GMOD Projects at the Center for Genomics and Bioinformatics

Chris Hemmerich - Indiana University, Bloomington

A Simple Web Interface for Configuring GBrowse: WebGBrowse

Ram Podicheti

WebGBrowse

- A web interface for configuring GBrowse installations
  - Upload GFF file
  - Upload optional config file to use as starting point
- Add, edit, and remove new tracks using web forms
  - Extensive help embedded in forms and includes tutorial
  - Preview your changes at any point in GBrowse
- Makes GBrowse more feasible for small projects
  - We host the GBrowse server, so no installation is required
  - Configuration is done online through form
  - Use one configuration for multiple GFF files

WebGBrowse

- http://webgbrowse.cgb.indiana.edu/
- Available for download and local installation
- gmod-webgbrowse@lists.sourceforge.net
  - Support, make feature requests, contribute
  - We want to help you help us add support for more features
- Pending GMOD component
  - Migration of development environment

Podicheti, R., Gollapudi, R. & Dong, Q*. WebGbrowse – a web server for GBrowse *Bioinformatics*, 2009

Web-based Bioinformatics Pipelines for Biologists: ISGA

Chris Hemmerich, Aaron Buechlein

Ram Podicheti, Jeong-Hyeon Choi, Boshu Liu

ISGA: Driving Forces

- Workflow Management system that can meet the needs of a small sequencing center.

- Flexible pipeline definition
  - Design new pipelines
  - Incorporate new programs as components
- Support distributed computing environments
  - Potential need to grow beyond local computing resources
- Minimize CGB staff involvement in pipeline running
  - Free resources for building new pipelines

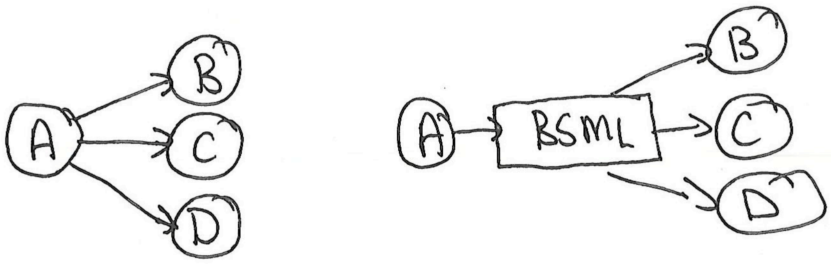
Workflow Management

- Ergatis (http://ergatis.sourceforge.net)
- Institute for Genome Sciences, U. Maryland
- Build pipelines from existing programs
- Supports distributed computing environments
- Robust monitoring of pipeline execution

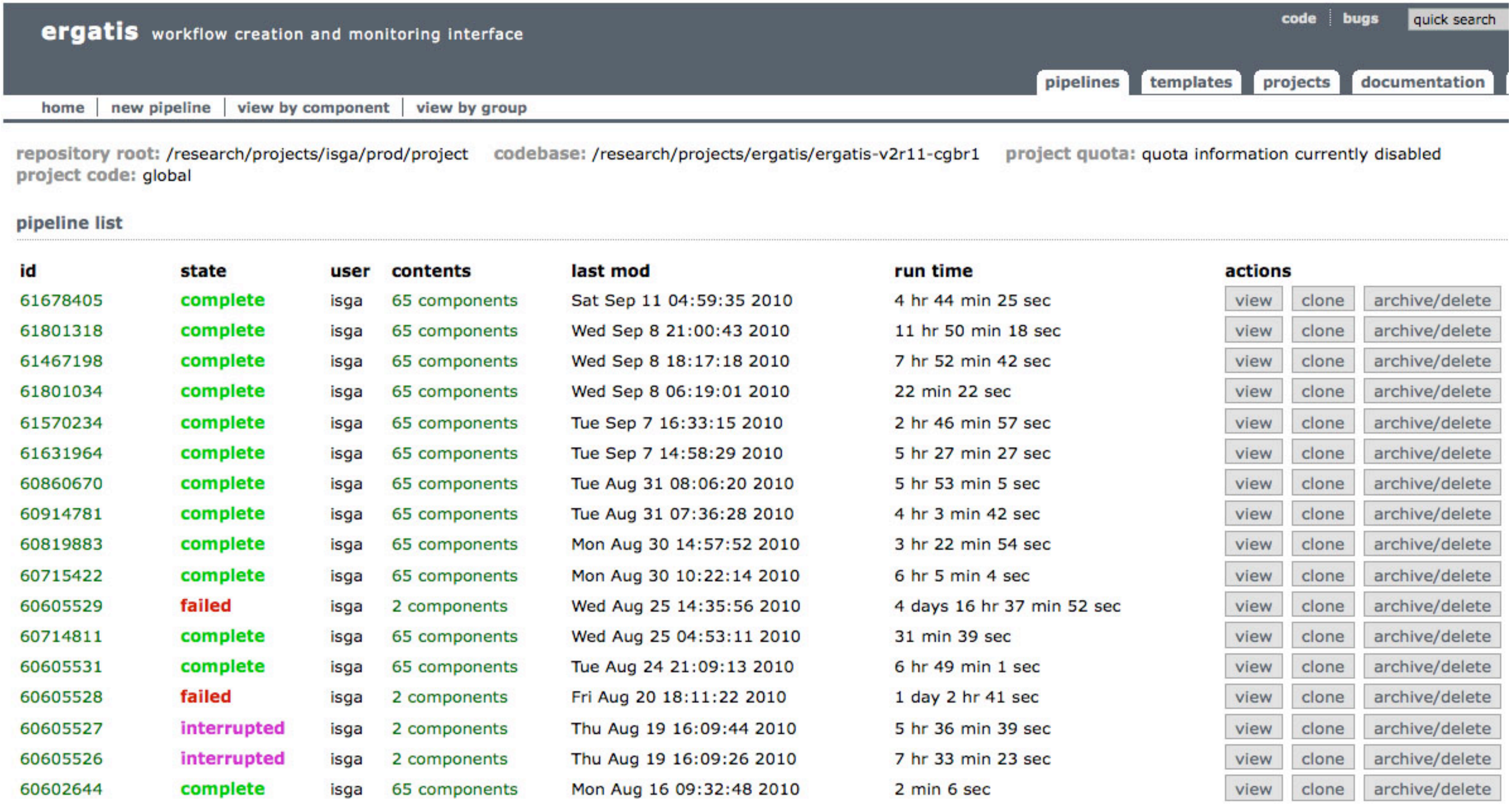
Orvis J, Crabtree J, Galens K, Gussman A, Inman JM, Lee E, Nampally S, Riley D, Sundaram JP, Felix V, Whitty B, Mahurkar A, Wortman J, White O, Angiuoli SV. Ergatis: A web interface and scalable software system for bioinformatics workflows. *Bioinformatics*. 2010 Jun 15;26(12).

Ergatis Workflow

- 10+ readily available pipelines, more in the community
- 220 components in svn, more in the community
- XML component and pipeline definition
- XML/BSML common data exchange format
  - Optional, but recommended for reusable components
  - Conversion tools for FASTA, GFF, Chado, etc…
  - Isolates format changes from other programs



