



**Open Science Grid**

**Galaxy based BLAST submission to distributed  
high throughput computing resources**

Rob Quick

Slides Prepared by Soichi Hayashi

Open Science Grid Operations  
Indiana University / Research Technologies

# Topics

- What is BLAST / Galaxy?
- Why BLAST on OSG?
- How to run BLAST on HTC?
- Conclusion and future TODO...

# NCBI-BLAST

**NCBI** (National Center for Biotechnology Information)

**BLAST** (Basic Local Alignment Search Tool)

Popular application for Bioinformaticists

Compares biological sequences

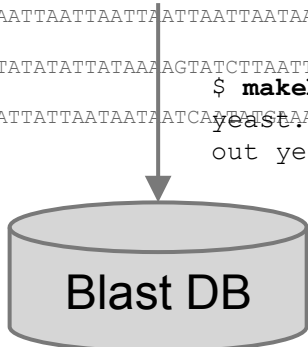
- Identify unknown sequences
- Discover related organism

# Database Source fasta

```
>gi|6226515|ref|NC_001224.1| Saccharomyces cerevisiae mitochondrion
TTCATAATAAATTTTTTATATATATATTATATTATAAATTAATTTATATTATAAAAAATAATTTATTATTTAAAAATAT
T
TATTCTCCTTTCGGGGTTCGGGCTCCCGTGGCCGGGCCCCGGAATTATTAATTAATAATAAATTATTATTAATAAATTAT
T
TATTATTTTATCATTAAAAATATAAAATAAAAAATATTAAGATAAAAAAATAATGTTTATTCTTTATATAAAATTA
T
ATATATATATATAATTAATTAATTAATTAATTAATTAATTAATAATAAAAAATAAATATAAAATAATATAAAATATTATTCTTT
A
TTAATAAATATATATTTTATATATATAAAAGTATCTTAATTAATAAAAAATAAACATTTAATAATATGAATTATATATTA
T
TATTATTATTAATAAAATTATTAATAATTAATCAATTAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAA
... (150,000 lines)
```

# Input Query (Unknown Organism)

```
>CHR1.19971009 Chromosome I Sequence
CCACACCACACCCACACCCACACACCACACCACACACCACACCACACCCACACACACA
CATCCTAACACTACCCTAACACAGCCCTAATCTAACCCCTGGCCAACCTGTCTCTCAACTT
ACCCTCCATTACCCTGCCTCCACTCGTTACCCTGTCCCATTCAACCATAACCACTCCGAAC
CACCATCCATCCCTCTACTTACTACCCTCACCACCGTTACCCTCCAATTACCCATATC
CAACCCACTGGCACTTACCCTACCATTACCCTACCATCCACCATGACCTACTCACCATAC
TGTTCTTCTACCCACCATATTTGAAACGCTAACAAATGATCGTAAATAACACACACGCTGCT
TACCCTACCCTTTTATACCACCACCATGCCATACTCACCTCACTTGTATACTGATTT
TACGTACGCACACGGATGCTACAGTATATACCATCTCAAACCTACCCTACTCTCAGATTC
CACTTCACTCCATGGCCATCTCTCACTGAATCAGTACCAAATGCCTCACATCATTATG
CAGGGCACTTGCCCTCAGCGGTCTATACCCTGTGCCATTTACCATAACGCCCATCATTAT
CCACATTTTGATATCTATATCTCATTTCGGCGGTCCCAAATATTGTATAACTGCCCTTAAT
ACATACGTTATACCCTTTTGCCACATATACTTACCCTCCATTTATATACACTTATGTC
AATATTACAGAAAAATCCCCACAAAAATCACCTAAACATAAAAAATATTCTACTTTTCAAC
```



```
$ makeblastdb -in
yeast.fasta -dbtype nucl
out yeast
```

```
$ blastn -db mydb -query input_query.fasta -out output.txt -outfmt 1
```

comp10597_c0_seq1	Uextra	100.00	28 0 0 168	195	3953904	3953931	4e-06	52.8
comp10597_c0_seq1	Uextra	100.00	28 0 0 168	195	28550642	28550615	4e-06	52.8
comp12438_c0_seq1	2L	100.00	29 0 0 116	144	8509466	8509494	2e-06	54.7
comp12438_c0_seq2	2L	100.00	29 0 0 134	162	8509466	8509494	2e-06	54.7

