Enabling Phylogenetic Research via the CIPRES Science Gateway

Wayne Pfeiffer

SDSC/UCSD

August 5, 2013

In collaboration with
Mark A. Miller, Terri Schwartz, & Bryan Lunt
SDSC/UCSD

Supported by NSF

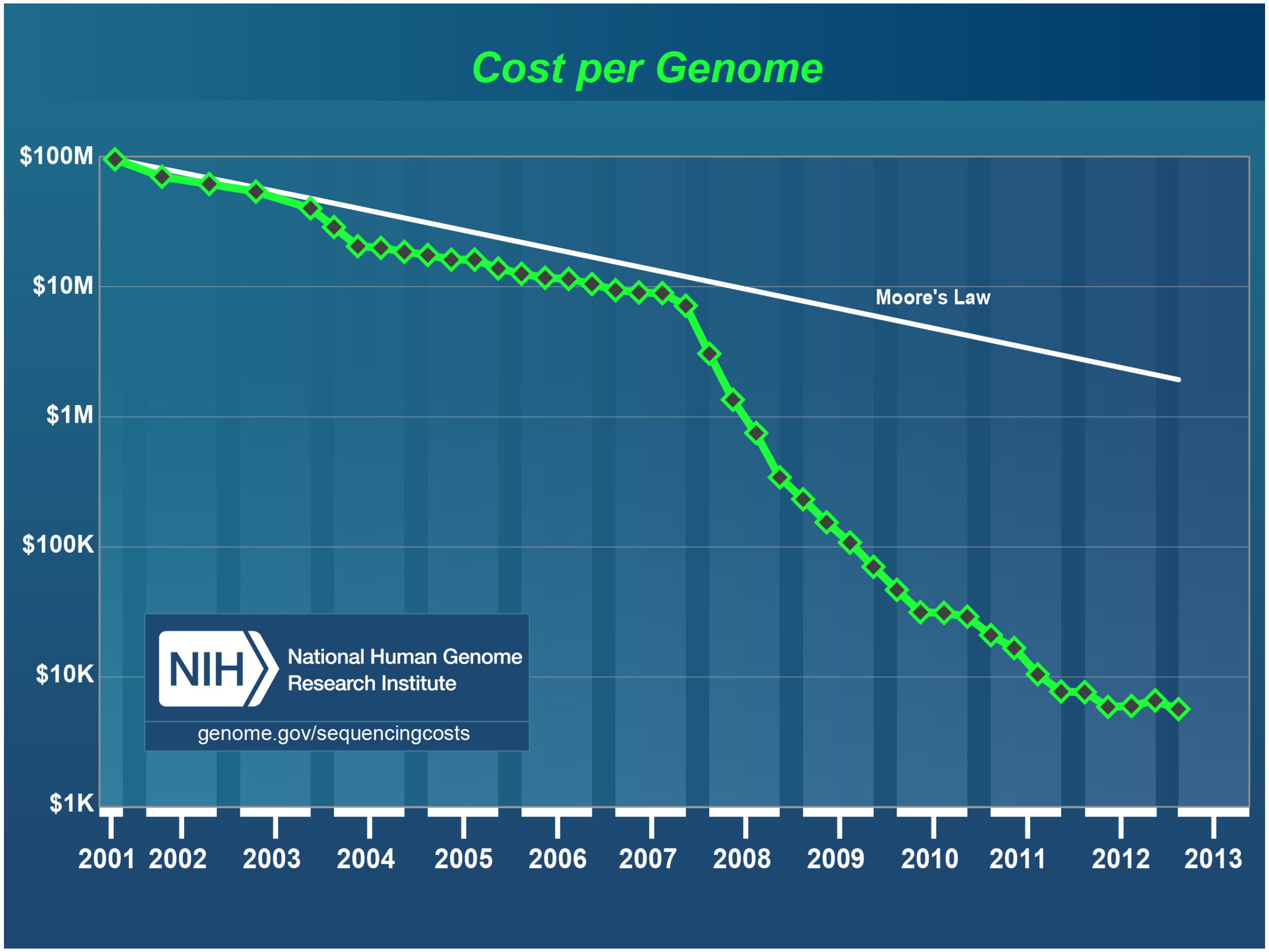
Phylogenetics is the study of evolutionary relationships among groups of organisms called taxa (typically species)

- **The result of a phylogenetic analysis is a phylogeny, most often represented as a tree**

```
/-------- Human
|
|---------- Chimpanzee
+
|   /---------- Gorilla
|   |
\---+             /-------------------------------- Orangutan
    \-------------+
                  \----------------------------------------------- Gibbon
```

