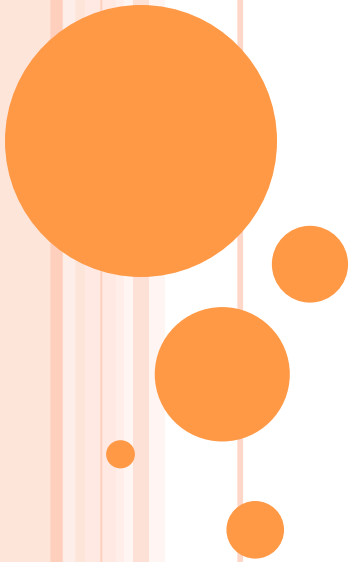
WEB-BASED BIOINFORMATICS PIPELINES FOR BIOLOGISTS

Integrative Services for Genomic Analysis (ISGA)

Chris Hemmerich

Center for Genomics and Bioformatics

CONTACT: biohelp@cgb.indiana.edu



JUSTIFICATION AND HISTORY

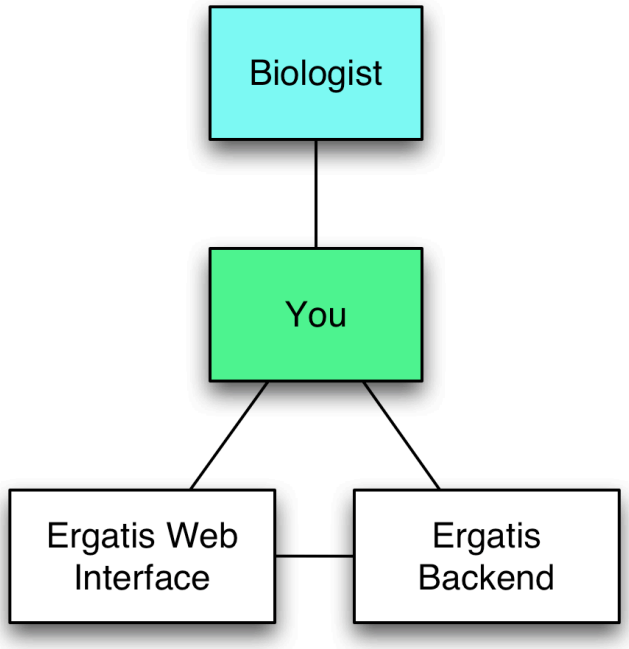
ISGA BACKGROUND

- Provide a high-throughput microbial annotation service to local biologists
  - Reliable and pipelined execution
  - Efficient maintenance
  - Provide privacy and security for data
- High-quality (automated) annotation
  - Biologists able to customize parameters
  - Able to incorporate new programs and pipelines

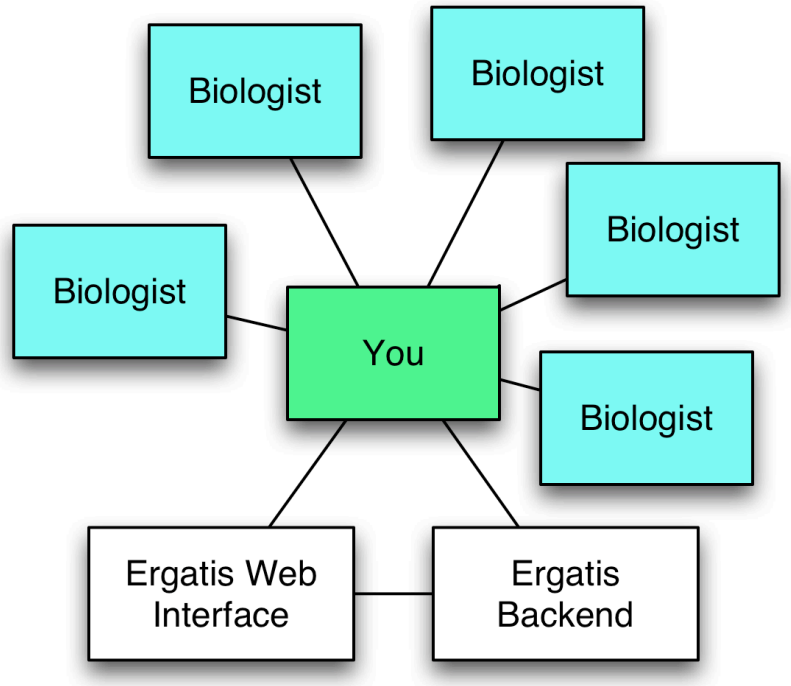
ERGATIS (ERGATIS.SOURCEFORGE.NET)