# Common Blast Databases

## NCBI RefSeq Databases

NT/NR (10-20 parts 400-800M each compressed)

Collection of taxonomically diverse, non-redundant and richly annotated sequences.

\* plasmids, organelles, viruses, archaea, bacteria, and eukaryotes.

patnt/pataa (1-4 parts 1G each)

Patent database from USPTO or from EU/Japan Patent Agencies via EMBL/DDBJ

## Flybase Databases

dmel-all-chromosome

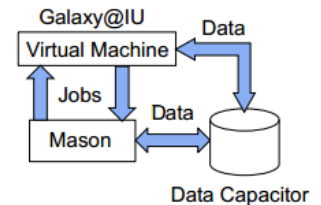
# Galaxy

A popular Web-based platform for data intensive biomedical research

**NCGAS** (National Center for Genome Analysis Support) hosts an instance of Galaxy portal

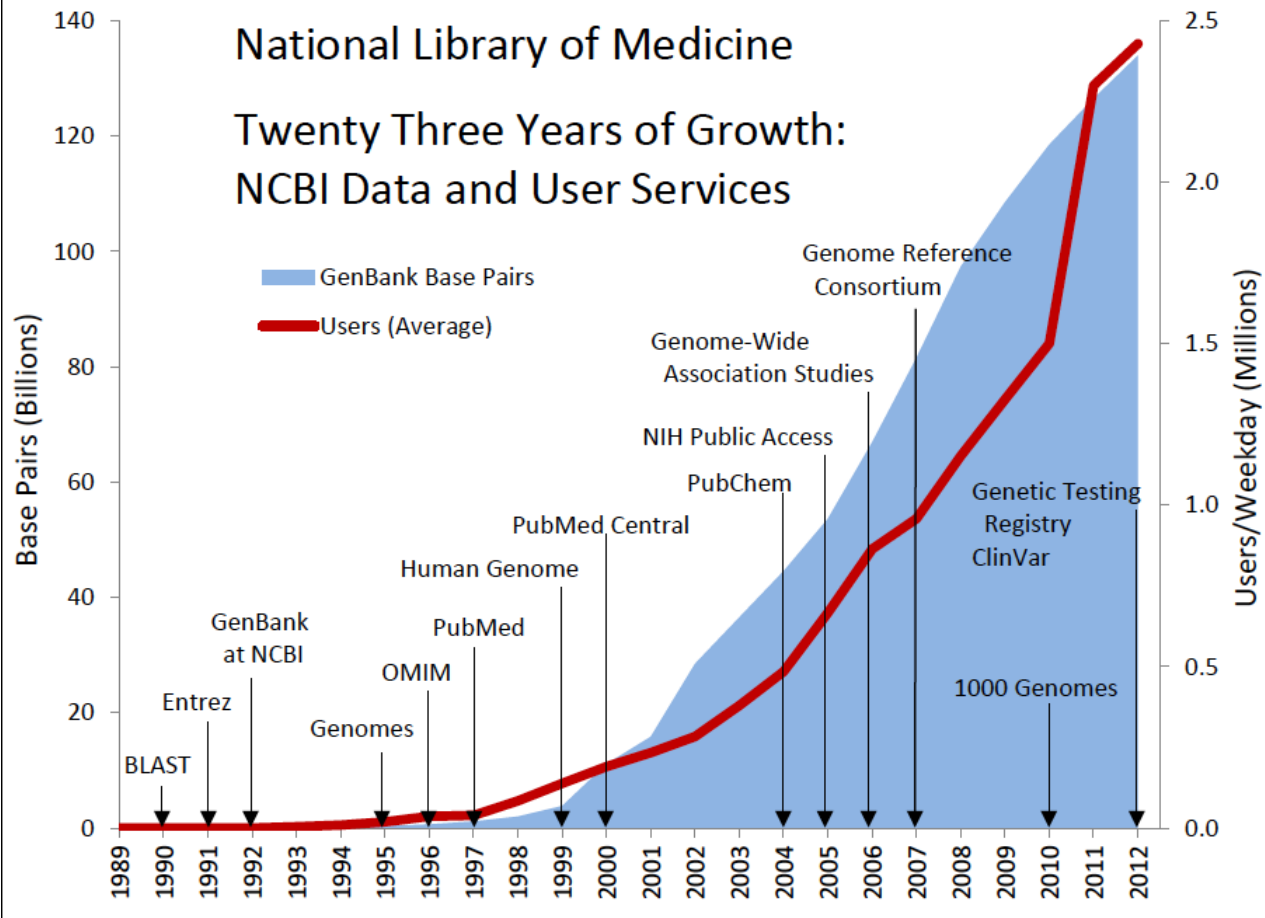
- IU Mason Cluster (8TB-memory)
- Access to IU DC2 (3.5PB)
- Genome assembly
- Large-scale phylogenetic software
- Blast

