INSTITUTE FOR GENOME SCIENCES
University of Maryland School of Medicine

DIAG: Data Intensive Academic Grid

A computational platform for bioinformatics analyses and training

Owen White
Anup Mahurkar

Outline

- Who we are
- Motivation for MRI-R^2
- System Description
- Highlight of users and applications
- Timeline
- Challenges

Who we are

- Part of the University of Maryland School of Medicine in Baltimore
- Quasi-independent institute founded about 2 ½ years ago
- IGS has ~20 faculty members
- Total size of ~100
- Small to medium size sequencing center
- Mix of sequencing platforms including Sanger, Illumina, and Roche 454
- Areas of focus: microbial genomics, human genomics, and metagenomics
- http://www.igs.umaryland.edu

Motivation

- Sequencing technology improvements
  - 1st Generation
    - Sanger-based capillary sequencing with throughput of ~500 Kb/run
  - Lots of players on the horizon
    - 454, Illumina, Solid, Helicos, Visigen, PacBio, Complete Genomics, Oxford Molecular, NABsys, IBM, Life Technologies
  - 2nd Generation
    - 454 pyrosequencing – 160-320 Mb/run (8 hour run)
    - Illumina HiSeq2000 200 Gb/run (8 day run) or 25 Gb/day
    - ABI SOLiD 200-300 Gb/run
  - 3rd Generation
    - Pacific Biosciences,
      - Human genome < 15 mins for $100 by 2014
    - Complete Genomics
      - 50 human genomes completed
      - 500 on order to be completed in 2010
      - 18 genomes per 11 day run will rise to 120 genomes per run
      - Roadmap calls for 1 million genomes/year capacity in next few years

Motivation

- Democratization of sequencing with 100s if not 100s of these sequencers being sold
- Burgeoning data sets
  - 1000 human genomes data
  - 1000s of cancer genomes
  - 2000 bacterial genomes
  - 10s of plant genomes
  - Metagenomes for various environments including soil, ocean, air, human body, etc

Motivation

- Newer applications/projects
  - 1000 genomes project
  - Cancer Genomics
  - Transcriptomics
  - Epigenetics
  - Population genomics
  - Metegenomics
    - 16S rRNA based community classification
    - Whole genome sequencing of metagenomes
    - Metatranscriptomics
- Computationally intensive
  - A *de novo* assembly will take order of a day or two on a 128 node cluster using tools like *ABySS*
  - Metagenomic annotation using Blast, and HMM search will take over 9000 hours on a single core for data generated by a 454 in 8 hours
- Lack of proximity between reference data sets and computational resources

DIAG Highlights

- Over 20 users from 13 US and international institutions
- Diverse applications
  - Microbial genomics
    - Annotation
    - Comparative Genomics
  - Plant Genomics
    - Assembly
    - Annotation
    - Transcriptomics
  - Metagenomics
    - Marine
    - Human
    - Plant
    - Environment
  - Proteomics
  - Livestock Research

