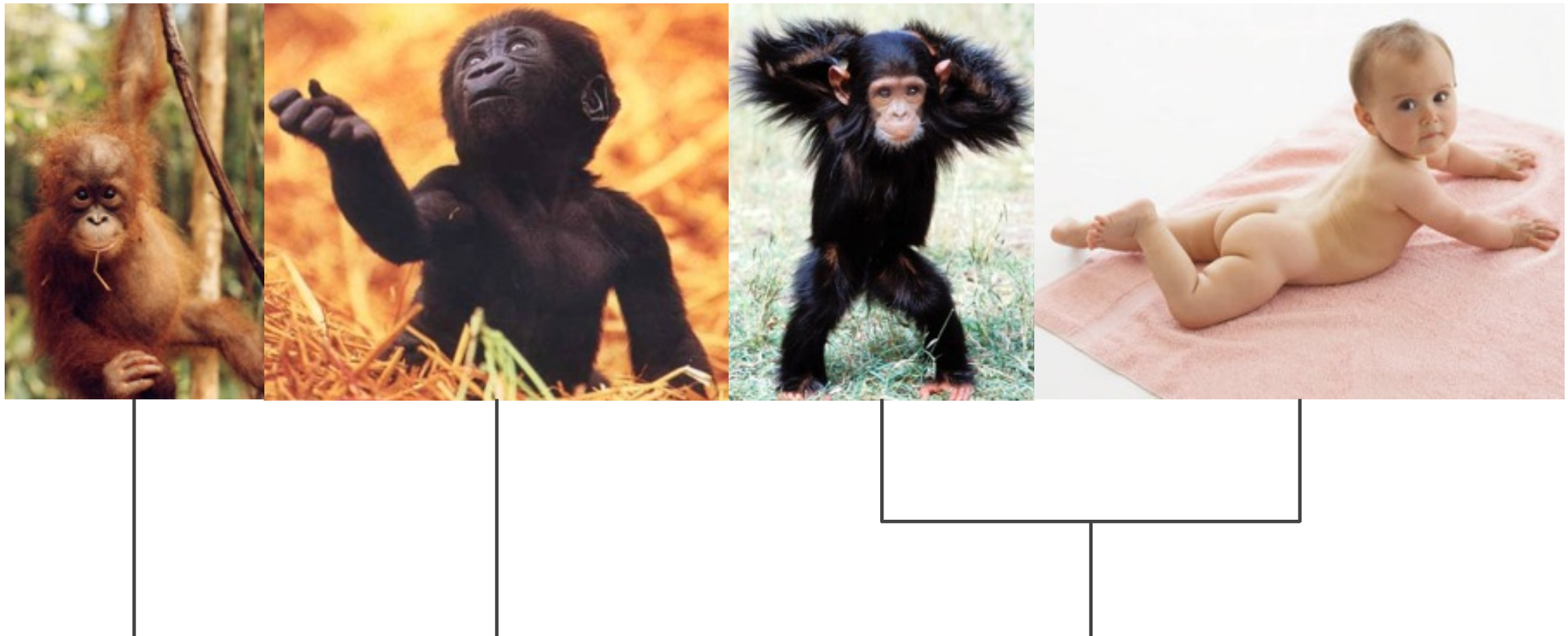
# the great apes



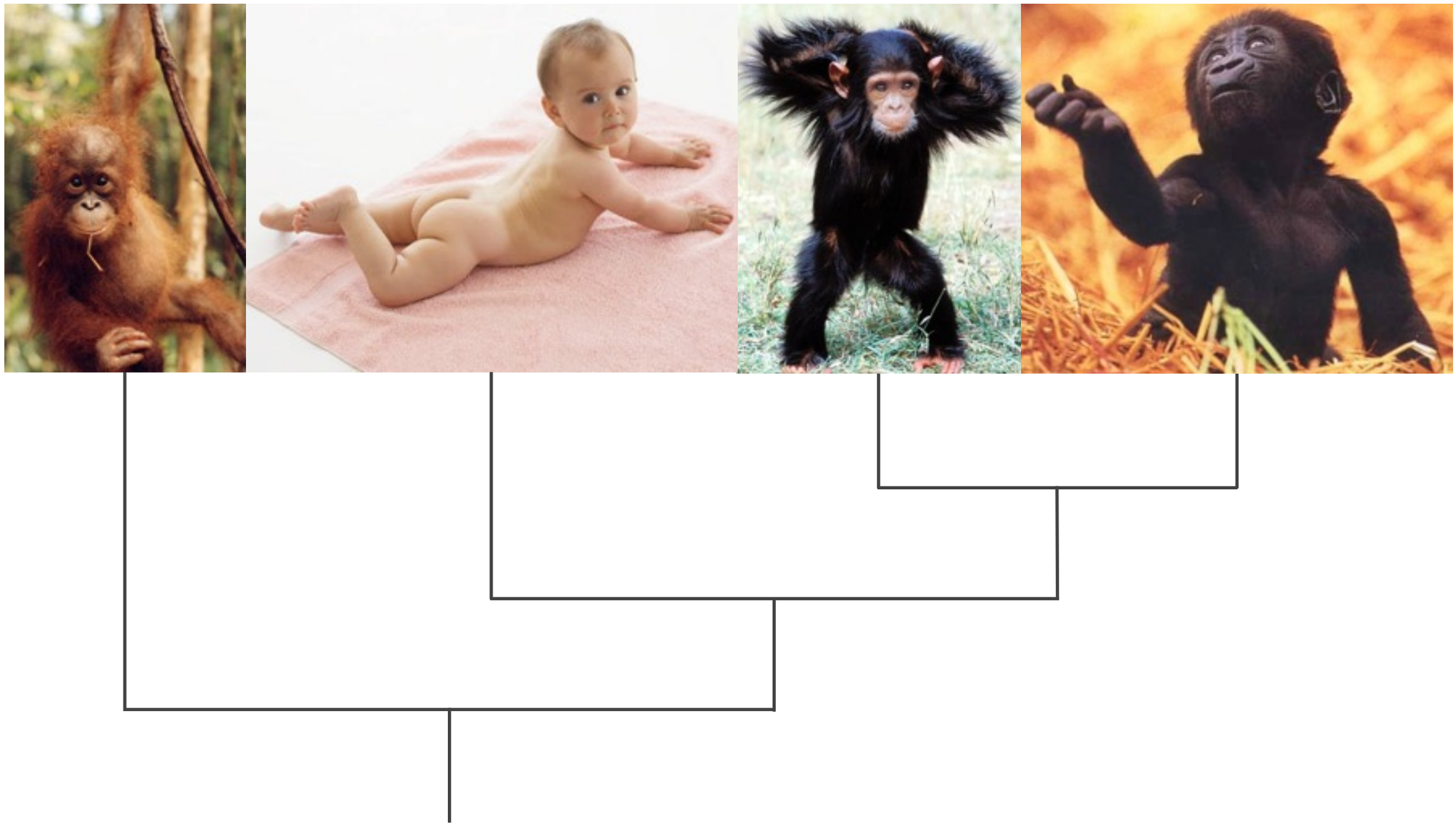

# great apes: phylogenetic relationships?



# great apes: phylogenetic relationships?

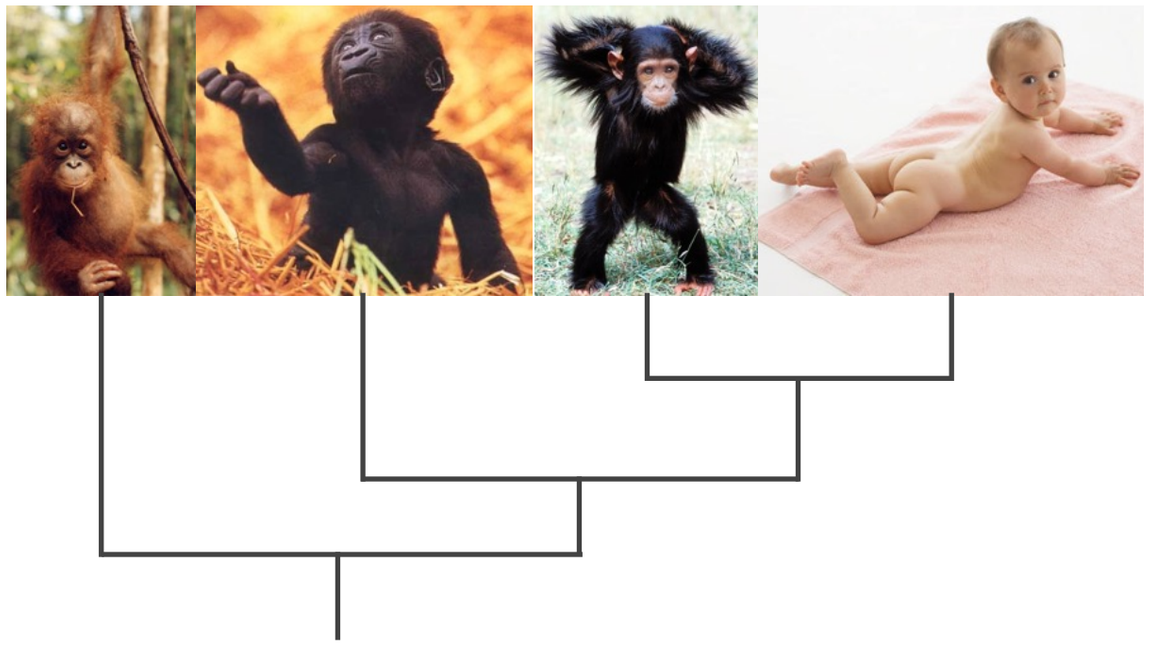


# great apes: phylogenetic relationships?



# phylogenetic relationships of great apes

when subjected to phylogenetic analysis overwhelming evidence supports chimps and humans being each others most closest relatives



# number of possible topologies

| tips | unrooted trees |
|---|---|
| 3 | 1 |
| 4 | 3 |
| 5 | 15 |
| 6 | 105 |
| 7 | 945 |
| 8 | 10,395 |
| 9 | 135,135 |
| 10 | 2,027,025 |
| 11 | 34,459,425 |
| 12 | 654,729,075 |
| 13 | 13,749,310,575 |
| 14 | 316,234,143,225 |
| 15 | 7,905,853,580,625 |
| 20 | 213,643,476,699,771,875 |