- Web-based analysis pipeline tool
- Wraps tools and utilities in “components”
- Ability to add new components
- Build new and customize existing pipelines
- In-depth monitoring of pipelines
- Underlying Workflow package supports SGE
- XML/BSML common data exchange format
- Includes prokaryotic annotation pipeline

ERGATIS WORKFLOW

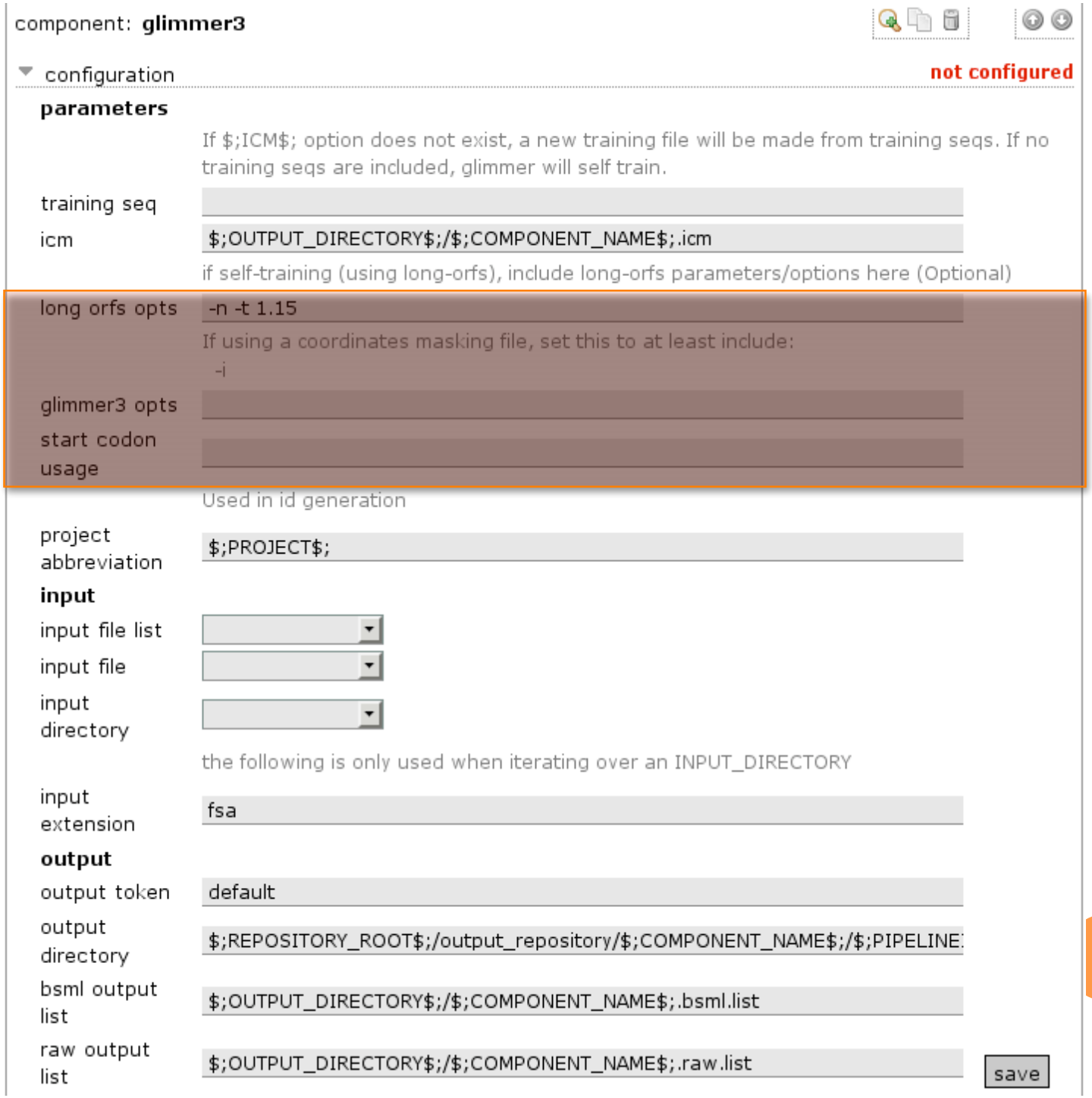


A SLIGHT CORRECTION



WHY NOT EXPOSE ERGATIS?

