Simulation, Modeling, and Benchmarks

U Penn: Junhyong Kim, Sampath Kannan, Susan Davidson
U Texas : David Hillis, Lauren Meyers
NC State: Spencer Muse
Florida State: Mark Holder
Yale: Paul Turner

Goal: Develop validated datasets of sufficient complexity and scale to realistically benchmark latest tree algorithms

Or,

seriously kick some algorithmic b*%t

Rationale: Current approaches for tree method validation has some important limitations

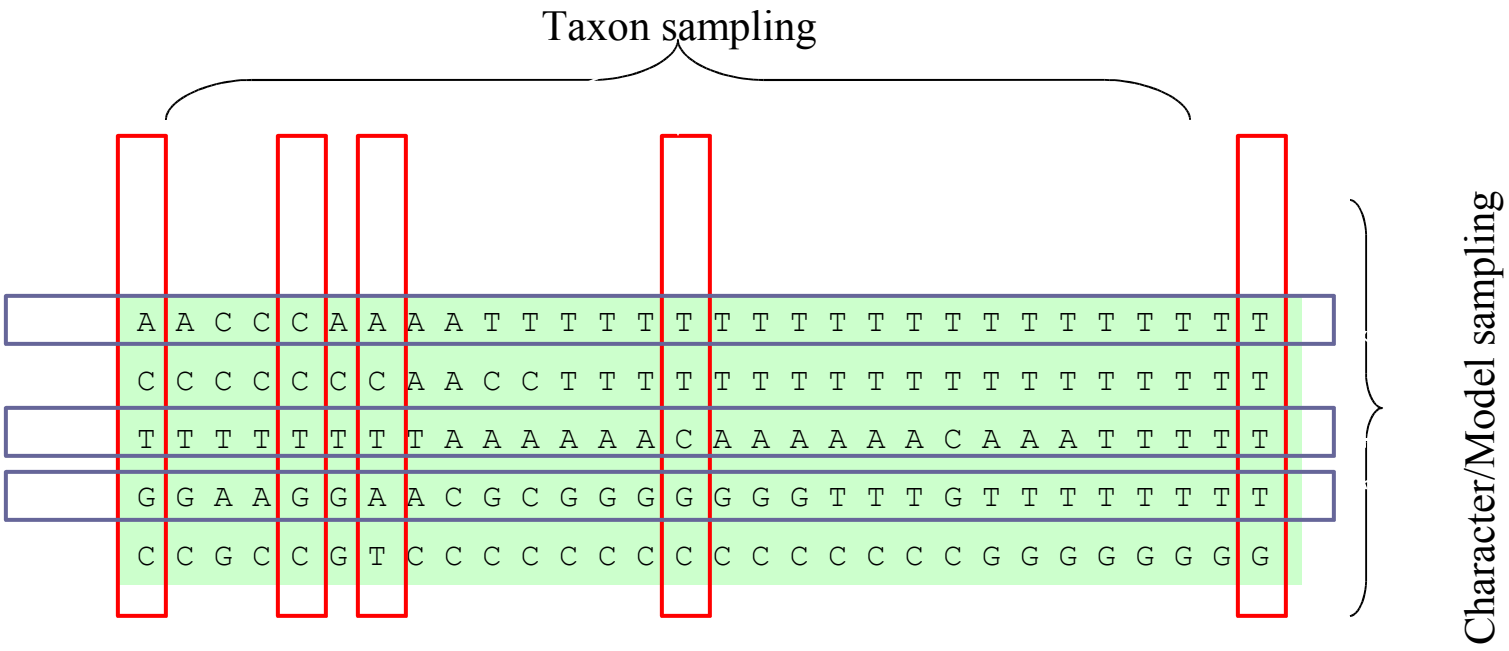
- Too small scale: We want to provide trees of millions of taxa
- Too simple: Time homogeneous, simple rate mixture, independent site, simple stochastic tree generation model
- **Everybody does their own thing, algorithms are not tested on the same dataset.**

