Galaxy / CEGR: Pugh/Mahony Labs, Center for Eukaryotic Gene Regulation

Greg consulted with Dr. Frank Pugh and Dr. Shaun Mahony in the Center for Eukaryotic Gene Regulation to build the Galaxy / CEGR instance (see Figure 1). This environment is hosted within a virtual machine (VM) on ICDS-ACI and has access to cluster nodes for job execution. The environment enables researchers in the Center for Eukaryotic Gene Regulation (CEGR) to perform ChIP-exo analyses on very large datasets produced by the Center’s sequencer.

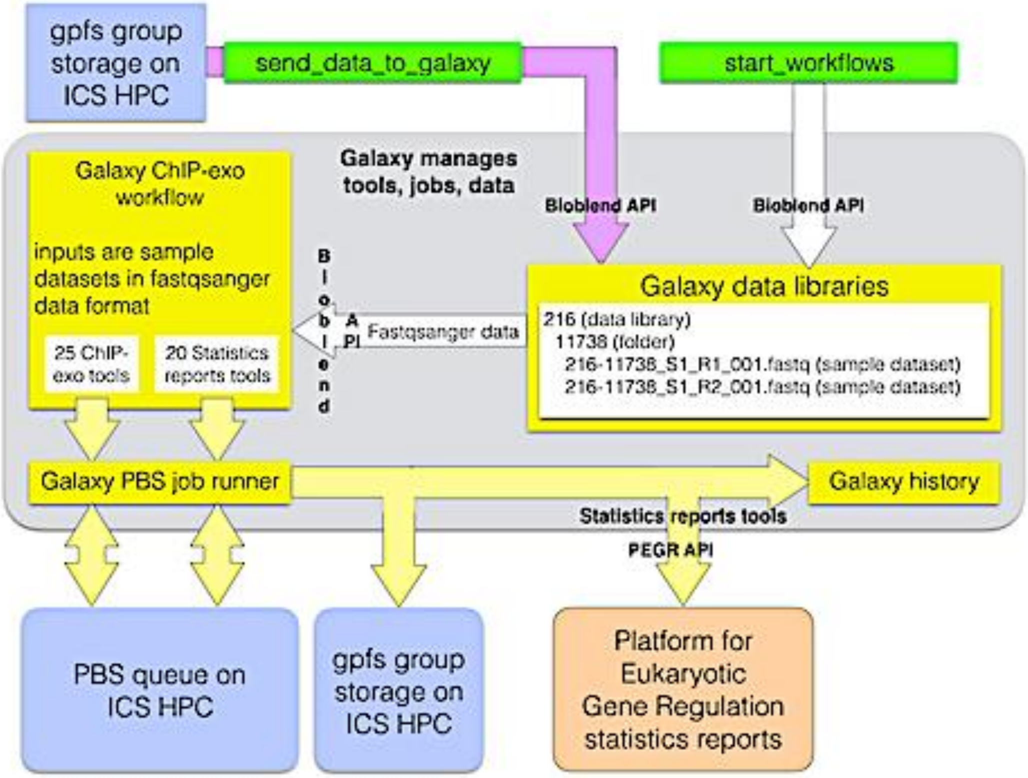

gpfs group storage on ICS HPC
send_data_to_galaxy
start_workflows
Galaxy manages tools, jobs, data
Bioblend API
Bioblend API
Galaxy ChIP-exo workflow
inputs are sample datasets in fastqsanger data format
25 ChIP-exo tools
20 Statistics reports tools
Bioblend
API
Fastqsanger data
Galaxy data libraries
216 (data library)
11738 (folder)
216-11738_S1_R1_001.fastq (sample dataset)
216-11738_S1_R2_001.fastq (sample dataset)
Galaxy PBS job runner
Galaxy history
Statistics reports tools
PEGR API
PBS queue on ICS HPC
gpfs group storage on ICS HPC
Platform for Eukaryotic Gene Regulation statistics reports

Figure 1: Galaxy / CEGR Process Flow

At the time that it was built, the environment consisted of four primary components, an Illumina NextSeq 500 sequencer owned by the Pugh/Mahony labs, the CEGR pre-processing pipeline, the Galaxy ChIP-exo environment and the Platform for Eukaryotic Gene Regulation (PEGR), a web application developed by the labs.