Ergatis: Pipeline List

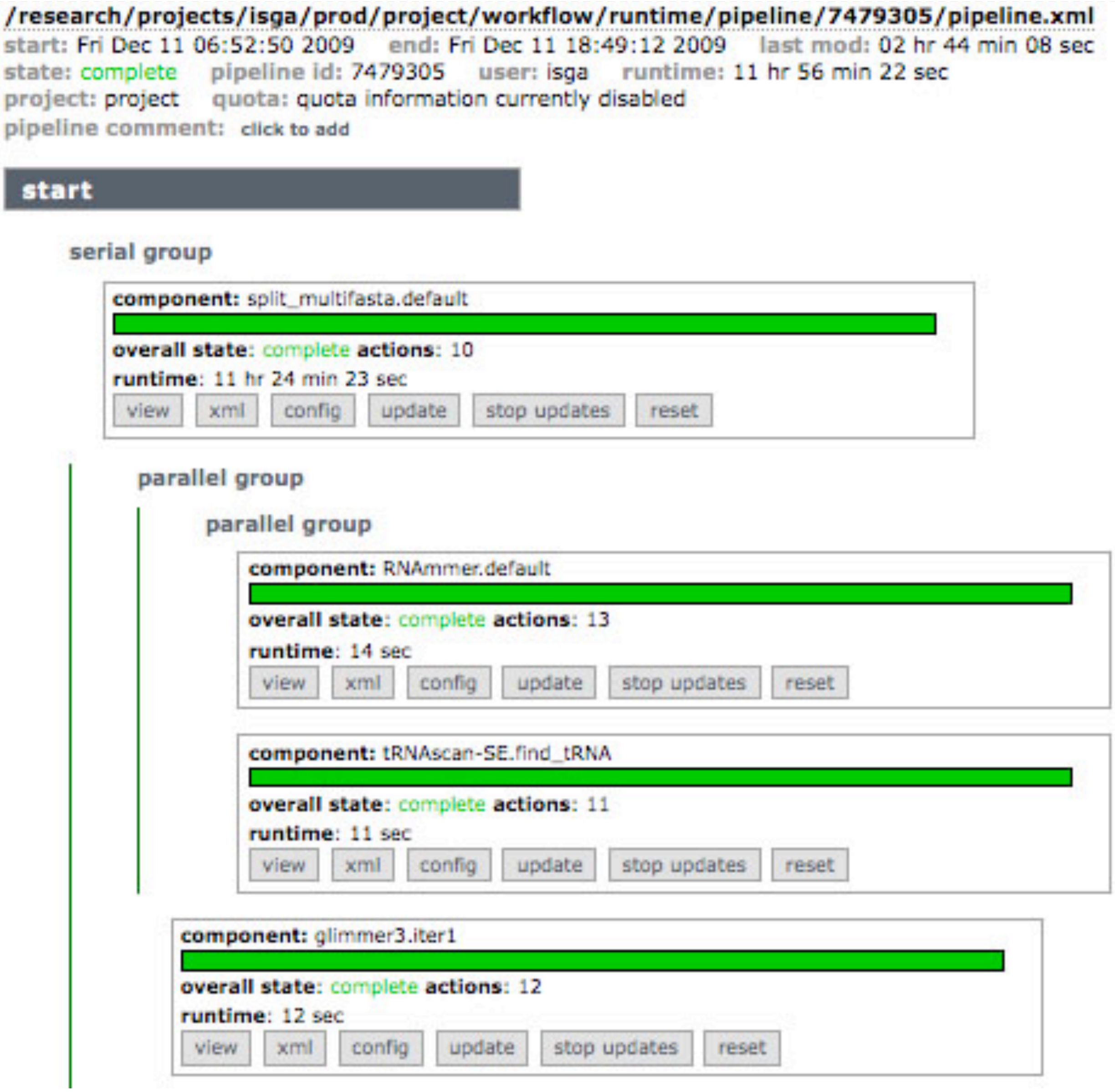


repository root: /research/projects/isga/prod/project **codebase:** /research/projects/ergatis/ergatis-v2r11-cgbr1 **project quota:** quota information currently disabled
project code: global

pipeline list

id	state	user	contents	last mod	run time	actions		
61678405	complete	isga	65 components	Sat Sep 11 04:59:35 2010	4 hr 44 min 25 sec	view	clone	archive/delete
61801318	complete	isga	65 components	Wed Sep 8 21:00:43 2010	11 hr 50 min 18 sec	view	clone	archive/delete
61467198	complete	isga	65 components	Wed Sep 8 18:17:18 2010	7 hr 52 min 42 sec	view	clone	archive/delete
61801034	complete	isga	65 components	Wed Sep 8 06:19:01 2010	22 min 22 sec	view	clone	archive/delete
61570234	complete	isga	65 components	Tue Sep 7 16:33:15 2010	2 hr 46 min 57 sec	view	clone	archive/delete
61631964	complete	isga	65 components	Tue Sep 7 14:58:29 2010	5 hr 27 min 27 sec	view	clone	archive/delete
60860670	complete	isga	65 components	Tue Aug 31 08:06:20 2010	5 hr 53 min 5 sec	view	clone	archive/delete
60914781	complete	isga	65 components	Tue Aug 31 07:36:28 2010	4 hr 3 min 42 sec	view	clone	archive/delete
60819883	complete	isga	65 components	Mon Aug 30 14:57:52 2010	3 hr 22 min 54 sec	view	clone	archive/delete
60715422	complete	isga	65 components	Mon Aug 30 10:22:14 2010	6 hr 5 min 4 sec	view	clone	archive/delete
60605529	failed	isga	2 components	Wed Aug 25 14:35:56 2010	4 days 16 hr 37 min 52 sec	view	clone	archive/delete
60714811	complete	isga	65 components	Wed Aug 25 04:53:11 2010	31 min 39 sec	view	clone	archive/delete
60605531	complete	isga	65 components	Tue Aug 24 21:09:13 2010	6 hr 49 min 1 sec	view	clone	archive/delete
60605528	failed	isga	2 components	Fri Aug 20 18:11:22 2010	1 day 2 hr 41 sec	view	clone	archive/delete
60605527	interrupted	isga	2 components	Thu Aug 19 16:09:44 2010	5 hr 36 min 39 sec	view	clone	archive/delete
60605526	interrupted	isga	2 components	Thu Aug 19 16:09:26 2010	7 hr 33 min 23 sec	view	clone	archive/delete
60602644	complete	isga	65 components	Mon Aug 16 09:32:48 2010	2 min 6 sec	view	clone	archive/delete

Ergatis: Pipeline Monitor



/research/projects/isga/prod/project/workflow/runtime/pipeline/7479305/pipeline.xml

start: Fri Dec 11 06:52:50 2009 end: Fri Dec 11 18:49:12 2009 last mod: 02 hr 44 min 08 sec