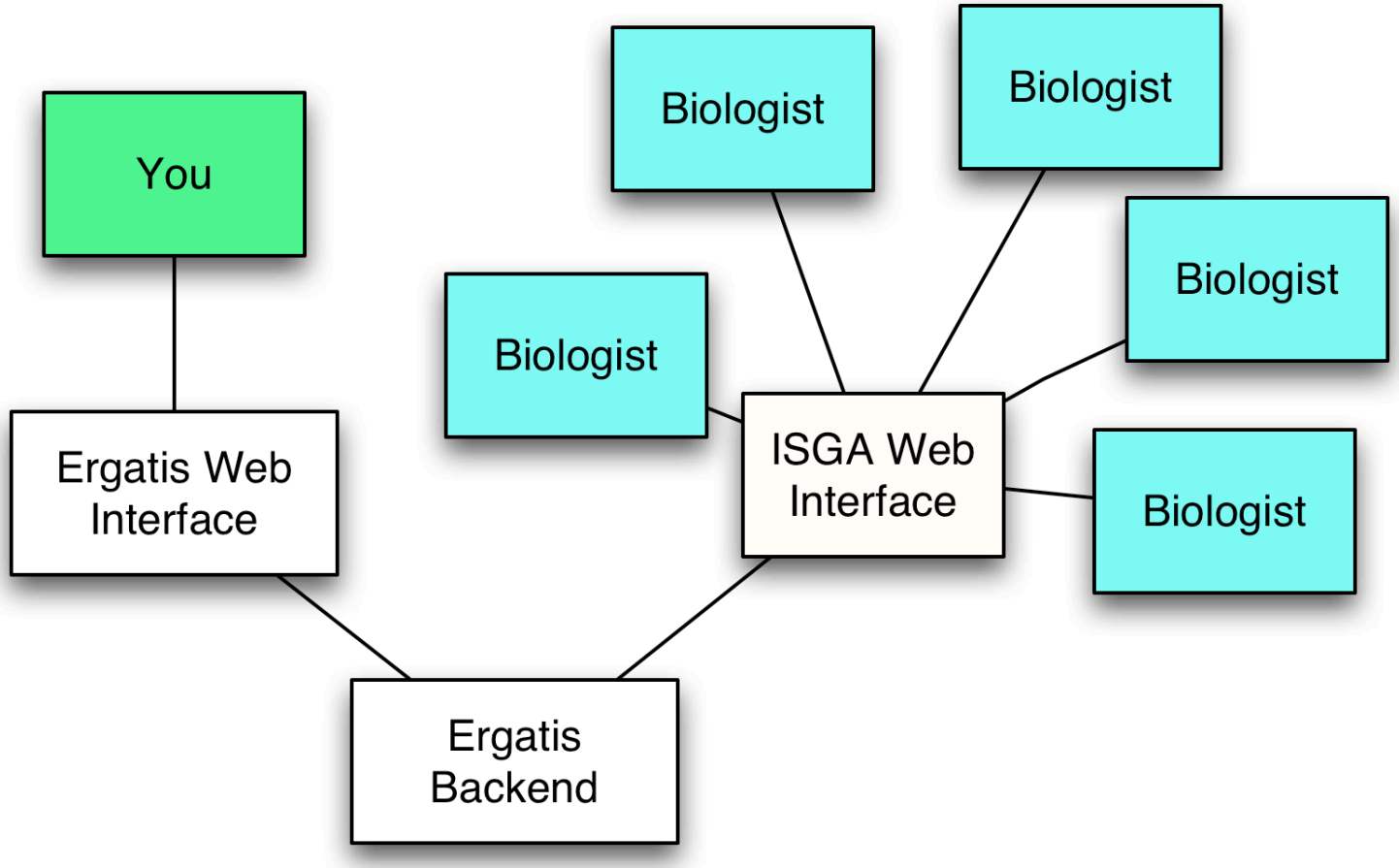
- Insufficient accounts and permissions
- Shared interface for building and customizing pipelines
- Users must submit and retrieve results through filesystem
- Pipeline monitoring interface is slow and complex.
- Information of use to biologists is lost in “noise”
  - High umber of components in a pipeline
  - Complexity of configuration interface