The sequencer produces raw datasets that are converted into the fastqsanger data format and imported into Galaxy data libraries (Galaxy's hierarchical container for datasets). The Galaxy ChIP-exo workflow is automatically executed for each set of samples, with each tool in the workflow submitting jobs to the HPC cluster nodes, producing datasets that are then used as inputs to the next tool in the workflow chain. The Galaxy ChIP-exo workflow includes tools that generate metadata (fine-grained statistics) about datasets produced by specified tools within the workflow. These statistics are sent to PEGR for review.

The CEGR pre-processing pipeline consists of four custom programs, developed to automate the processes used by the labs, ultimately preparing the data for analysis within the Galaxy ChIP-exo instance (the green boxes in Figure 2 below depict these custom programs). Each of these programs includes quality assurance components that automatically halt processing if errors occur, logging the details for review and correction. Each program can be executed independently (assuming that the previous program in the pipeline has completed successfully) allowing for a certain step to be re-executed after corrections are made.

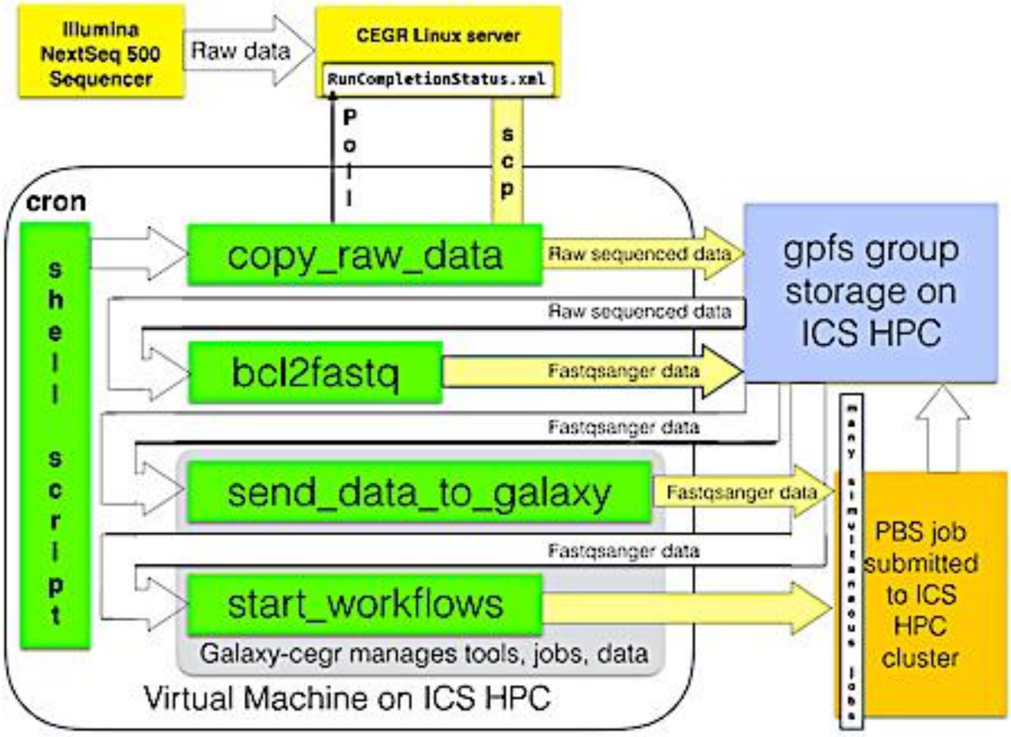


Figure 2: Galaxy / CEGR Pre-processing Pipeline

- **copy_raw_data** - Polls the lab's server that contains the raw datasets produced by the sequencer to determine when the sequencing run is complete. The raw datasets are then copied from the server to the Science Gateway's file store.
- **bcl2fastq** – Converts the raw datasets into the fastqsanger data format required by the initial tools within the Galaxy ChIP-exo workflow.
- **send_data_to_galaxy** – Creates a Galaxy data library for the run's sample datasets. The sample datasets are imported into appropriate folders from which they can be retrieved for analysis. The program submits a PBS job to the ICS HPC cluster to import each sample dataset, storing them on the Science Gateway's file store.
- **start_workflows** – Retrieves sample datasets from the run's Galaxy data library folders and provides them as input to the Galaxy ChIP-exo workflow. Each tool in the workflow performs its function by submitting a PBS job to the HPC cluster, storing the resulting datasets on the Science Gateway's file store.