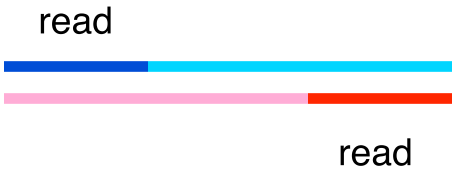
- **In olden times, phylogenies were based on morphology**
- **Now phylogenies are usually based on DNA sequences**

Cost of DNA sequencing has dropped much faster than cost of computing in recent years, producing a flood of data for biological analysis



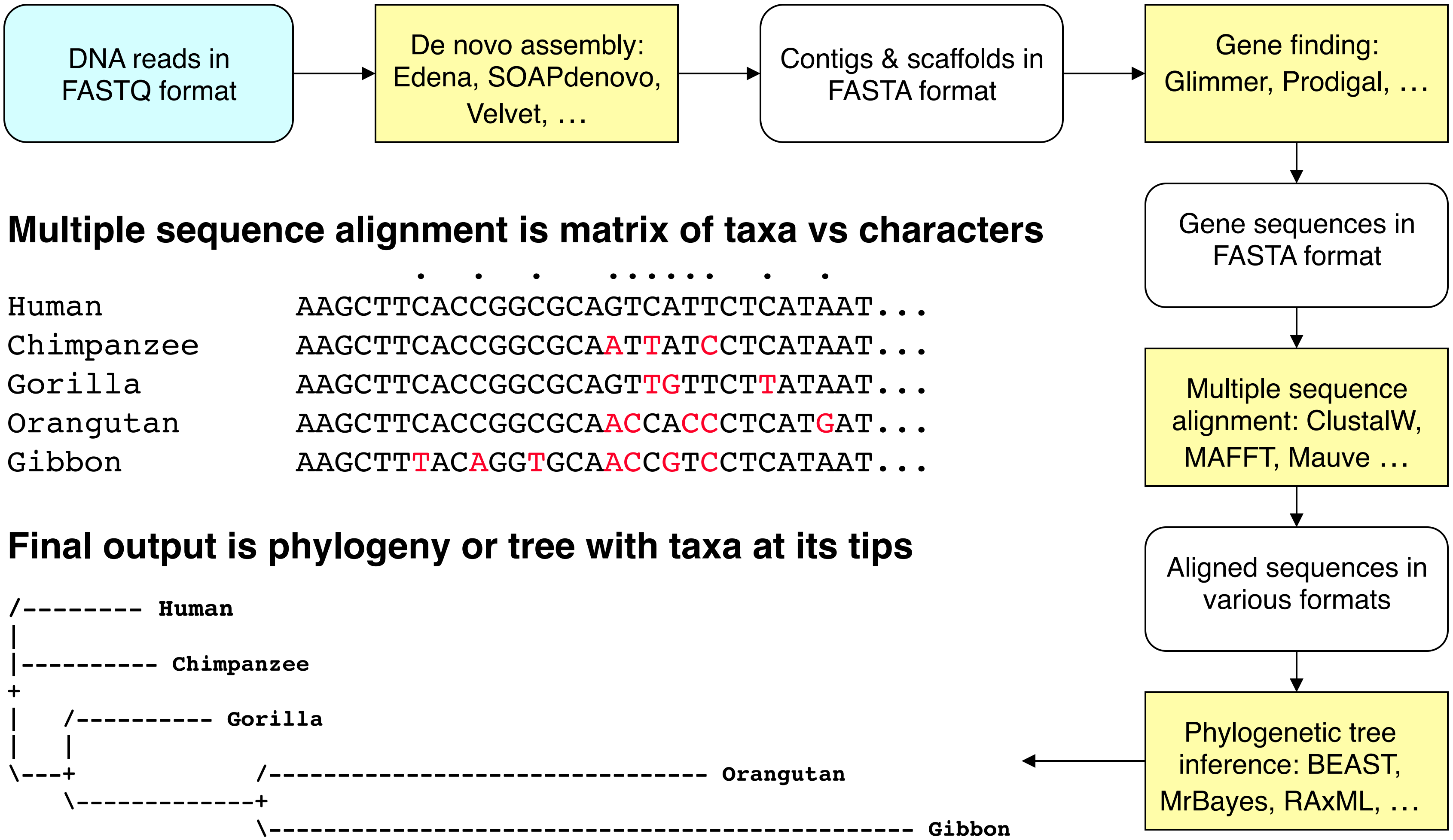
Market-leading DNA sequencers come from Illumina & Life Technologies (both SD County companies)

- **Illumina HiSeq 2500**
  - Big; $740,000 list price
  - High throughput
  - Low error rate
  - 150-bp paired-end reads



- **Life Technologies Ion Proton**
  - Small; $243,000 list price
  - Medium throughput
  - Modest error rate
  - 200-bp reads