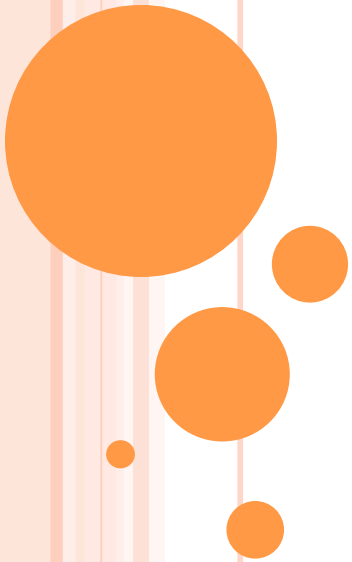

component: glimmer3
configuration
not configured
parameters
If $;ICM$; option does not exist, a new training file will be made from training seqs. If no training seqs are included, glimmer will self train.
training seq
icm
$;OUTPUT_DIRECTORY$;/$;COMPONENT_NAME$;.icm
if self-training (using long-orfs), include long-orfs parameters/options here (Optional)
long orfs opts
-n -t 1.15
If using a coordinates masking file, set this to at least include:
-i
glimmer3 opts
start codon usage
Used in id generation
project abbreviation
$;PROJECT$;
input
input file list
input file
input directory
the following is only used when iterating over an INPUT_DIRECTORY
input extension
fsa
output
output token
default
output directory
$;REPOSITORY_ROOT$;/output_repository/$;COMPONENT_NAME$;/$;PIPELINE
bsml output list
$;OUTPUT_DIRECTORY$;/$;COMPONENT_NAME$;.bsml.list
raw output list
$;OUTPUT_DIRECTORY$;/$;COMPONENT_NAME$;.raw.list
save

OUR SOLUTION

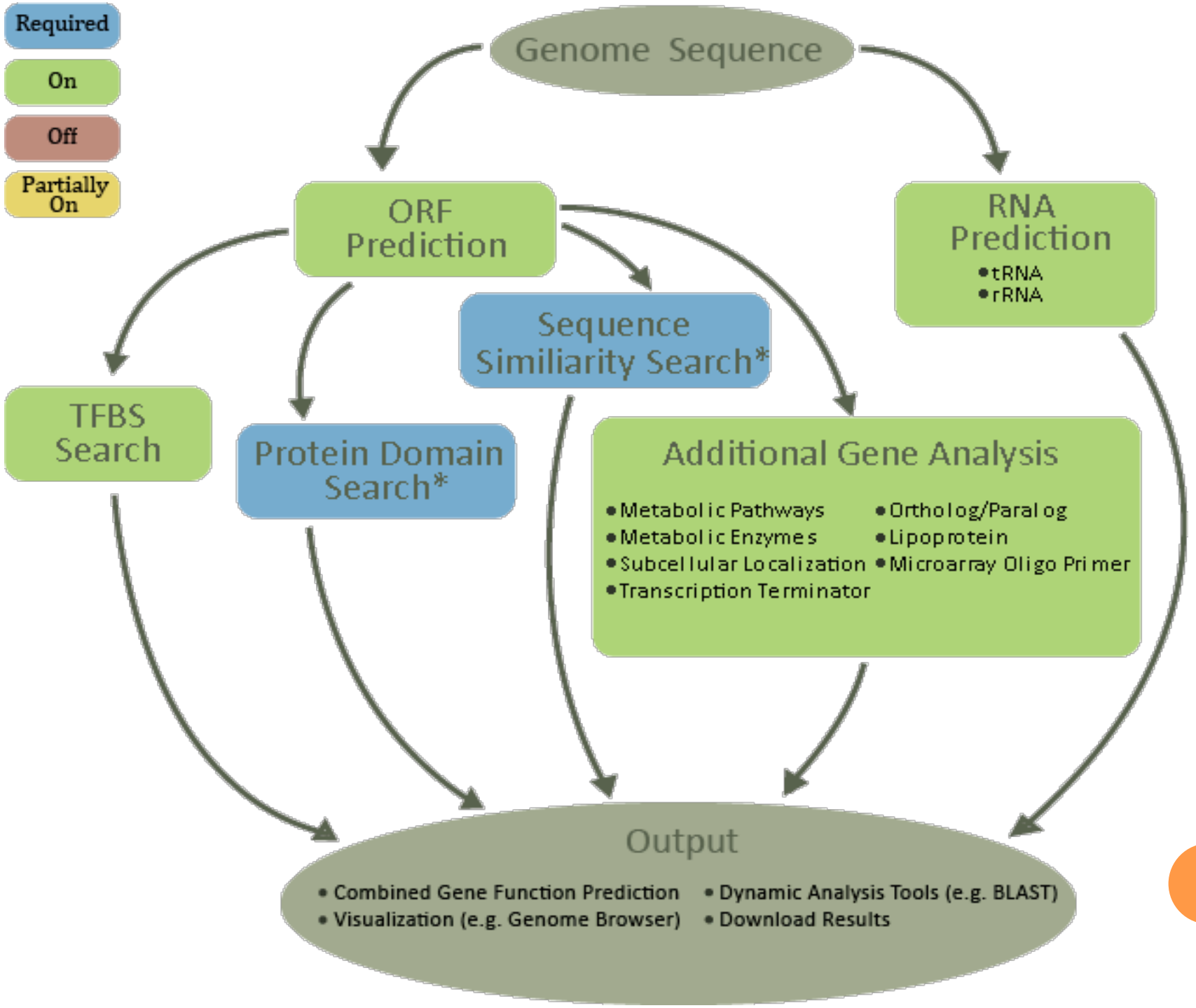
- Develop an alternative interface for biologists that uses the Ergatis backend
  - Administrators also use Ergatis
- New interface features
  - Accounts and permission system
  - File management
  - Simplify pipelines and component management by reducing functionality
  - Provide form validation, documentation and other features to improve usability