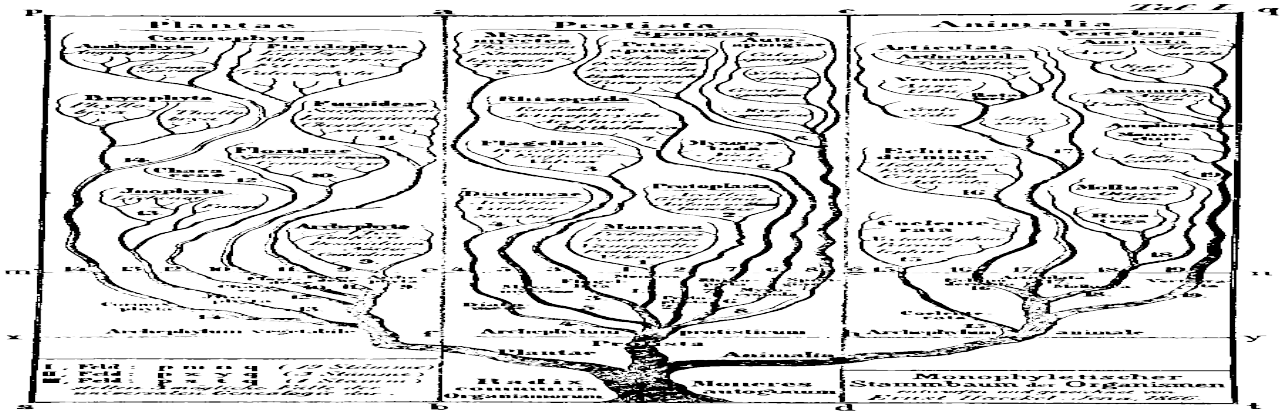
Problems and Approaches:

- Basic infrastructure
  - Data management support
  - Computational infrastructure
  - Benchmark Criteria, evaluation systems
- Benchmark data and tree
  - Data Simulators
  - Tree Simulators
  - Empirical Data

