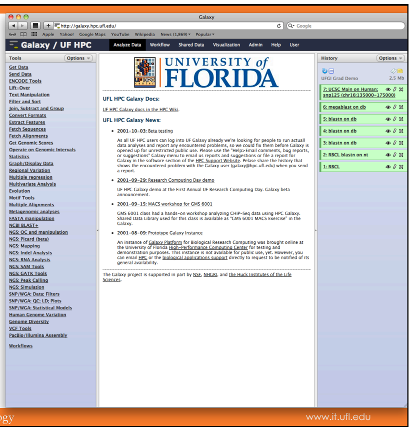
Phylogenetic Analyses at the HPC Center

Matt Gitzendanner magitz@ufl.edu

Oleksandr Moskalenko om@hpc.ufl.edu

Galaxy



Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NCBI BLAST+
NGS: QC and manipulation
NGS: Picard (beta)
NGS: Mapping
NGS: Indel Analysis
NGS: RNA Analysis
NGS: SAM Tools
NGS: GATK Tools
NGS: Peak Calling
NGS: Simulation
SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models
Human Genome Variation
Genome Diversity
VCF Tools
PacBio/Illumina Assembly

Workflows

UFL HPC Galaxy Docs:

UF HPC Galaxy docs in the HPC Wiki.

UFL HPC Galaxy News:

- **2001-10-03:** Beta testing

  As all UF HPC users can log into UF Galaxy already we're looking for people to run actuall data analyses and report any encountered problems, so we could fix them before Galaxy is opened up for unrestricted public use. Please use the "Help>Email comments, bug reports, or suggestions" Galaxy menu to email us reports and suggestions or file a report for Galaxy in the software section of the HPC Support Website. Pelase share the history that shows the encountered problem with the Galaxy user (galaxy@hpc.ufl.edu) when you send a report.

- **2001-09-29:** Research Computing Day demo

  UF HPC Galaxy demo at the First Annual UF Research Computing Day. Galaxy beta announcement.

- **2001-09-15:** MACS workshop for GMS 6001

  GMS 6001 class had a hands-on workshop analyzing CHiP-Seq data using HPC Galaxy. Shared Data Library used for this class is available as "GMS 6001 MACS Exercise" in the Galaxy.

- **2001-08-09:** Prototype Galaxy Instance

  An instance of Galaxy Platform for Biological Research Computing was brought online at the University of Florida High-Performance Computing Center for testing and demonstration purposes. This instance is not available for public use, yet. However, you can email HPC or the biological applications support directly to request to be notified of its general availability.

The Galaxy project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

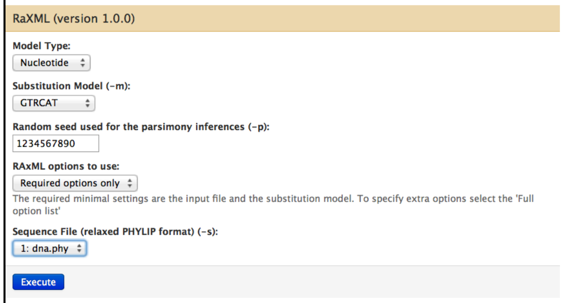
7: UCSC Main on Human: snp125 (chr16:135000-175000)
6: megablast on db
5: blastn on db
4: blastn on db
3: blastn on db
2: RBCL blastn on nt
1: RBCL

Galaxy

- RAxML
- Garli
- Beast
  - TreeAnnotator

Phylogenetics

- RaXML – Maximum Likelihood based inference of large phylogenetic trees
- Garli phylogenetic inference using the maximum-likelihood
- Beast Bayesian MCMC analysis of molecular sequences.
- TreeAnnotator BEAST tree annotator.



