Distributed Workflow-Driven Analysis of Large-Scale Biological Data using bioKepler

Ilkay ALTINTAS, Ph.D.

Deputy Coordinator for Research, San Diego Supercomputer Center, UCSD

Lab Director, Scientific Workflow Automation Technologies

altintas@sdsc.edu

bioKepler.org

Welcome to SDSC!

– Workshop website

http://www.biokepler.org/workshops/2012-sep

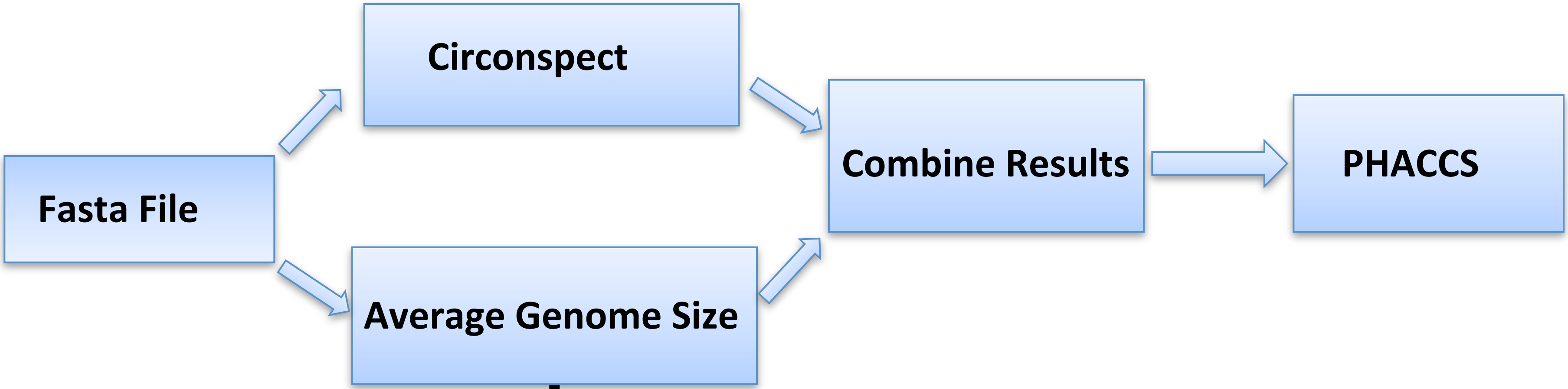
–Logistics for the next two days

So, what is a scientific workflow?

Scientific workflows emerged as an answer to the need to **combine** multiple Cyberinfrastructure components in **automated process networks**.

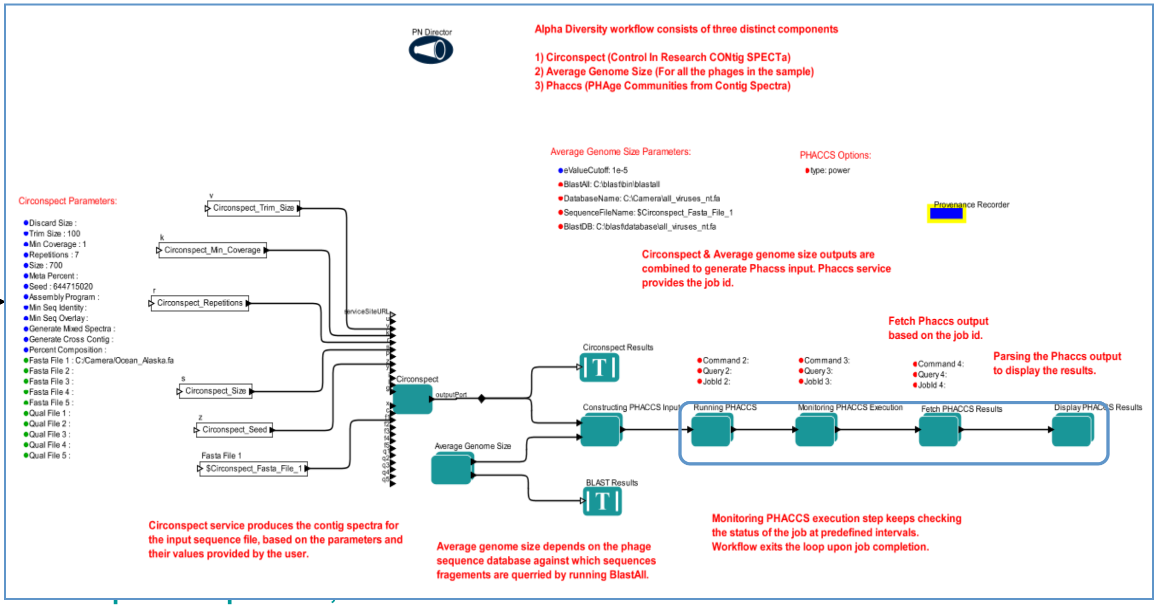
The Big Picture is Supporting the Scientist

From *"Napkin Drawings"* to *Executable Workflows*



Conceptual SWF

Executable SWF



Alpha Diversity workflow consists of three distinct components

1) Circonspect (Control In Research CONtig SPECTa)
2) Average Genome Size (For all the phages in the sample)
3) Phaccs (PHAge Communities from Contig Spectra)

Circonspect Parameters:

- Discard Size :
- Trim Size : 100
- Min Coverage : 1
- Repetitions : 7
- Size : 700
- Meta Percent :
- Seed : 644715020
- Assembly Program :
- Min Seq Identity :
- Min Seq Overlay :
- Generate Mixed Spectra :
- Generate Cross Contig :
- Percent Composition :
- Fasta File 1 : C:/Camera/Ocean_Alaska.fa
- Fasta File 2 :
- Fasta File 3 :
- Fasta File 4 :
- Fasta File 5 :
- Qual File 1 :
- Qual File 2 :
- Qual File 3 :
- Qual File 4 :
- Qual File 5 :

Average Genome Size Parameters:

- eValueCutoff: 1e-5
- BlastAll: C:\blast\bin\blastall
- DatabaseName: C:\Camera\all_viruses_nt.fa
- SequenceFileName: $Circonspect_Fasta_File_1
- BlastDB: C:\blast\database\all_viruses_nt.fa

PHACCS Options:

- type: power

Circonspect & Average genome size outputs are combined to generate Phacss input. Phaccs service provides the job id.

Fetch Phaccs output based on the job id.

Parsing the Phaccs output to display the results.

- Command 2:
- Query 2:
- JobId 2:

- Command 3:
- Query 3:
- JobId 3:

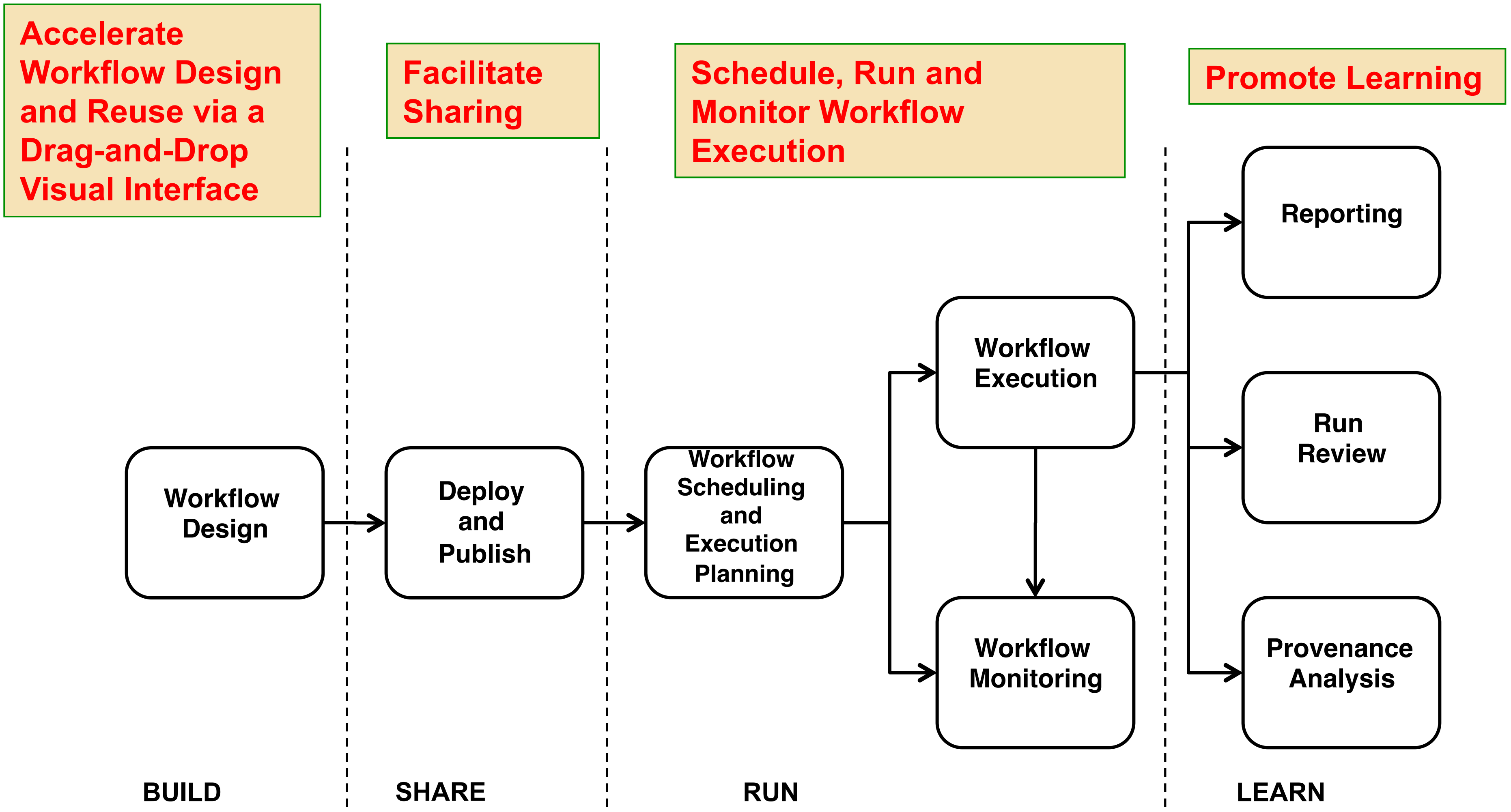
- Command 4:
- Query 4:
- JobId 4:

Circonspect service produces the contig spectra for the input sequence file, based on the parameters and their values provided by the user.

Average genome size depends on the phage sequence database against which sequences fragements are querried by running BlastAll.

Monitoring PHACCS execution step keeps checking the status of the job at predefined intervals. Workflow exits the loop upon job completion.