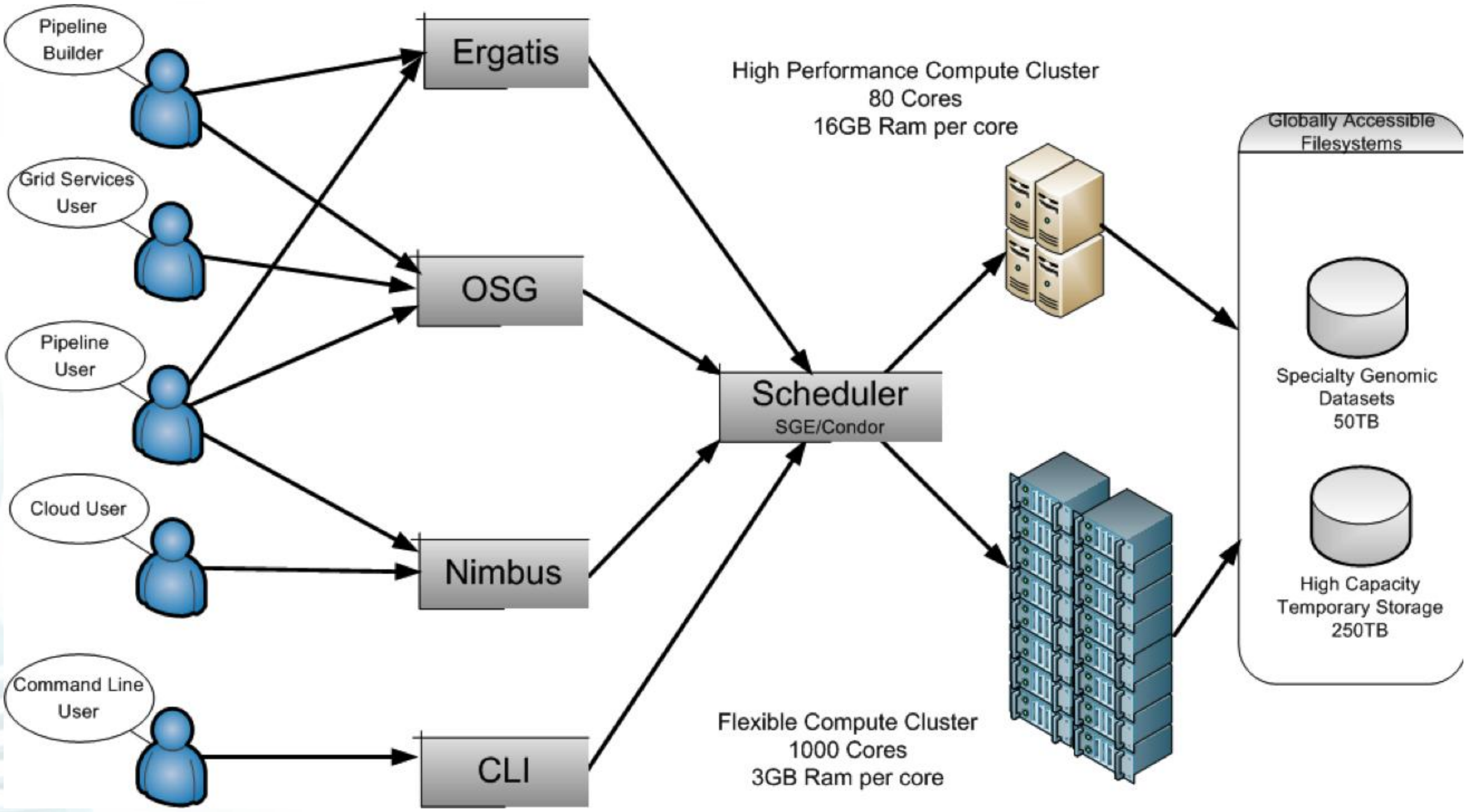
Data Sets

- Sources include: GenBank, RefSeq, Uniprot (including Swiss-Prot), UniProtKB, PDB, PIR, Ensembl, EMBL, CAMERA, MG-RAST, RAST sub-systems and Greengenes.
- Non-redundant protein and nucleotide data sets
  - Custom data sets for bacteria, viruses, eukaryotes, mammals generated by PANDA
- Data from 1000 genomes project
- Metagenomics data sets
  - CAMERA
  - Virome
  - Human Microbiome Project
- Transcriptomes for various model organisms including human, mouse, drosophila, Arabidopsis, etc.

System Highlights

- 100-125 high-throughput compute nodes
  - 400 GB local storage per node
  - 48 GB RAM per node
  - Intel/AMD processors
- 5 high-performance compute nodes
  - 1 TB local storage per node
  - 12-16 cores per node
  - ~12-18 GB RAM per core
  - InfiniBand QDR interconnect
- 400-600 Tera Bytes (TB) storage
  - High-performance, grid-attached parallel file system (GPFS, Lustre, Isilon, Panasas)
  - Possibly hierarchical storage
  - Archival upon request

