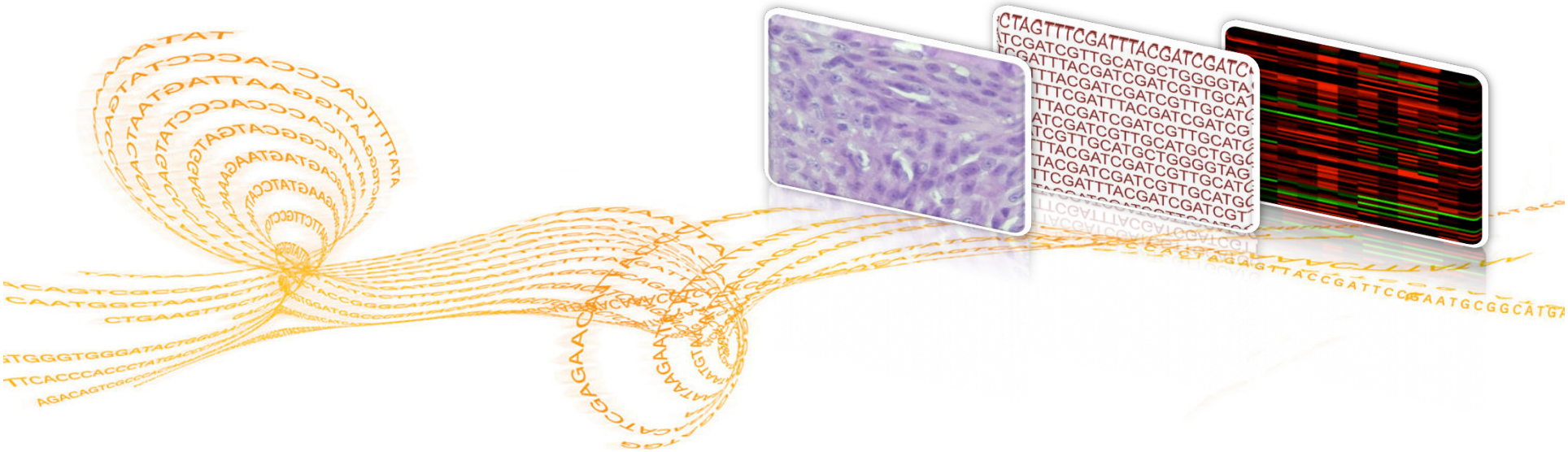
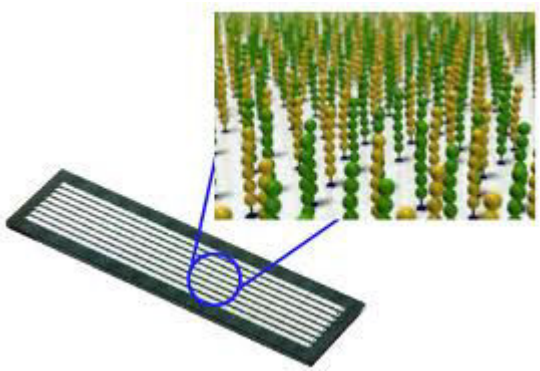