# phylogenetic analysis

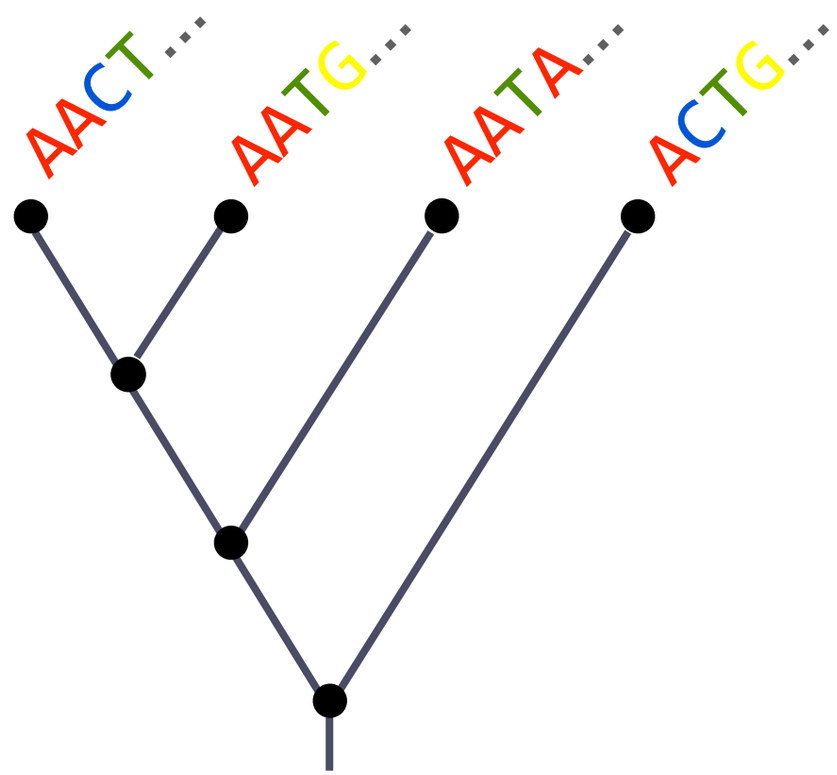
the most accurate methods are model-based and involve likelihood calculations

- maximum likelihood estimation

- Bayesian analysis

$$\text{Prob}(H|D) = \frac{\text{Prob}(D|H)\ \text{Prob}(H)}{\text{Prob}(D)}$$

we can only directly calculate $\text{Prob}(D|H)$

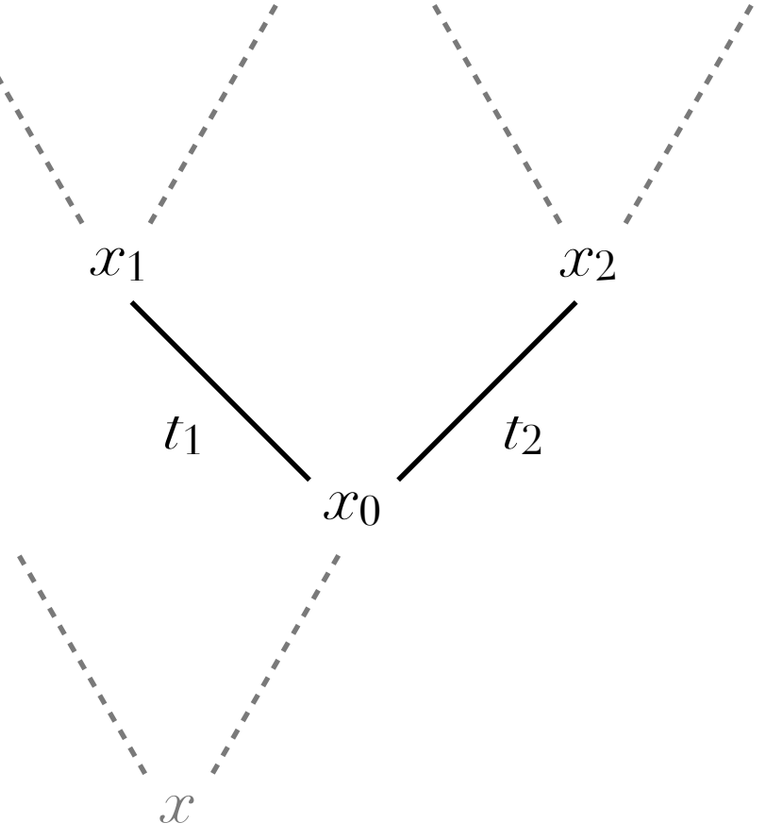



-1574.63624
(log likelihood)

# likelihood calculation

peeling algorithm (Felsenstein 1981) does post-order traversal with calculation of partial likelihoods at each node that depend only on its immediate children

$$L_0^{(i)}(x_0) = \left(\sum_{x_1} \text{Prob}(x_1|x_0,t_1) L_1^{(i)}(x_1)\right)\left(\sum_{x_2} \text{Prob}(x_2|x_0,t_2) L_2^{(i)}(x_2)\right)$$




nonetheless, likelihood calculations are very computationally intensive -

*O(taxa* x *sites* x *rates* x *states²)*

# likelihood calculations: majority of computation

| | likelihood related calculations |
|---|---|
| nucleotide | 94.69% |
| amino acid | 95.72% |
| codon | 81.24% |

GARLI profiling; 11 taxa; 2178 characters

# BEAGLE: broad-platform evolutionary analysis general likelihood evaluator

an application programming interface (API) and high-performance computing library for statistical phylogenetics

emphasis is evaluating phylogenetic likelihoods of biomolecular sequence evolution

aim is to provide high performance evaluation 'services' to a wide range of phylogenetic software, both Bayesian samplers and maximum likelihood optimizers

allows phylogenetic software using the library to make use of optimized hardware such as GPUs

# BEAGLE library design goals

open-source (LGPL)

multi-platform support (i.e., Linux, OS X, Windows)

low level
  C API
  does not explicitly have concept of tree

minimize transfer of data

support multiple implementations (e.g., CPU, SSE, CUDA, OpenCL)
  uses dynamic plug-in system

support both single and double precision

# GPU implementation

CPU-side code only used to manage GPU
- memory allocations and transfers, kernel launches
- allows client to use CPU in parallel to GPU