THE GOAL


You
Ergatis Web Interface
Ergatis Backend
ISGA Web Interface
Biologist
Biologist
Biologist
Biologist
Biologist
Biologist



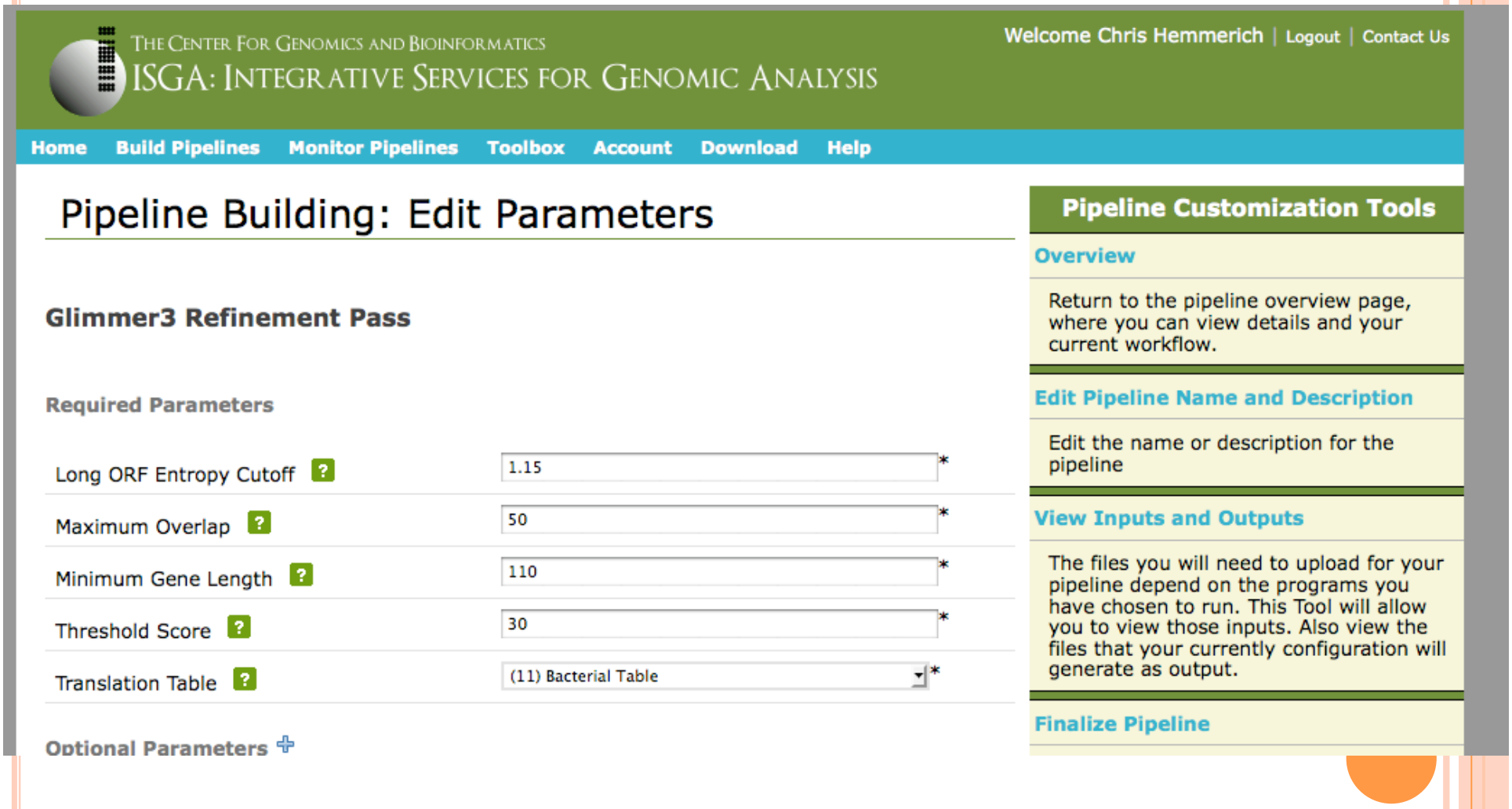
ISGA: WHIRLWIND TOUR


Required
On
Off
Partially On
Genome Sequence
ORF Prediction
RNA Prediction
•tRNA
•rRNA
Sequence Similiarity Search*
TFBS Search
Protein Domain 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 CUSTOMIZATION

- Ability to toggle some clusters on/off.
- Some clusters contain parallel programs that can be independently toggled.
- Ability to edit component parameters
- Ability to save customizations to use with later data sets

PIPELINE BUILDER



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 PIPELINE EXECUTION

- ISGA writes configuration and pipeline definition files to the Ergatis installation
- ISGA then triggers execution through Ergatis and receives the pipeline id in return
- Status is updated directly from Ergatis XML files
- Selected output is copied to ISGA, and the rest is available for download if needed