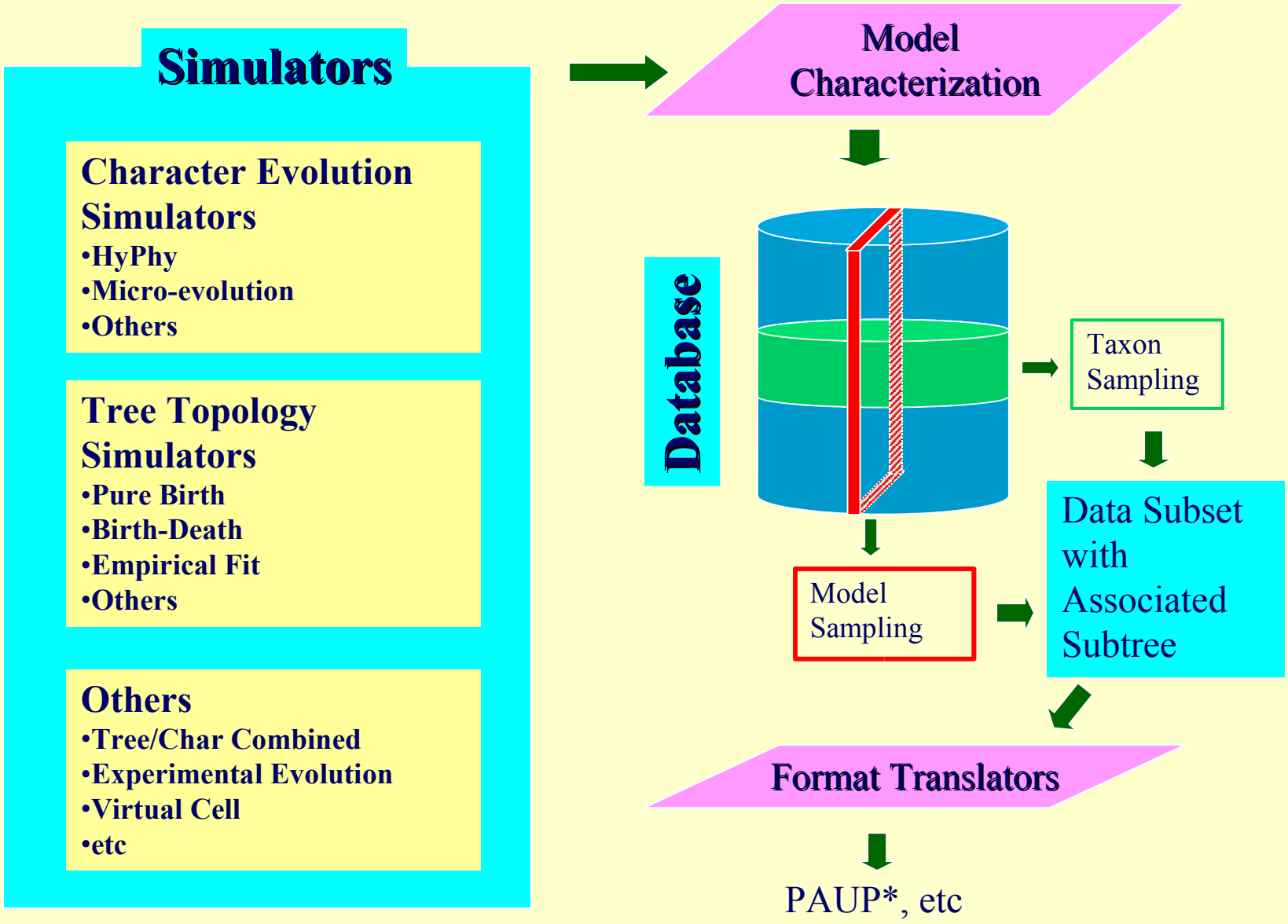
- Basic infrastructure
  - Simulation database
  - Parallelization
  - Tree comparison methods, protocols
- Benchmark data and tree
  - Multi-layered simulation models
  - Complex tree simulation
  - Experimental evolution using viral systems