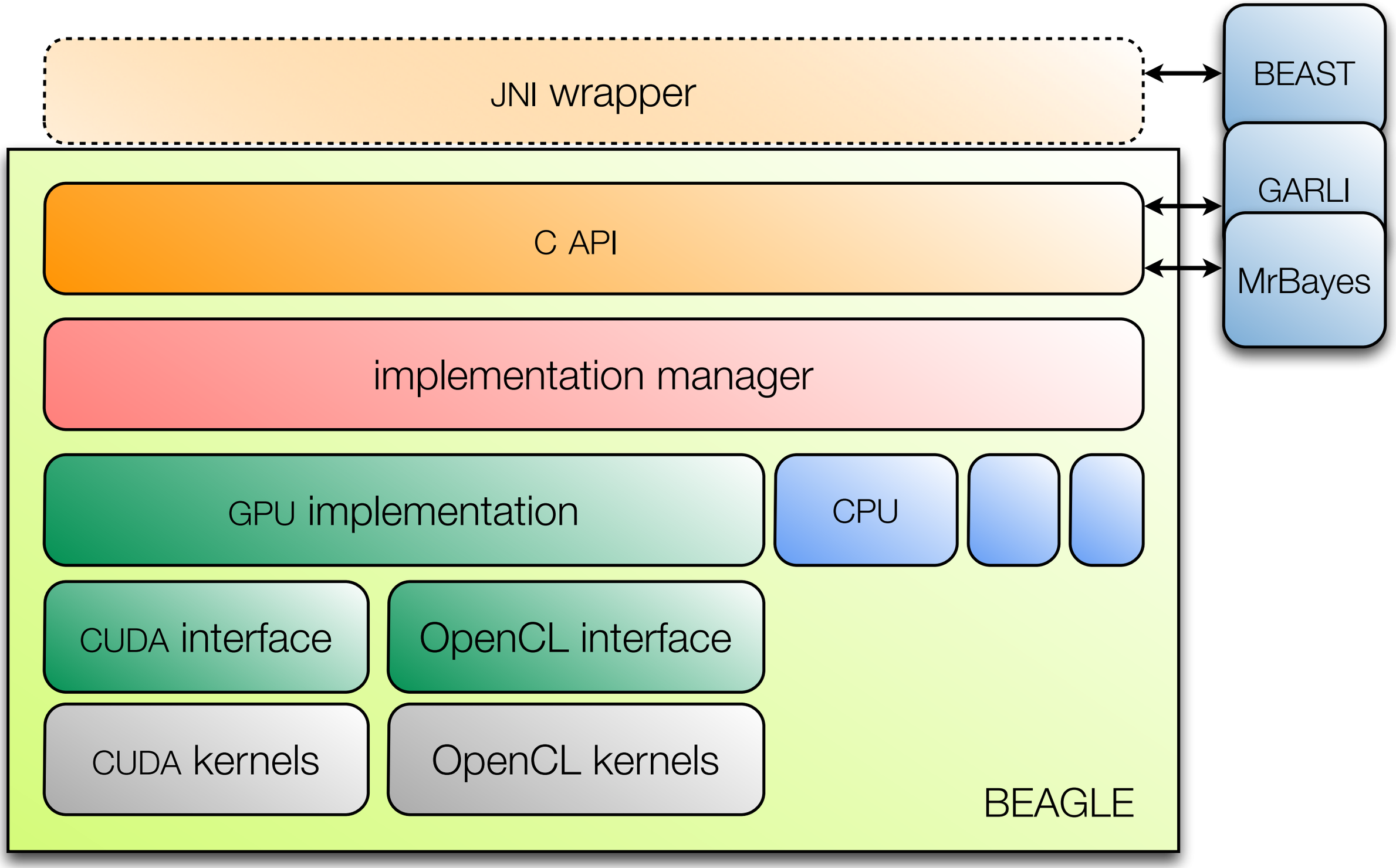
GPU interface abstraction layer
- CUDA and OpenCL implementations share same CPU-side code

CUDA implementation uses the driver API
- Parallel Thread Execution (PTX) kernels
- Java Native Interface (JIT) compilation