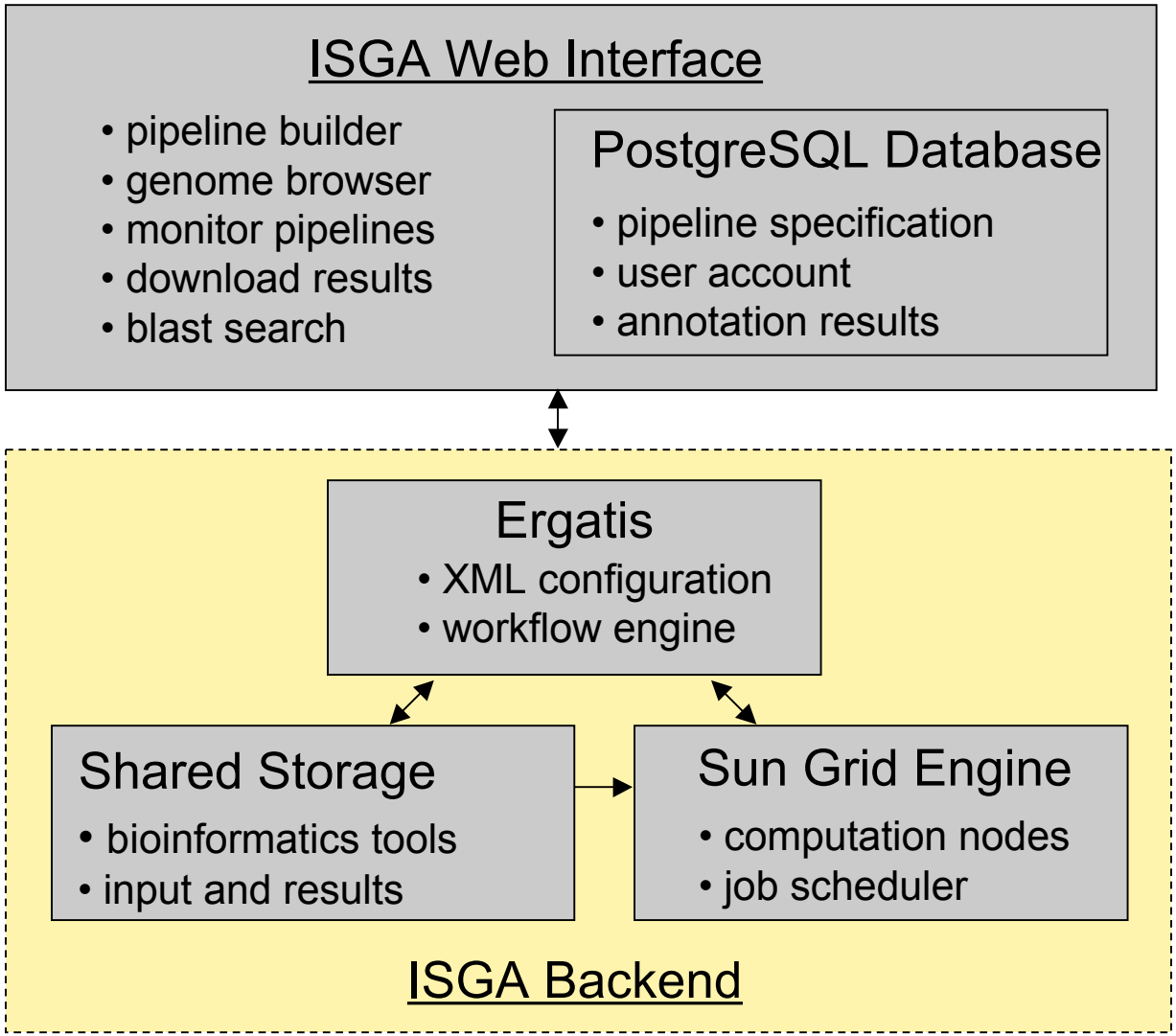
ISGA TOOLBOX

- Includes a GBrowse instance for visualizing annotation results
- BLAST support for pipeline results as query or database
- Text search against annotation results
- Tools can be executed over SGE and monitored

ADMINISTRATIVE TOOLS

- Lightly monitor status in ISGA w/ link to Ergatis page
- Notification when pipeline fails, ISGA will pick up a resumed pipeline
- Ability to redirect ISGA to a cloned Ergatis pipeline or cancel (w/ user notification)
- Disable new job submissions