Computational workflow for phylogenetic analysis using DNA sequence data


DNA reads in FASTQ format
De novo assembly: Edena, SOAPdenovo, Velvet, …
Contigs & scaffolds in FASTA format
Gene finding: Glimmer, Prodigal, …
Gene sequences in FASTA format
Multiple sequence alignment: ClustalW, MAFFT, Mauve …
Aligned sequences in various formats
Phylogenetic tree inference: BEAST, MrBayes, RAxML, …

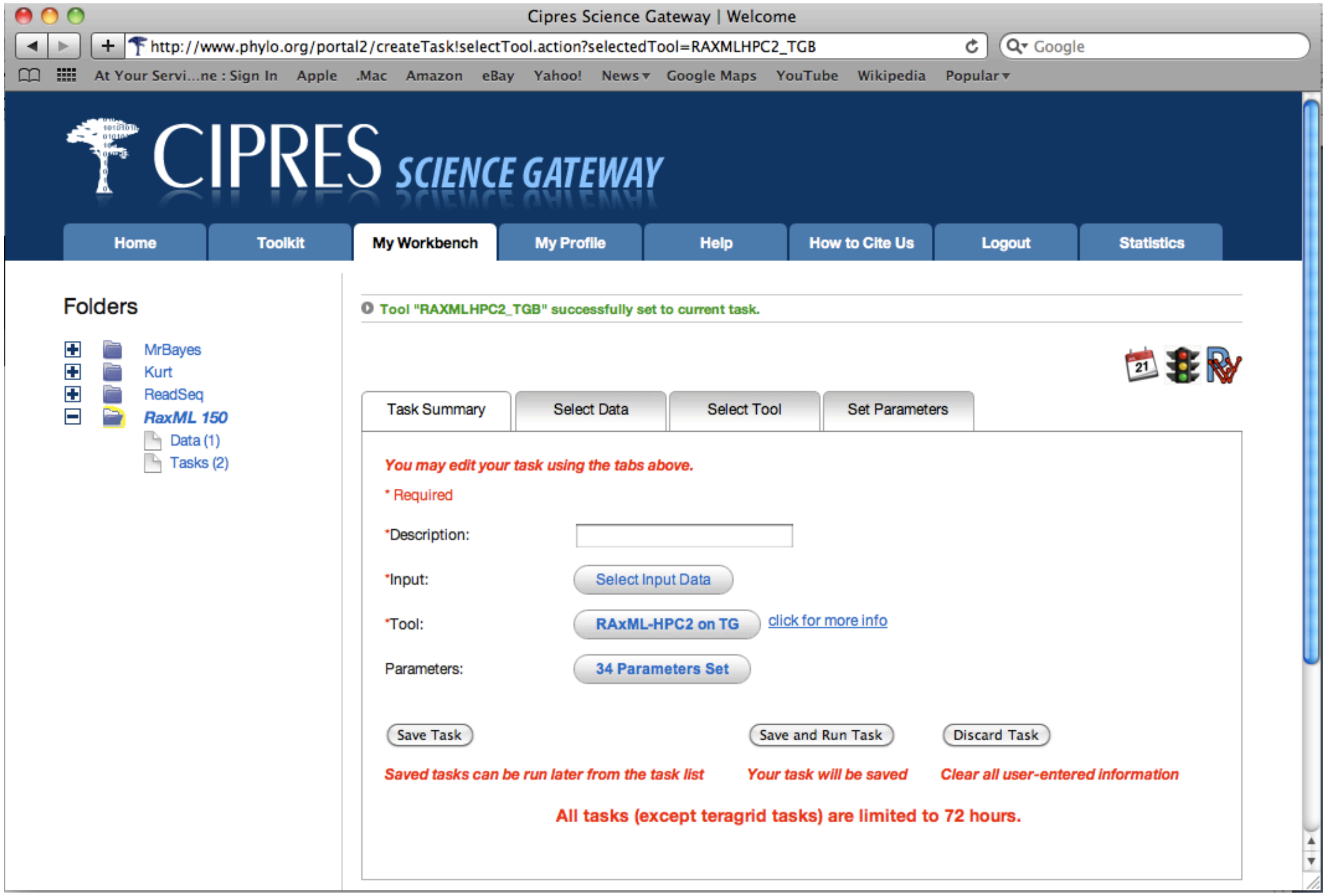
Multiple sequence alignment is matrix of taxa vs characters

```
                     .  .  .   ......  .  .
Human          AAGCTTCACCGGCGCAGTCATTCTCATAAT...
Chimpanzee     AAGCTTCACCGGCGCAATTATCCTCATAAT...
Gorilla        AAGCTTCACCGGCGCAGTTGTTCTTATAAT...
Orangutan      AAGCTTCACCGGCGCAACCACCCTCATGAT...
Gibbon         AAGCTTTACAGGTGCAACCGTCCTCATAAT...
```