Workflows are a Part of Cyberinfrastructure



Accelerate Workflow Design and Reuse via a Drag-and-Drop Visual Interface

Facilitate Sharing

Schedule, Run and Monitor Workflow Execution

Promote Learning

Workflow Design

Deploy and Publish

Workflow Scheduling and Execution Planning

Workflow Execution

Workflow Monitoring

Reporting

Run Review

Provenance Analysis

BUILD

SHARE

RUN

LEARN

Support for end-to-end computational scientific process

Kepler is a Scientific Workflow System

www.kepler-project.org

- A cross-project collaboration

  … initiated August 2003
- 2.3 release released 01/2012

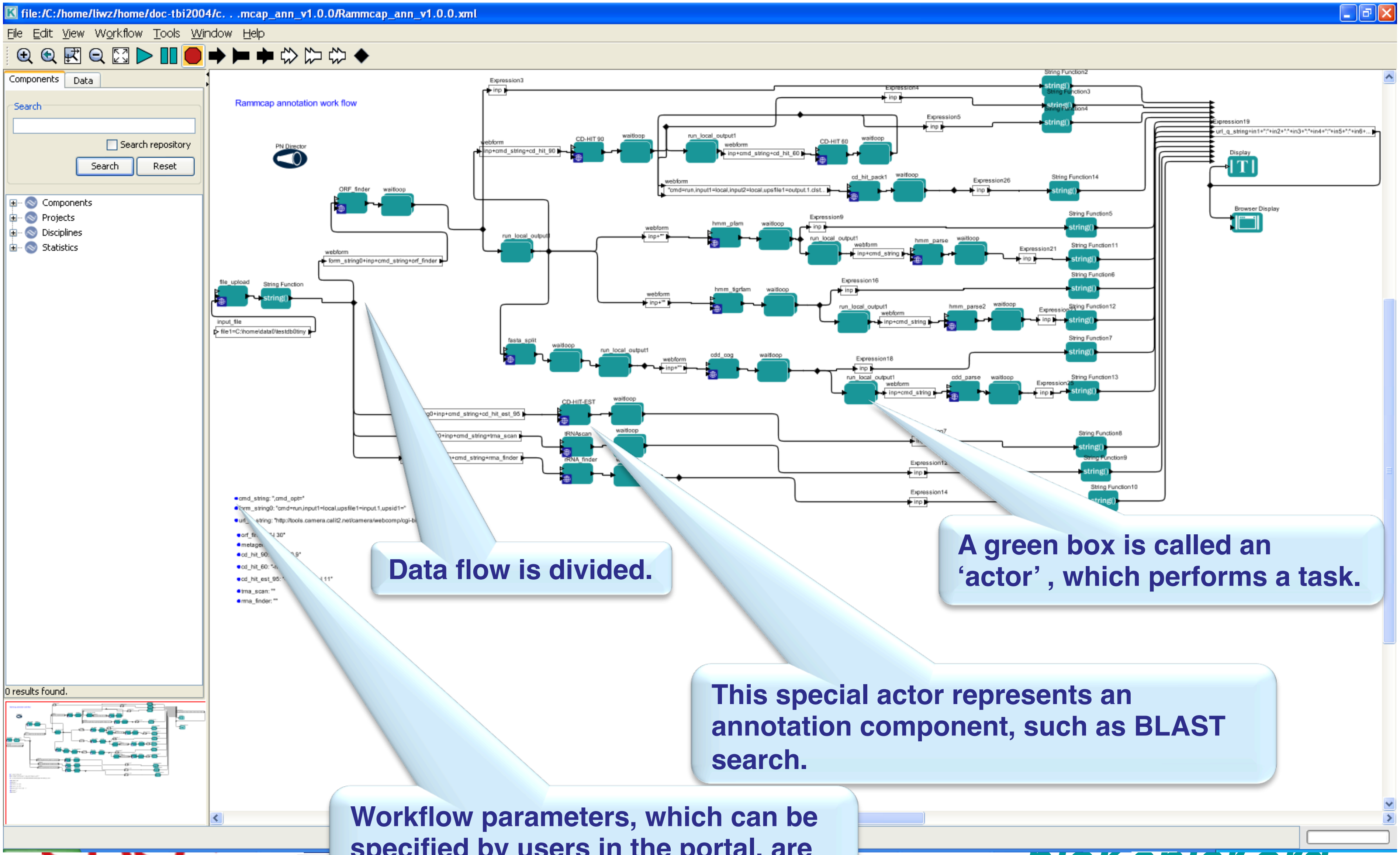
- Builds upon the open-source Ptolemy II framework

Ptolemy II: A laboratory for investigating design

KEPLER: A problem-solving environment for Scientific Workflow

KEPLER = "Ptolemy II + X" for Scientific Workflows

A Typical Kepler Workflow



- cmd_string: ",cmd_opt="
- form_string0: "cmd=run,input1=local,upsfile1=input.1,upsid1="
- url_ string: "http://tools.camera.calit2.net/camera/webcomp/cgi-bi
- orf_fi -l 30"
- metagen
- cd_hit_90: 0.9"
- cd_hit_60: "-
- cd_hit_est_95: " 11"
- trna_scan: ""
- rrna_finder: ""

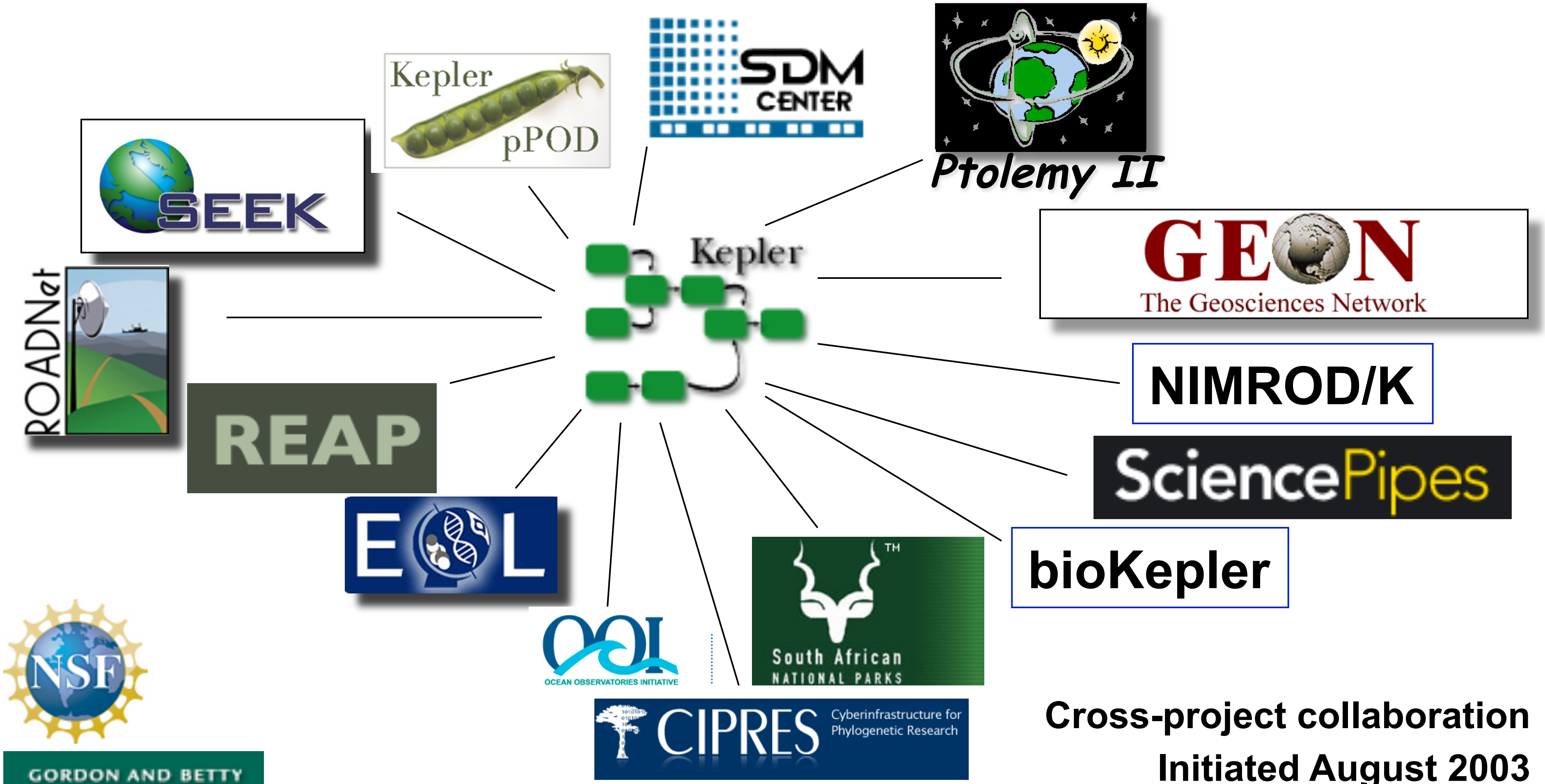
Data flow is divided.

A green box is called an ‘actor’ , which performs a task.

This special actor represents an annotation component, such as BLAST search.

Workflow parameters, which can be specified by users in the portal, are passed to workflow components.

Kepler is a Team Effort and Modular


Kepler
Kepler pPOD
SDM CENTER
Ptolemy II
SEEK
GEON The Geosciences Network
ROADNet
NIMROD/K
REAP
SciencePipes
EOL
bioKepler
OOI OCEAN OBSERVATORIES INITIATIVE
South African NATIONAL PARKS
CIPRES Cyberinfrastructure for Phylogenetic Research

NSF

GORDON AND BETTY MOORE FOUNDATION

SciDAC Scientific Discovery through Advanced Computing

Cross-project collaboration

Initiated August 2003

Kepler 2.3 release: January, 2012

Kepler Your Science. Enabled.

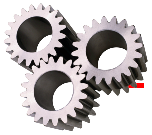
Full list of contributors, projects, individuals and funding info are at the Kepler website!!

Requirements are similar for many domains

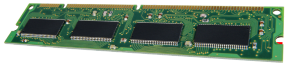
-- with slight variations --

Facilitating and Accelerating XXX-Info or Comp-XXX Research using Scientific Workflows

- Important Attributes



- Assemble complex processing easily



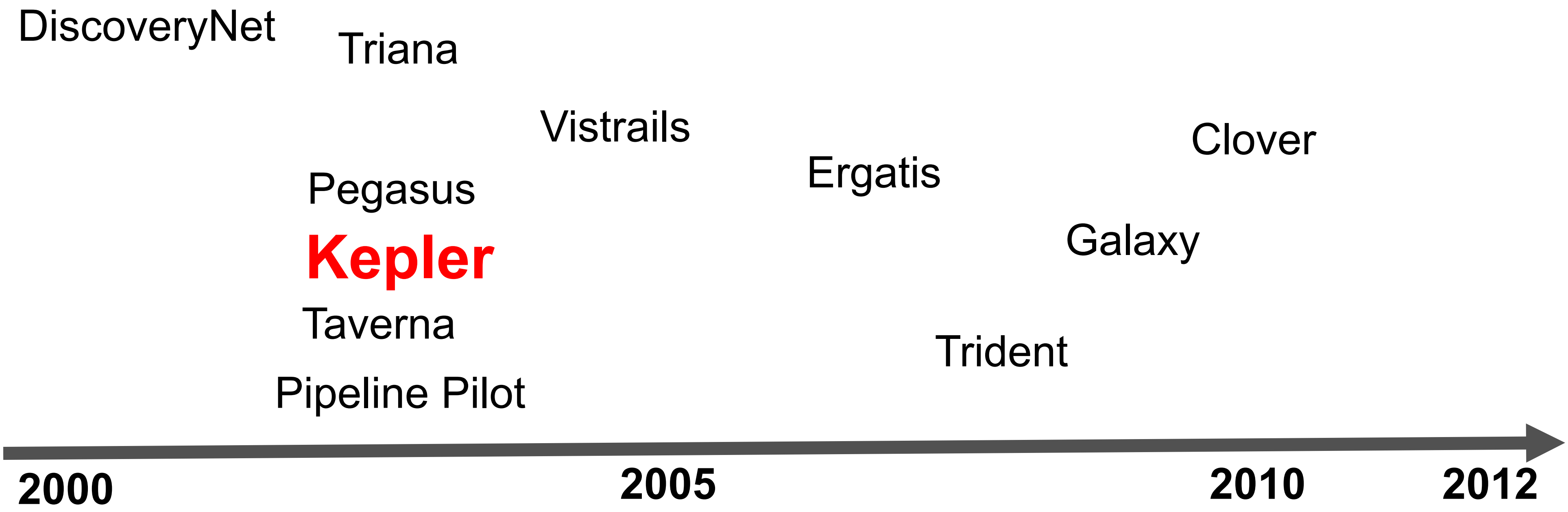
- Access transparently to diverse resources