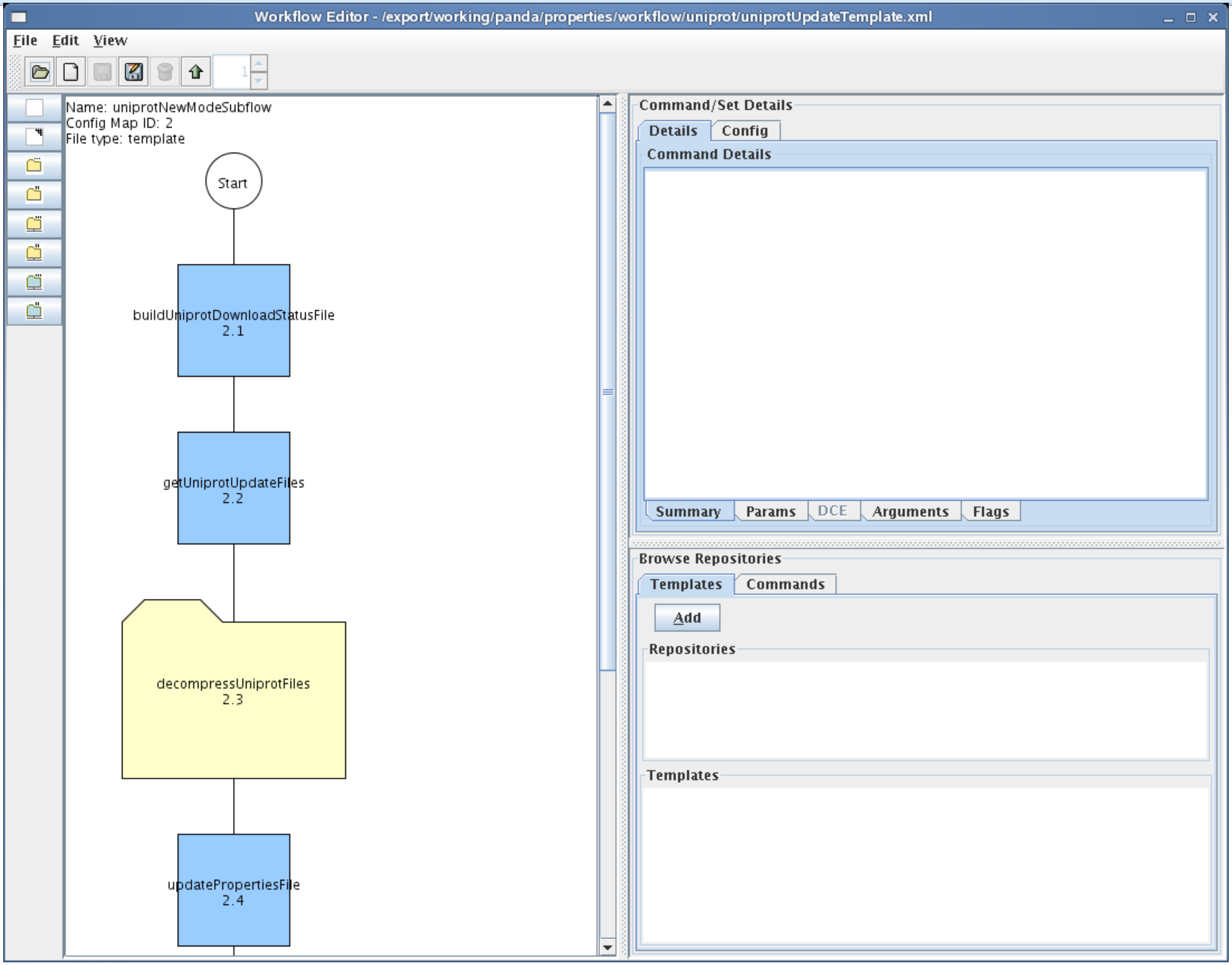
System Highlights



TIGR Workflow

- Open source workflow management software written in Java at
- Written at The Institute for Genomic Research/J Craig Venter Institute
- Currently maintained by our group at IGS
- http://tigr-workflow.sourceforge.net/
- In use for over 8 years
- A GUI for authoring pipelines
- Multi-threaded workflow execution engine
- A GUI for monitoring pipeline execution
- Job execution on a single host or on a grid
- Built in support for Condor and Sun Grid Engine
- Working on support for PBS and a generic grid through DRMAA