Final output is phylogeny or tree with taxa at its tips

```
/--------- Human
|
|---------- Chimpanzee
+
|   /---------- Gorilla
|   |
\---+             /-------------------------------- Orangutan
    \-------------+
                  \----------------------------------------------- Gibbon
```

The CIPRES gateway (or portal) lets biologists run phylogenetics codes at SDSC via a browser interface; http://www.phylo.org/index.php/portal



Folders

- MrBayes
- Kurt
- ReadSeq
- *RaxML 150*
  - Data (1)
  - Tasks (2)

Tool "RAXMLHPC2_TGB" successfully set to current task.

You may edit your task using the tabs above.

* Required

*Description:	
*Input:	Select Input Data
*Tool:	RAxML-HPC2 on TG click for more info
Parameters:	34 Parameters Set

Saved tasks can be run later from the task list

Your task will be saved

Clear all user-entered information

All tasks (except teragrid tasks) are limited to 72 hours.

Browser interface simplifies access to community codes, especially for users who only occasionally compute

- **Users do not log onto HPC systems & so do not need to learn about Linux, parallelization, or job scheduling**
- **Users simply use browser interface to**
  - pick code, select options, & set parameters
  - upload sequence data
- **Numbers of cores, processes, & threads are selected automatically based on**
  - input options & parameters
  - rules developed from benchmarking
- **Occasionally we make special runs not allowed by rules**
- **In most cases, users do not need individual allocations**
- **Users still need to understand code options!**

Parallel versions of six phylogenetics codes are available via the CIPRES gateway

Code & version	Parallelization	Cores	Computer
MAFFT 7.037	Pthreads	8	Trestles
BEAST 1.7.5	Pthreads/Pthreads	8	Trestles
GARLI 2.0	MPI	≤32	Trestles
MrBayes 3.1.2h	MPI/OpenMP	10 to 32	Gordon
MrBayes 3.2.1	MPI	8 to 16	Gordon
RAxML 7.6.6	MPI/Pthreads	8, 30, or 60	Trestles
RAxML-Light 1.0.9	bash/Pthreads	≤1,000	Trestles

Run times for some analyses are substantial

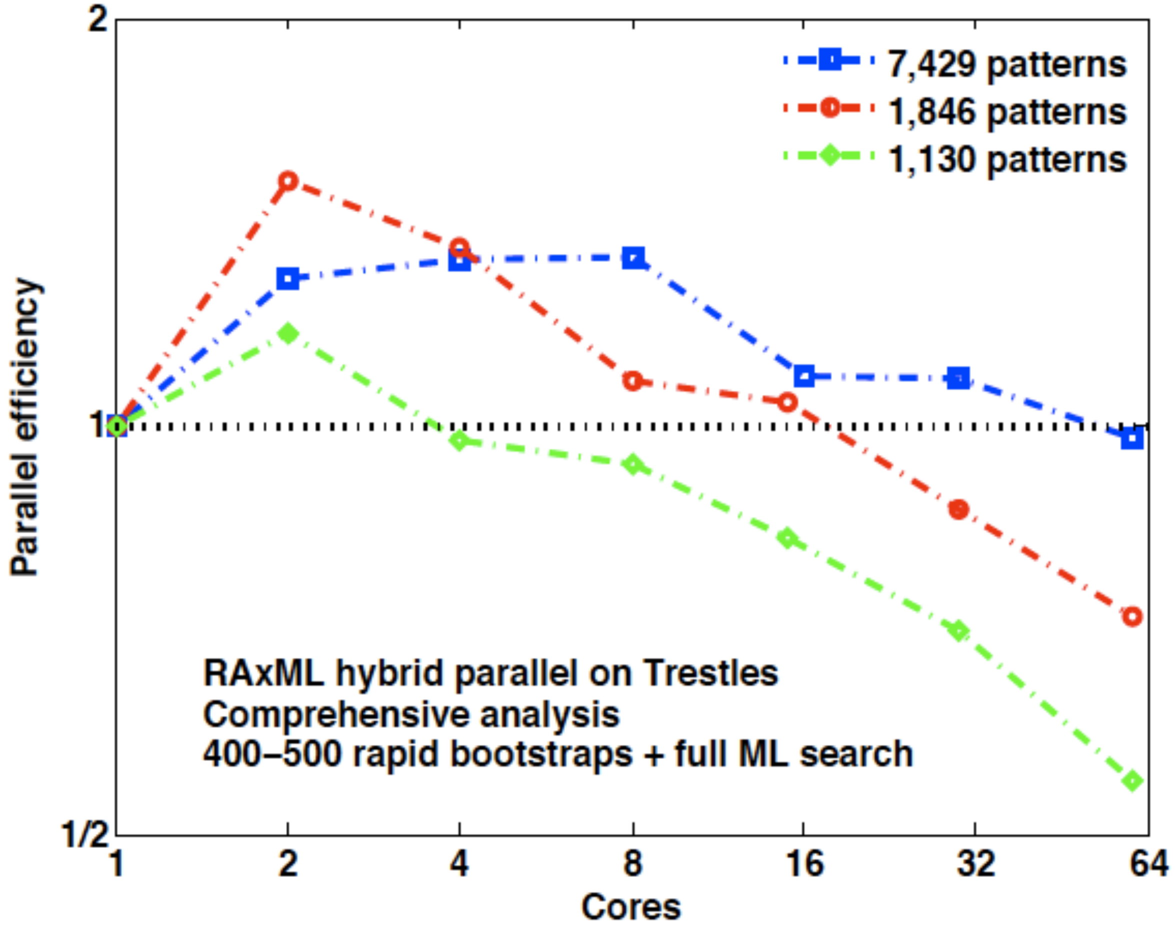
Code & data set	Time (h)	Cores	Computer
MrBayes 3.1.2h, AA data, 73 taxa, 10.4k patterns*, 3M generations (HL)	194	32	Gordon
MrBayes 3.2.1, DNA data, 40 taxa, 16k patterns*, 100M generations (NJ)	155	8	Gordon
RAxML 7.2.7, AA data, 1.6k taxa, 8.8k patterns*, 160 bootstraps+ (JG)	106	160	Trestles