Basic Infrastructure (yr 1 and 2): Simulation Database

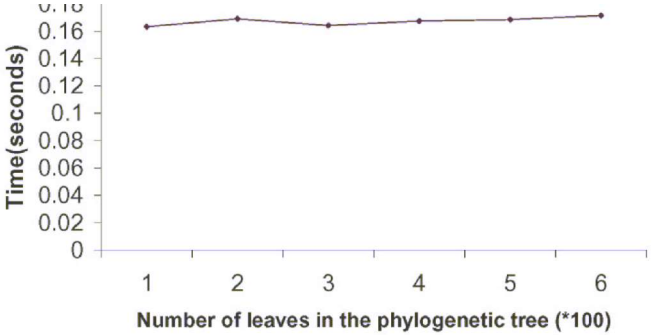




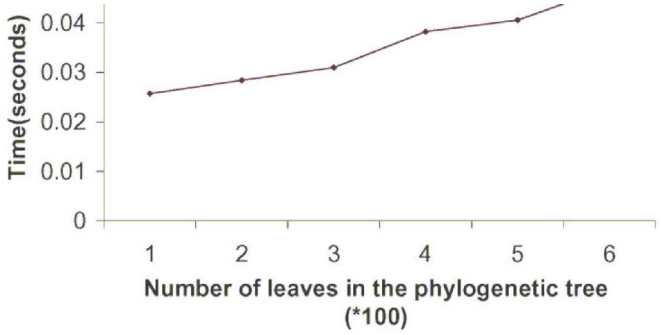
Simulation and Data Access



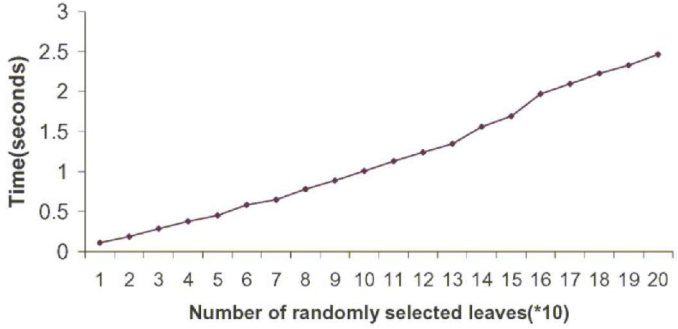
Database Performance: Constant or Linear Time Queries

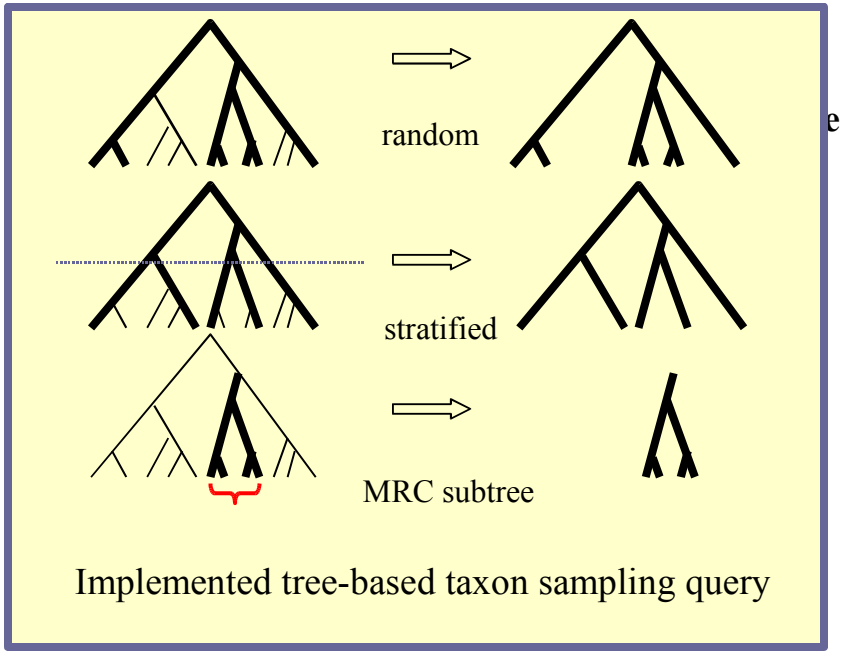


Select 20 fixed taxa from tree of size *t* (100 to 600)



Select 20 ***random* taxa from tree of size *t* (100 to 600)**





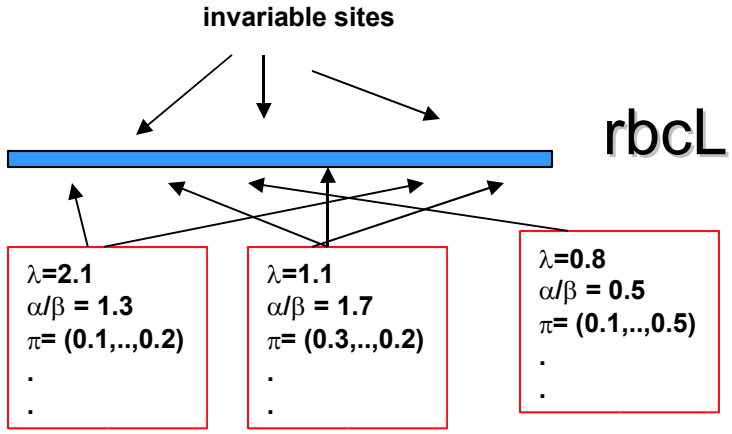
Implemented tree-based taxon sampling query