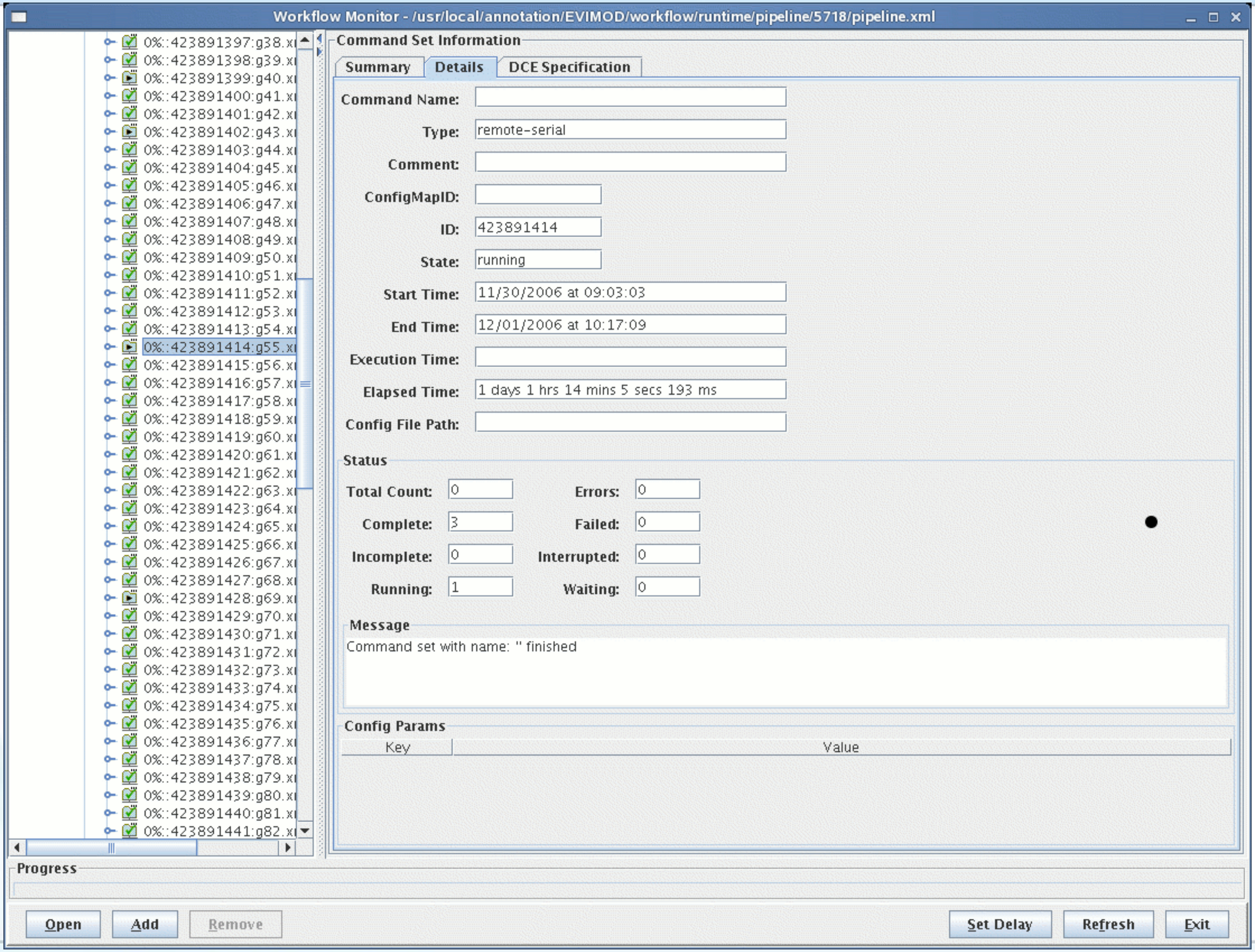
TIGR Workflow – Authoring Tool



Workflow Editor - /export/working/panda/properties/workflow/uniprot/uniprotUpdateTemplate.xml

Name: uniprotNewModeSubflow
Config Map ID: 2
File type: template

TIGR Workflow – Monitoring Tool



Workflow Monitor - /usr/local/annotation/EVIMOD/workflow/runtime/pipeline/5718/pipeline.xml

- ☑ 0%:423891397:g38.xı
- ☑ 0%:423891398:g39.xı
- ▶ 0%:423891399:g40.xı
- ☑ 0%:423891400:g41.xı
- ☑ 0%:423891401:g42.xı
- ▶ 0%:423891402:g43.xı
- ☑ 0%:423891403:g44.xı
- ☑ 0%:423891404:g45.xı
- ☑ 0%:423891405:g46.xı
- ☑ 0%:423891406:g47.xı
- ☑ 0%:423891407:g48.xı
- ☑ 0%:423891408:g49.xı
- ☑ 0%:423891409:g50.xı
- ☑ 0%:423891410:g51.xı
- ☑ 0%:423891411:g52.xı
- ☑ 0%:423891412:g53.xı
- ☑ 0%:423891413:g54.xı
- ▶ 0%:423891414:g55.xı
- ☑ 0%:423891415:g56.xı
- ☑ 0%:423891416:g57.xı
- ☑ 0%:423891417:g58.xı
- ☑ 0%:423891418:g59.xı
- ☑ 0%:423891419:g60.xı
- ☑ 0%:423891420:g61.xı
- ☑ 0%:423891421:g62.xı
- ☑ 0%:423891422:g63.xı
- ☑ 0%:423891423:g64.xı
- ☑ 0%:423891424:g65.xı
- ☑ 0%:423891425:g66.xı
- ☑ 0%:423891426:g67.xı
- ☑ 0%:423891427:g68.xı
- ▶ 0%:423891428:g69.xı
- ☑ 0%:423891429:g70.xı
- ☑ 0%:423891430:g71.xı
- ☑ 0%:423891431:g72.xı
- ☑ 0%:423891432:g73.xı
- ☑ 0%:423891433:g74.xı
- ☑ 0%:423891434:g75.xı
- ☑ 0%:423891435:g76.xı
- ☑ 0%:423891436:g77.xı
- ☑ 0%:423891437:g78.xı
- ☑ 0%:423891438:g79.xı
- ☑ 0%:423891439:g80.xı
- ☑ 0%:423891440:g81.xı
- ☑ 0%:423891441:g82.xı

Command Set Information

Command Name:

Type: remote-serial

Comment:

ConfigMapID:

ID: 423891414

State: running

Start Time: 11/30/2006 at 09:03:03

End Time: 12/01/2006 at 10:17:09

Execution Time:

Elapsed Time: 1 days 1 hrs 14 mins 5 secs 193 ms

Config File Path:

Status

Total Count:	0	Errors:	0
Complete:	3	Failed:	0
Incomplete:	0	Interrupted:	0
Running:	1	Waiting:	0