Our instance at IU



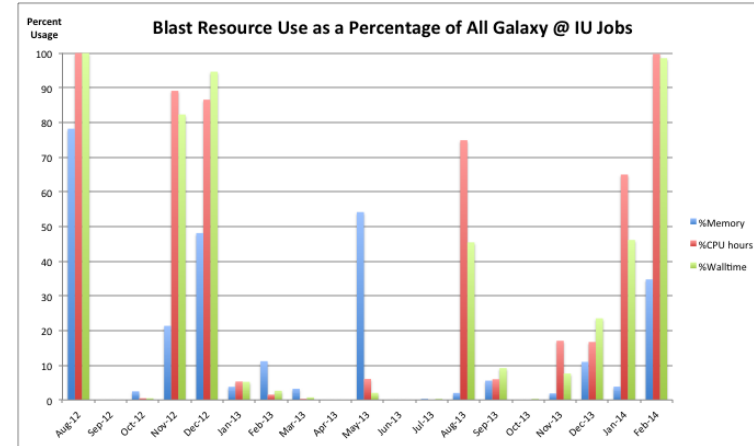
# National Library of Medicine

## Twenty Three Years of Growth: NCBI Data and User Services



# Why BLAST on OSG?

- BLAST is CPU intensive (not memory)
- IU/Mason is not an optimal resource to run BLAST
- Growth in data volume will squeeze available resource capacity at NCGAS in coming years.
- OSG's opportunistic resource could be used as an alternative for Mason and can provide necessary resource capacity.





Galaxy

galaxy.ncgas-globus.indiana.edu/root

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 35.2 GB

Tools

search tools

[Import Data](#)

[Data Manipulation](#)

[Quality Control](#)

[De novo Assembly](#)

[Mapping and Alignments](#)

[Run Blast+](#)

**[Run Blast+ on Open Science Grid](#)**

NCBI BLAST+ Search database with query sequence(s)

[Annotation](#)

[Statistics](#)

[Variants](#)

[Clustering/Phylogeny](#)

[Visualization](#)

[Workflows](#)

NCBI BLAST+ (version 0.0.17)

**Choose which Blast+ program to run::**

blastn - search nucleotide databases using a nucleotide query.