* Number of patterns = number of unique columns in multiple sequence alignment

+ 20 thorough searches were also done

		Cores/	Memory/
Computer	Processors	node	node (GB)
Gordon	2.6-GHz Intel Sandy Bridge	16	64
Trestles	2.4-GHz AMD Magny-Cours	32	64

RAxML parallel efficiency is >0.5 up to 60 cores for >1,000 patterns; speedup is superlinear for comprehensive analysis at some core counts; scalability generally improves with number of patterns*



* Number of patterns = number of unique columns in multiple sequence alignment

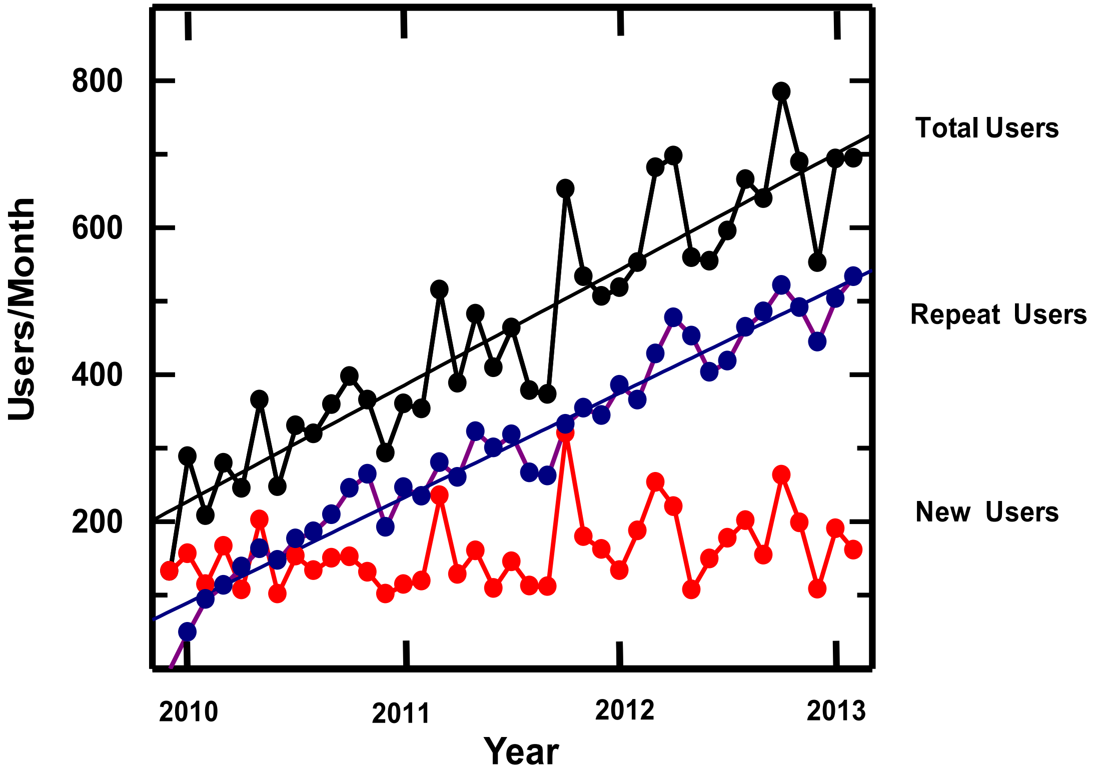
Rules for running RAxML on Trestles were developed based on benchmarking