- Incorporate multiple software tools

- Assure reproducibility

- Build around community development model

Many Bioinformatics Workflow Systems



DiscoveryNet

Triana

Vistrails

Clover

Ergatis

Pegasus

Galaxy

Kepler

Taverna

Trident

Pipeline Pilot

2000

2005

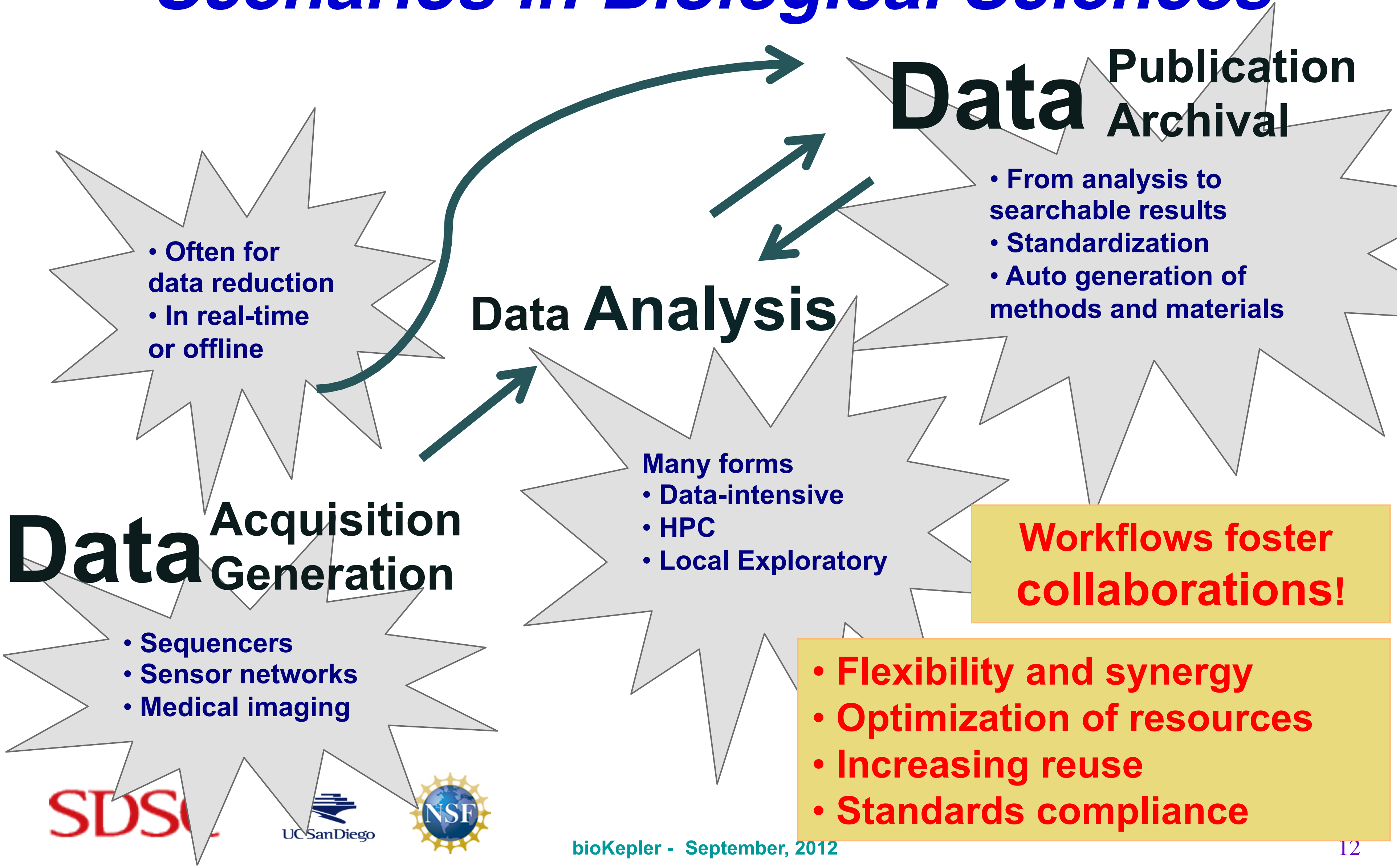
2010

2012

Kepler

- A diverse library of scientific components and usecases
- Transparent support for multiple workflow engines
- Used by many communities, specialized gateways and individuals

Workflows are Used in These Diverse Scenarios in Biological Sciences



Data Acquisition Generation

- Sequencers
- Sensor networks
- Medical imaging

- Often for data reduction
- In real-time or offline

Data Analysis

Many forms

- Data-intensive
- HPC
- Local Exploratory

Data Publication Archival

- From analysis to searchable results
- Standardization
- Auto generation of methods and materials

Workflows foster collaborations!

- Flexibility and synergy
- Optimization of resources
- Increasing reuse
- Standards compliance

A Toolbox with Many Tools

- **Data**
  - Search, database access, IO operations, streaming data in real-time…
- **Compute**
  - Data-parallel patterns, external execution, …
- **Network operations**
- **Provenance and fault tolerance**

Need expertise to identify which tool to use when and how!
Require computation models to schedule and optimize execution!

CAMERA Example:

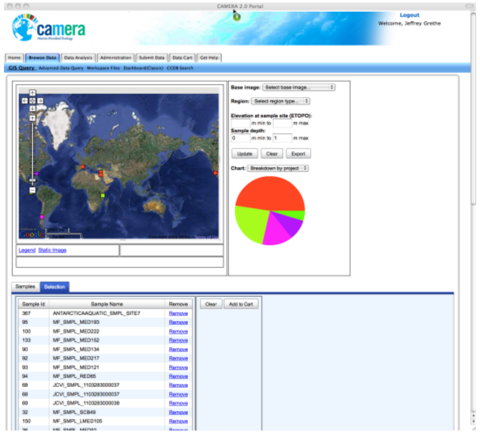
Using Scientific Workflows and Related Provenance for Collaborative Metagenomics Research

Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis

(CAMERA)

http://camera.calit2.net

CAMERA is a Collaborative Environment



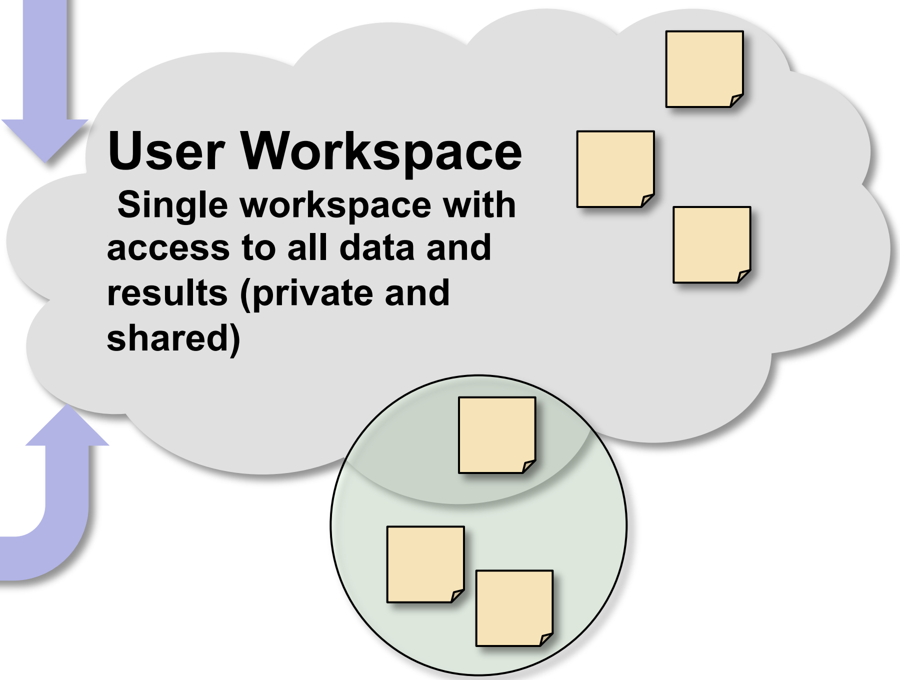
Sample Id	Sample Name	Remove
367	ANTARCTICAAQUATIC_SMPL_SITE7	Remove
95	MF_SMPL_MED193	Remove
100	MF_SMPL_MED222	Remove
133	MF_SMPL_MED152	Remove
90	MF_SMPL_MED134	Remove
92	MF_SMPL_MED217	Remove
93	MF_SMPL_MED121	Remove
94	MF_SMPL_RED65	Remove
68	JCVI_SMPL_1103283000037	Remove
68	JCVI_SMPL_1103283000037	Remove
69	JCVI_SMPL_1103283000038	Remove
32	MF_SMPL_SCB49	Remove
150	MF_SMPL_LMED105	Remove

Data Discovery

GIS and Advanced query options

Data Cart

Multiple Available Mixed collections of CAMERA Data (e.g. projects, samples)

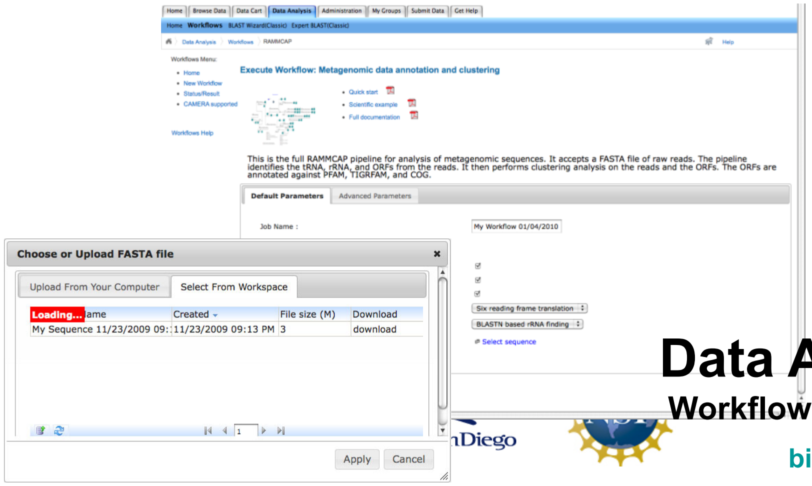


User Workspace

Single workspace with access to all data and results (private and shared)

Group Workspace

Share specified User Workspace data with



Execute Workflow: Metagenomic data annotation and clustering

This is the full RAMMCAP pipeline for analysis of metagenomic sequences. It accepts a FASTA file of raw reads. The pipeline identifies the tRNA, rRNA, and ORFs from the reads. It then performs clustering analysis on the reads and the ORFs. The ORFs are annotated against PFAM, TIGRFAM, and COG.