Galaxy includes a feature rich REST API which is used by the pre-processing pipeline for all direct interaction with Galaxy.

The Galaxy environment is configured with 8 front-end web server processes. It contains the ChIP-exo workflows for both single and paired reads. These workflows consist of the chain of tools that perform the ChIP-exo analysis and send the statistics to PEGR (see Figure 3).

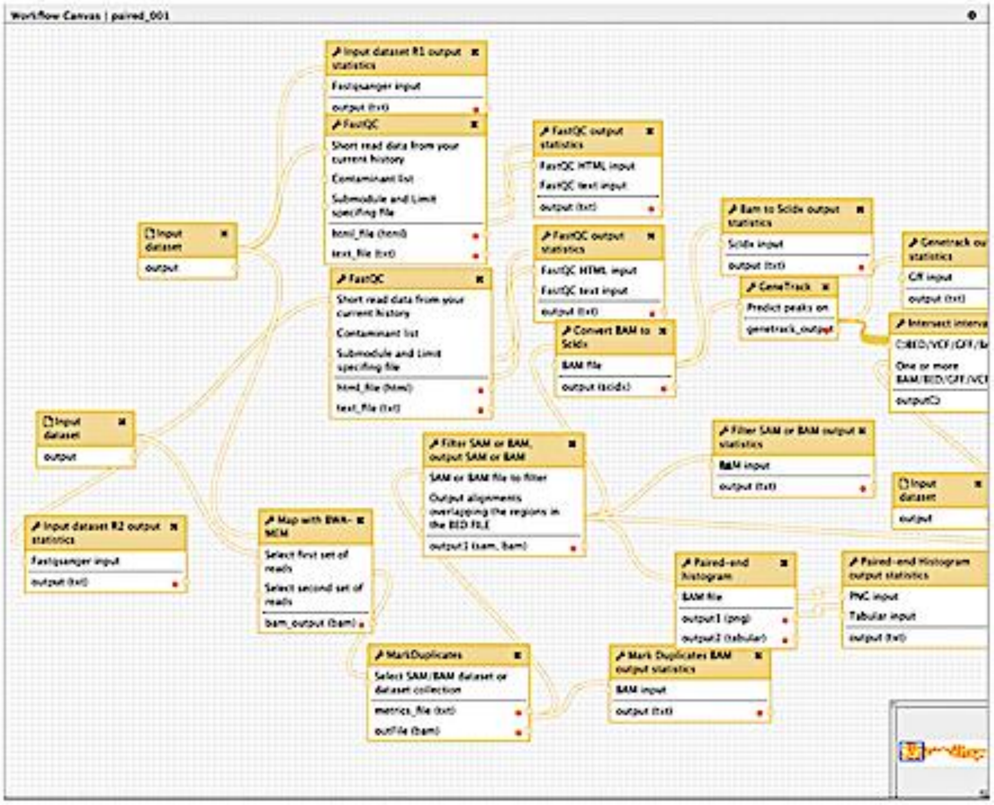


Figure 3: Galaxy / CEGR ChIP-exo Workflow

The Galaxy environment is also configured with 8 job handler processes that submit PBS jobs to the ICS HPC cluster nodes, producing datasets that are stored on the Science Gateway's file store. This configuration allows for load balancing on the web front-end and the many simultaneous PBS jobs submitted to the HPC cluster.