Message

Command set with name: '' finished

Config Params

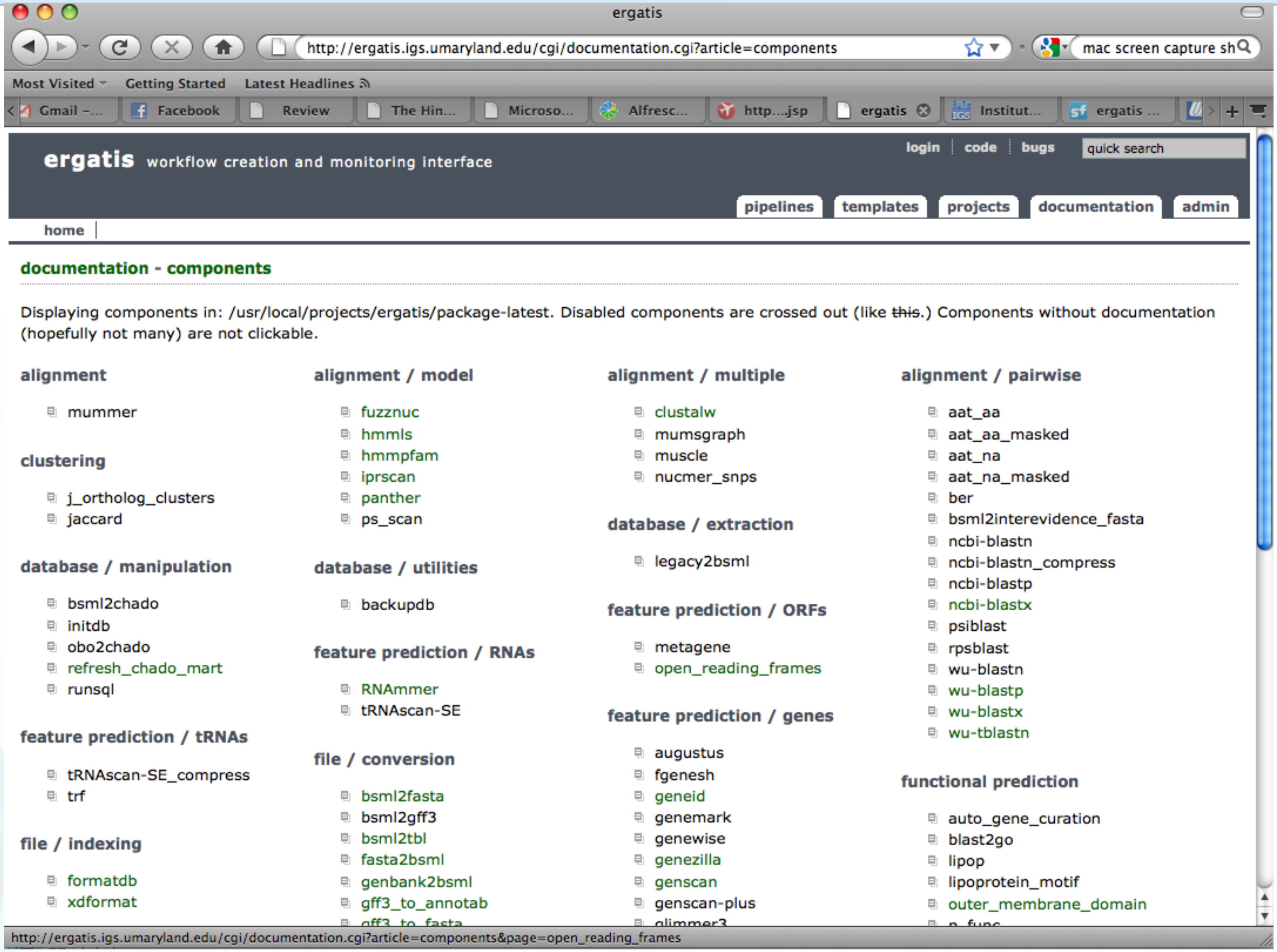
Key	Value

Progress

Ergatis

- Open source web-based pipeline creation and monitoring platform
- http://sourceforge.net/projects/ergatis/
- Uses TIGR Workflow as job execution engine
- Generates workflow XML directly
- Has over 50 commonly used bioinformatics applications used for genome assembly, sequence searches
- In use for over 8 years at a number of institutions
- Used in a number of major research projects including the analysis of Annotation of Plasmodium Genomes, Global Ocean Survey Metagenomics

Ergatis – Component List



http://ergatis.igs.umaryland.edu/cgi/documentation.cgi?article=components

ergatis workflow creation and monitoring interface

documentation - components

Displaying components in: /usr/local/projects/ergatis/package-latest. Disabled components are crossed out (like ~~this~~.) Components without documentation (hopefully not many) are not clickable.

alignment

- mummer

clustering

- j_ortholog_clusters
- jaccard

database / manipulation

- bsml2chado
- initdb
- obo2chado
- refresh_chado_mart
- runsql

feature prediction / tRNAs

- tRNAscan-SE_compress
- trf

file / indexing

- formatdb
- xdformat

alignment / model

- fuzznuc
- hmmls
- hmmpfam
- iprscan
- panther
- ps_scan

database / utilities

- backupdb

feature prediction / RNAs

- RNAmmer
- tRNAscan-SE

file / conversion

- bsml2fasta
- bsml2gff3
- bsml2tbl
- fasta2bsml
- genbank2bsml
- gff3_to_annotab
- gff3_to_fasta

alignment / multiple

- clustalw
- mumsgraph
- muscle
- nucmer_snps

database / extraction

- legacy2bsml

feature prediction / ORFs

- metagene
- open_reading_frames

feature prediction / genes

- augustus
- fgenesh
- geneid
- genemark
- genewise
- genezilla
- genscan
- genscan-plus
- glimmer3