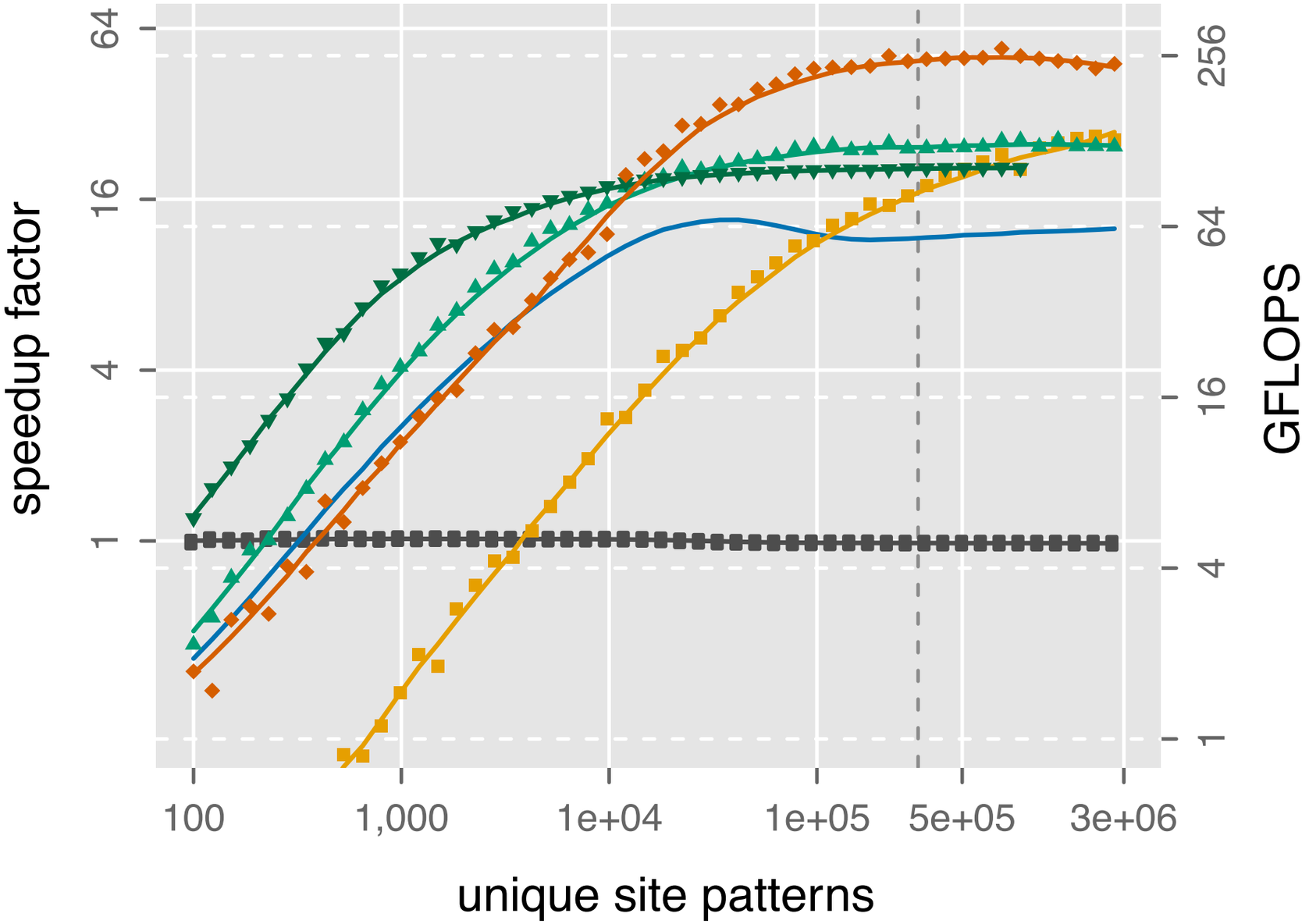
templated kernels support arbitrary number of states

multiple GPUs supported via client-side partitioning (scales linearly)