The Platform for Eukaryotic Gene Regulation (PEGR) is a web-based sample tracking application developed by the Pugh/Mahony labs. Lab technicians enter sample information into PEGR that is used by the CEGR pre-processing pipeline to prepare the data for analysis within Galaxy. PEGR interacts with Galaxy via the Galaxy REST API.

Galaxy / CoralSNP: Baums lab, Department of Biology

Greg consulted with Dr. Iliana Baums in the Department of Biology to develop the Galaxy / CoralSNP analysis environment. Reliable and standardized identification of genotypes is a critical need for basic research and restoration planning, especially in plants and animals that reproduce asexually and form large clonal populations like aspens, seagrasses, waterfleas and corals. Figure 4 provides the general process.

- The scientist collects the coral, extracts the DNA and submits it to the processing facility
- The raw sequenced samples and sample metadata are uploaded to the Galaxy / CoralSNP environment for analysis
- The analysis produces new sample MLGs and genotype information that can be accessed via the Galaxy / CoralSNP interface

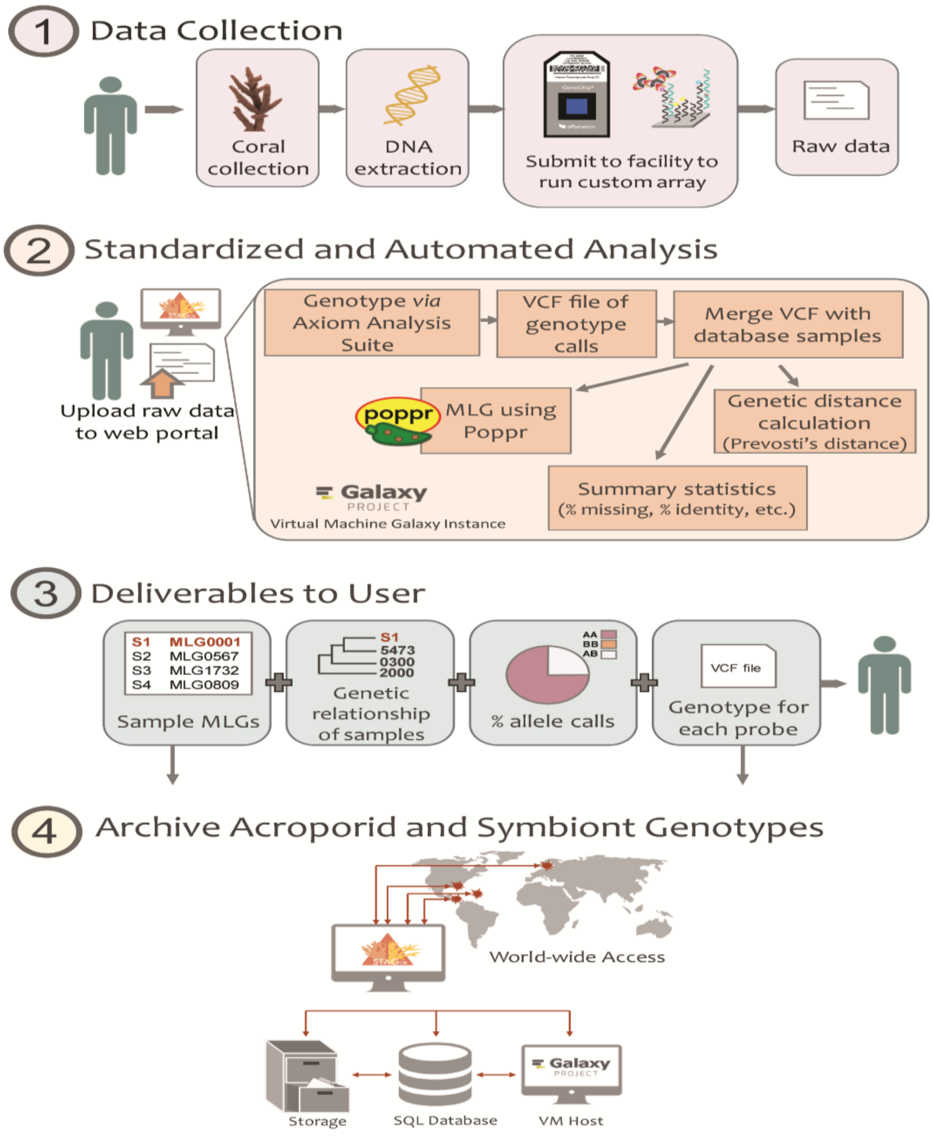

1
Data Collection
Coral collection
DNA extraction
Submit to facility to run custom array
Raw data
2
Standardized and Automated Analysis
Upload raw data to web portal
Genotype *via* Axiom Analysis Suite
VCF file of genotype calls
Merge VCF with database samples
poppr
MLG using Poppr
Genetic distance calculation (Prevosti's distance)
Galaxy PROJECT
Virtual Machine Galaxy Instance
Summary statistics (% missing, % identity, etc.)
3
Deliverables to User
S1 MLG0001
S2 MLG0567
S3 MLG1732
S4 MLG0809
Sample MLGs
S1
5473
0300
2000
Genetic relationship of samples
AA
BB
AB
% allele calls
VCF file
Genotype for each probe
4
Archive Acroporid and Symbiont Genotypes
World-wide Access
Storage
SQL Database
Galaxy PROJECT
VM Host

Figure 4: General process for standardized identification of genotypes

Galaxy / CoralSNP provides a high-resolution hybridization-based genotype array coupled with a standardized analysis workflow and database for the most speciose genus of coral, Acropora, and their symbionts (see figure 5).