state: complete pipeline id: 7479305 user: isga runtime: 11 hr 56 min 22 sec

project: project quota: quota information currently disabled

pipeline comment: click to add

component: split_multifasta.default

overall state: complete **actions:** 10

runtime: 11 hr 24 min 23 sec

component: RNAmmer.default

overall state: complete **actions:** 13

runtime: 14 sec

component: tRNAscan-SE.find_tRNA

overall state: complete **actions:** 11

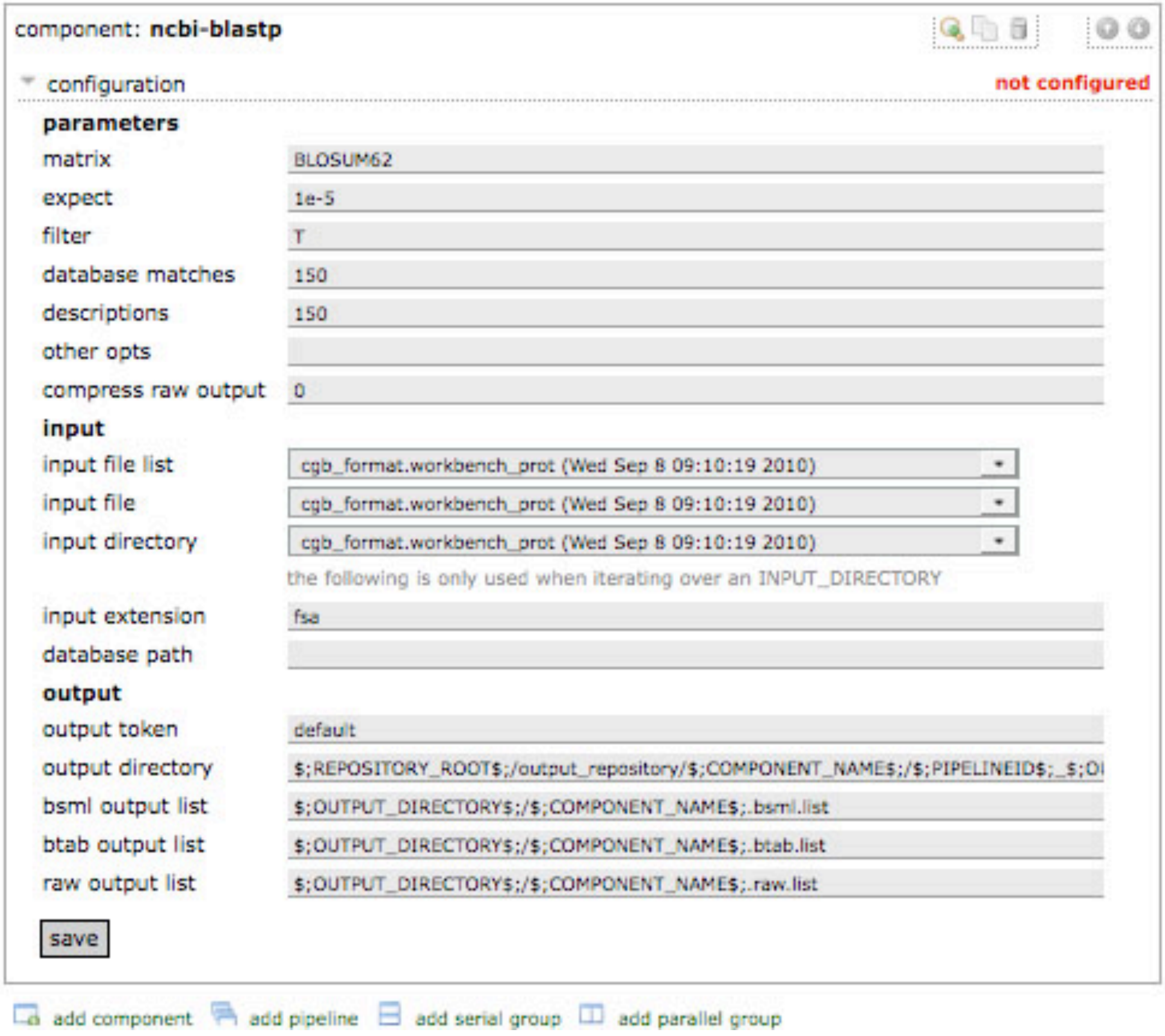
runtime: 11 sec

component: glimmer3.iter1

overall state: complete **actions:** 12

runtime: 12 sec

Ergatis: Configure Component

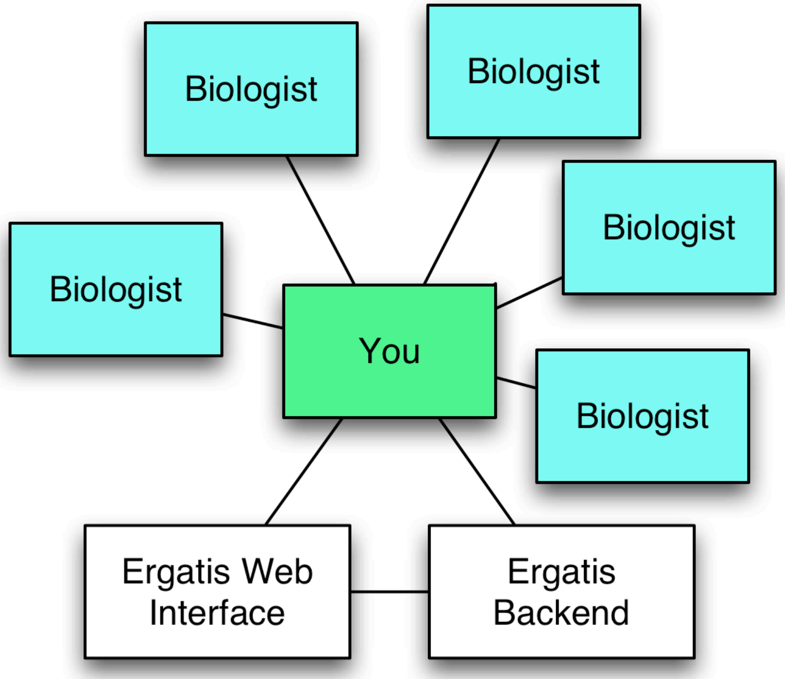


component: **ncbi-blastp**

configuration — not configured

parameters	
matrix	BLOSUM62
expect	1e-5
filter	T
database matches	150
descriptions	150
other opts	
compress raw output	0
input	
input file list	cgb_format.workbench_prot (Wed Sep 8 09:10:19 2010)
input file	cgb_format.workbench_prot (Wed Sep 8 09:10:19 2010)
input directory	cgb_format.workbench_prot (Wed Sep 8 09:10:19 2010)
	the following is only used when iterating over an INPUT_DIRECTORY
input extension	fsa
database path	
output	
output token	default
output directory	$;REPOSITORY_ROOT$;/output_repository/$;COMPONENT_NAME$;/$;PIPELINEID$;_$;O
bsml output list	$;OUTPUT_DIRECTORY$;/$;COMPONENT_NAME$;.bsml.list
btab output list	$;OUTPUT_DIRECTORY$;/$;COMPONENT_NAME$;.btab.list
raw output list	$;OUTPUT_DIRECTORY$;/$;COMPONENT_NAME$;.raw.list

Ergatis Architecture



Biologist Interface Requirements

- Support single-lab biologists
  - Self-sufficient but have limited bioinformatics resources
  - Embrace tools that don’t require extensive training
- Ability to run pre-configured pipelines quickly
- Option to customizing specific tools in a pipeline
- Interface that encourages exploration
  - Remove complexity and information they don’t need
  - Inline help
  - Immediately detect errors and allow them to correct them
  - Return output in useful formats
  - Simple tools for visualizing and searching large result sets