Job Name : My Workflow 01/04/2010

☑
☑
☑

Six reading frame translation

BLASTN based rRNA finding

Choose or Upload FASTA file

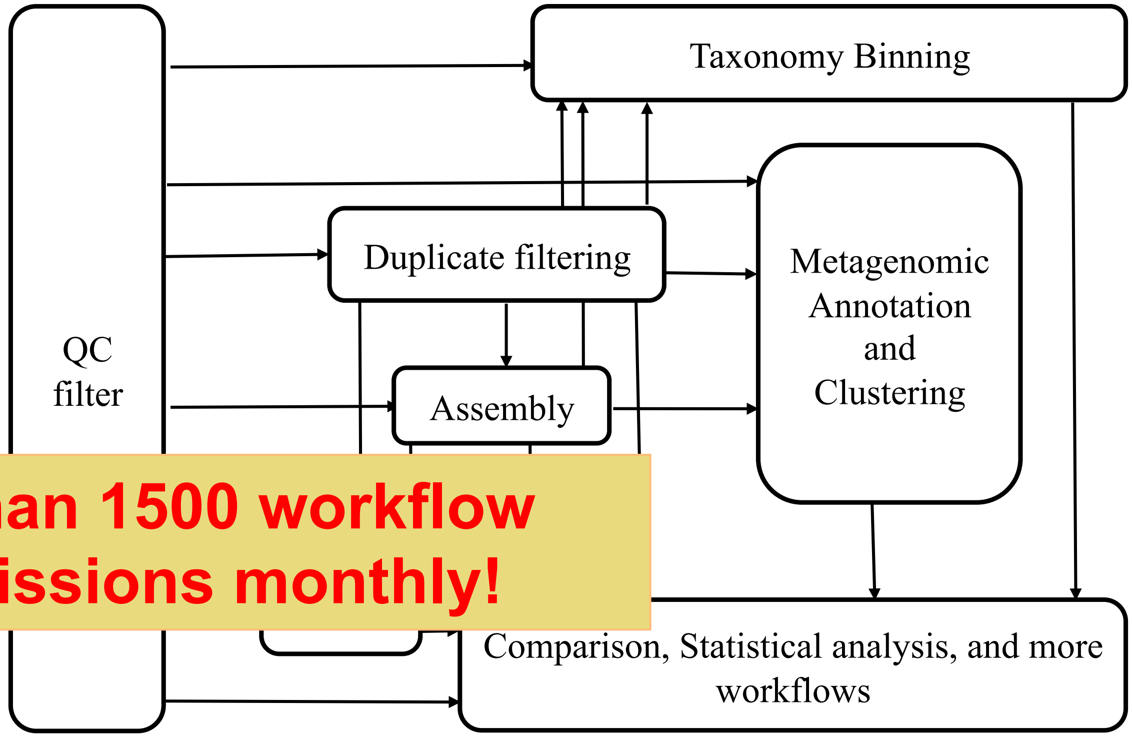
Loading... lame	Created	File size (M)	Download
My Sequence 11/23/2009 09:	11/23/2009 09:13 PM	3	download

Data Analysis

Workflow based analysis

Workflows are a Central Part of CAMERA

- **CAMERA-supported**
  - 28 existing workflows
- **Workflows under development**
  - Fragment Recruitment Viewer
  - Next Generation Sequencing
  - VIROME Pipeline
  - Standalone b tools
  - National Cent Research
  - Joint Genome Institute
- **User built**
  - Currently running in a sandbox
  - Will be ported to a virtual cloud environment

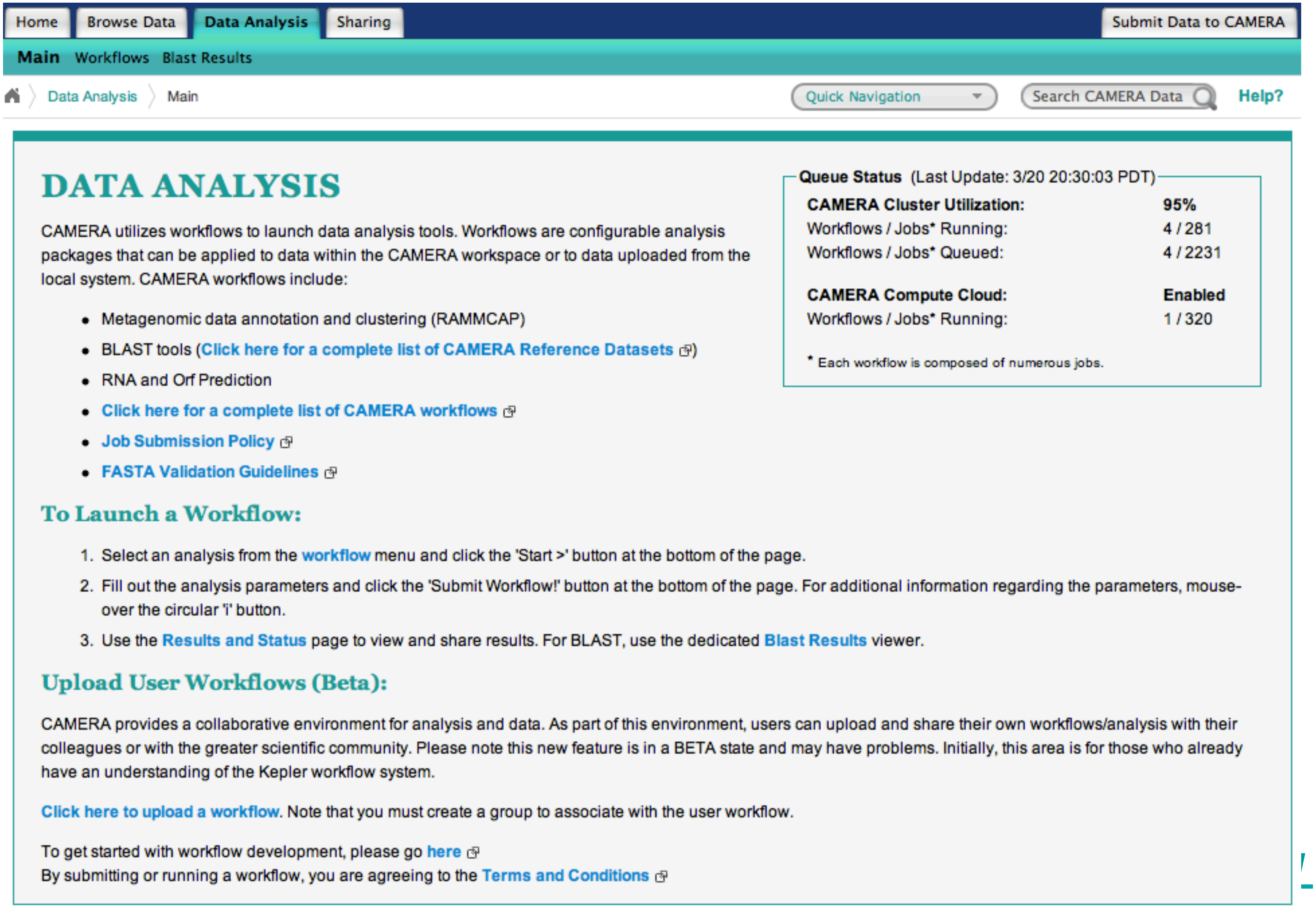

QC filter
Taxonomy Binning
Duplicate filtering
Assembly
Metagenomic Annotation and Clustering
Comparison, Statistical analysis, and more workflows

More than 1500 workflow submissions monthly!

- **Inputs:** from local or CAMERA file systems; user-supplied parameters
- **Outputs:** sharable with a group of users and links to the semantic database

All can be reached through the CAMERA portal at:http://portal.camera.calit2.net

CAMERA Portal - Workflows



DATA ANALYSIS

CAMERA utilizes workflows to launch data analysis tools. Workflows are configurable analysis packages that can be applied to data within the CAMERA workspace or to data uploaded from the local system. CAMERA workflows include:

- Metagenomic data annotation and clustering (RAMMCAP)
- BLAST tools (Click here for a complete list of CAMERA Reference Datasets)
- RNA and Orf Prediction
- Click here for a complete list of CAMERA workflows
- Job Submission Policy
- FASTA Validation Guidelines

Queue Status (Last Update: 3/20 20:30:03 PDT)

CAMERA Cluster Utilization:	**95%**
Workflows / Jobs* Running:	4 / 281
Workflows / Jobs* Queued:	4 / 2231
CAMERA Compute Cloud:	**Enabled**
Workflows / Jobs* Running:	1 / 320

* Each workflow is composed of numerous jobs.

To Launch a Workflow:

1. Select an analysis from the workflow menu and click the 'Start >' button at the bottom of the page.
2. Fill out the analysis parameters and click the 'Submit Workflow!' button at the bottom of the page. For additional information regarding the parameters, mouse-over the circular 'i' button.
3. Use the Results and Status page to view and share results. For BLAST, use the dedicated Blast Results viewer.

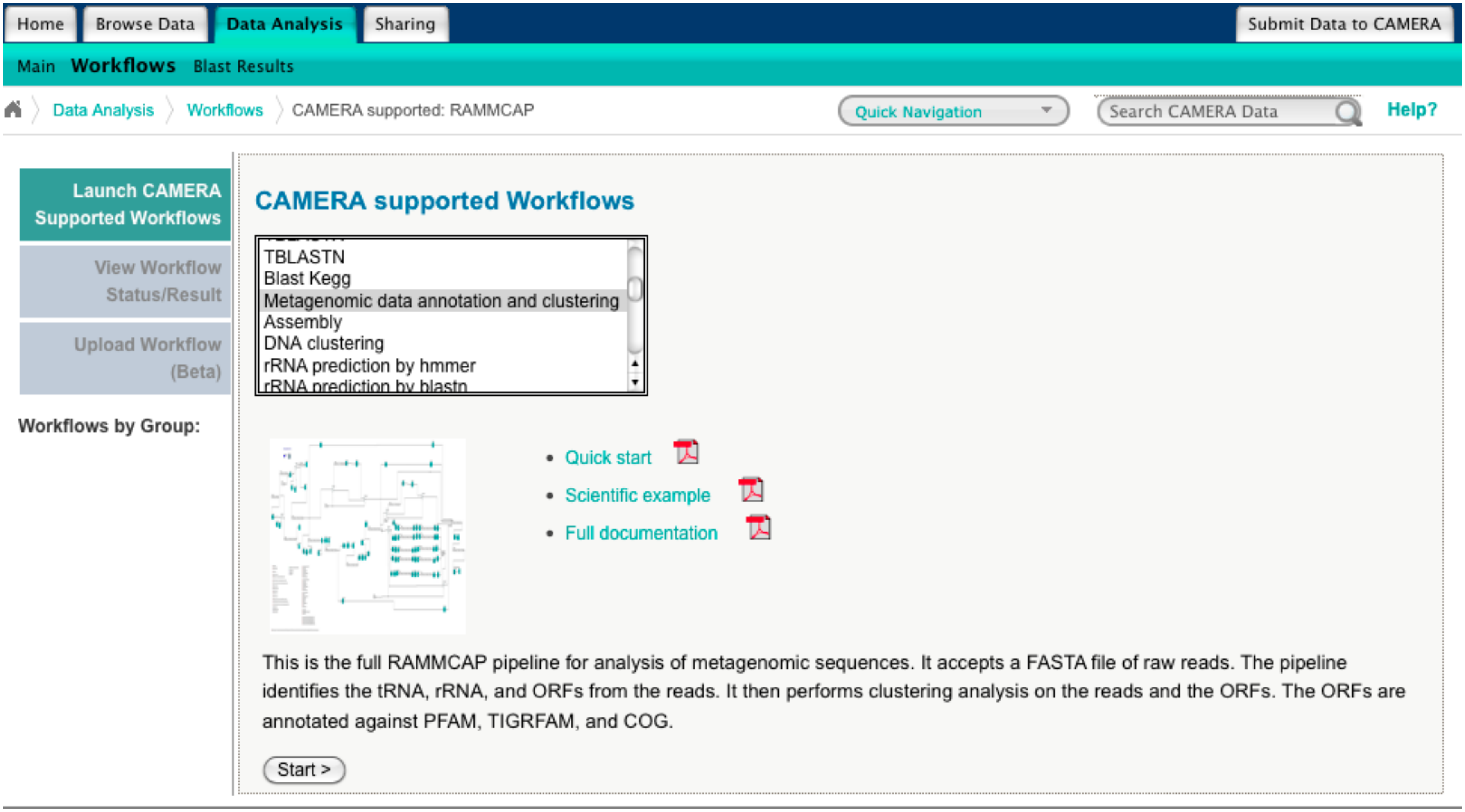
Upload User Workflows (Beta):

CAMERA provides a collaborative environment for analysis and data. As part of this environment, users can upload and share their own workflows/analysis with their colleagues or with the greater scientific community. Please note this new feature is in a BETA state and may have problems. Initially, this area is for those who already have an understanding of the Kepler workflow system.