alignment / pairwise

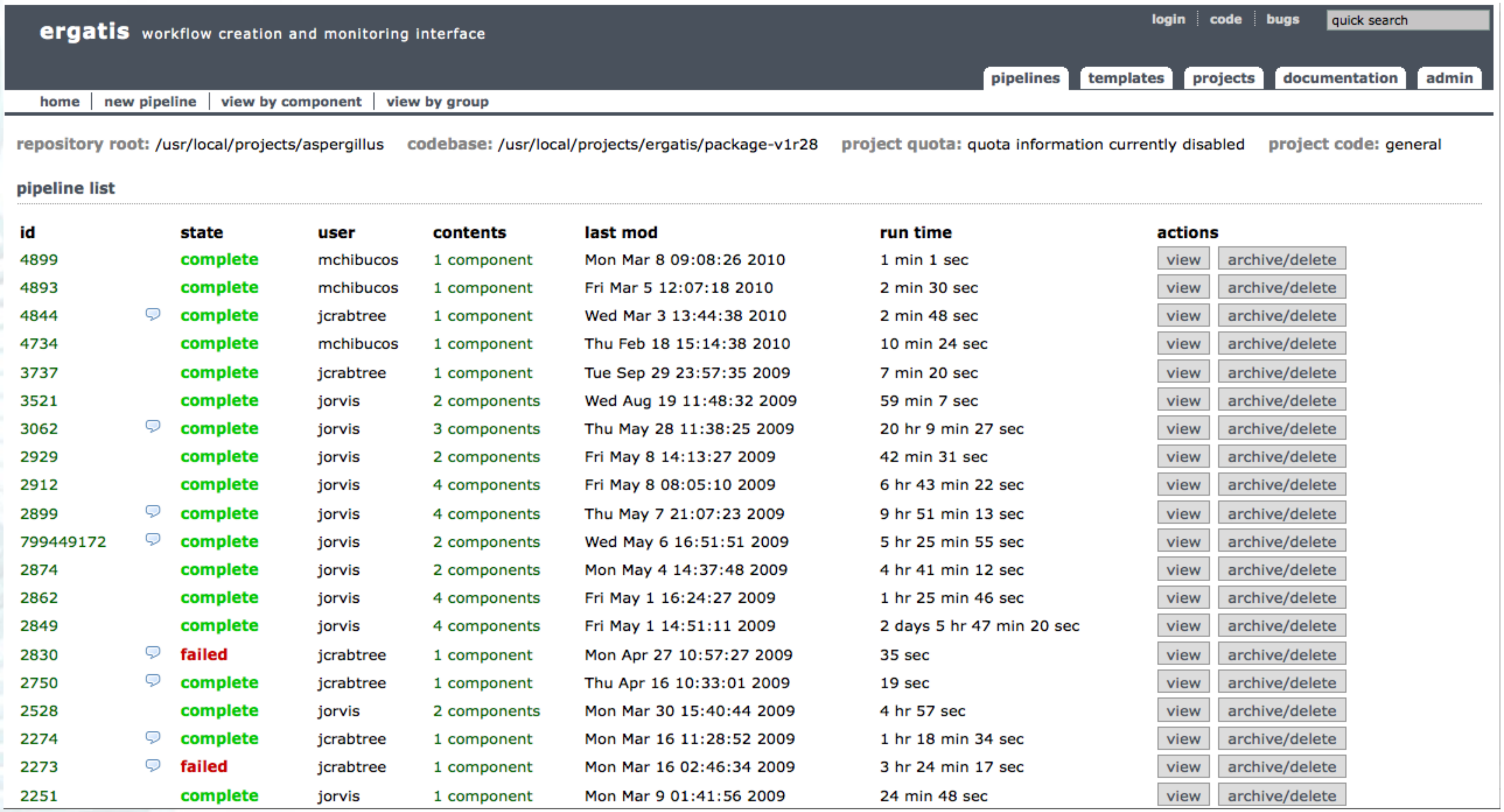
- aat_aa
- aat_aa_masked
- aat_na
- aat_na_masked
- ber
- bsml2interevidence_fasta
- ncbi-blastn
- ncbi-blastn_compress
- ncbi-blastp
- ncbi-blastx
- psiblast
- rpsblast
- wu-blastn
- wu-blastp
- wu-blastx
- wu-tblastn

functional prediction

- auto_gene_curation
- blast2go
- lipop
- lipoprotein_motif
- outer_membrane_domain
- p_func

http://ergatis.igs.umaryland.edu/cgi/documentation.cgi?article=components&page=open_reading_frames