RaXML (version 1.0.0)

Model Type:
Nucleotide

Substitution Model (-m):
GTRCAT

Random seed used for the parsimony inferences (-p):
1234567890

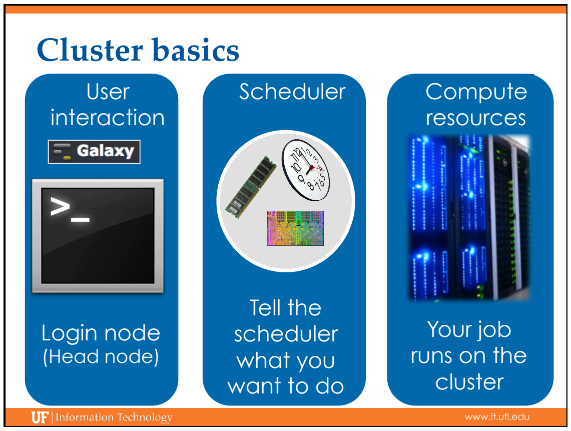
RAxML options to use:
Required options only
The required minimal settings are the input file and the substitution model. To specify extra options select the 'Full option list'

Sequence File (relaxed PHYLIP format) (-s):
1: dna.phy

The full RAxML options list is LONG!

Galaxy: Data intensive biology *for everyone*

- Accessible, reproducible, transparent computational biology
- galaxy.hpc.ufl.edu
  - Local instance of Galaxy
    - Faster access to storage, easier upload
    - Local compute resources
    - Local control



Cluster basics

User interaction

Galaxy

Login node (Head node)

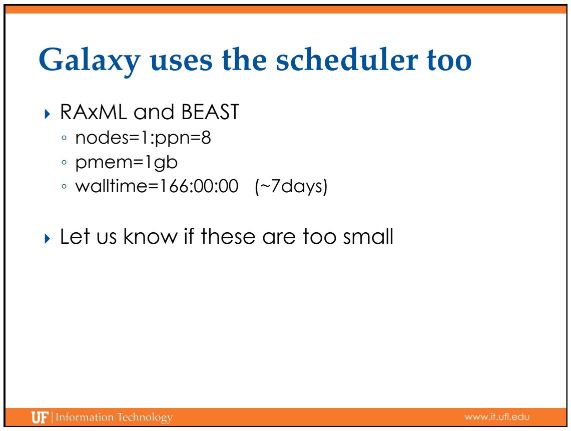
Scheduler

Tell the scheduler what you want to do

Compute resources

Your job runs on the cluster

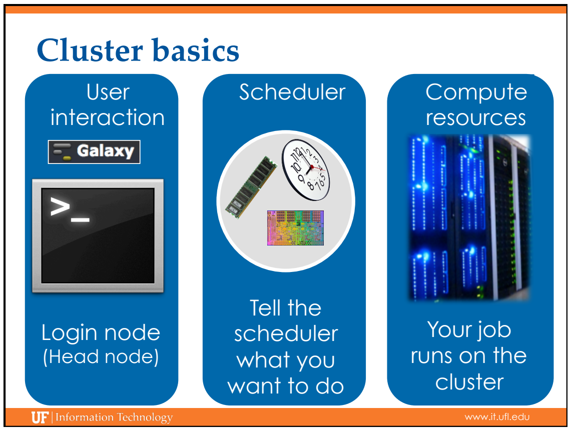
UF | Information Technology www.it.ufl.edu



Galaxy uses the scheduler too

- RAxML and BEAST
  - nodes=1:ppn=8
  - pmem=1gb
  - walltime=166:00:00 (~7days)
- Let us know if these are too small

UF | Information Technology www.it.ufl.edu



Cluster basics

User interaction

Galaxy

Login node (Head node)