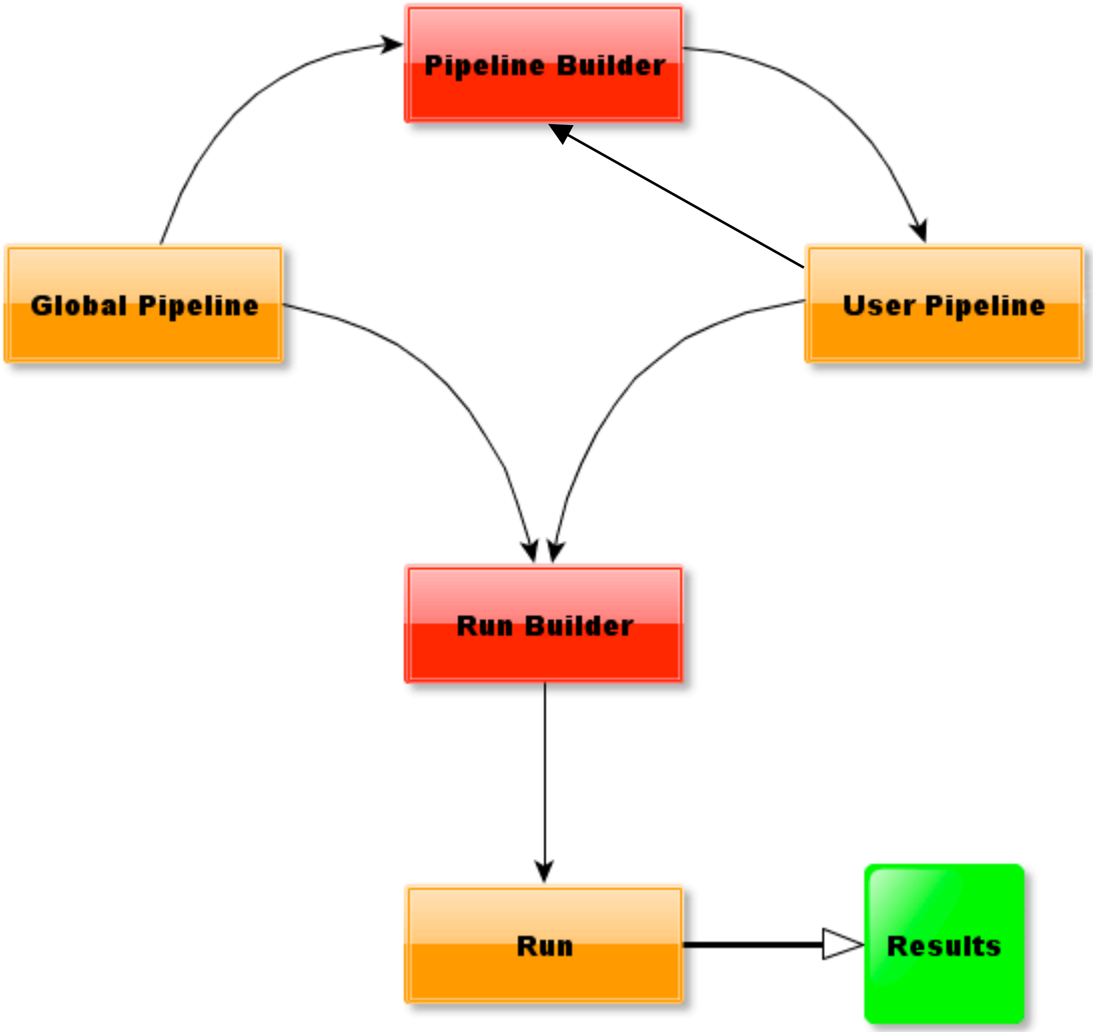
ISGA Design

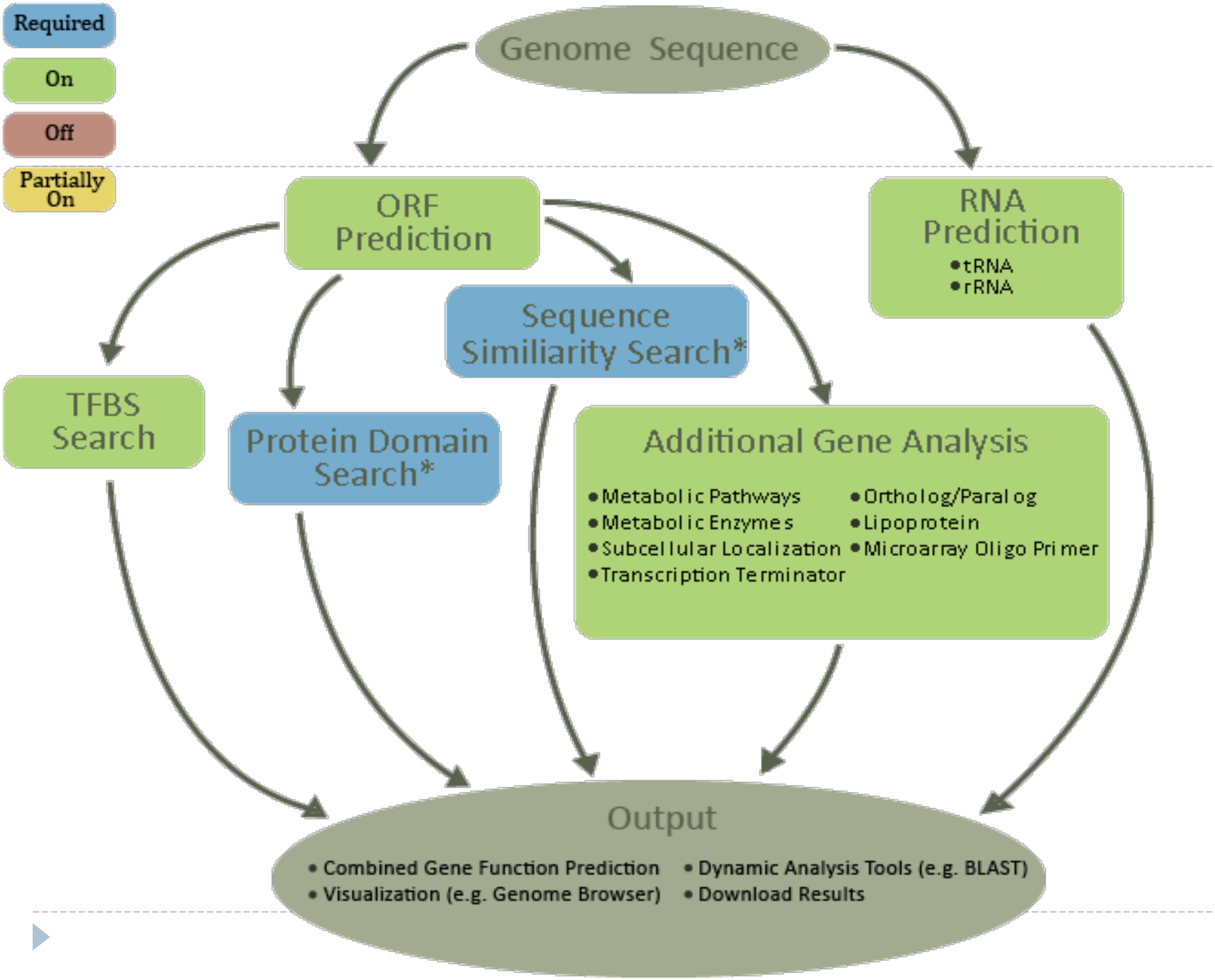
- Simplify pipelines
  - Hide housekeeping components
  - Group components into clusters representing processes
- Support customization
  - Disable components where possible
  - Replace components with pre-computed data where possible
  - Edit scientifically-active program parameters
- Help and validation for all forms
- Users and data privacy
- Provide download and upload
- Incorporate visualization & analysis tools

Why develop ISGA as a separate package?

- ISGA only re-implements the web interface of Ergatis
  - Ergatis libraries, component definitions, and method of running and monitoring pipelines is used by ISGA as-is
- ISGA adds and removes Ergatis features
  - Accessing component information
  - Building pipelines from components
- A hybrid ISGA/Ergatis interface wouldn't serve anyone
  - ISGA biologist users need to be given limited functionality for simplicity and security
  - Ergatis bioinformatician users need full functionality and a complex interface to work efficiently

Workflow


Pipeline Builder
Global Pipeline
User Pipeline
Run Builder
Run
Results



🟦 Required

🟩 On

🟥 Off

🟨 Partially On

Genome Sequence

ORF Prediction

RNA Prediction
- tRNA
- rRNA

TFBS Search

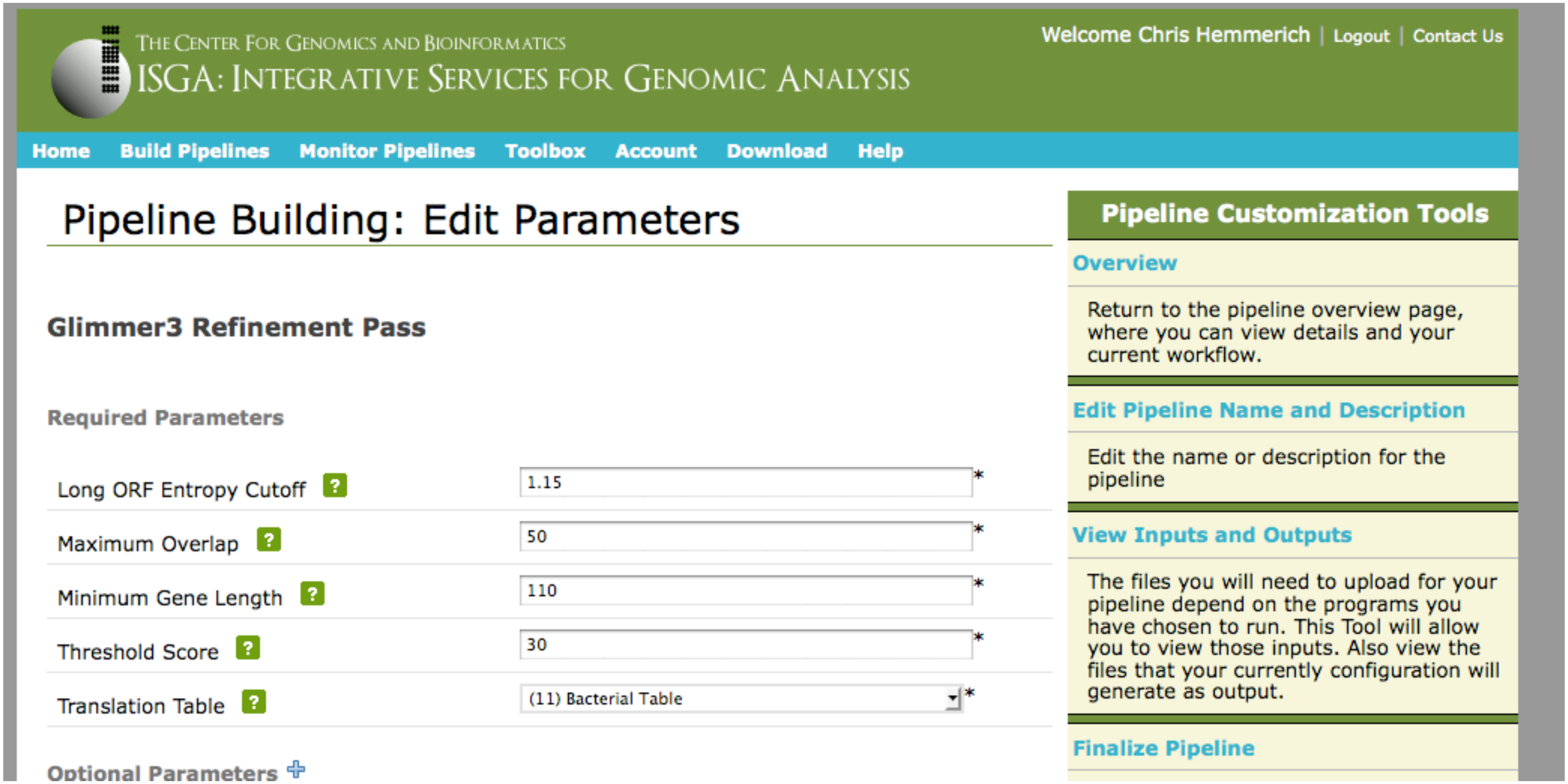
Protein Domain Search*

Sequence Similiarity Search*

Additional Gene Analysis
- Metabolic Pathways
- Metabolic Enzymes
- Subcellular Localization
- Transcription Terminator
- Ortholog/Paralog
- Lipoprotein
- Microarray Oligo Primer