Benchmark Data: Multi-layered simulations

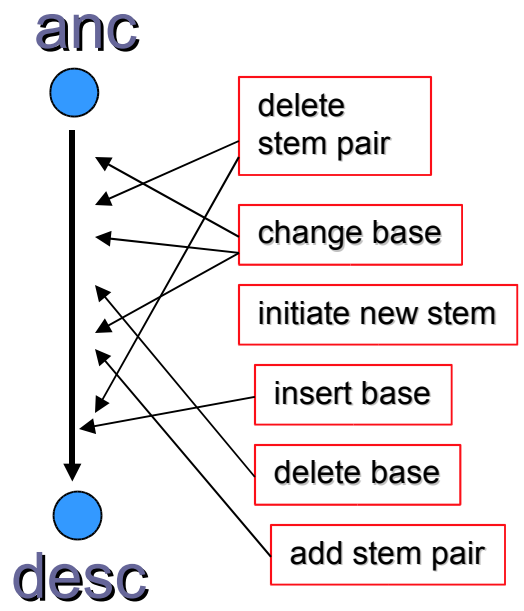
- Key molecule simulation (Muse, Hillis)
- General mutation simulation (Kim)

- Micro-Macro simulation (Kim, Meyers)
- Experimental viral evolution (Turner)

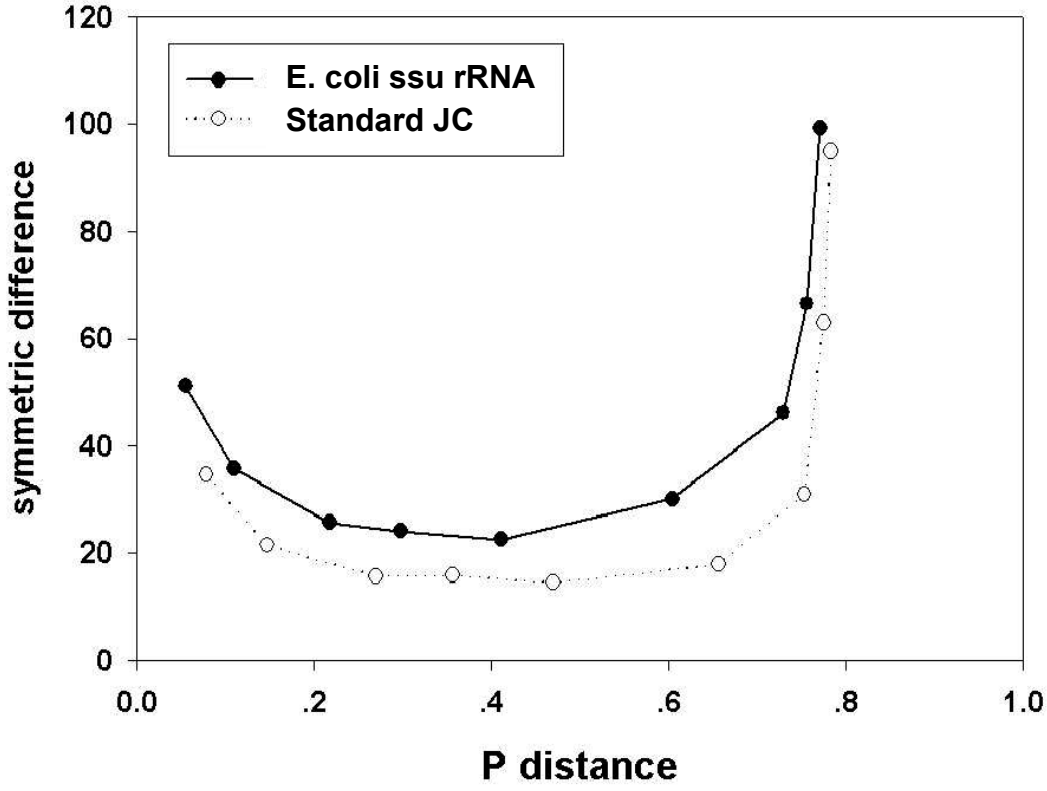
- Key molecule simulation (Muse, Hillis, Holder)
  - Estimate statistical parameters for real molecules (e.g., rbcL) using HyPhy, extend model family to include more discrete rate distribution and positional dependencies, and finally generate a very large tree of 10^6~10^7 taxa using the key molecule models as its basis.