UNDER THE HOOD

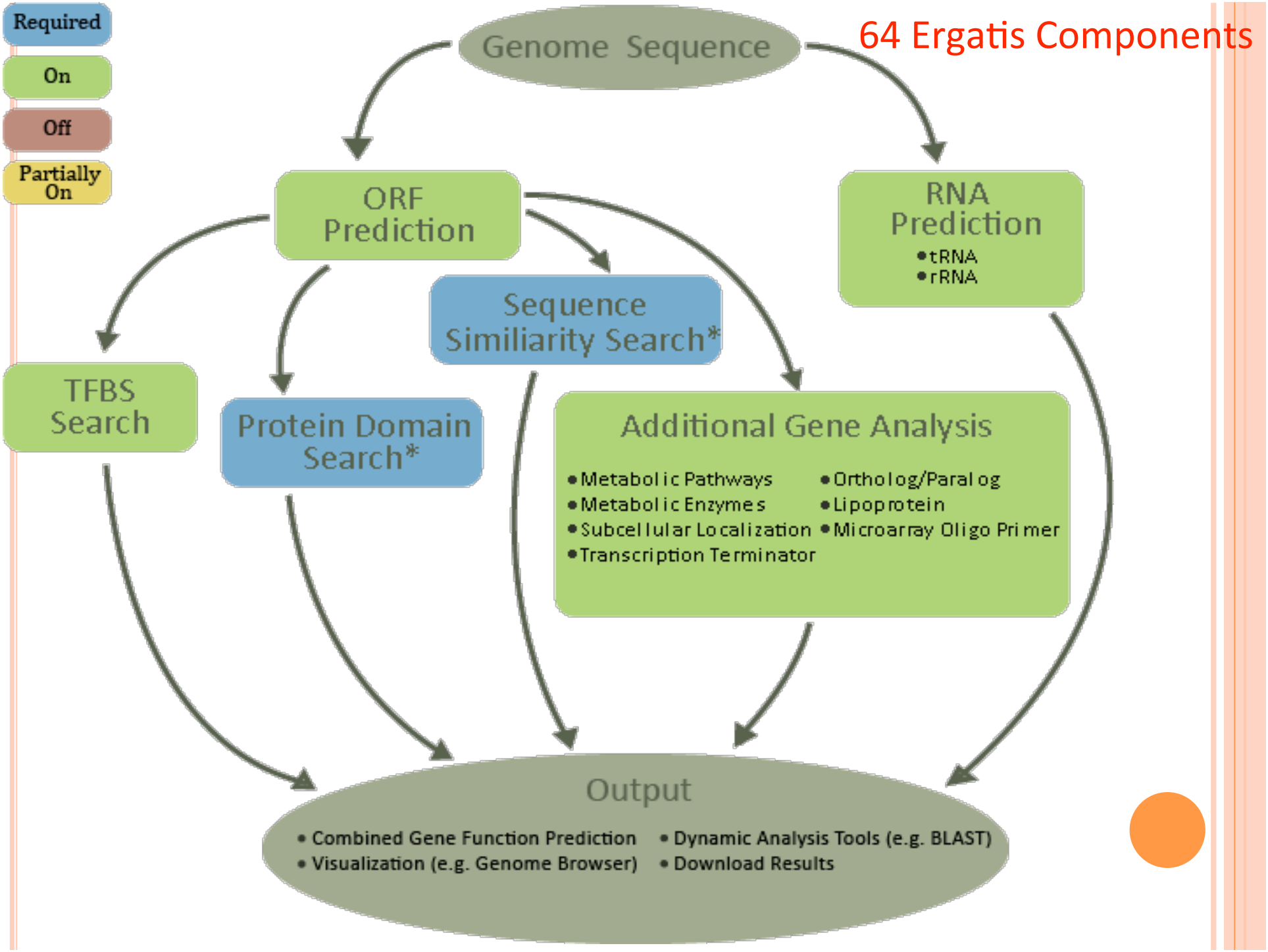


UNDER THE HOOD (CONTINUED)

- Perl & jQuery
- Persistence = PostgreSQL & YAML & XML
- Mason
- MasonX::WebApp
- Hacked up HTML::FormEngine

ADDING AN ERGATIS PIPELINE TO ISGA

64 Ergatis Components



FIRST: UNDERSTAND THE PIPELINE

- ISGA takes a description of an Ergatis pipeline
  - YAML
  - Database Schema
  - Ergatis component .config files
- Document input and output of all components
- Which components are optional?
  - The user can upload previously generated data in their stead?
  - Alternative data from the pipeline can be used?
  - The pipeline is still useful without this functionality

SIMPLIFICATION

- Our microbial annotation pipeline is composed of 64 Ergatis components
  - Impossible to diagram for you on a slide or for a biologist on our web page
- Many of these components are file format conversions, program iterations, database preparation, etc…
  - They are not relevant to a high level view of the pipeline and offer no useful parameters for a biologist to customize

CLUSTERS OF ERGATIS COMPONENTS

- Break the pipeline into biologically meaningful clusters of one or more components
  - This is as much art as science, may depend on your audience
  - Example: 'Alternative Start Site Analysis'

- overlap_analysis.default
- start_site_curation.default
- translate_sequence.translate_new_model
- parse_evidence.hypothetical
- hmmpfam.post_overlap_analysis
- parse_evidence.hmmpfam_post
- wu-blastp.post_overlap_analysis
- bsml2fasta.post_overlap_analysis