Ergatis – Project Summary View

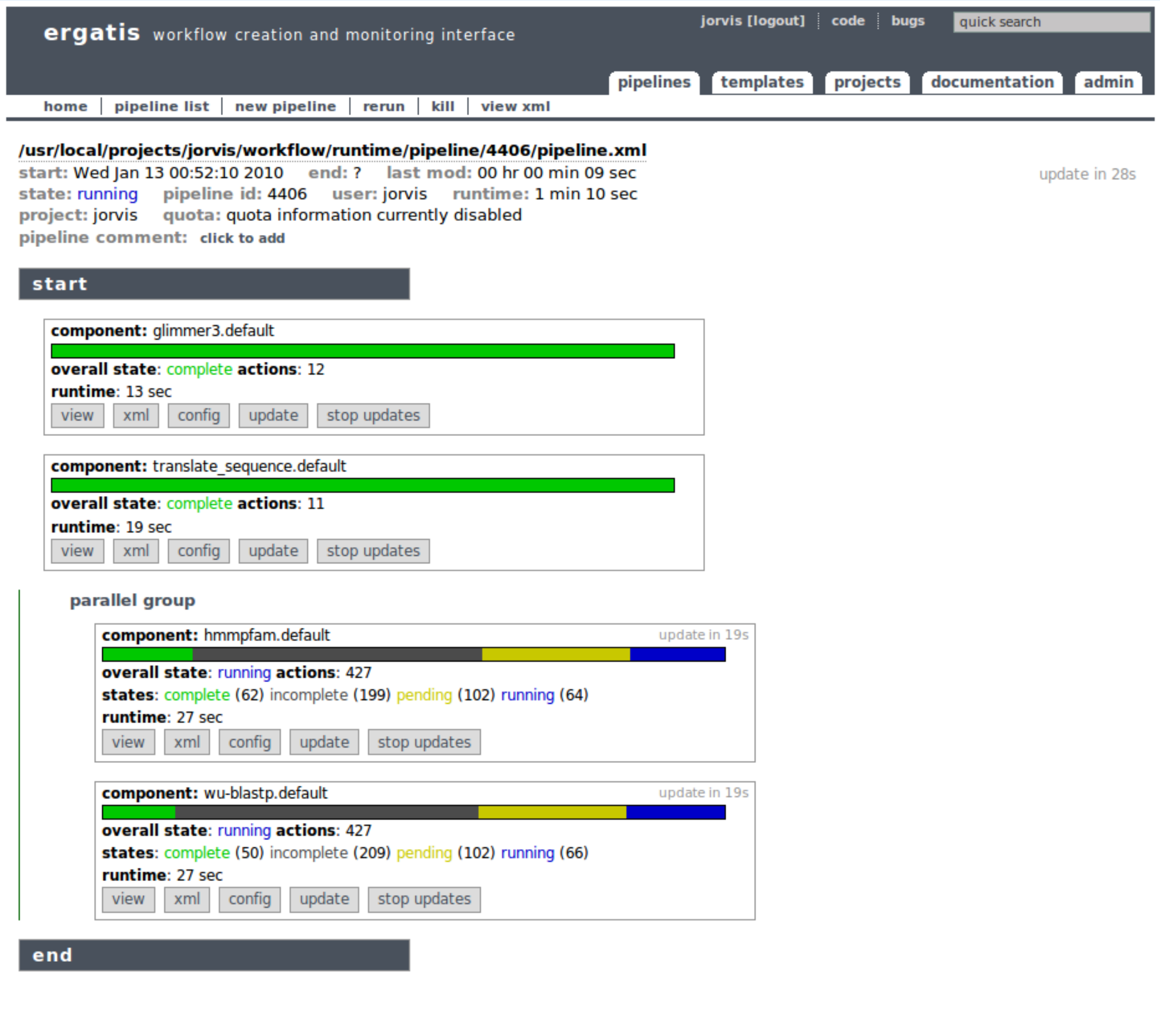


repository root: /usr/local/projects/aspergillus **codebase:** /usr/local/projects/ergatis/package-v1r28 **project quota:** quota information currently disabled **project code:** general

pipeline list

id		state	user	contents	last mod	run time	actions	
4899		complete	mchibucos	1 component	Mon Mar 8 09:08:26 2010	1 min 1 sec	view	archive/delete
4893		complete	mchibucos	1 component	Fri Mar 5 12:07:18 2010	2 min 30 sec	view	archive/delete
4844	💬	complete	jcrabtree	1 component	Wed Mar 3 13:44:38 2010	2 min 48 sec	view	archive/delete
4734		complete	mchibucos	1 component	Thu Feb 18 15:14:38 2010	10 min 24 sec	view	archive/delete
3737		complete	jcrabtree	1 component	Tue Sep 29 23:57:35 2009	7 min 20 sec	view	archive/delete
3521		complete	jorvis	2 components	Wed Aug 19 11:48:32 2009	59 min 7 sec	view	archive/delete
3062	💬	complete	jorvis	3 components	Thu May 28 11:38:25 2009	20 hr 9 min 27 sec	view	archive/delete
2929		complete	jorvis	2 components	Fri May 8 14:13:27 2009	42 min 31 sec	view	archive/delete
2912		complete	jorvis	4 components	Fri May 8 08:05:10 2009	6 hr 43 min 22 sec	view	archive/delete
2899	💬	complete	jorvis	4 components	Thu May 7 21:07:23 2009	9 hr 51 min 13 sec	view	archive/delete
799449172	💬	complete	jorvis	2 components	Wed May 6 16:51:51 2009	5 hr 25 min 55 sec	view	archive/delete
2874		complete	jorvis	2 components	Mon May 4 14:37:48 2009	4 hr 41 min 12 sec	view	archive/delete
2862		complete	jorvis	4 components	Fri May 1 16:24:27 2009	1 hr 25 min 46 sec	view	archive/delete
2849		complete	jorvis	4 components	Fri May 1 14:51:11 2009	2 days 5 hr 47 min 20 sec	view	archive/delete
2830	💬	failed	jcrabtree	1 component	Mon Apr 27 10:57:27 2009	35 sec	view	archive/delete
2750	💬	complete	jcrabtree	1 component	Thu Apr 16 10:33:01 2009	19 sec	view	archive/delete
2528		complete	jorvis	2 components	Mon Mar 30 15:40:44 2009	4 hr 57 sec	view	archive/delete
2274	💬	complete	jcrabtree	1 component	Mon Mar 16 11:28:52 2009	1 hr 18 min 34 sec	view	archive/delete
2273	💬	failed	jcrabtree	1 component	Mon Mar 16 02:46:34 2009	3 hr 24 min 17 sec	view	archive/delete
2251		complete	jorvis	1 component	Mon Mar 9 01:41:56 2009	24 min 48 sec	view	archive/delete