- General mutation simulation (Kim)
  - Incorporate structural constraints, indel, functional constraints, etc. using a simulator based on edit mutations. A set of edit operators are implemented, such as stem-loop edit, each of which operate on evolving strings with a characteristic wait time.

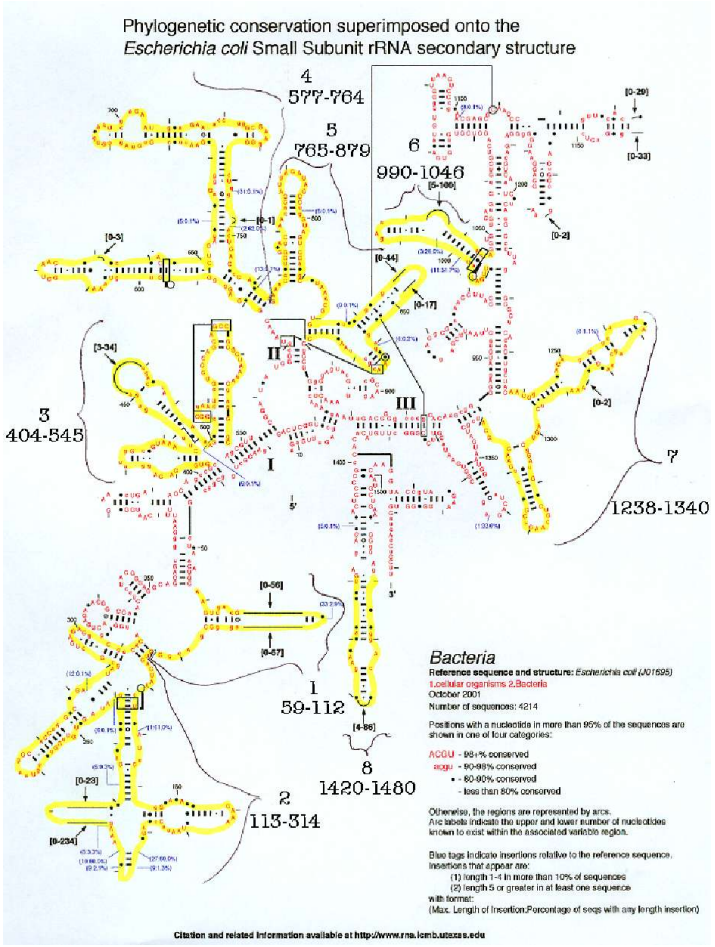


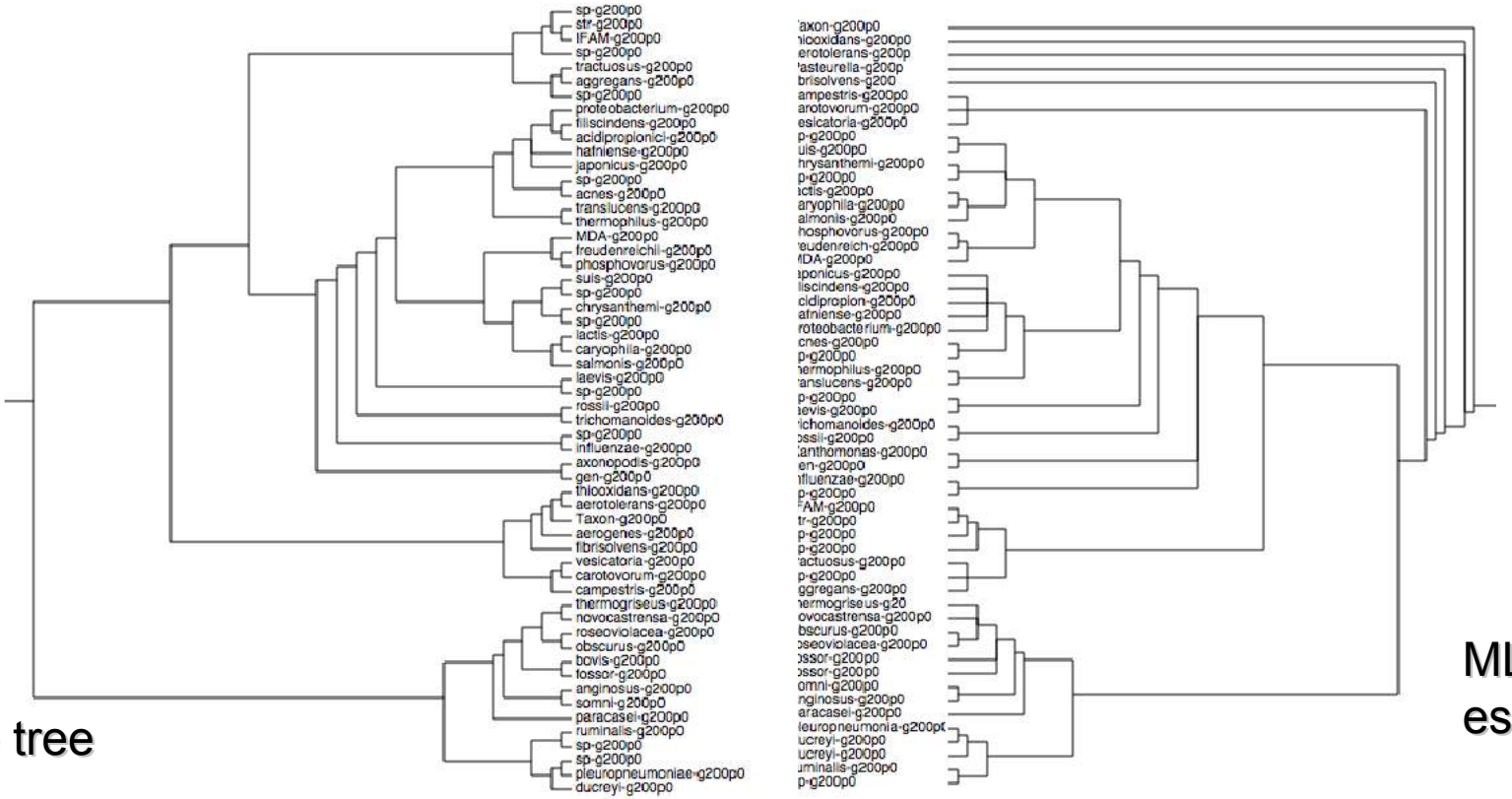
General mutation model based on E. coli ssu rRNA (~1.5kb). 99-taxon beta-splitting model tree, 9 different rates, 50 replicates, ClustalW default alignment



- Micro-Macro simulation model (Meyers, Kim)
  - Generate a population of molecules incorporating a fitness model and speciation process based on RNA folding. Fitness from (1) similarity to known 16S RNA (~67k seqs); (2) similarity to known 16S structure (~200 crystal structure); (3) folding stability

- Experimental viral evolution (Turner; non-ITR funding for empirical work)
  - Use the RNA bacteriophage phi-6 system to generate an experimental phylogeny (~64-taxon tree with host switching and horizontal transfer)