**Subject database/sequences:**

BLAST Database

**BLAST database - make sure you are using the correct type (nuc/prot)!:**

NCBI NT 01-22-2014

**Nucleotide query sequence(s):**

3: nucl.2000.fasta

**Type of BLAST:**

normal blastn

blastn-short

megablast

dc-megablast

**Set expectation value cutoff:**

10 - Blast default

**Output format is currently set to xml.:**

BLAST XML

Execute

History

Unnamed history

2.2 GB

5: normal blastn on db

4: normal blastn on db

3: nucl.2000.fasta

2: normal blastn on db

1: nucl.2000.fasta

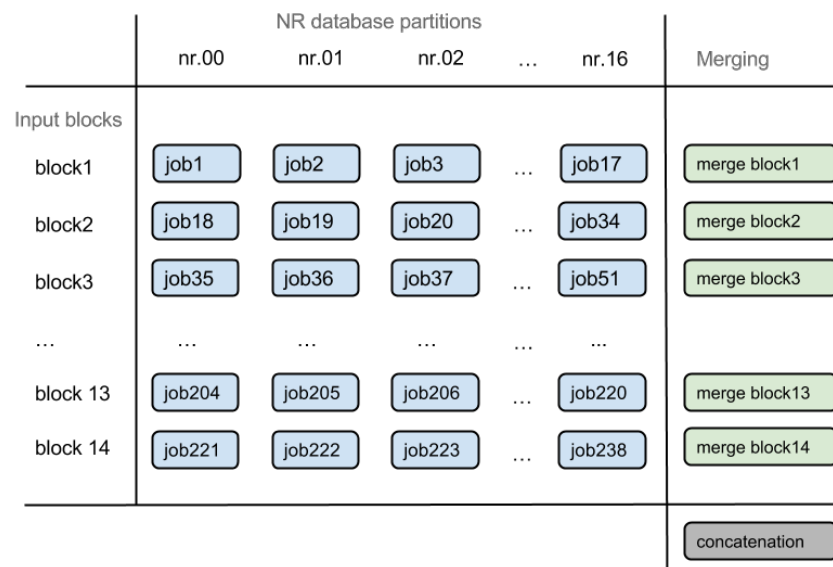
**Note.** Our current implementation only supports the nr database. More to be added soon!

# osg-blast (v2)

- Written in nodejs / node-osg & node-htcondor modules
- Can be installed on any OSG submit hosts via “npm install osg-blast”
- Hosted databases (NT/NR) distributed via OASIS (CVMFS)
- Needs to be highly reliable and autonomous
  - Handle unexpected issues well
  - Needs to figure out the best configuration by itself.
  - Report site specific issues to GOC (and recover)
  - Cleanup after itself (removing temp files, canceling jobs)

# osg-blast (v2)

- Splits both input queries / databases and run all jobs in parallel.
- Results are merged to create a single output sorted by e-value.