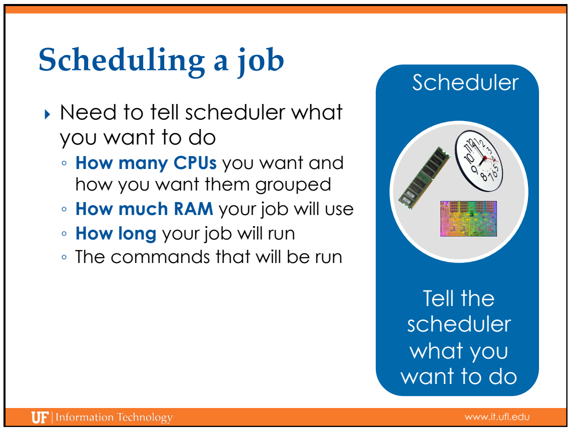
Scheduler

Tell the scheduler what you want to do

Compute resources

Your job runs on the cluster

UF | Information Technology www.it.ufl.edu



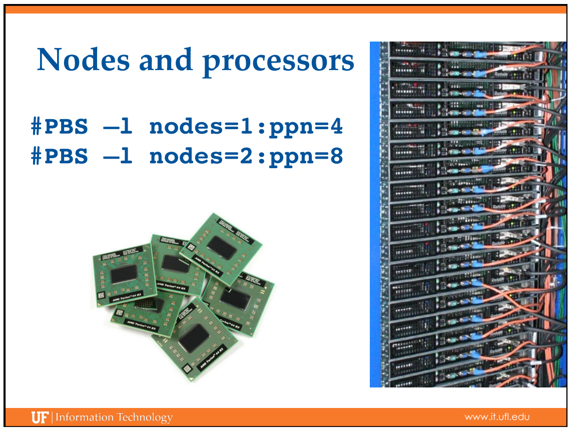
Scheduling a job

- Need to tell scheduler what you want to do
  - **How many CPUs** you want and how you want them grouped
  - **How much RAM** your job will use
  - **How long** your job will run
  - The commands that will be run

Scheduler

Tell the scheduler what you want to do

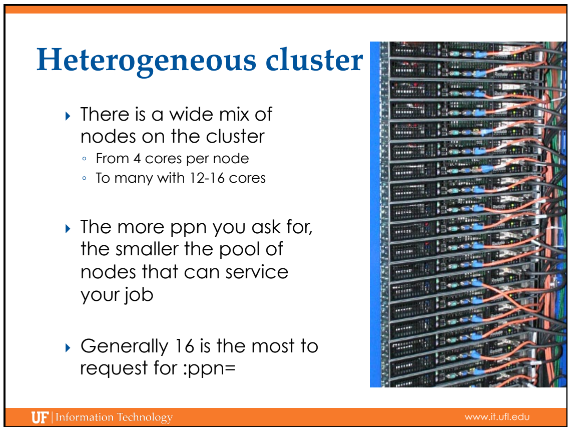
UF | Information Technology www.it.ufl.edu



Nodes and processors

```
#PBS -l nodes=1:ppn=4
#PBS -l nodes=2:ppn=8
```

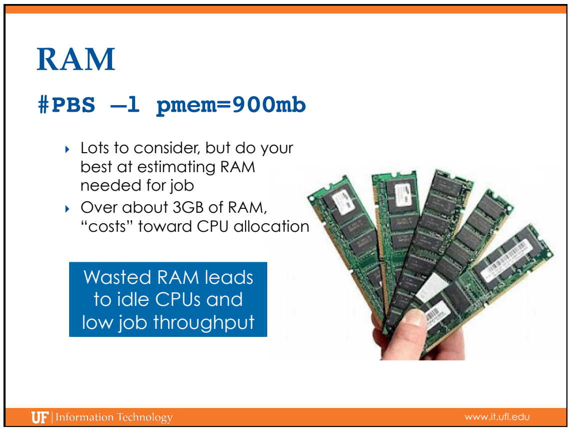
UF | Information Technology www.it.ufl.edu



Heterogeneous cluster

- There is a wide mix of nodes on the cluster
  - From 4 cores per node
  - To many with 12-16 cores
- The more ppn you ask for, the smaller the pool of nodes that can service your job
- Generally 16 is the most to request for :ppn=

UF | Information Technology www.it.ufl.edu

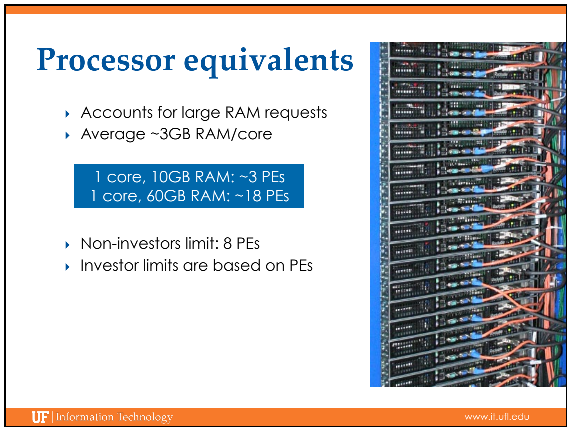


RAM

#PBS –l pmem=900mb

- Lots to consider, but do your best at estimating RAM needed for job
- Over about 3GB of RAM, "costs" toward CPU allocation