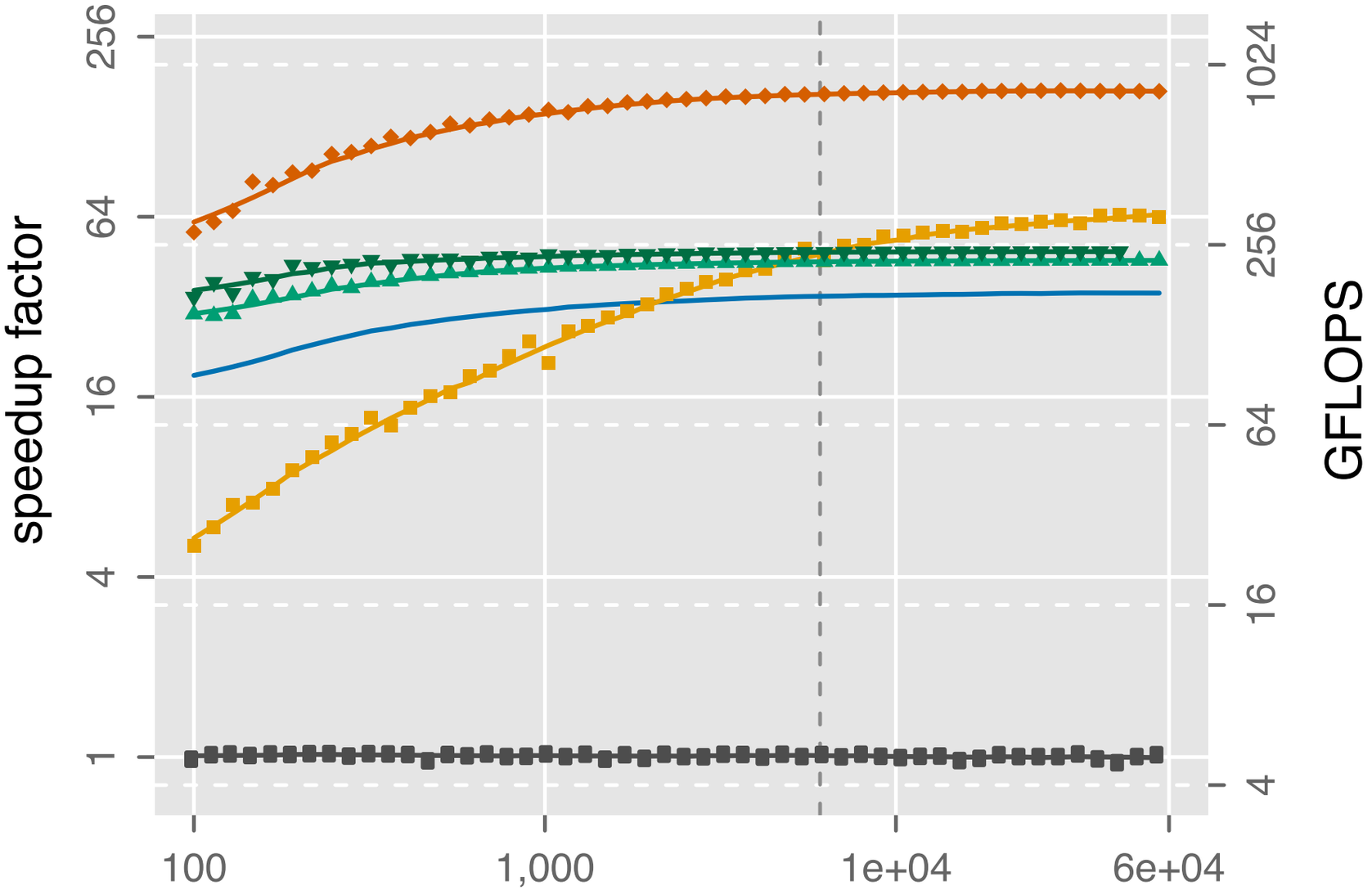
# gross structure of BEAGLE



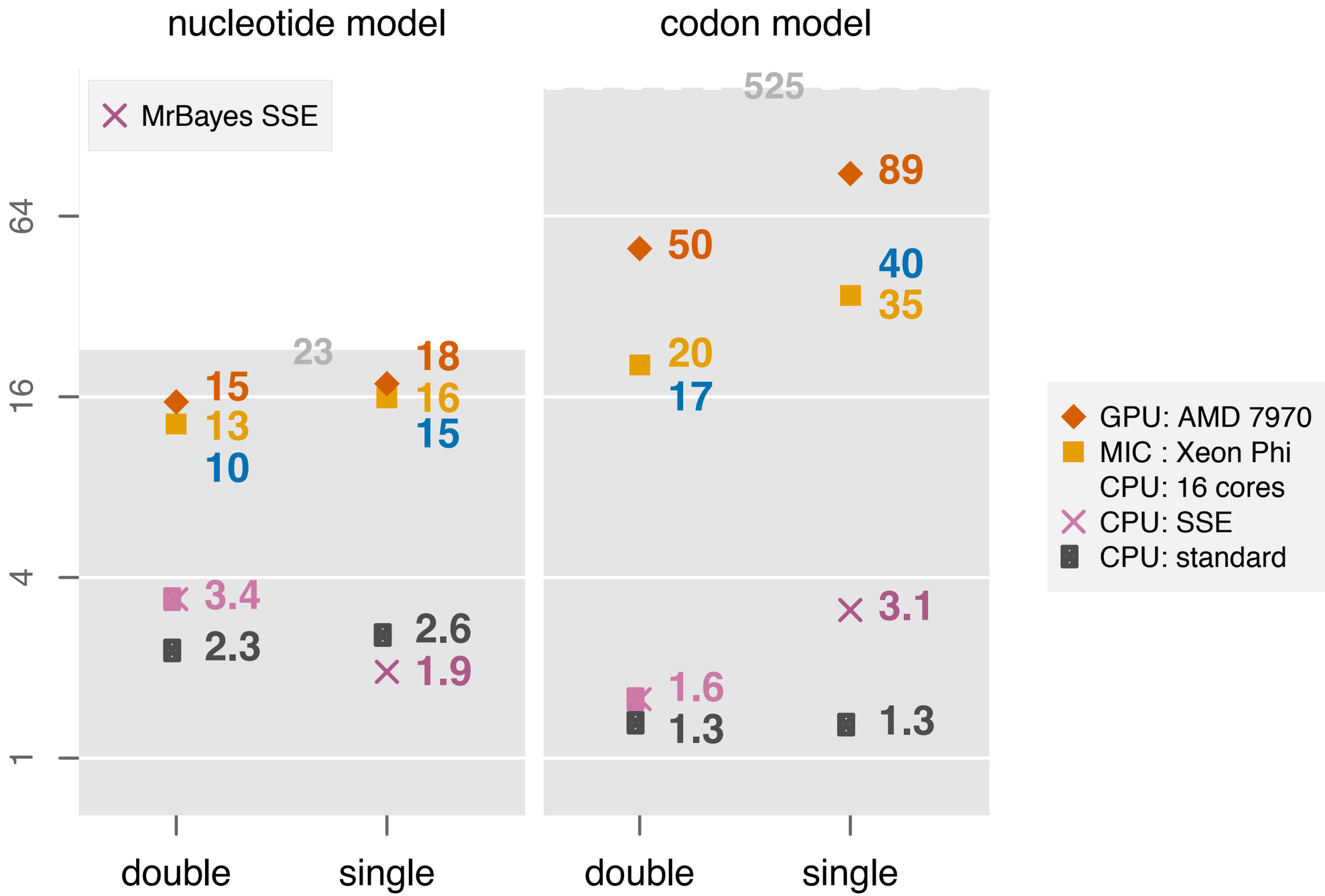
# throughput for nucleotide data (4 states)