Output
- Combined Gene Function Prediction
- Visualization (e.g. Genome Browser)
- Dynamic Analysis Tools (e.g. BLAST)
- Download Results

Pipeline Builder



THE CENTER FOR GENOMICS AND BIOINFORMATICS
ISGA: INTEGRATIVE SERVICES FOR GENOMIC ANALYSIS
Welcome Chris Hemmerich | Logout | Contact Us
Home
Build Pipelines
Monitor Pipelines
Toolbox
Account
Download
Help

Pipeline Building: Edit Parameters

Glimmer3 Refinement Pass

Required Parameters

Long ORF Entropy Cutoff ?	1.15 *
Maximum Overlap ?	50 *
Minimum Gene Length ?	110 *
Threshold Score ?	30 *
Translation Table ?	(11) Bacterial Table *

Optional Parameters ✚

Pipeline Customization Tools

Overview

Return to the pipeline overview page, where you can view details and your current workflow.

Edit Pipeline Name and Description

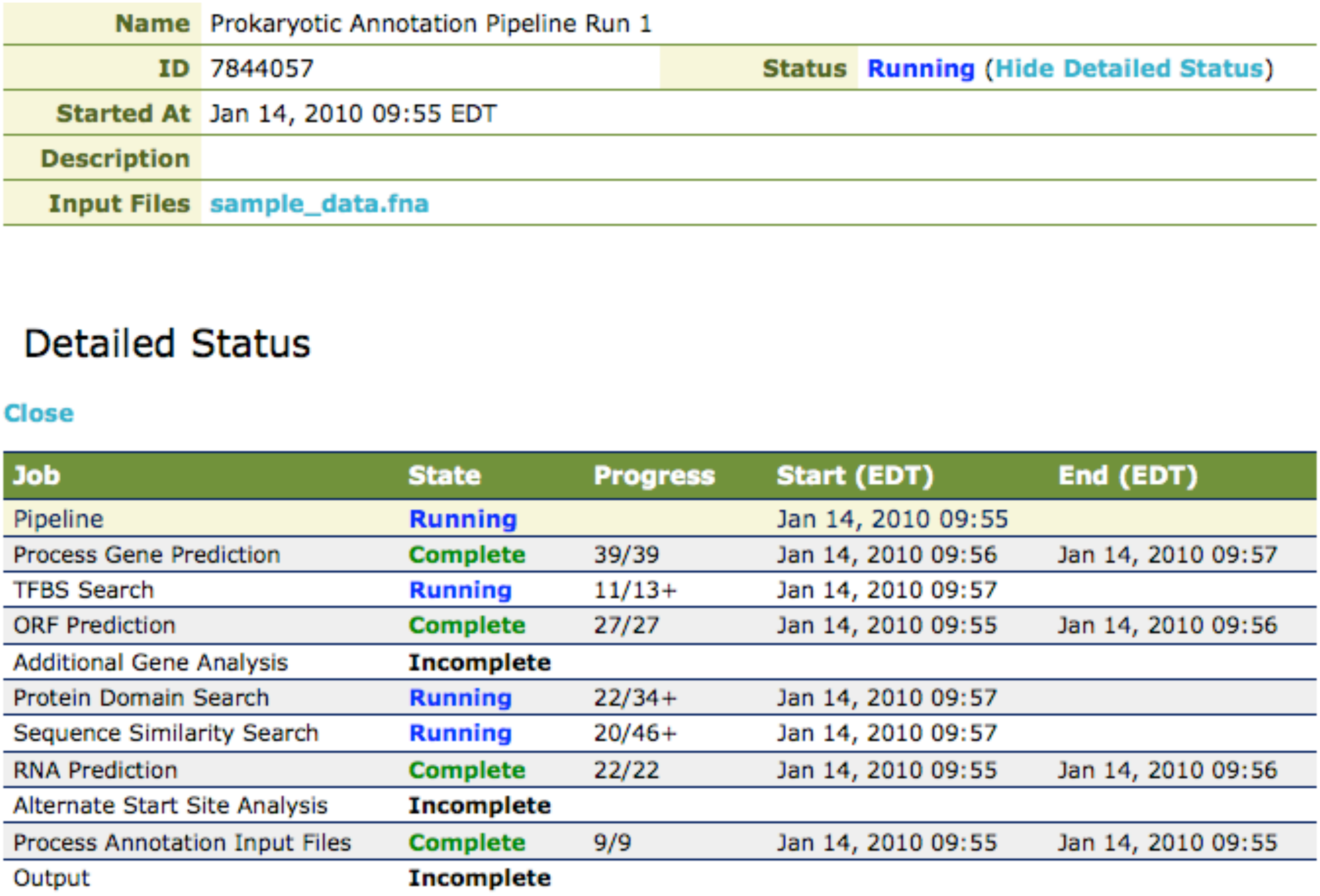
Edit the name or description for the pipeline

View Inputs and Outputs

The files you will need to upload for your pipeline depend on the programs you have chosen to run. This Tool will allow you to view those inputs. Also view the files that your currently configuration will generate as output.

Finalize Pipeline

Run Status



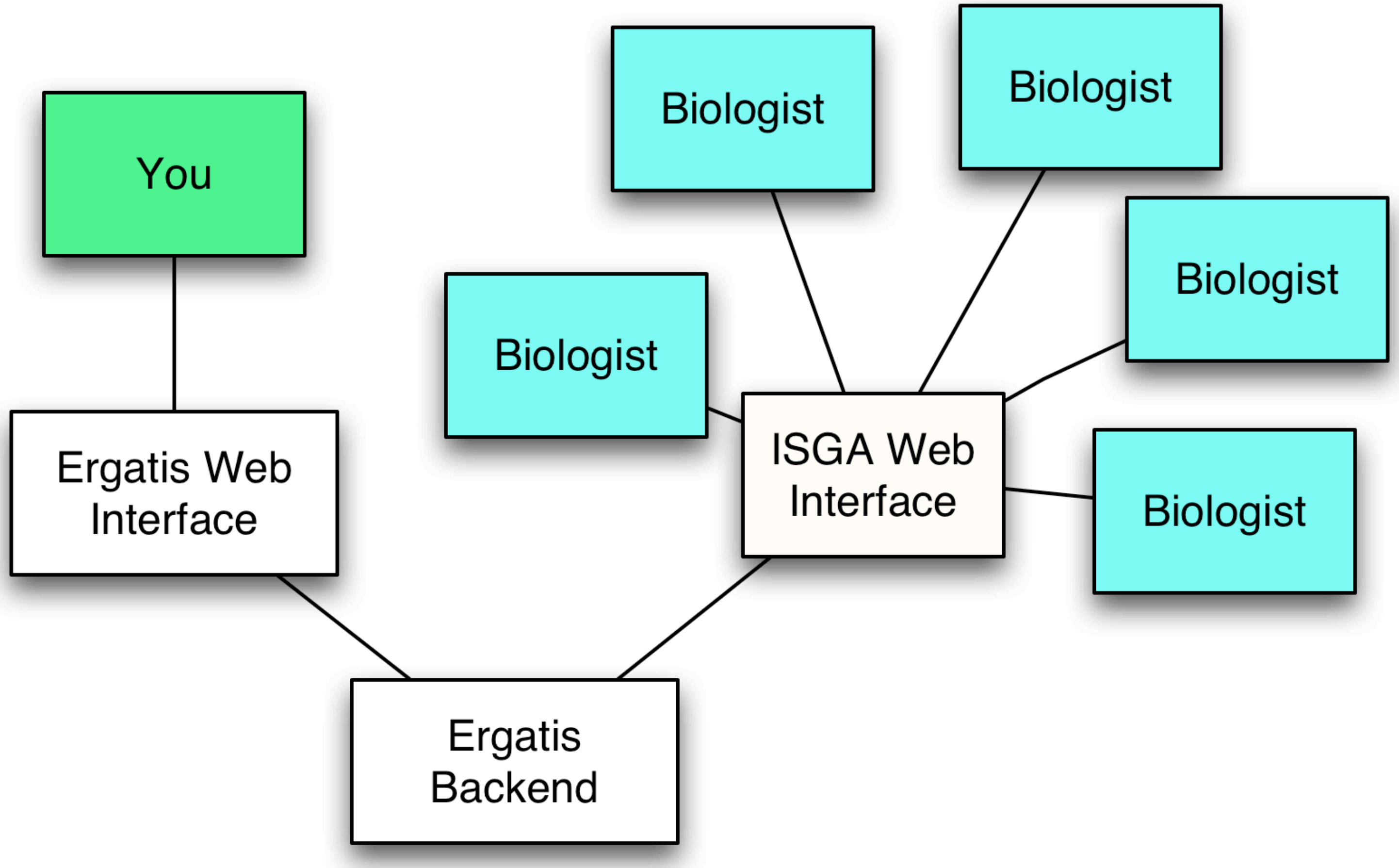
Name	Prokaryotic Annotation Pipeline Run 1		
ID	7844057	**Status**	**Running** (**Hide Detailed Status**)
Started At	Jan 14, 2010 09:55 EDT		
Description			
Input Files	**sample_data.fna**		