- **Check number of searches specified by -N option**
- **If -N is not specified,**
  - Run with 8 Pthreads on 8 cores of a single node in shared queue
- **If -N n is specified with n < 50,**
  - Run with 5 MPI processes & 6 Pthreads on 30 cores of a single node in normal queue
- **If -N n is specified with n ≥ 50 or n = auto,**
  - Run with 10 MPI processes & 6 Pthreads on 60 cores of two nodes in normal queue

Some operational facts & considerations

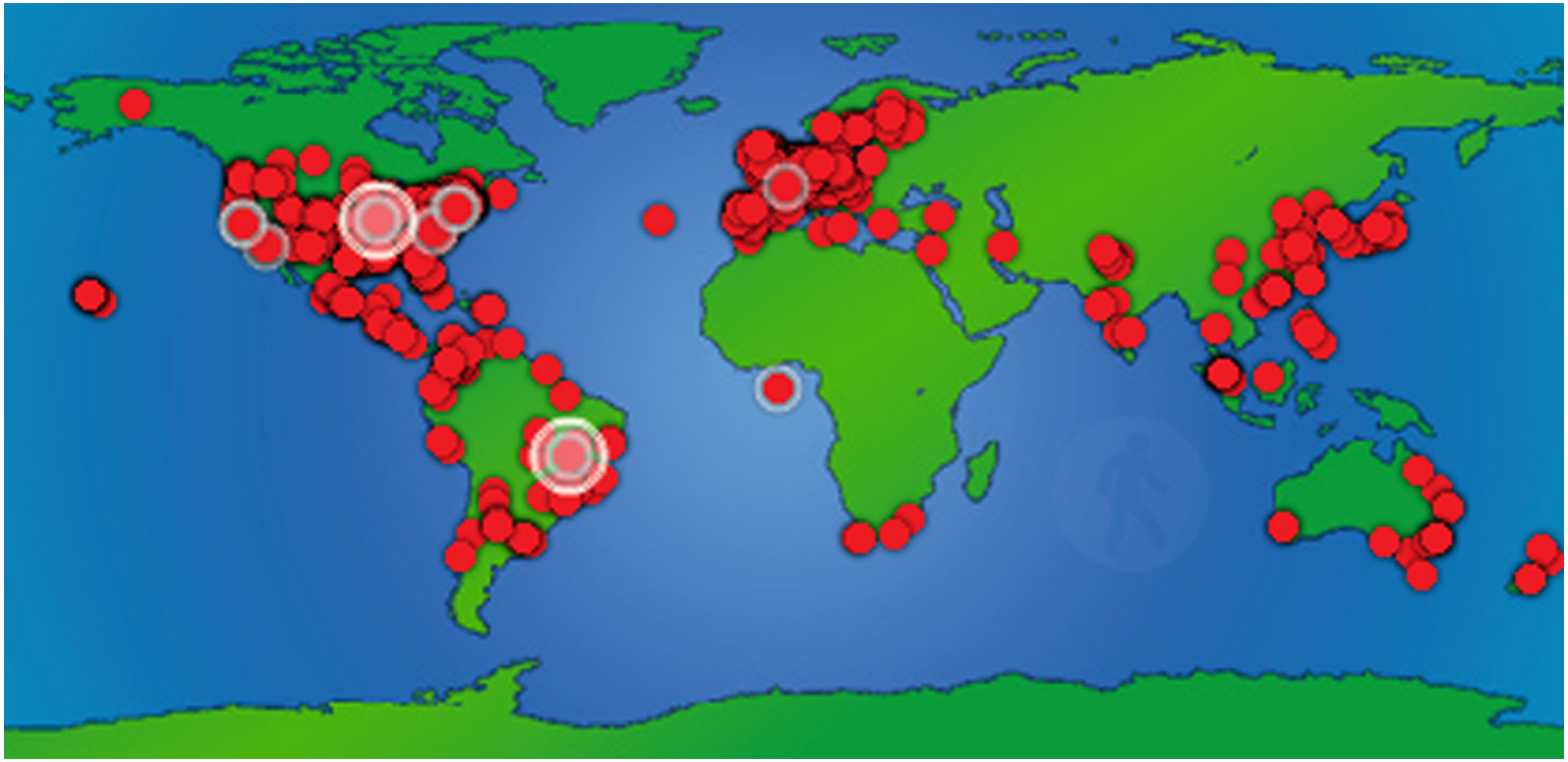
- **>100 jobs are usually running; a July 3 snapshot showed**
  - 66 MrBayes jobs using 920 cores on Gordon
  - 79 BEAST jobs using 632 cores on Trestles
  - 14 RAxML jobs using 896 cores on Trestles
  - 1 GARLI job using 32 cores on Trestles
- **Jobs are run on both systems to distribute load**
  - ~15% of load on Trestles is from CIPRES gateway jobs
- **Jobs can run a long time; allowable limits are**
  - 168 hours (1 week) on Gordon
  - 334 hours (2 weeks) on Trestles
- **I/O is done via ZFS (/projects), not Luster (/oasis)**
  - BEAST & MrBayes output frequent, small updates to log files
  - This can overwhelm the Lustre metadata servers