Ergatis – Pipeline Execution View


ergatis workflow creation and monitoring interface
jorvis [logout] | code | bugs
quick search
pipelines
templates
projects
documentation
admin
home | pipeline list | new pipeline | rerun | kill | view xml

/usr/local/projects/jorvis/workflow/runtime/pipeline/4406/pipeline.xml

start: Wed Jan 13 00:52:10 2010 end: ? last mod: 00 hr 00 min 09 sec
state: running pipeline id: 4406 user: jorvis runtime: 1 min 10 sec
project: jorvis quota: quota information currently disabled
pipeline comment: click to add

update in 28s

start

component: glimmer3.default
overall state: complete actions: 12
runtime: 13 sec

component: translate_sequence.default
overall state: complete actions: 11
runtime: 19 sec

parallel group

component: hmmpfam.default update in 19s
overall state: running actions: 427
states: complete (62) incomplete (199) pending (102) running (64)
runtime: 27 sec

component: wu-blastp.default update in 19s
overall state: running actions: 427
states: complete (50) incomplete (209) pending (102) running (66)
runtime: 27 sec

end

VMs -- CloVR

- A genomics tool for automated and portable sequence analysis using Virtual Machines and Cloud computing
- PI: Florian Fricke
- Tools for microbial assembly, annotation, comparative analysis
- Tools for microbial and viral metagenomic community profiling
- Tools for whole metagenome analysis and annotation
- http://clovr.igs.umaryland.edu

Selects Users and Applications

PI	Institution	Application
Fricke	IGS	VM based pipelines
Giglio	IGS	Microbial Annotation
Silva	IGS	Evolutionary Analysis
Meyer	Univ. Chicago/ANL	Metagenomics
Andersen	LBNL	Metagenomics
Wommack	Univ. Delaware	Viral Metagenomics
Buell	Michigan State	Plant Genomics
Tettelin	IGS	Comparative Pangenome Analysis

Select Contributors of Tools/Pipelines/Data

PI	Institution	Contribution
Fricke	IGS	Analysis VMs (CloVR)
Wommack	Univ. Delaware	Virome
Pearson	Univ. Virginia	FASTA
Tettelin	IGS	Sybil
Giglio	IGS	Microbial Annotation Pipeline
Field		Bio-Linux
Ellisman	Univ. California	CAMERA data sets, API
Meyer	Univ. Chicago/ANL	MG-RAST
Andersen	LBNL	Metagenomics classification tools

Timeline

- Prototype system in Spring 2010
  - ~100 cores, 10TB storage
  - Direct access
  - Ergatis
  - Nimbus
  - OSG compute element
- Hardware Acquisition Fall 2010
- Deployment and testing Winter 2010
- Fully operational Early 2011

Challenges

- Deploying the entire cluster as a Cloud computing platform
- Authentication and authorization for multi-institution user base
- Dynamic scheduling of various access methods
- Identifying optimal data sets
- Bandwidth from institutions to upload data
- Space management

Acknowledgements

- All the users and contributors
- PI and Co-PIs
  - Owen White (PI)
  - Eric Wommack
  - Sam Angiuoli
  - Jacques Ravel
- Institute for Genome Sciences
  - Engineering Team
    - Victor Felix
    - Joshua Orvis
    - Sam Angiuoli
  - IT Team
    - Dave Kemeza
    - Brian Cotton
    - Daniel Strassler
- University of Maryland School of Medicine
    - James McNamee
    - Scott Hunsinger
- University of Maryland Institute for Advanced Computer Studies
  - Fritz McCall
- RENCI
  - John McGee
- University of Chicago/Argonne National Lab
  - Kate Keahey

Funding provided by:

ARRA funded project

NSF National Science Foundation
WHERE DISCOVERIES BEGIN