Detailed Status

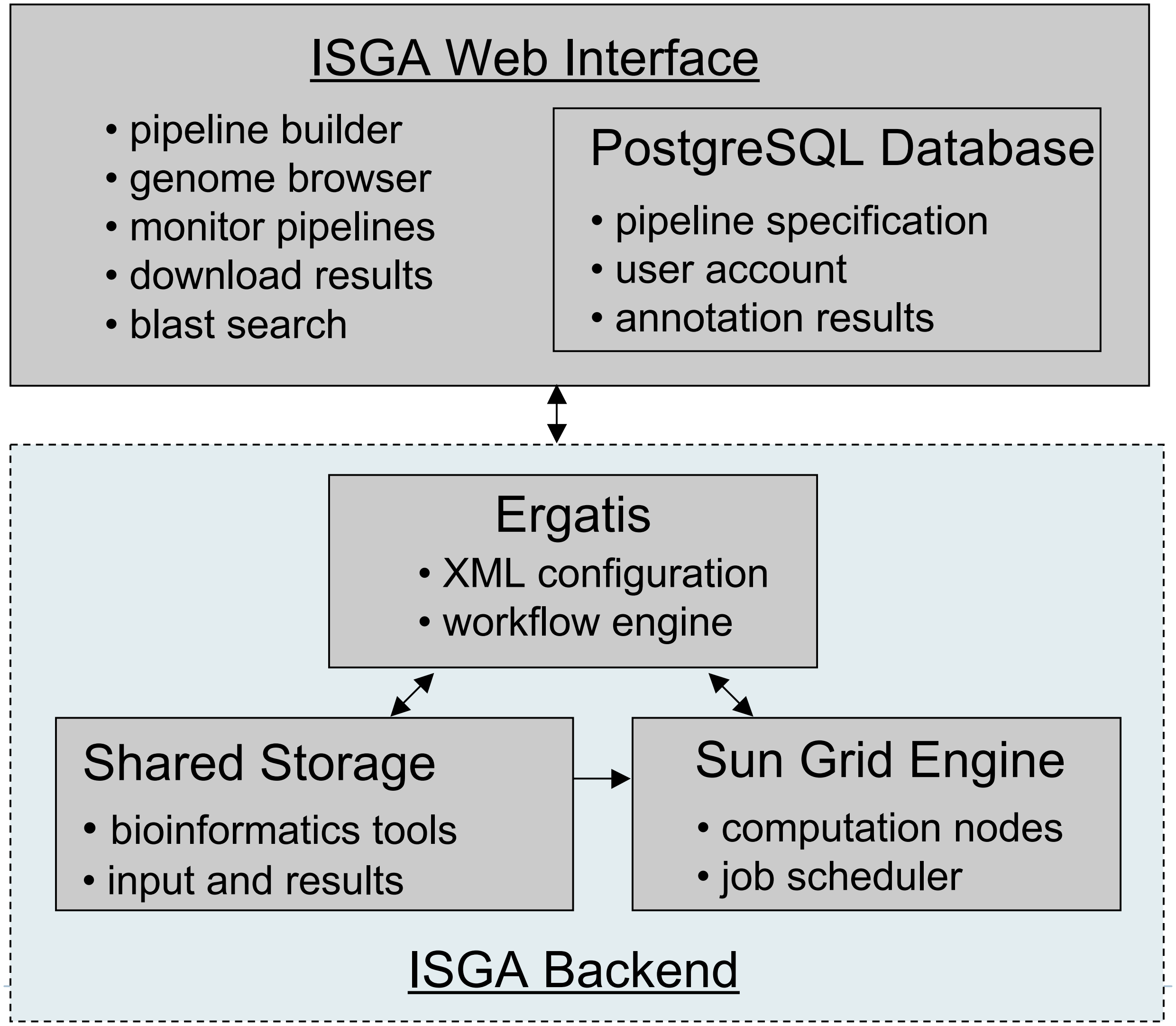
Close

Job	State	Progress	Start (EDT)	End (EDT)
Pipeline	**Running**		Jan 14, 2010 09:55	
Process Gene Prediction	**Complete**	39/39	Jan 14, 2010 09:56	Jan 14, 2010 09:57
TFBS Search	**Running**	11/13+	Jan 14, 2010 09:57	
ORF Prediction	**Complete**	27/27	Jan 14, 2010 09:55	Jan 14, 2010 09:56
Additional Gene Analysis	**Incomplete**			
Protein Domain Search	**Running**	22/34+	Jan 14, 2010 09:57	
Sequence Similarity Search	**Running**	20/46+	Jan 14, 2010 09:57	
RNA Prediction	**Complete**	22/22	Jan 14, 2010 09:55	Jan 14, 2010 09:56
Alternate Start Site Analysis	**Incomplete**			
Process Annotation Input Files	**Complete**	9/9	Jan 14, 2010 09:55	Jan 14, 2010 09:55
Output	**Incomplete**			