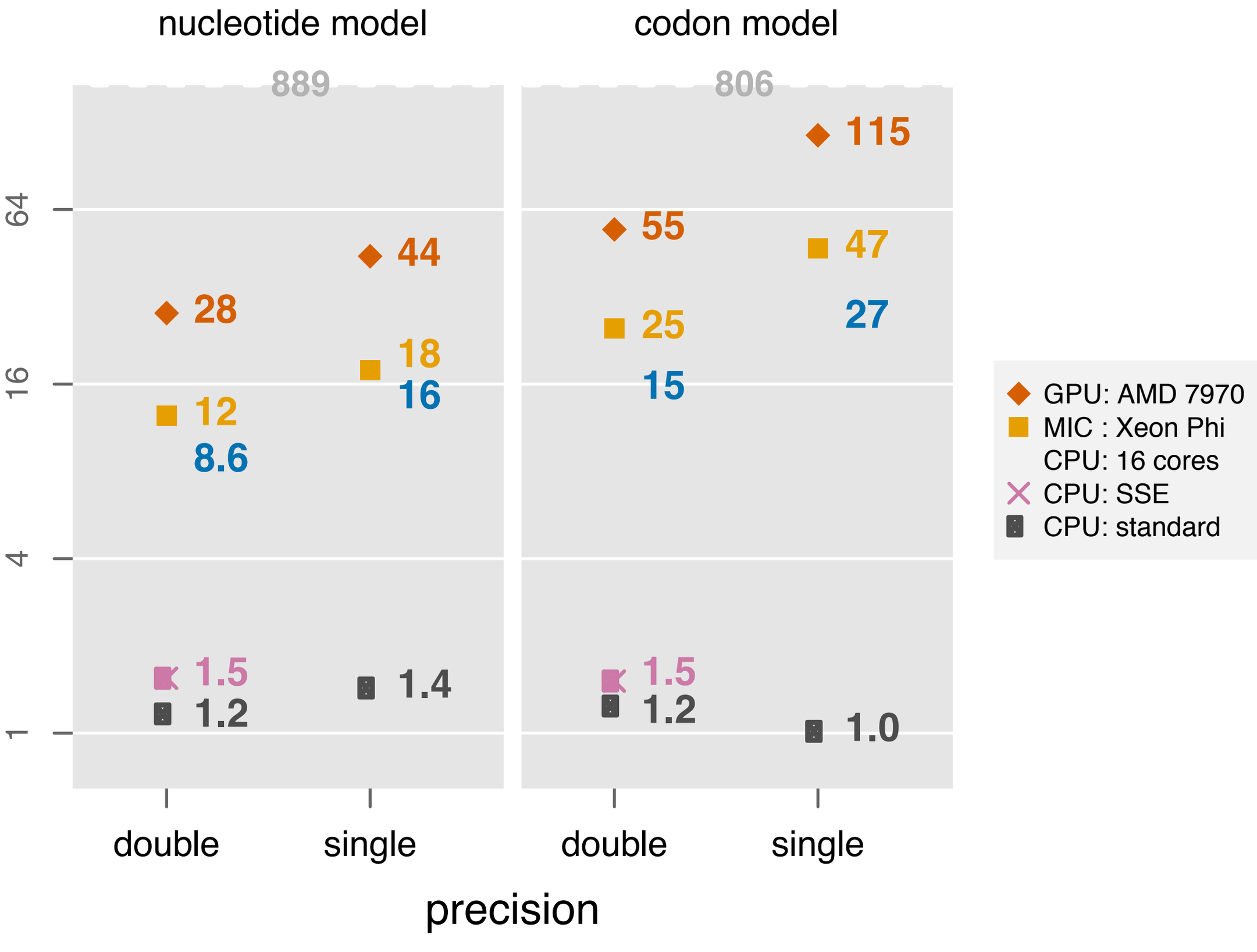
# throughput for codon data (64 states)