The CIPRES gateway has been extremely popular



- >6,000 users have run on TeraGrid/XSEDE supercomputers
- ~173,000 jobs were run & ~29M Trestles SUs were used thru Feb 2013
- >600 publications have been enabled by CIPRES use

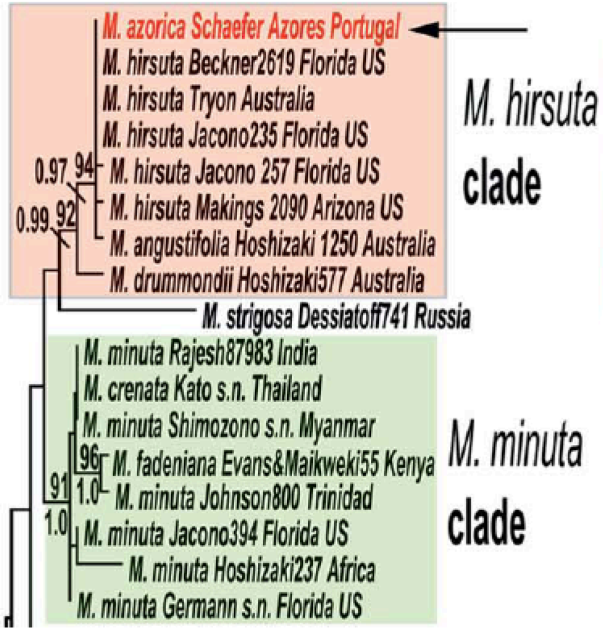
Most CIPRES gateway jobs are submitted from US, but many come from elsewhere

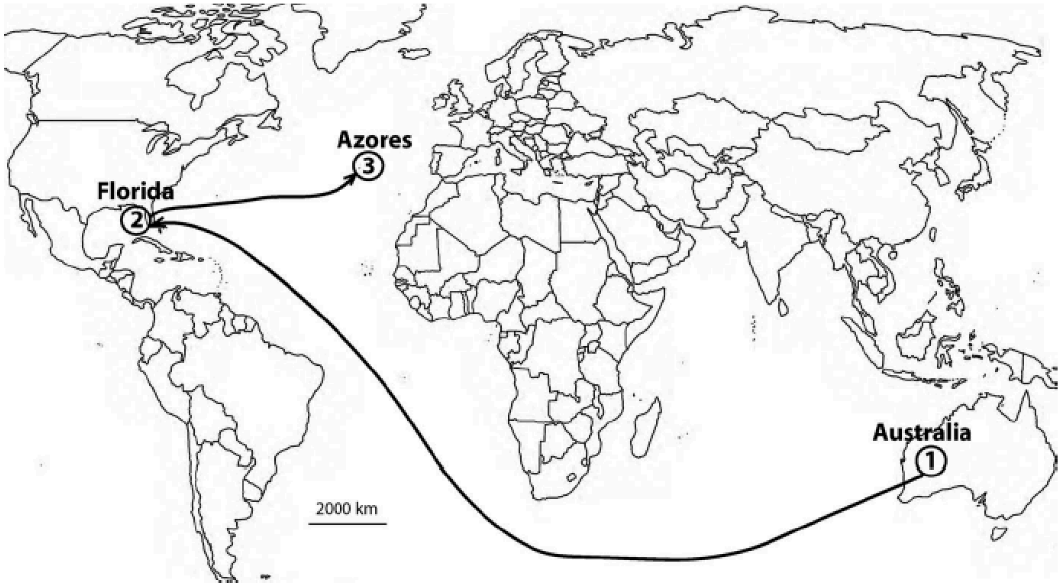


- Screen shot shows locations of 1,000 consecutive user logons as of April 20, 2011
- Highlighted dots show users online

Protected clover fern in Azores was shown to be an invasive species from Australia introduced from the US