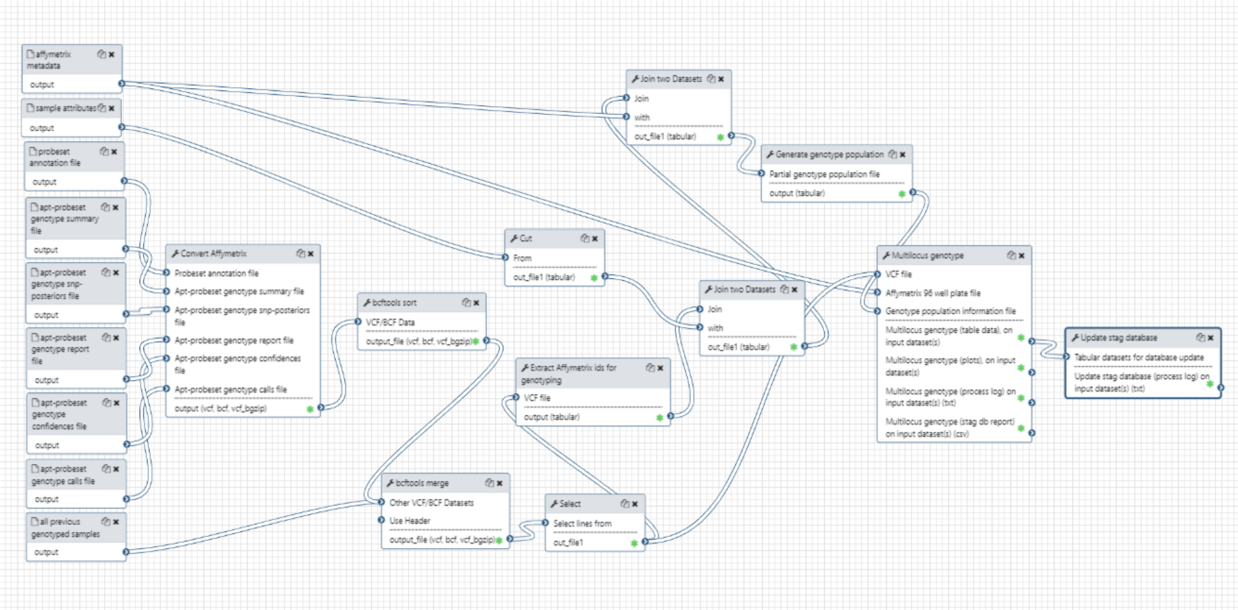