ISGA Architecture



Under the Hood



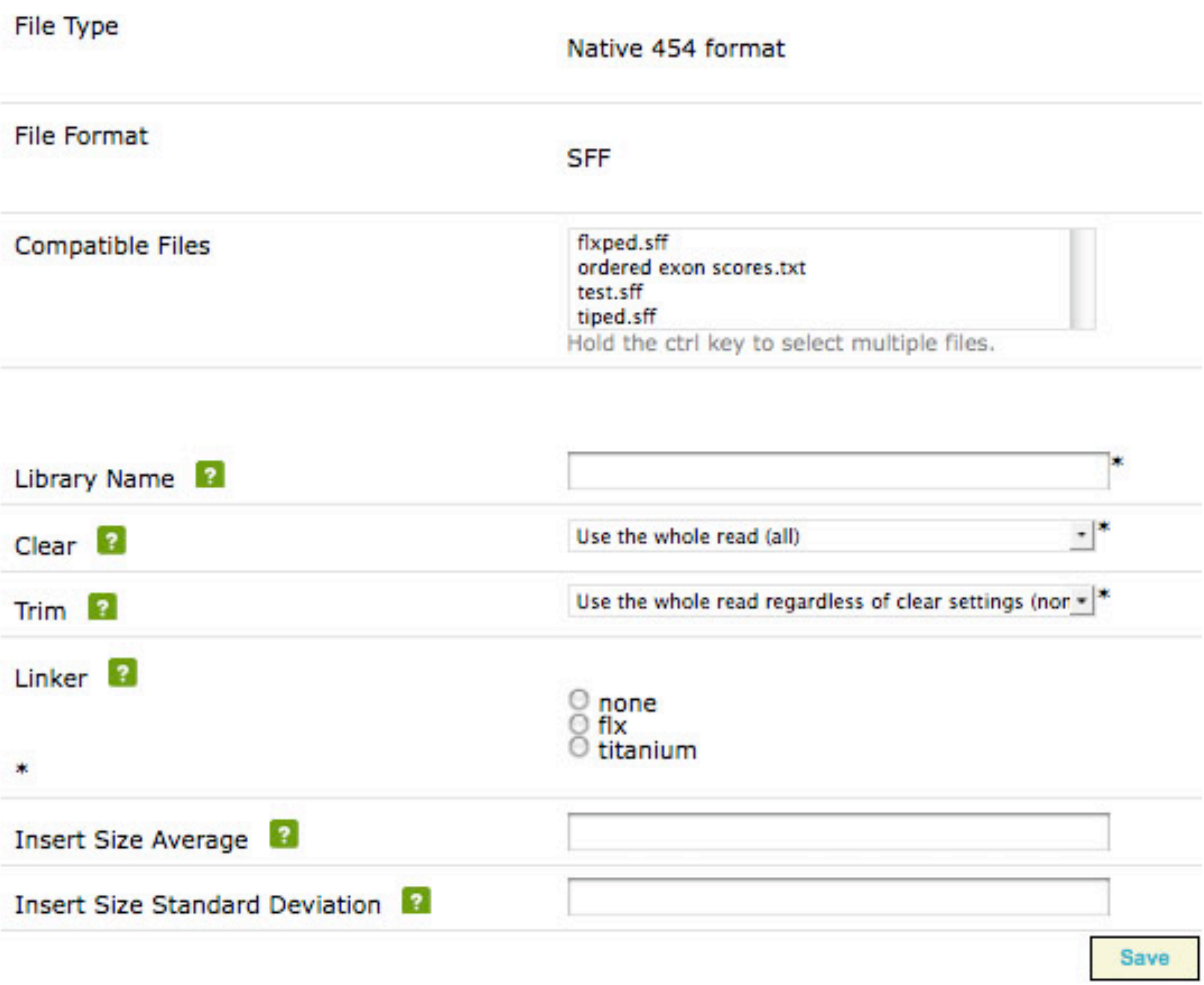
Usage

- > 100 pipelines run
- > 60 users
- Two external sites evaluating local ISGA installations that we know of

What’s new?

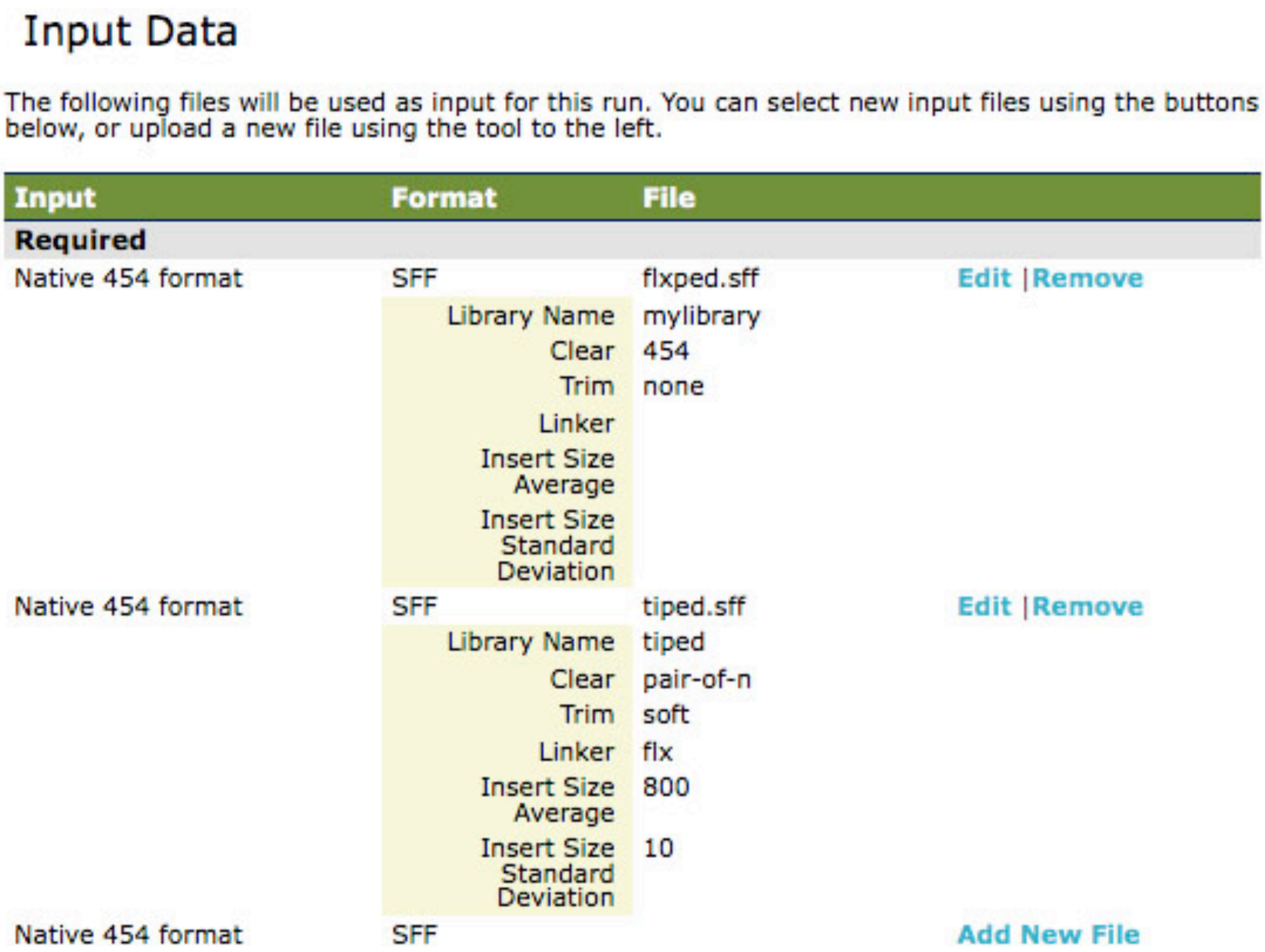
- Celera assembly pipeline
  - Ability to accept parameters with pipeline inputs
  - Ability to iterate components over a list of pipeline inputs
  - Conversion scripts for Hawkeye visualization
- Installation instructions :shame
- isga-users@lists.sourceforge.net
- Administration improvements
  - Online configuration
  - User classes and pipeline quotas

Parameterized Inputs



File Type	Native 454 format
File Format	SFF
Compatible Files	flxped.sff ordered exon scores.txt test.sff tiped.sff Hold the ctrl key to select multiple files.
Library Name ?	*
Clear ?	Use the whole read (all) *
Trim ?	Use the whole read regardless of clear settings (nor *
Linker ? *	○ none ○ flx ○ titanium
Insert Size Average ?	
Insert Size Standard Deviation ?	

Input Iterator



Input Data

The following files will be used as input for this run. You can select new input files using the buttons below, or upload a new file using the tool to the left.

Input	Format	File	
Required			
Native 454 format	SFF	flxped.sff	Edit |Remove
	Library Name	mylibrary	
	Clear	454	
	Trim	none	
	Linker		
	Insert Size Average		
	Insert Size Standard Deviation		
Native 454 format	SFF	tiped.sff	Edit |Remove
	Library Name	tiped	
	Clear	pair-of-n	
	Trim	soft	
	Linker	flx	
	Insert Size Average	800	
	Insert Size Standard Deviation	10	
Native 454 format	SFF		Add New File