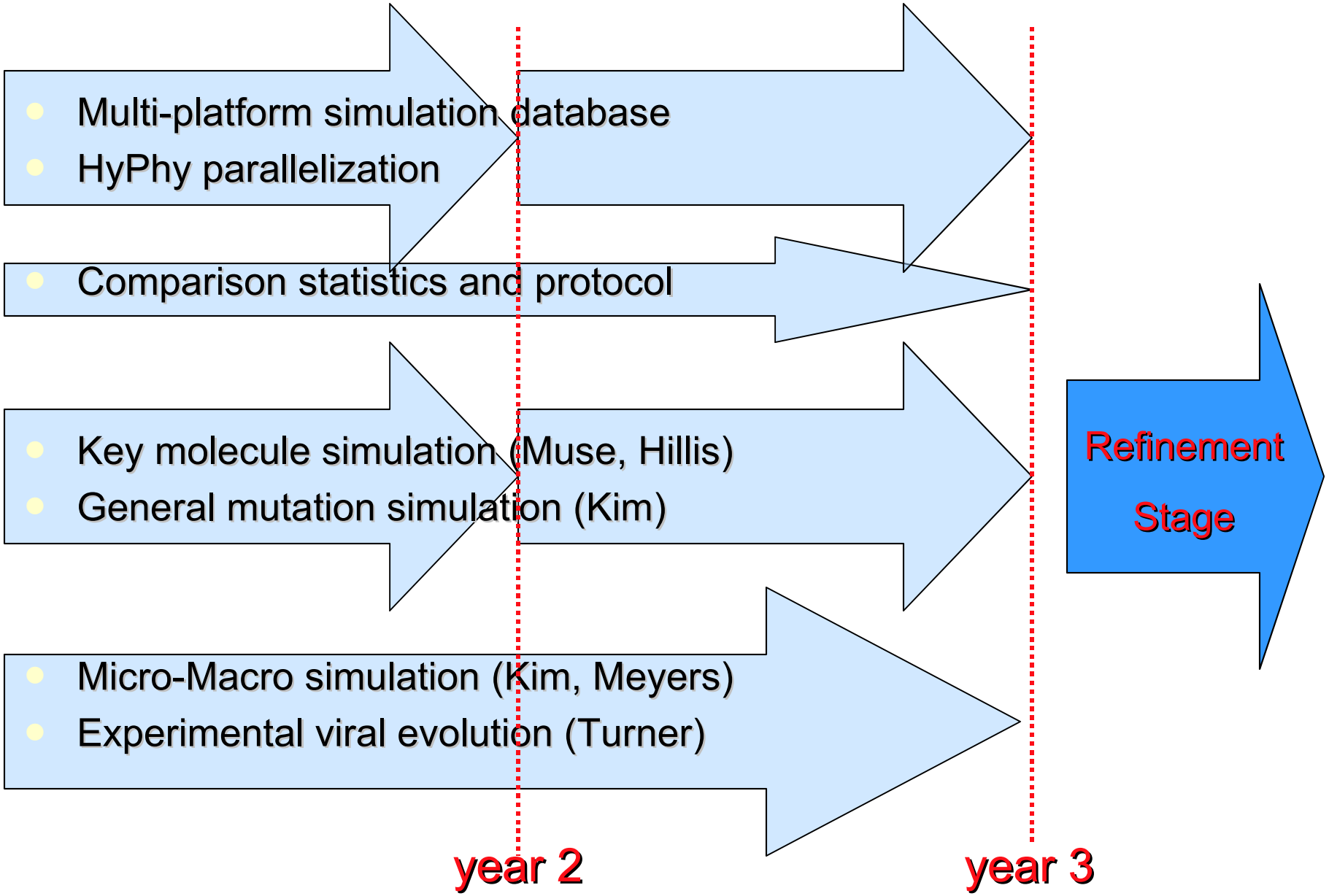


True tree

ML estimate

ssu RNA micro-evolution simulation:

200 generation simulation with population size 1000 per species, speciation when the sequence best matches a different ssu RNA in database, indel/point mutation model


Multi-platform simulation database
HyPhy parallelization
Comparison statistics and protocol
Key molecule simulation (Muse, Hillis)
General mutation simulation (Kim)
Micro-Macro simulation (Kim, Meyers)
Experimental viral evolution (Turner)
Refinement Stage
year 2
year 3