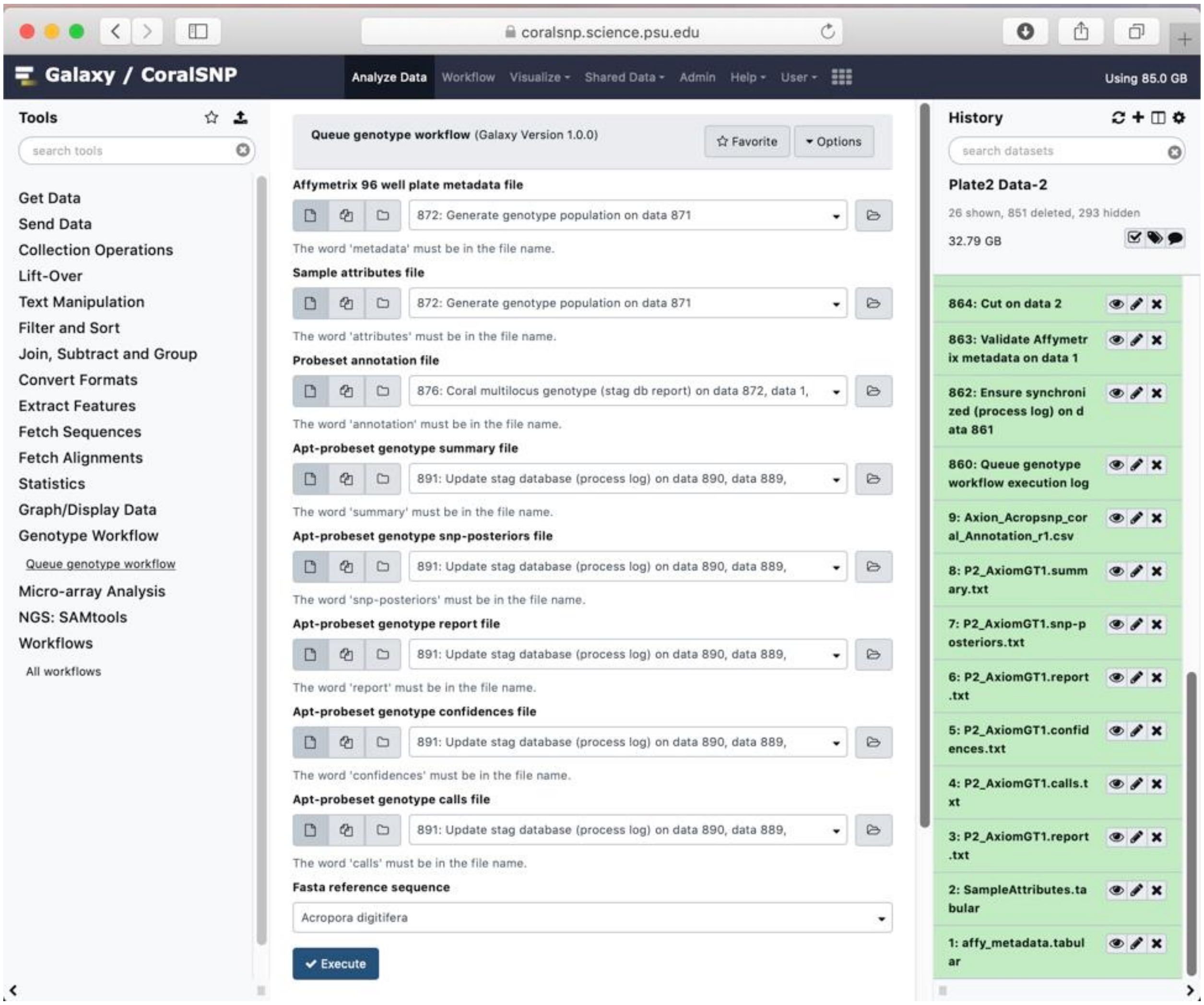