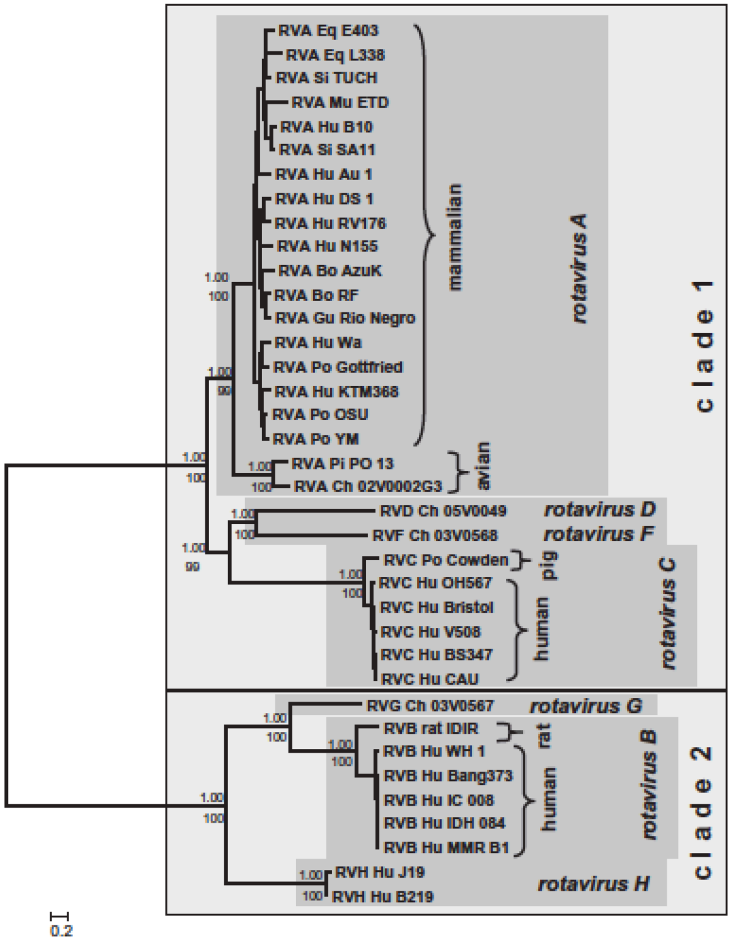
- **RAxML & MrBayes analyses were done via CIPRES gateway**
- **H. Schaefer, M.A. Carine, & F.J. Rumsey, “From European Priority Species to Invasive Weed: *Marsilea azorica* (Marsileaceae) is a Misidentified Alien,” *Systematic Biology*, v. 36, pp. 845-853 (2011)**


M. azorica Schaefer Azores Portugal
M. hirsuta Beckner2619 Florida US
M. hirsuta Tryon Australia
M. hirsuta Jacono235 Florida US
0.97
94
M. hirsuta Jacono 257 Florida US
M. hirsuta Makings 2090 Arizona US
0.99
92
M. angustifolia Hoshizaki 1250 Australia
M. drummondii Hoshizaki577 Australia
M. hirsuta clade
M. strigosa Dessiatoff741 Russia
M. minuta Rajesh87983 India
M. crenata Kato s.n. Thailand
M. minuta Shimozono s.n. Myanmar
96
M. fadeniana Evans&Maikweki55 Kenya
91
1.0
M. minuta Johnson800 Trinidad
1.0
M. minuta Jacono394 Florida US
M. minuta Hoshizaki237 Africa
M. minuta Germann s.n. Florida US
M. minuta clade



Most studies using CIPRES gateway involve basic research, but some have biomedical implications

- Humans are much more likely to infect apes with malaria than the reverse (J.C. Silva, et al., *Parasitology*, 2011)
- Graph-theory method allows viral variants to be ranked for effectiveness in vaccines (T.K. Anderson, et al., *Bioinformatics*, 2012)
- Rotoviruses are in two major clades: rotovirus A/C/D/F and rotovirus B/G/H (E. Kindler, et al., *Infection, Genetics and Evolution*, 2013) –>



Recently received NSF grant is supporting enhancements to user interface & codes

- **Enhancements in progress**
  - A RESTful web services interface in addition to browser interface is being developed
  - Additional codes are being installed
  - Hybrid parallel version of MrBayes 3.2.1 is being developed
- **Enhancements planned**
  - Input files will be checked before submission
  - Smarter rules will be developed
  - A guide will be prepared to help users decide which code(s) to use

Summary

- The CIPRES gateway has been extremely effective at providing access to HPC resources for phylogenetics researchers
- Planned enhancements will improve the gateway’s accessibility and utility

Questions?

Moki Dugway in Utah

Part of 107-mi bicycle ride
on May 10, 2013

Mt Haeckel in the Sierra

First ascent of year
on May 27, 2013