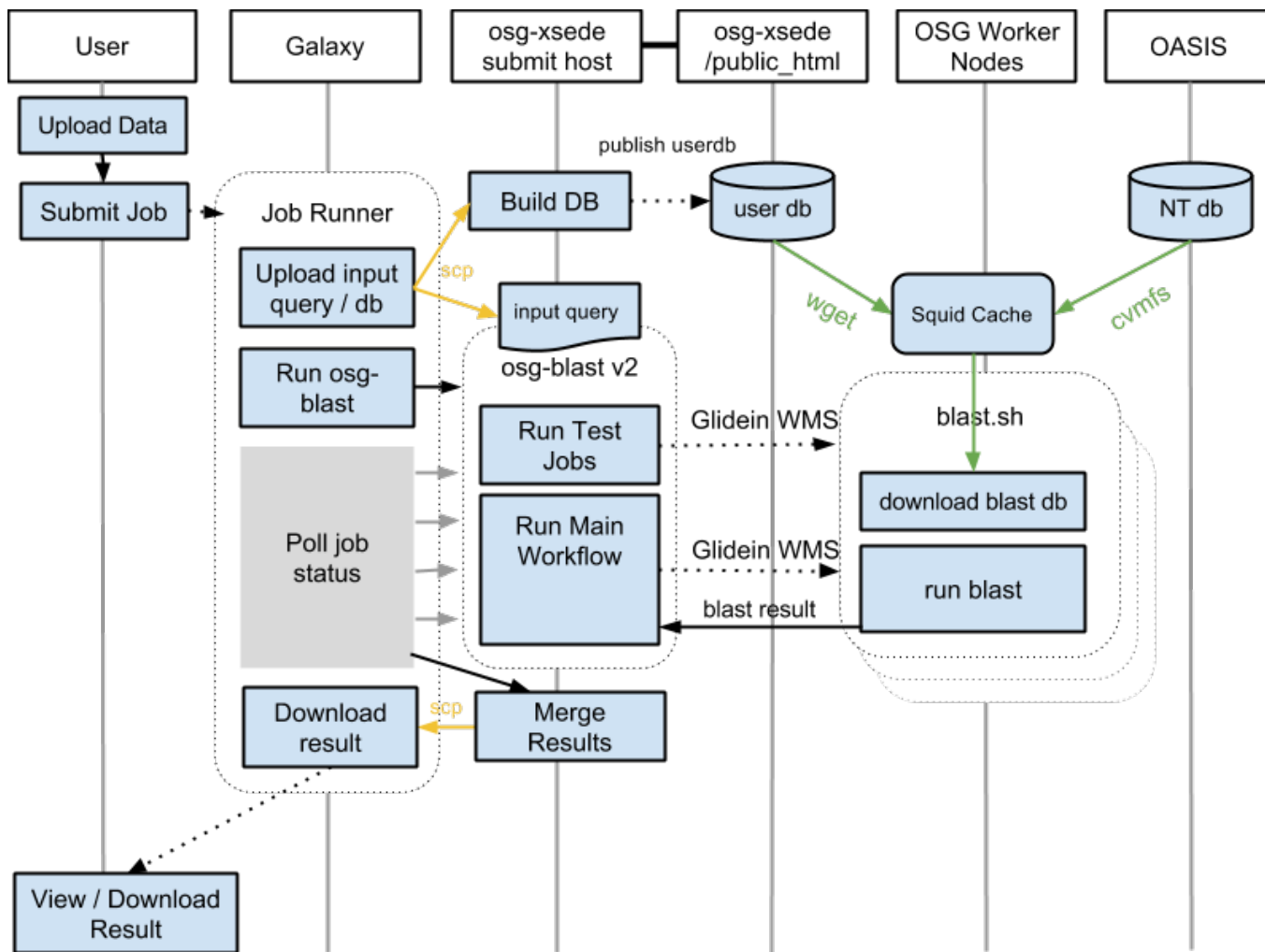


## Test Stage

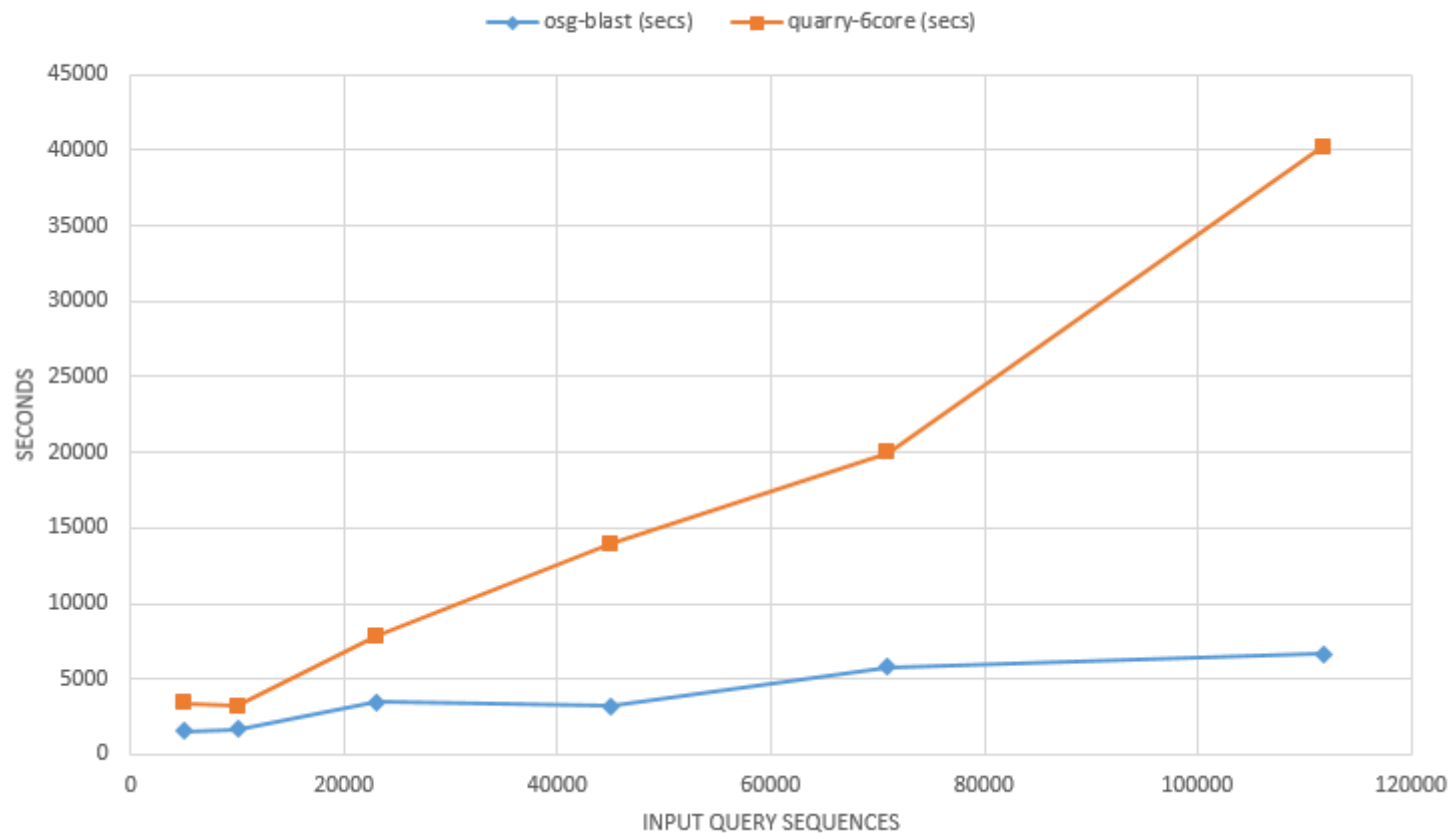
- Determine best input block size
- Detects issue with user input / OSG environment.

## Main Stage

- Submit all jobs using information gathered during the test stage.
- Use -dbsize to correct e-value



## BLASTN / NT DATABASE



# Conclusions

- Clearly, we will need more computing resources to run BLAST in coming years, and OSG's opportunistic environment can provide that need.
- Galaxy allows bioinformatics community to use existing UI to submit BLAST jobs.
- BLAST works well in HTC environment, and it seems to scale as expected using OSG's opportunistic resources.

## Challenges / Future Goal

- osg-blast workflow needs to be highly robust (error-tolerant), reliable, and self-diagnosing to be practical (can't rely on users to fix problems)
- osg-blast output merger needs to be implemented for other output formats.
- Might need to explore alternative to CVMFS for hosting BLAST DBs.

# Acknowledgements

Bill Barnett, Tom Doak, Rich LeDuc (SCT @ IU)

Ruth Pordes, Chander Seghal (Fermilab)

Derek Weitzel (UNL)

Mats Rynge (Information Science Institute @ USC)

Alain Deximo, Kyle Gross, Tom Lee, Vince Neal, Chris Pipes, and  
Michel Tavares (OSG Operations Center @ IU)

## Contacts

Soichi Hayashi [hayashis@iu.edu](mailto:hayashis@iu.edu) @soichih | soichi.us

Rob Quick [rquick@iu.edu](mailto:rquick@iu.edu)