Click here to upload a workflow. Note that you must create a group to associate with the user workflow.

To get started with workflow development, please go here
By submitting or running a workflow, you are agreeing to the Terms and Conditions

CAMERA Workflows



CAMERA supported Workflows

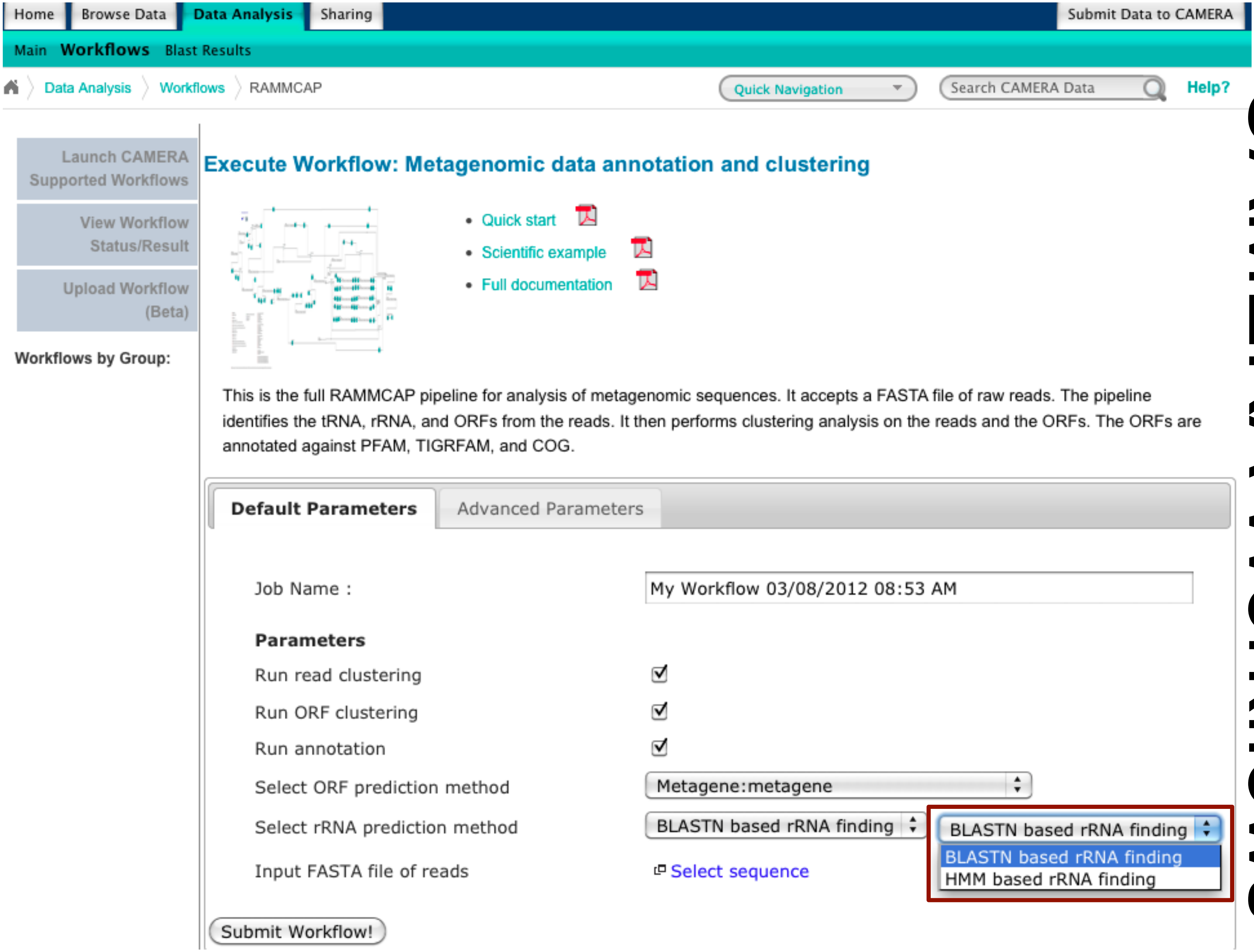
TBLASTN
Blast Kegg
Metagenomic data annotation and clustering
Assembly
DNA clustering
rRNA prediction by hmmer
rRNA prediction by blastn

- Quick start
- Scientific example
- Full documentation

This is the full RAMMCAP pipeline for analysis of metagenomic sequences. It accepts a FASTA file of raw reads. The pipeline identifies the tRNA, rRNA, and ORFs from the reads. It then performs clustering analysis on the reads and the ORFs. The ORFs are annotated against PFAM, TIGRFAM, and COG.

RAMMCAP

CAMERA Workflows



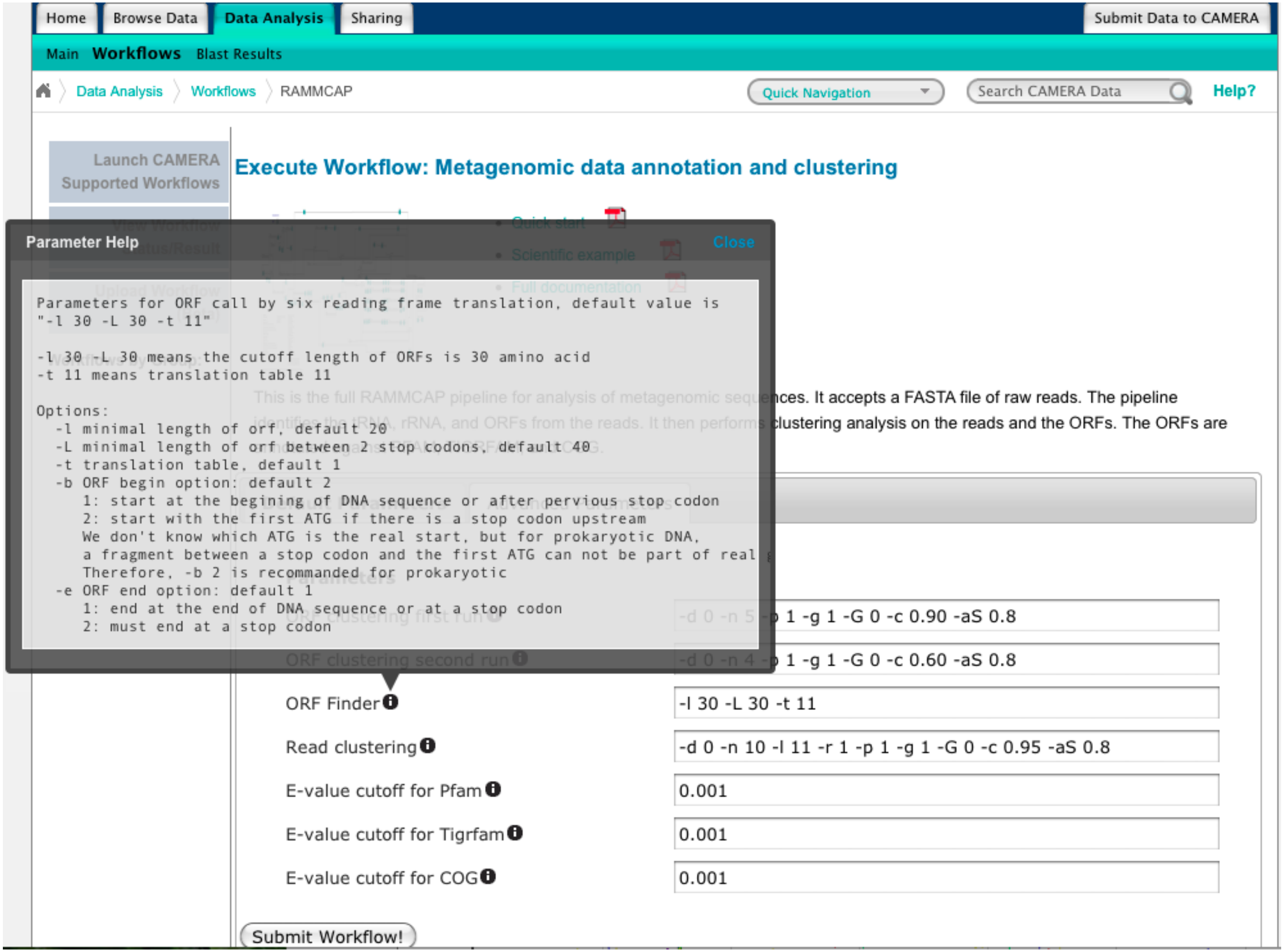
Execute Workflow: Metagenomic data annotation and clustering

- Quick start
- Scientific example
- Full documentation

This is the full RAMMCAP pipeline for analysis of metagenomic sequences. It accepts a FASTA file of raw reads. The pipeline identifies the tRNA, rRNA, and ORFs from the reads. It then performs clustering analysis on the reads and the ORFs. The ORFs are annotated against PFAM, TIGRFAM, and COG.

Job Name :	My Workflow 03/08/2012 08:53 AM
Parameters	
Run read clustering	☑
Run ORF clustering	☑
Run annotation	☑
Select ORF prediction method	Metagene:metagene
Select rRNA prediction method	BLASTN based rRNA finding
Input FASTA file of reads	Select sequence

CAMERA Workflows



Execute Workflow: Metagenomic data annotation and clustering

Parameter Help

```
Parameters for ORF call by six reading frame translation, default value is
"-l 30 -L 30 -t 11"

-l 30 -L 30 means the cutoff length of ORFs is 30 amino acid
-t 11 means translation table 11

Options:
  -l minimal length of orf, default 20
  -L minimal length of orf between 2 stop codons, default 40
  -t translation table, default 1
  -b ORF begin option: default 2
     1: start at the begining of DNA sequence or after pervious stop codon
     2: start with the first ATG if there is a stop codon upstream
     We don't know which ATG is the real start, but for prokaryotic DNA,
     a fragment between a stop codon and the first ATG can not be part of real g
     Therefore, -b 2 is recommanded for prokaryotic
  -e ORF end option: default 1
     1: end at the end of DNA sequence or at a stop codon
     2: must end at a stop codon
```