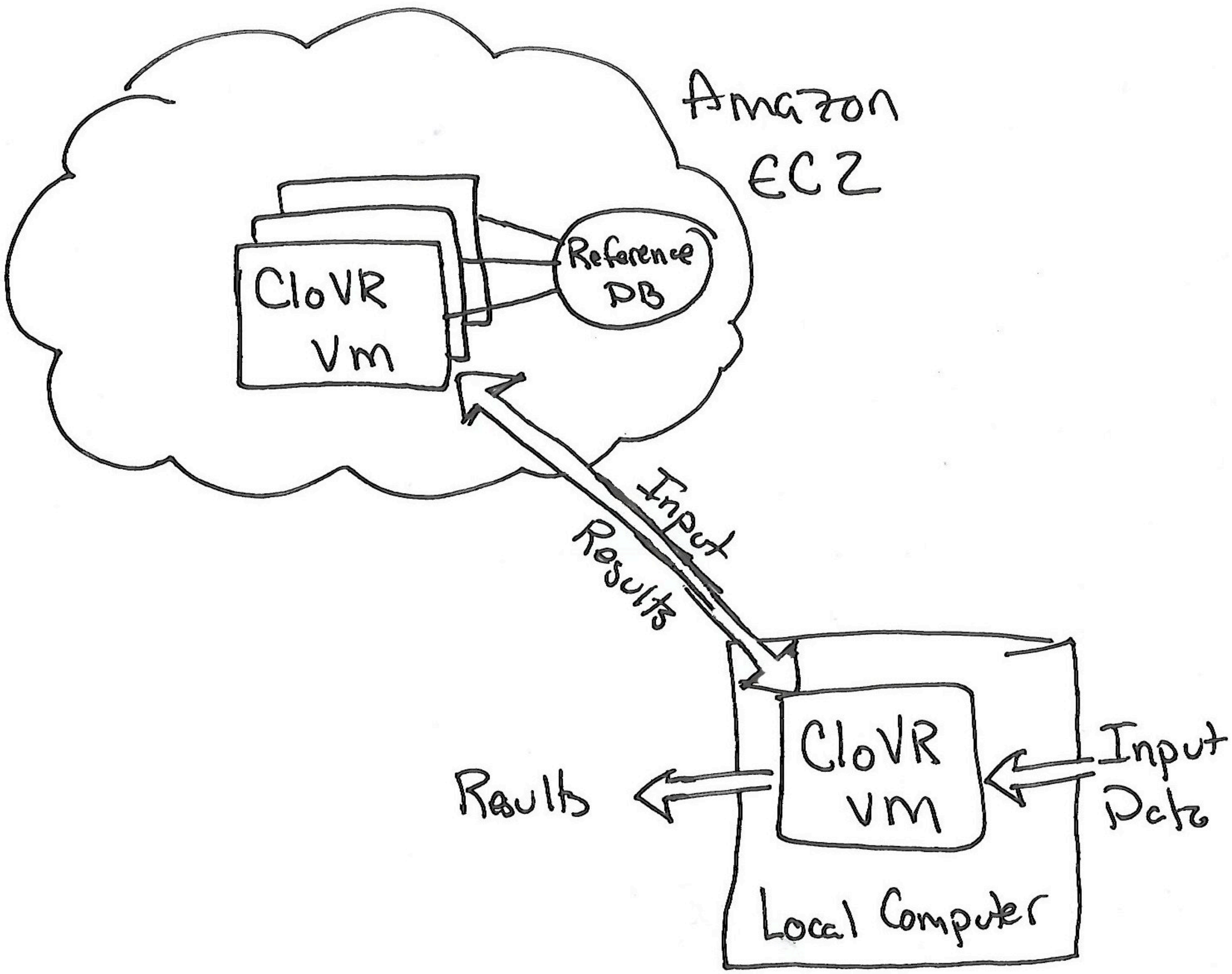
What's in the works?

- Pipelines
  - SHORE SNP Calling (ISGA)
  - Gene clustering over Microbial phylogenies (Ergatis)
  - Transcriptome annotation pipeline (Ergatis)
  - Methyl-seq (Ergatis)
- Features
  - Pipeline reproducibility and provenance
  - User groups and sharing
  - Modular pipeline and toolbox installation
    - ISGA pipelines as standalone Ergatis templates
  - ISGA pipeline over Amazon EC2 via CLoVR

Cloud Resources through CloVR

- Execute Ergatis Pipelines over an SGE instance hosted on Amazon EC2 machine images
- CloVR manages creation and shutdown of cloud images as part of pipeline
- Upload input as part of pipeline or access data hosted at Amazon
- Results are retrieved to local machine
- Ergatis assumes a shared filesystem, so some modification is required to manage file transfers

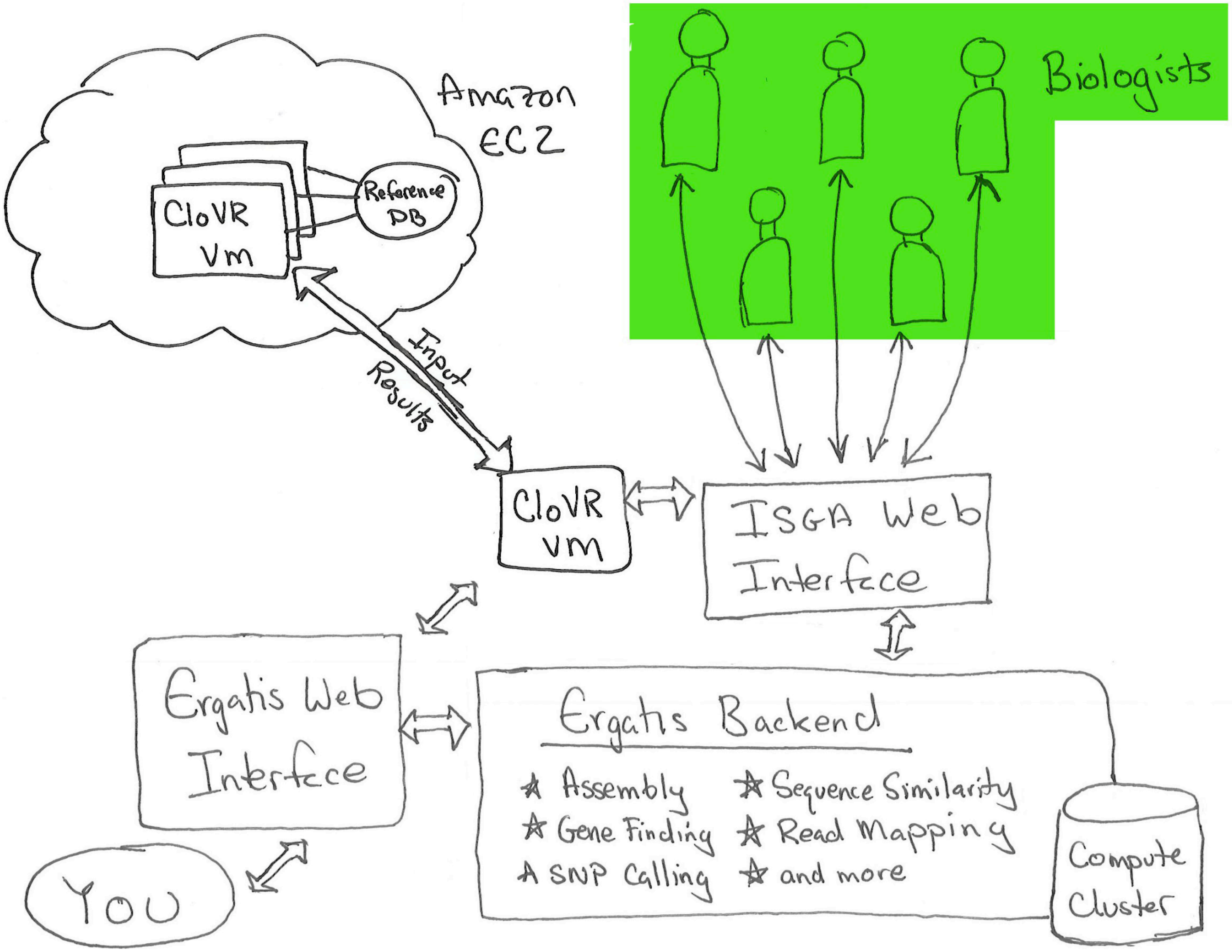
CloVR Architecture



Using CloVR with ISGA

- ISGA/Ergatis pipelines can be ported to ISGA/CloVR
- ISGA installation communicates with local Ergatis and CloVR
- EC2 presents challenges for billing customers

ISGA with CloVR Architecture



Acknowledgements

Funding

•Indiana Metabolomics and Cytomics Initiative(METACyt) – Lilly Endowment, Inc.

•National Institutes of Health under grant 5 RC2 HG005806-02.

CGB

Genomics

John Colbourne
Keithanne Mockaitis

Bioinformatics

Haixu Tang
Jeong-Hyeon Choi
Aaron Buechlein
Ram Podicheti

Computing

Phillip Steinbachs
Jon Burgoyne

ISGA Aumni

Qunfeng Dong
Kashi Revanna

External Projects

Joshua Orvis & Ergatis team
Sam Angiuoli & CLoVR team
Anup Mahurkar & Workflow team