- bsml2featurerelationships.post_overlap
- xdformat.post_overlap_analysis
- ber.post_overlap_analysis
- parse_evidence.ber_post
- translate_sequence.final_polypeptides
- bsml2fasta.final_cds

COMPONENT CUSTOMIZATION

- Scripts and XML files are unchanged
- ISGA stores the configuration template for each component
- Components with editable parameters have a YAML definition that is used to build the web form
- These values are incorporated into the configuration template

COMPONENT TEMPLATE

--- !perl/ISGA::ComponentBuilder
Name: RNAmmer
Description: 'RNAmmerpredicts 5s/8s, 16s/18s, and …'
Params:
- { templ: 'select', NAME: 'molecules', TITLE: 'rRNA Molecules', REQUIRED: 1, OPTION: ['ssu (5/8s rRNA)', 'lsu (16 /18s rRNA)', 'tsu (23/28s rRNA)', 'ssu and lsu', …], OPT_VAL: ['ssu' , 'lsu', 'tsu', 'ssu,lsu', …], VALUE: 'ssu,lsu,tsu', DESCRIPTION: 'Declare what rRNA molecule types to search for.', CONFIGLINE: '___molecule___' }

RunBuilderParams:
- { templ: 'hidden', NAME: 'project_id_root', TITLE: 'Project Id Root', REQUIRED: 1, DESCRIPTION: 'The Id root used in bsml id generation', CONFIGLINE: '___project_id_root___' }

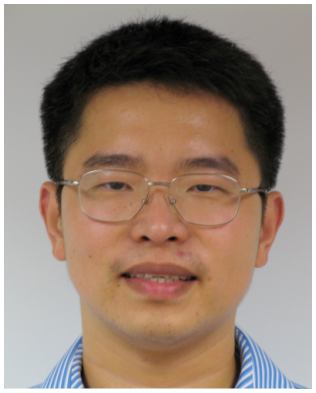
FUTURE ISGA WORK

- Incorporate additional pipelines
  - Small prokaryotic assembly pipeline
  - Comparative genomics
  - Functional genomics
- Add additional features
  - Make pipelines modular components of ISGA
  - Implement pipeline versioning
  - Pipeline and data sharing
- Ergatis Cloud Support?

ISGA







Aaron Buechlein

Chris Hemmerich

Kashi Revanna

Ram Podicheti

Qunfeng Dong