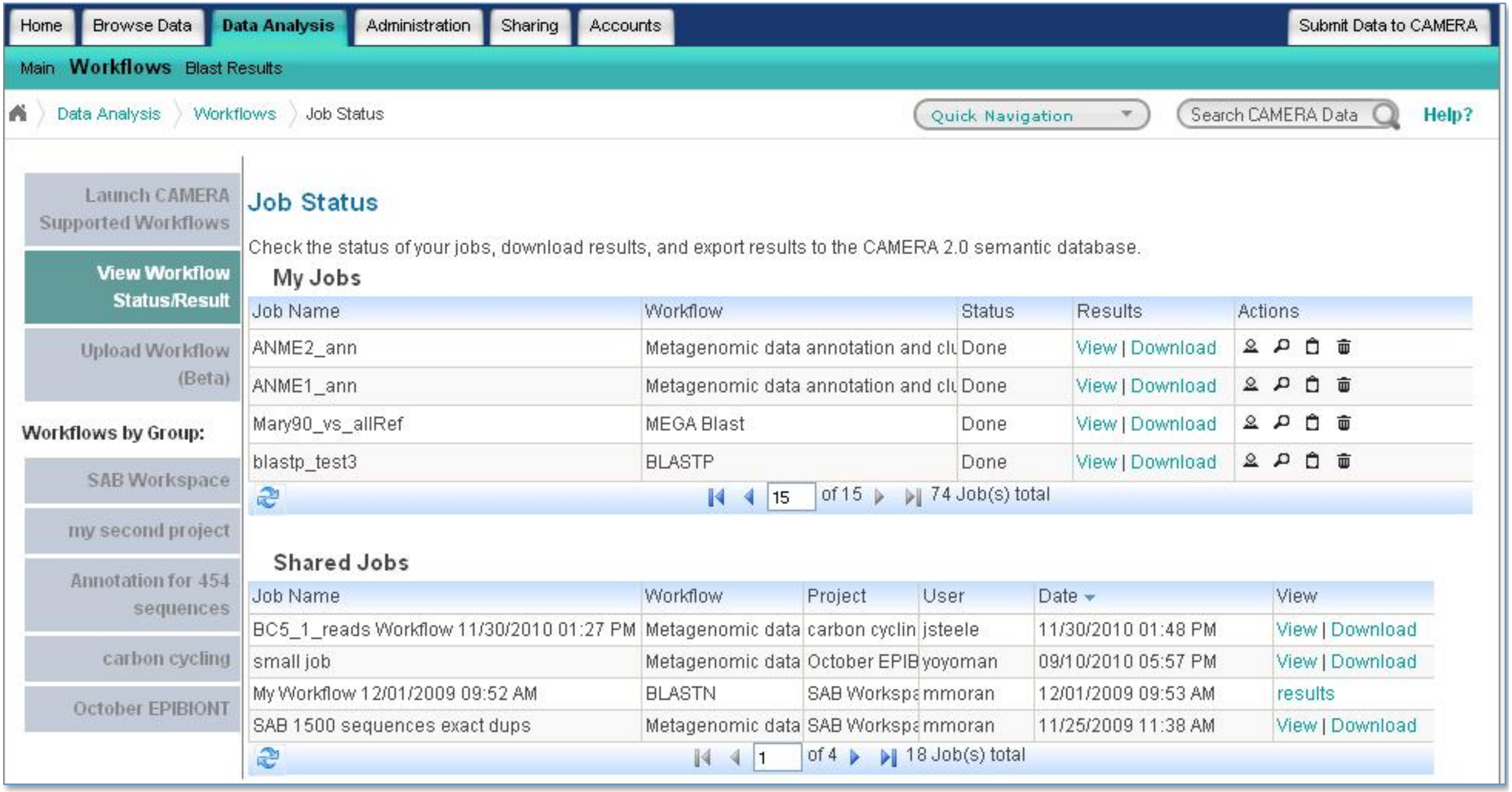
This is the full RAMMCAP pipeline for analysis of metagenomic sequences. It accepts a FASTA file of raw reads. The pipeline identifies the tRNA, rRNA, and ORFs from the reads. It then performs clustering analysis on the reads and the ORFs. The ORFs are ... against PFAM, TIGRFAM, COG.

Parameters

ORF clustering first run	-d 0 -n 5 -p 1 -g 1 -G 0 -c 0.90 -aS 0.8
ORF clustering second run	-d 0 -n 4 -p 1 -g 1 -G 0 -c 0.60 -aS 0.8
ORF Finder	-l 30 -L 30 -t 11
Read clustering	-d 0 -n 10 -l 11 -r 1 -p 1 -g 1 -G 0 -c 0.95 -aS 0.8
E-value cutoff for Pfam	0.001
E-value cutoff for Tigrfam	0.001
E-value cutoff for COG	0.001

Submit Workflow!

CAMERA Job Status



Job Status

Check the status of your jobs, download results, and export results to the CAMERA 2.0 semantic database.

My Jobs

Job Name	Workflow	Status	Results	Actions
ANME2_ann	Metagenomic data annotation and clu	Done	View | Download	
ANME1_ann	Metagenomic data annotation and clu	Done	View | Download	
Mary90_vs_allRef	MEGA Blast	Done	View | Download	
blastp_test3	BLASTP	Done	View | Download	

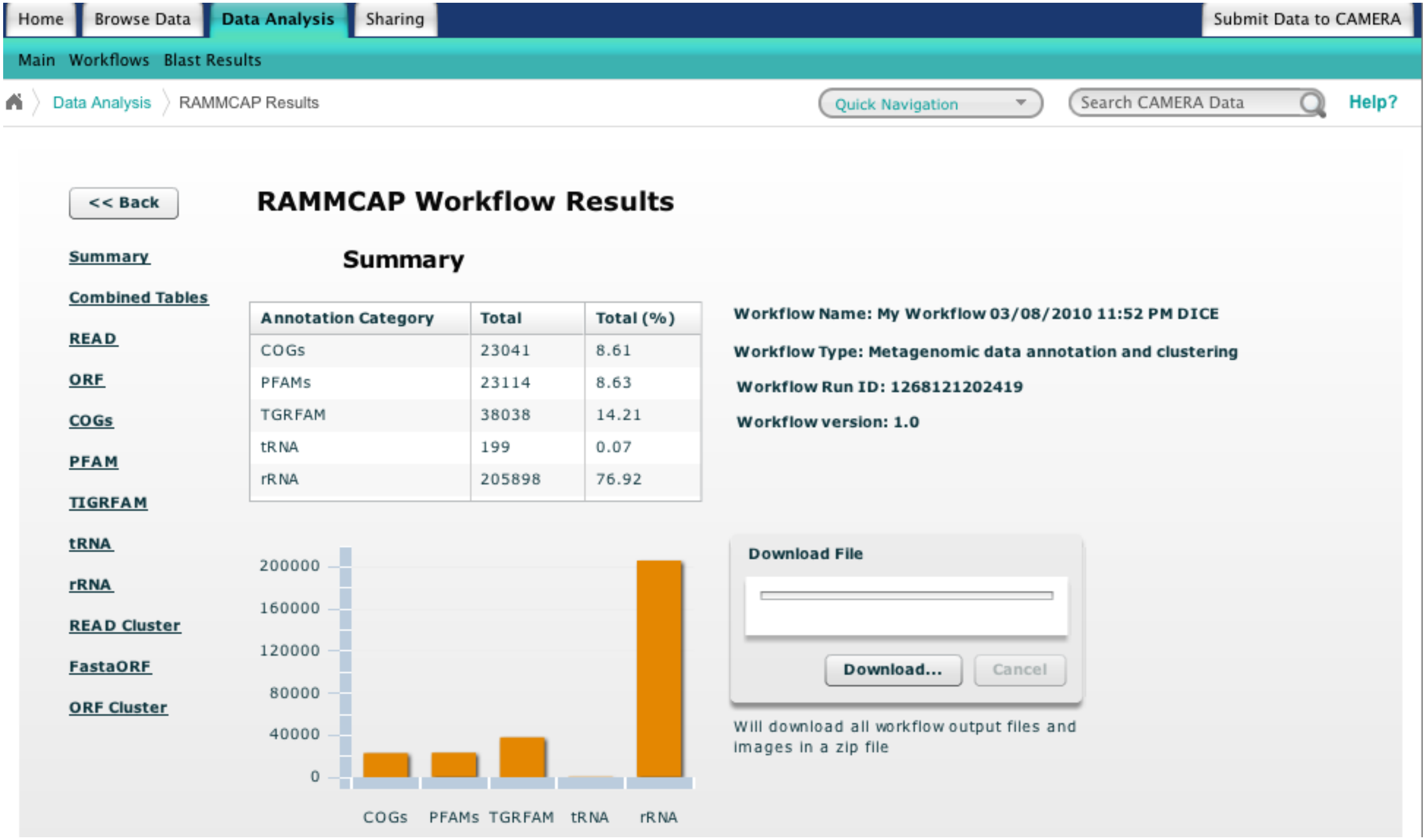
15 of 15 74 Job(s) total

Shared Jobs

Job Name	Workflow	Project	User	Date	View
BC5_1_reads Workflow 11/30/2010 01:27 PM	Metagenomic data	carbon cyclin	jsteele	11/30/2010 01:48 PM	View | Download
small job	Metagenomic data	October EPIB	yoyoman	09/10/2010 05:57 PM	View | Download
My Workflow 12/01/2009 09:52 AM	BLASTN	SAB Workspa	mmoran	12/01/2009 09:53 AM	results
SAB 1500 sequences exact dups	Metagenomic data	SAB Workspa	mmoran	11/25/2009 11:38 AM	View | Download

1 of 4 18 Job(s) total

CAMERA Workflow Results



RAMMCAP Workflow Results

Summary

Annotation Category	Total	Total (%)
COGs	23041	8.61
PFAMs	23114	8.63
TGRFAM	38038	14.21
tRNA	199	0.07
rRNA	205898	76.92

Workflow Name: My Workflow 03/08/2010 11:52 PM DICE

Workflow Type: Metagenomic data annotation and clustering

Workflow Run ID: 1268121202419

Workflow version: 1.0

Will download all workflow output files and images in a zip file

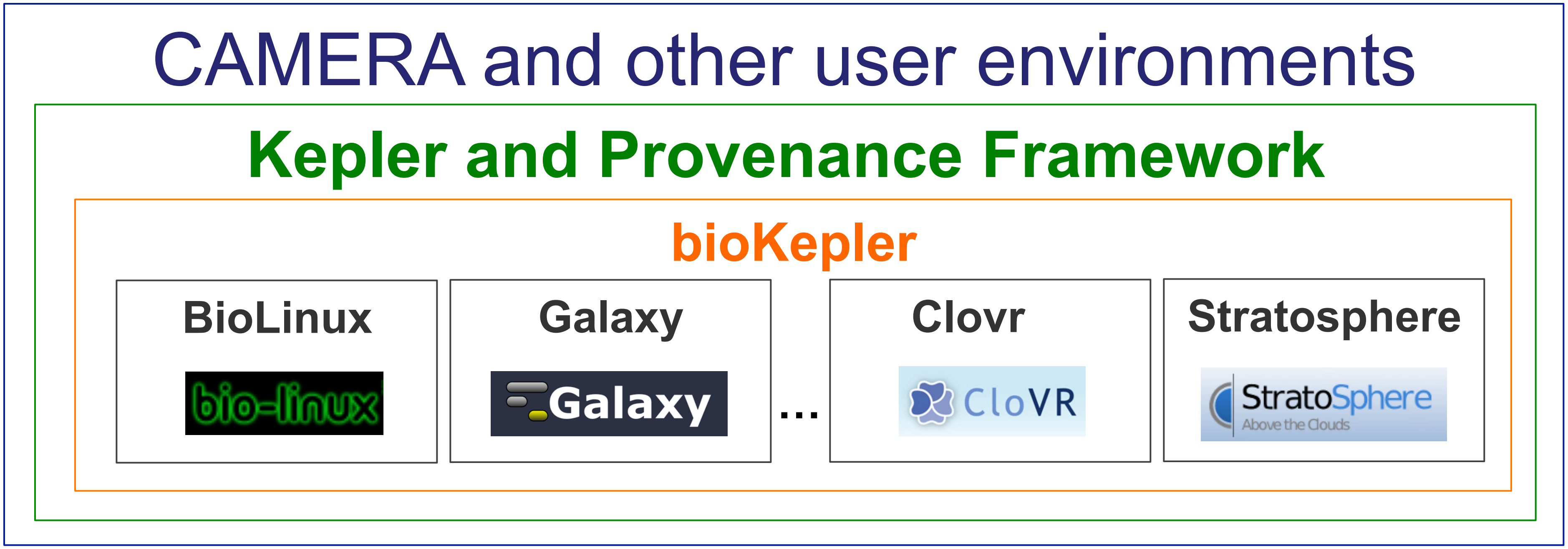
Pushing the boundaries of existing infrastructure and workflow system capabilities

Requirements from the User Community

- Increase reuse
  - best development practices by the scientific community
  - other bio packages
- Increase programmability by end users
  - users with various skill levels
  - to formulate actual domain specific workflows
- Increase resource utilization
  - optimize execution across available computing resources
  - in an efficient, transparent and intuitive manner
- Make analysis a part of the end-to-end scientific model from data generation to publication

bioKepler responds to these requirements!