# MrBayes speedup



# BEAST speedup



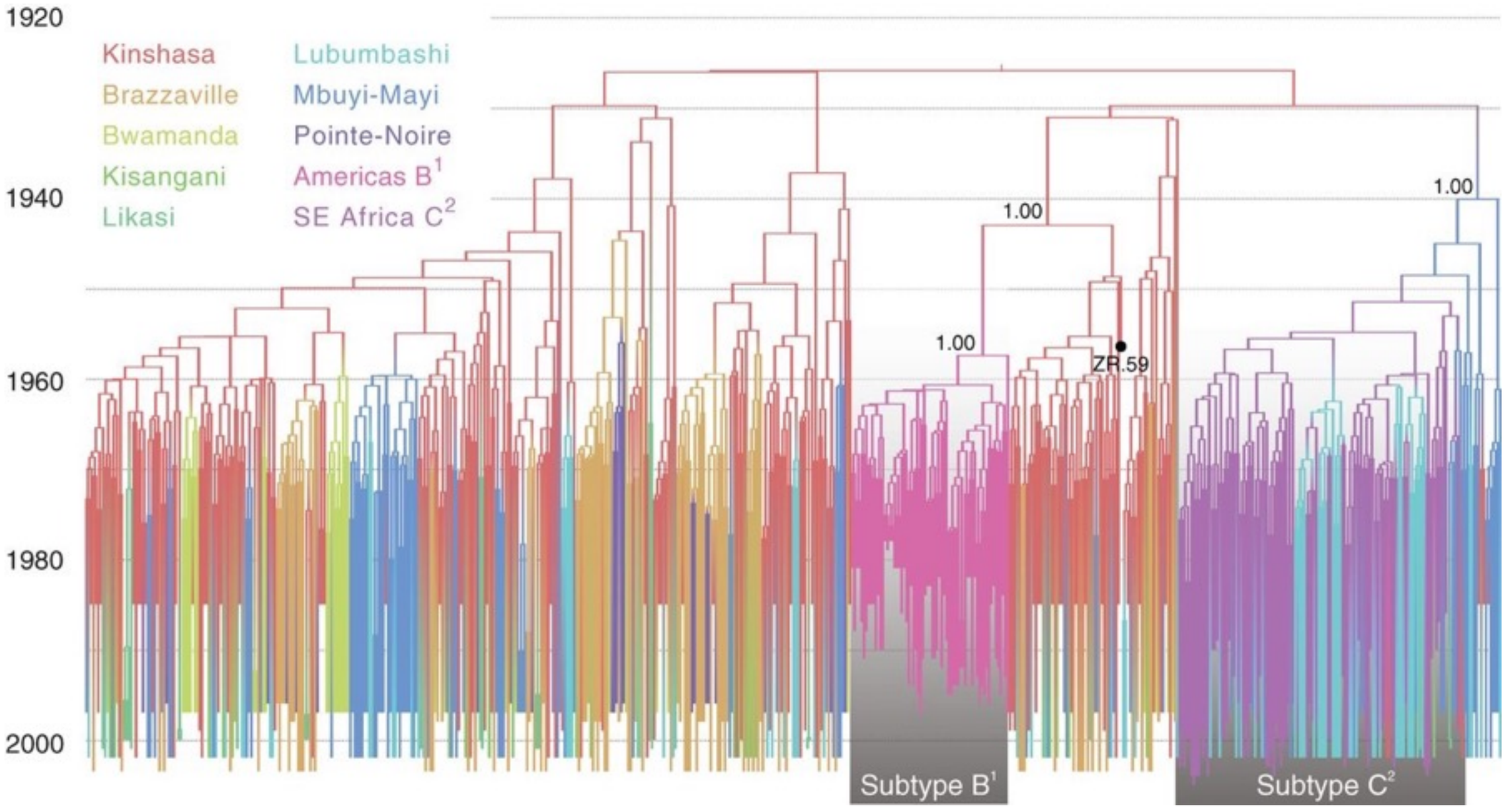
# more than academic

**academic**: having no practical or useful significance

*Webster's New Collegiate Dictionary*

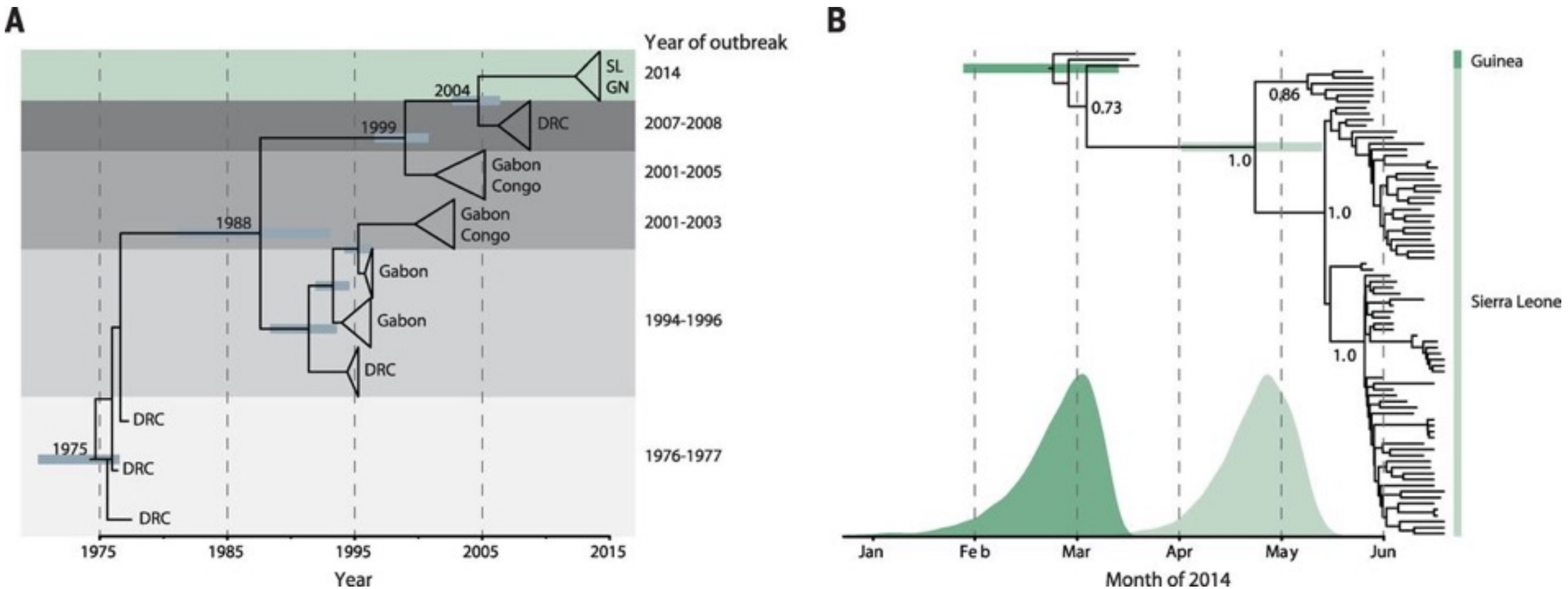
two recent studies using BEAGLE library

# phylogenetics in use: early spread of HIV-1




Faria et al. 2014 The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346:56-61

# phylogenetics in use: 2014 Ebola outbreak




Gire et al. 2014 Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 345:1369-1372

# acknowledgements

**Daniel Ayres, University of Maryland**
Peter Beerli, Florida State University
Aaron Darling, University of Technology Sydney
Mark Holder, University of Kansas
John Huelsenbeck, University of California, Berkeley
Paul Lewis, University of Connecticut
Andrew Rambaut, University of Edinburgh
Fredrik Ronquist, Swedish National Museum of Natural History
Marc Suchard, University of California, Los Angeles
David Swofford, Duke University
Derrick Zwickl, University of Arizona

Dan Stanzione, Texas Advanced Computing Center

Yariv Aridor and Arik Narkis, Intel Israel

Altera University Donation Program

National Science Foundation