affymetrix metadata
output
sample attributes
output
probeset annotation file
output
apt-probeset genotype summary file
output
apt-probeset genotype snp-posteriors file
output
apt-probeset genotype report file
output
apt-probeset genotype confidences file
output
apt-probeset genotype calls file
output
all previous genotyped samples
output
Convert Affymetrix
Probeset annotation file
Apt-probeset genotype summary file
Apt-probeset genotype snp-posteriors file
Apt-probeset genotype report file
Apt-probeset genotype confidences file
Apt-probeset genotype calls file
output (vcf, bcf, vcf_bgzip)
bcftools sort
VCF/BCF Data
output_file (vcf, bcf, vcf_bgzip)
bcftools merge
Other VCF/BCF Datasets
Use Header
output_file (vcf, bcf, vcf_bgzip)
Cut
From
out_file1 (tabular)
Extract Affymetrix ids for genotyping
VCF file
output (tabular)
Select
Select lines from
out_file1
Join two Datasets
Join
with
out_file1 (tabular)
Generate genotype population
Partial genotype population file
output (tabular)
Join two Datasets
Join
with
out_file1 (tabular)
Multilocus genotype
VCF file
Affymetrix 96 well plate file
Genotype population information file
Multilocus genotype (table data), on input dataset(s)
Multilocus genotype (plots), on input dataset(s)
Multilocus genotype (process log) on input dataset(s) (txt)
Multilocus genotype (stag db report) on input dataset(s) (csv)
Update stag database
Tabular datasets for database update
Update stag database (process log) on input dataset(s) (txt)

Figure 5: Galaxy / CoralSNP Workflow

The process is simple. A sample metadata file is created by the user from a template and uploaded along with their raw Affymetrix data files into the Galaxy CoralSNP environment. The appropriate files are selected as inputs to the Queue Genotype Workflow tool (Figure 6) which validates the metadata, executes the CoralSNP workflow (Figure 5) and updates a dataset that contains all previously genotyped samples as well as the STAG database with the samples in the current run. From the user's perspective, the entire analysis is as simple as uploading data and specifying it as the input to execute a tool.



Queue genotype workflow (Galaxy Version 1.0.0)

Affymetrix 96 well plate metadata file
872: Generate genotype population on data 871
The word 'metadata' must be in the file name.

Sample attributes file
872: Generate genotype population on data 871
The word 'attributes' must be in the file name.

Probeset annotation file
876: Coral multilocus genotype (stag db report) on data 872, data 1,
The word 'annotation' must be in the file name.

Apt-probeset genotype summary file
891: Update stag database (process log) on data 890, data 889,
The word 'summary' must be in the file name.

Apt-probeset genotype snp-posteriors file
891: Update stag database (process log) on data 890, data 889,
The word 'snp-posteriors' must be in the file name.

Apt-probeset genotype report file
891: Update stag database (process log) on data 890, data 889,
The word 'report' must be in the file name.

Apt-probeset genotype confidences file
891: Update stag database (process log) on data 890, data 889,
The word 'confidences' must be in the file name.

Apt-probeset genotype calls file
891: Update stag database (process log) on data 890, data 889,
The word 'calls' must be in the file name.

Fasta reference sequence
Acropora digitifera

Plate2 Data-2
26 shown, 851 deleted, 293 hidden
32.79 GB

864: Cut on data 2
863: Validate Affymetrix metadata on data 1
862: Ensure synchronized (process log) on data 861
860: Queue genotype workflow execution log
9: Axion_Acropsnp_coral_Annotation_r1.csv
8: P2_AxiomGT1.summary.txt
7: P2_AxiomGT1.snp-posteriors.txt
6: P2_AxiomGT1.report.txt
5: P2_AxiomGT1.confidences.txt
4: P2_AxiomGT1.calls.txt
3: P2_AxiomGT1.report.txt
2: SampleAttributes.tabular
1: affy_metadata.tabular

Figure 6: The Galaxy / CoralSNP Queue Genotype Workflow Tool

The Queue Genotype Workflow tool shields the complexity of the analysis from the user, and performs its functions via the Galaxy REST API to produce multi-locus genotypes for Caribbean or Pacific acroporids. The SNP array combined with the analysis allows for reliable, standardized identification of host genet and symbiont strains and serves as a template for the development of arrays for additional coral genera. This data can be used for downstream genomic analyses as well as restoration planning.