www.bioKepler.org



CAMERA and other user environments

Kepler and Provenance Framework

bioKepler

BioLinux | **Galaxy** | … | **Clovr** | **Stratosphere**

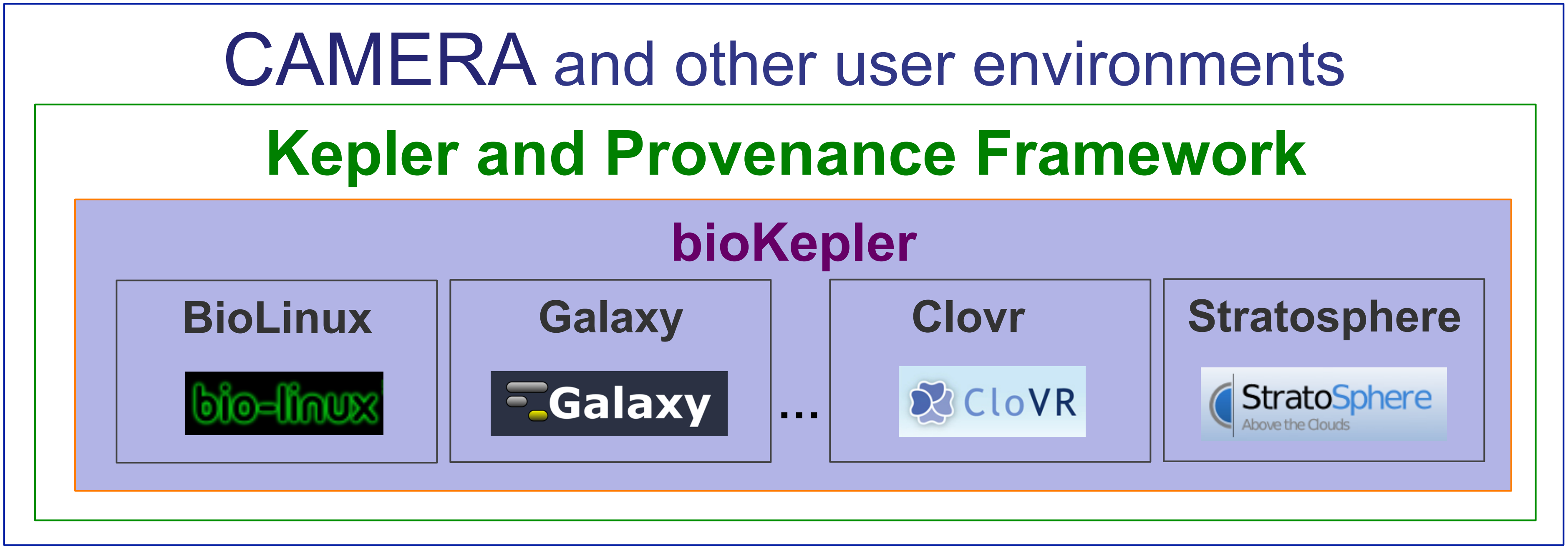
CLOUD and OTHER COMPUTING RESOURCES

e.g., SGE, Amazon, FutureGrid, XSEDE

A coordinated ecosystem of biological and technological packages for microbiology!

Reuse, Programmability, Execution

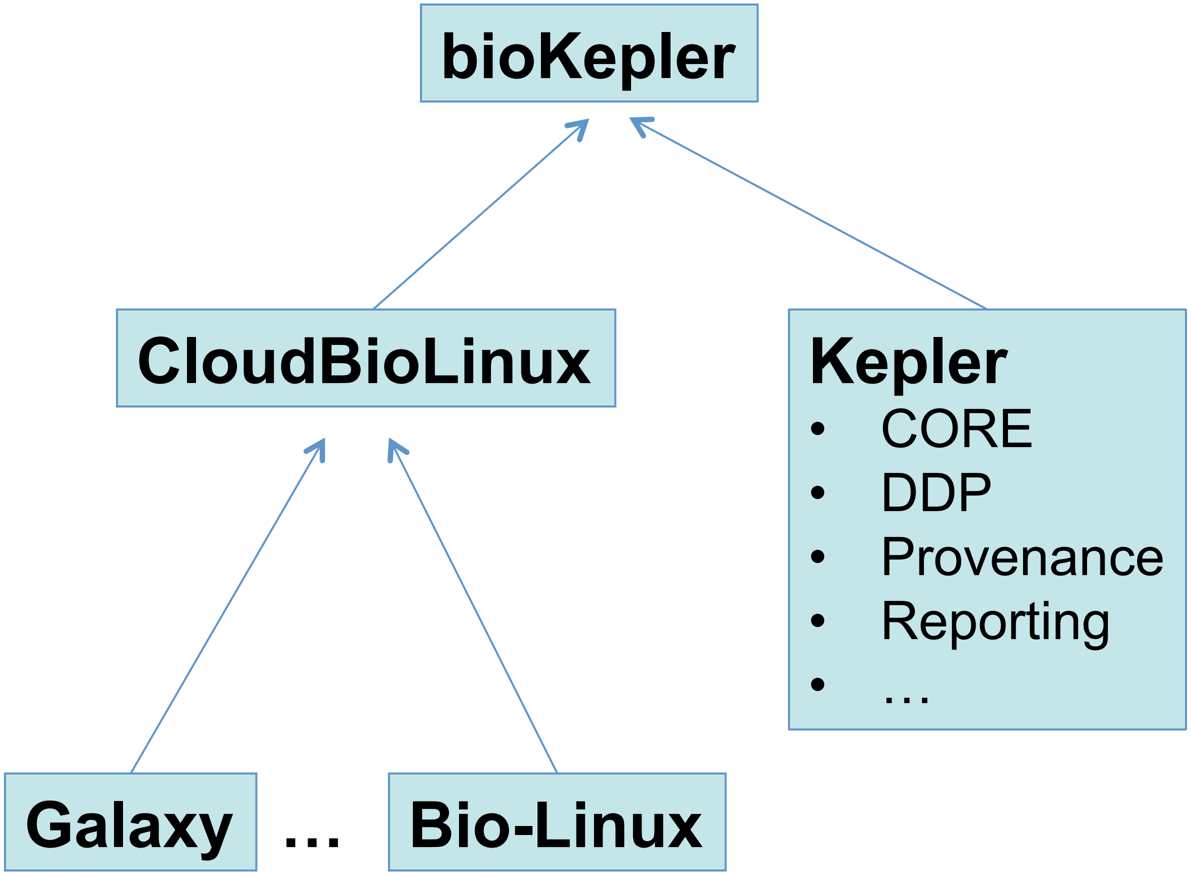
www.bioKepler.org



- Funded by NSF ABI & CI Reuse programs ($1.4M through 2015)
  - Ilkay Altintas (PI) and Weizong Li (Co-PI)
- Development of a comprehensive bioinformatics scientific workflow module for distributed analysis of large-scale biological data

Will be a huge improvement on usability and programmability by end users!

bioKepler and Other Related Systems



bioKepler

CloudBioLinux

Kepler
- CORE
- DDP
- Provenance
- Reporting
- …

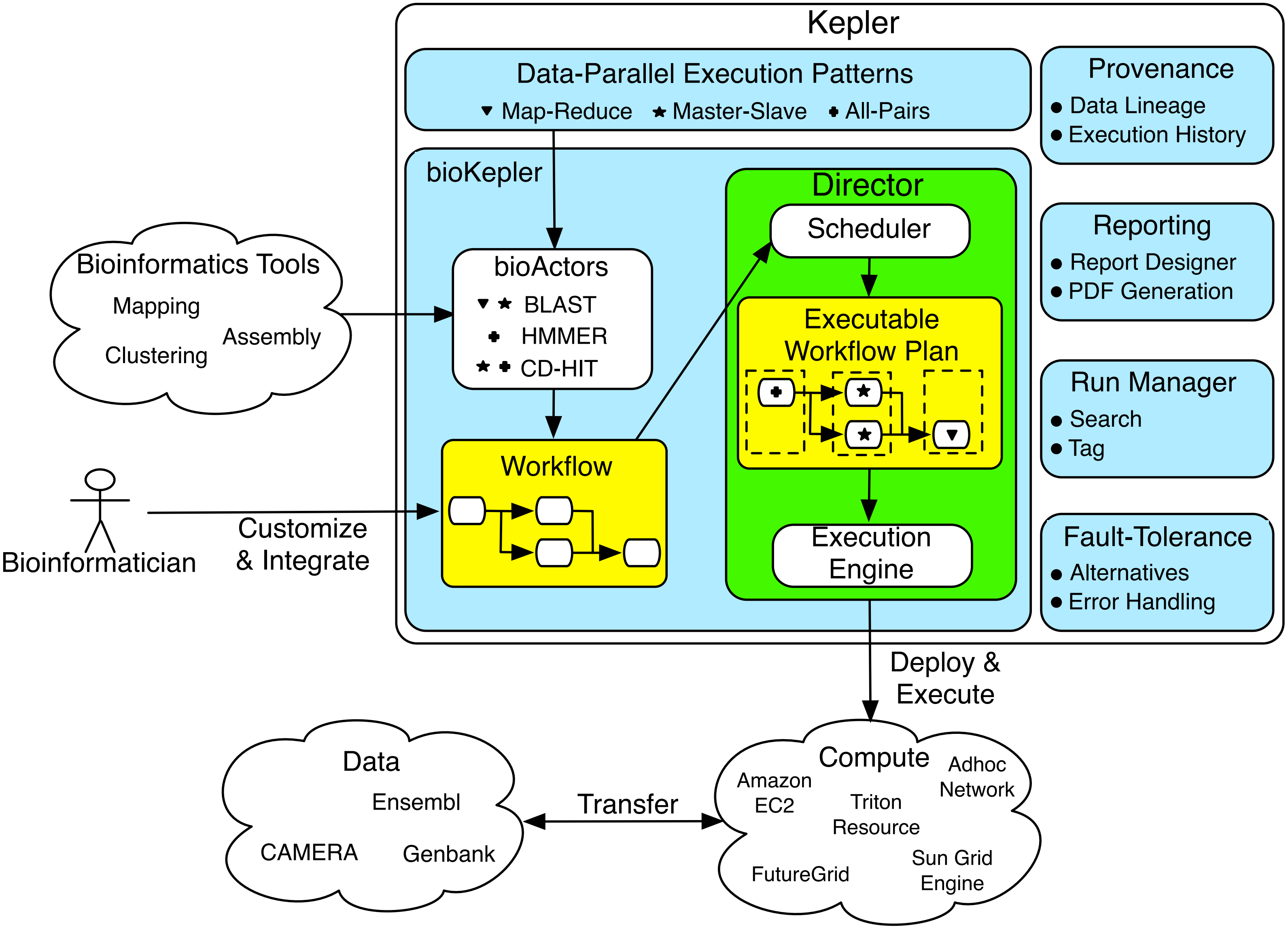
Galaxy **…** **Bio-Linux**

Kepler supports
- Workflows
- Other third party programming tools, e.g., R and Matlab
- Extensible task and data parallelization
- Service orientation
- Multiple engines, e.g., SDF, SGE, Hadoop