Wasted RAM leads to idle CPUs and low job throughput

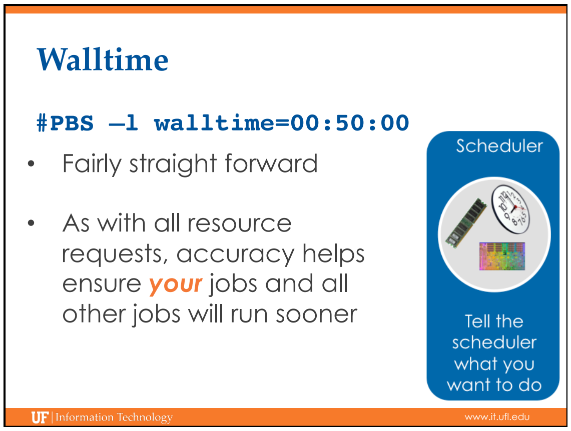


Processor equivalents

- Accounts for large RAM requests
- Average ~3GB RAM/core

1 core, 10GB RAM: ~3 PEs
1 core, 60GB RAM: ~18 PEs

- Non-investors limit: 8 PEs
- Investor limits are based on PEs



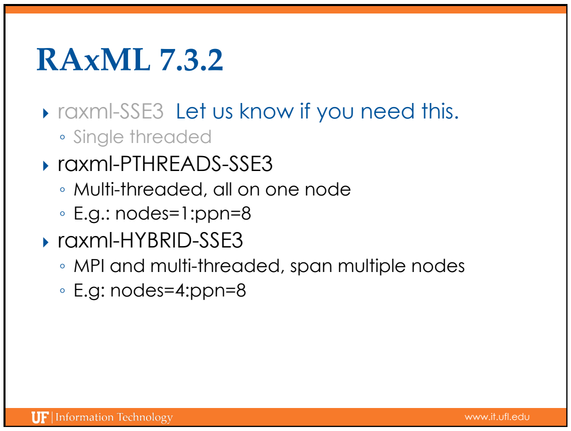
Walltime

#PBS –l walltime=00:50:00

- Fairly straight forward
- As with all resource requests, accuracy helps ensure ***your*** jobs and all other jobs will run sooner

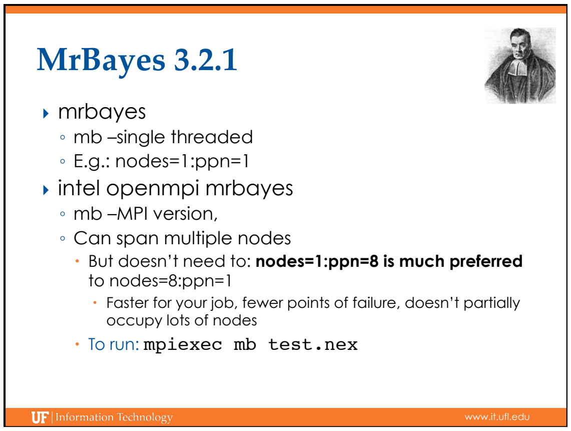
Scheduler

Tell the scheduler what you want to do



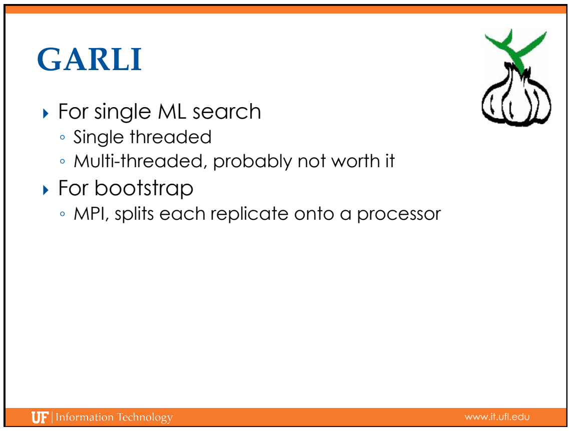
RAxML 7.3.2

- raxml-SSE3 Let us know if you need this.
  - Single threaded
- raxml-PTHREADS-SSE3
  - Multi-threaded, all on one node
  - E.g.: nodes=1:ppn=8
- raxml-HYBRID-SSE3
  - MPI and multi-threaded, span multiple nodes
  - E.g: nodes=4:ppn=8



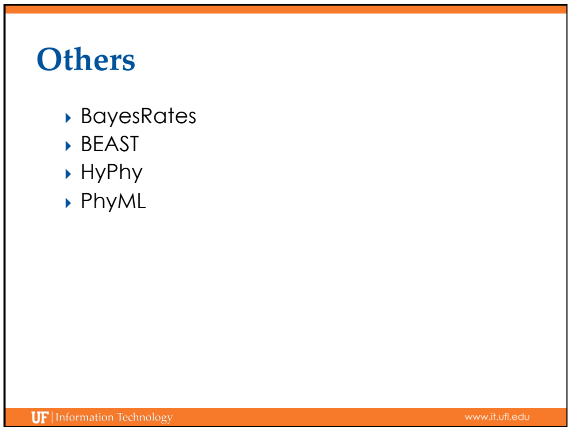
MrBayes 3.2.1

- mrbayes
  - mb –single threaded
  - E.g.: nodes=1:ppn=1
- intel openmpi mrbayes
  - mb –MPI version,
  - Can span multiple nodes
    - But doesn't need to: **nodes=1:ppn=8 is much preferred** to nodes=8:ppn=1
      - Faster for your job, fewer points of failure, doesn't partially occupy lots of nodes
    - To run: `mpiexec mb test.nex`



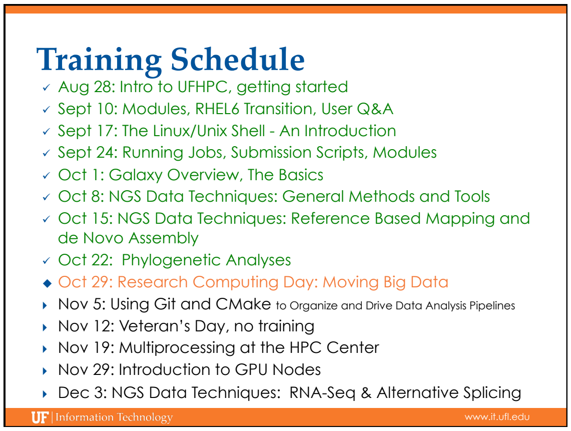
GARLI

- For single ML search
  - Single threaded
  - Multi-threaded, probably not worth it
- For bootstrap
  - MPI, splits each replicate onto a processor



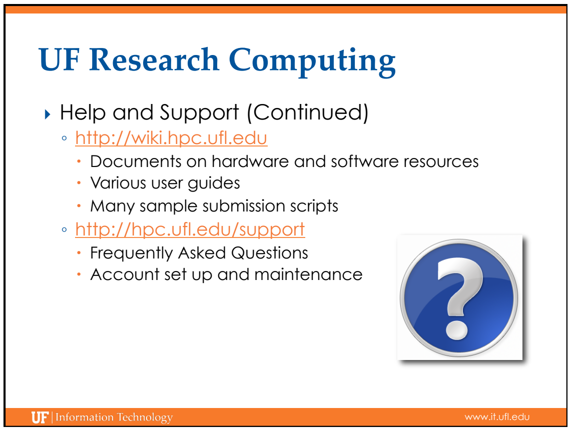
Others

- BayesRates
- BEAST
- HyPhy
- PhyML



Training Schedule

- ✓ Aug 28: Intro to UFHPC, getting started
- ✓ Sept 10: Modules, RHEL6 Transition, User Q&A
- ✓ Sept 17: The Linux/Unix Shell - An Introduction
- ✓ Sept 24: Running Jobs, Submission Scripts, Modules
- ✓ Oct 1: Galaxy Overview, The Basics
- ✓ Oct 8: NGS Data Techniques: General Methods and Tools
- ✓ Oct 15: NGS Data Techniques: Reference Based Mapping and de Novo Assembly
- ✓ Oct 22: Phylogenetic Analyses
- ◆ Oct 29: Research Computing Day: Moving Big Data
- Nov 5: Using Git and CMake to Organize and Drive Data Analysis Pipelines
- Nov 12: Veteran's Day, no training
- Nov 19: Multiprocessing at the HPC Center
- Nov 29: Introduction to GPU Nodes
- Dec 3: NGS Data Techniques: RNA-Seq & Alternative Splicing



UF Research Computing

- Help and Support (Continued)
  - http://wiki.hpc.ufl.edu
    - Documents on hardware and software resources
    - Various user guides
    - Many sample submission scripts
  - http://hpc.ufl.edu/support
    - Frequently Asked Questions
    - Account set up and maintenance