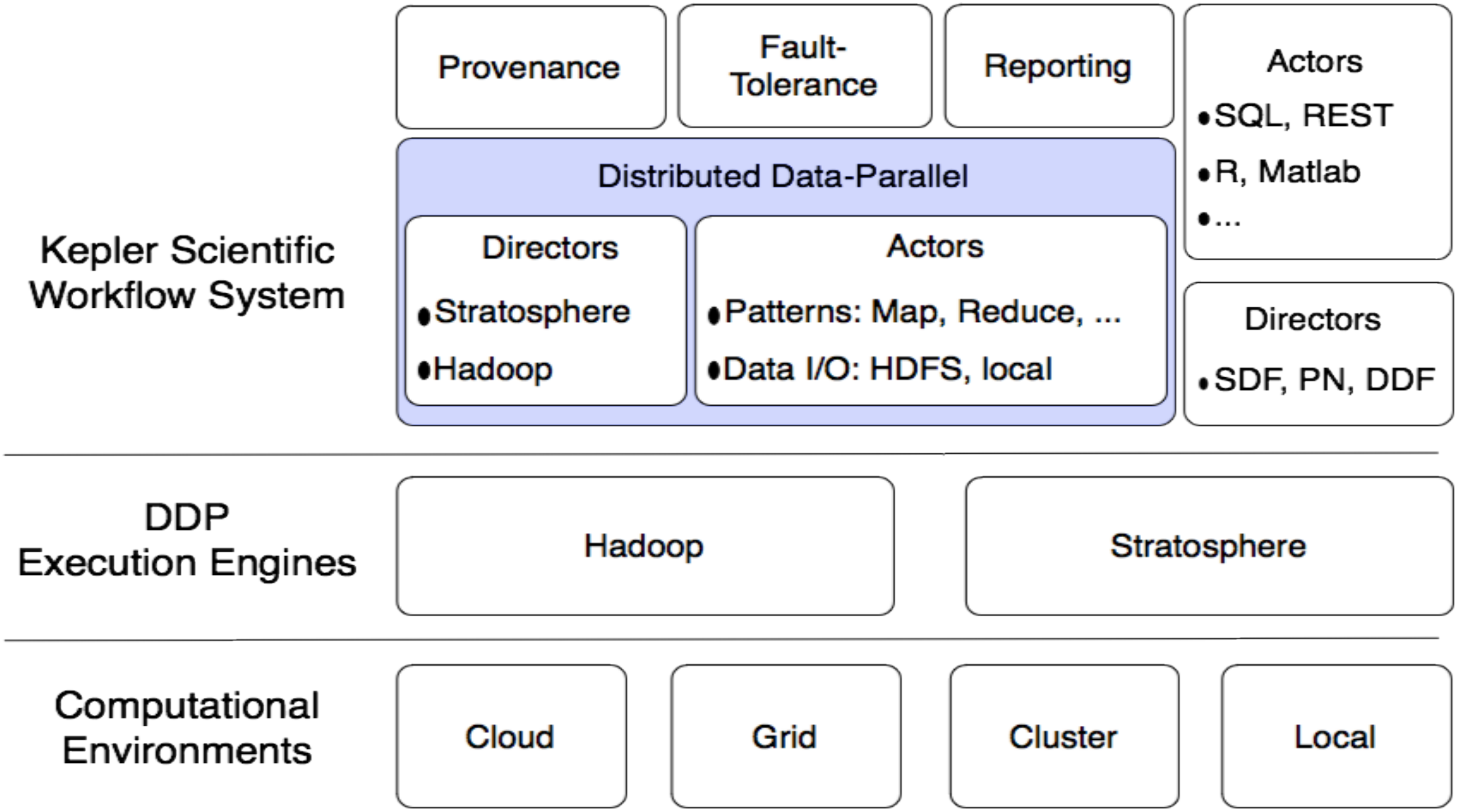
The bioKepler Approach

- Parallel Computation Framework
  - Use Distributed Data-Parallel (DDP) frameworks, e.g., MapReduce, and other parallelization methods to execute subworkflows
- bioActors
  - Configurable and reusable higher-order components for bioinformatics and computational biology
- Transparent support for different execution engines and computational environments
- Deployment on diverse environments

bioKepler's Conceptual Framework



bioKepler's Software Architecture



Kepler Scientific Workflow System

Provenance

Fault-Tolerance

Reporting

Distributed Data-Parallel

Directors
- Stratosphere
- Hadoop

Actors
- Patterns: Map, Reduce, ...
- Data I/O: HDFS, local

Actors
- SQL, REST
- R, Matlab
- ...

Directors
- SDF, PN, DDF

DDP Execution Engines

Hadoop

Stratosphere

Computational Environments

Cloud

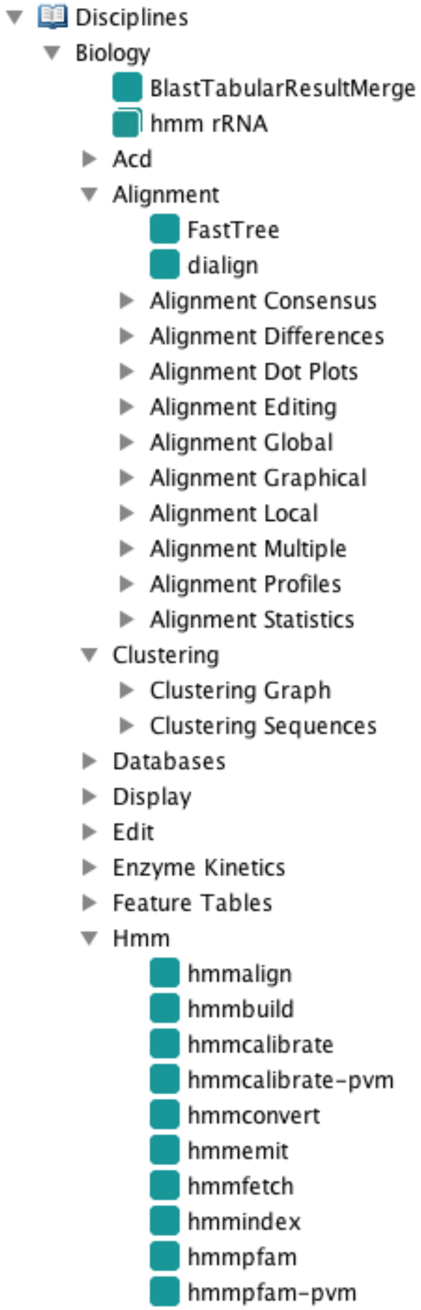
Grid

Cluster

Local

bioActors

- Set of steps to execute a bioinformatics tool locally or in an external environment
  - Locally executable
  - Parallelized external execution
- Customizable by the user based on external packages
  - Tools imported from CloudBioLinux
- Tools are evaluated on their computational requirements



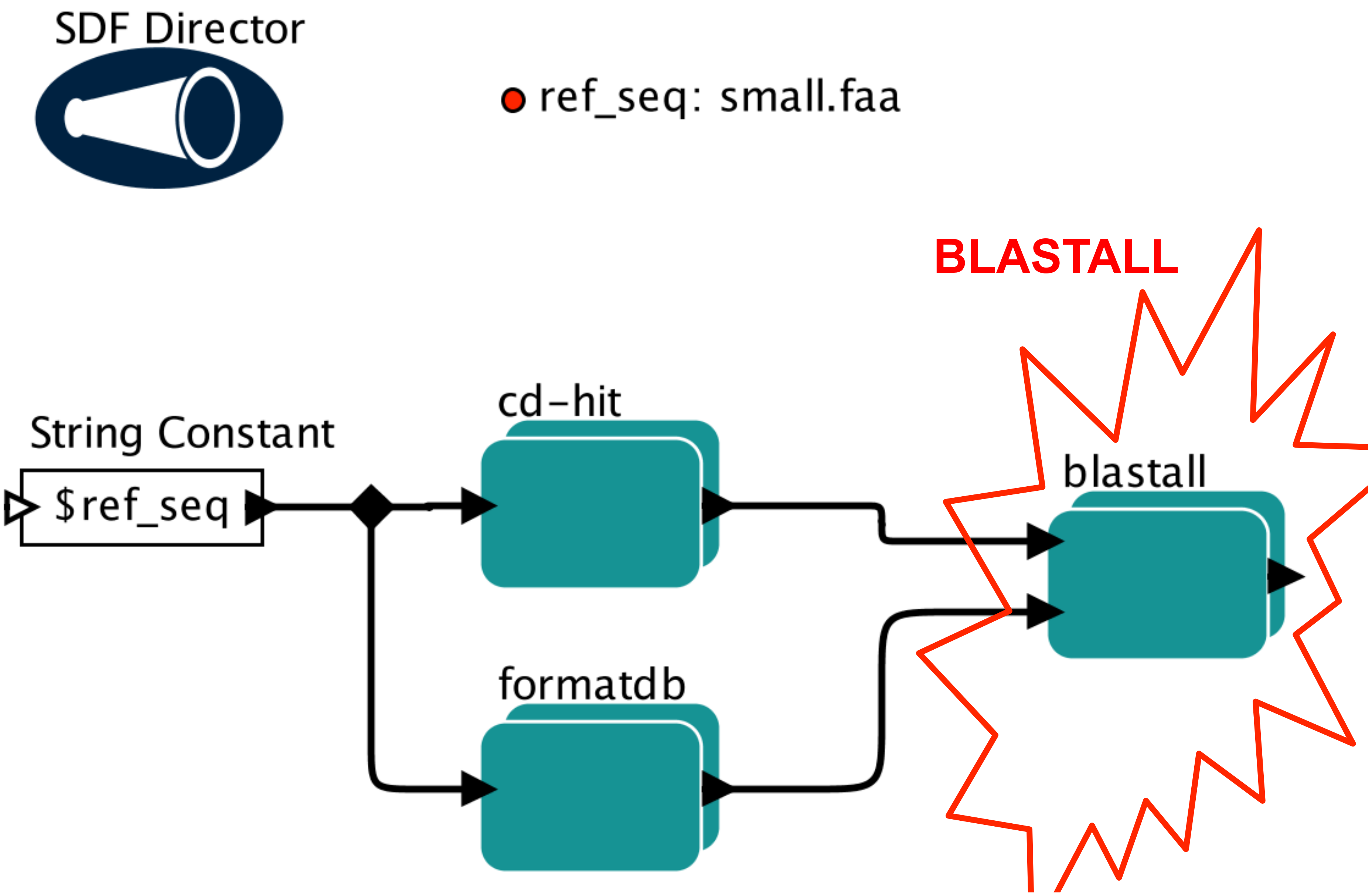
- Disciplines
  - Biology
    - BlastTabularResultMerge
    - hmm rRNA
    - Acd
    - Alignment
      - FastTree
      - dialign
      - Alignment Consensus
      - Alignment Differences
      - Alignment Dot Plots
      - Alignment Editing
      - Alignment Global
      - Alignment Graphical
      - Alignment Local
      - Alignment Multiple
      - Alignment Profiles
      - Alignment Statistics
    - Clustering
      - Clustering Graph
      - Clustering Sequences
    - Databases
    - Display
    - Edit
    - Enzyme Kinetics
    - Feature Tables
    - Hmm
      - hmmalign
      - hmmbuild
      - hmmcalibrate
      - hmmcalibrate-pvm
      - hmmconvert
      - hmmemit
      - hmmfetch
      - hmmindex
      - hmmpfam
      - hmmpfam-pvm

Example bioActors

- Alignment: BLAST, BLAT
- Profile-Sequence Alignment: PSI-BLAST
- Hidden Markov Model: HMMER
- Mapping: Bowtie, BWA, Samtools
- Multiple Alignment: ClustalW, Muscle
- Clustering: CD-HIT, Blastclust
- Gene Prediction: Glimmer, Genescan, Fraggenescan
- tRNA prediction: tRNA-scan, Meta-RNA
- Phylogeny: FastTree, RAxML

bioKepler.org

A Workflow with Three bioActors



Current Progress and Release

Downloadable as a package at: http://www.biokepler.org/releases

- A bioKepler VM executable on Amazon EC2 and FutureGrid
  - Builds upon CloudBioLinux including Bio-Linux and Galaxy
- A bioActor template that can be customized for different execution choices
  - e.g., local vs. Map/Reduce on a specific environment
- Example usecases

1st Workshop on bioKepler Tools and Its Applications

September 5-6, 2012

SDSC/UCSD La Jolla, CA

http://www.biokepler.org/workshops/2012-sep

Introductions

SDSC UC San Diego NSF

bioKepler.org

NEXT:

Introduction to bioActors

Weizhong Li

1st Workshop on bioKepler Tools and Its Applications

bioKepler.org