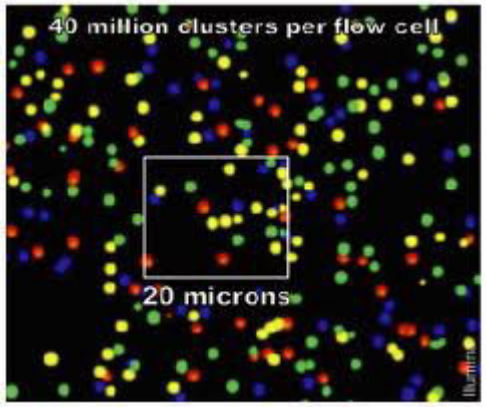
Galaxy for High Throughput Sequencing

illumina®

Agenda: Galaxy for High Throughput Sequencing

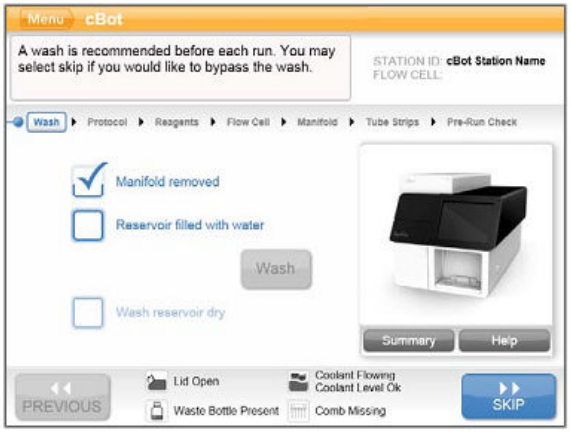
- Introduction: Speaker – Kirt Haden, Illumina, Software Engineering
- Key Requirements
- Why Galaxy?
- Significant Challenges
- System Architecture
- Ideas and Solutions
- Experiences with Galaxy
- Illumina's Vision
- Conclusion





Key Requirements

- Analysis of up to 100TB of sequencing data per month
- Ease of use – no informatics background required
- Parallel processing that results in equal or better performance to existing workflows
- Reproducibility of analysis across institutions – ability to share workflows
- Customization and easy integration of workflows
- Automation of workflows – offline and online



Why Galaxy?

- Ease of use and distribution
- Tool integration
- Availability of sequencing workflows
- Data and workflow sharing
- Tracking history
- DRMAA integration
- API support
  - Ability to automate repetitive tasks
- Ability to modify the presentation layer
  - Customization for specialized roles and reduce complexity
- Cloud implementation
- Community supported

Galaxy

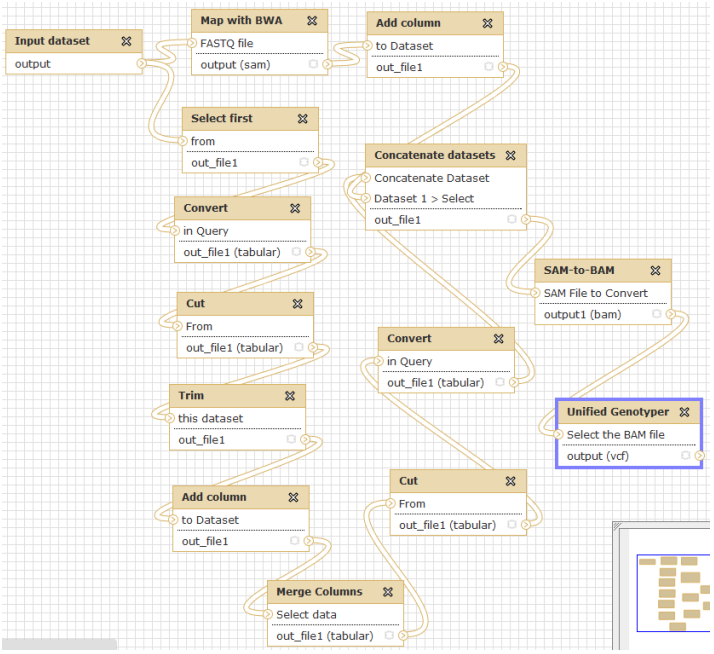
Saved Histories

search history names and tags

Advanced Search

Name	Datasets		Tags	Sharing	Created	Last Updated ↑
Unnamed history			0 Tags		Apr 20, 2011	Apr 20, 2011
Unnamed history	3		0 Tags		Apr 12, 2011	Apr 12, 2011
BCL to FASTQ	3		0 Tags		Apr 06, 2011	Apr 06, 2011
Testing BWA	31	1	0 Tags		Mar 21, 2011	Mar 28, 2011
Test Users	3		0 Tags		Mar 23, 2011	Mar 23, 2011
Unnamed history	32		0 Tags		Mar 21, 2011	Mar 22, 2011
Testing BWA2	1		0 Tags		Mar 21, 2011	Mar 21, 2011
References	3		0 Tags		Mar 21, 2011	Mar 21, 2011

For 0 selected histories: Rename Delete Undelete

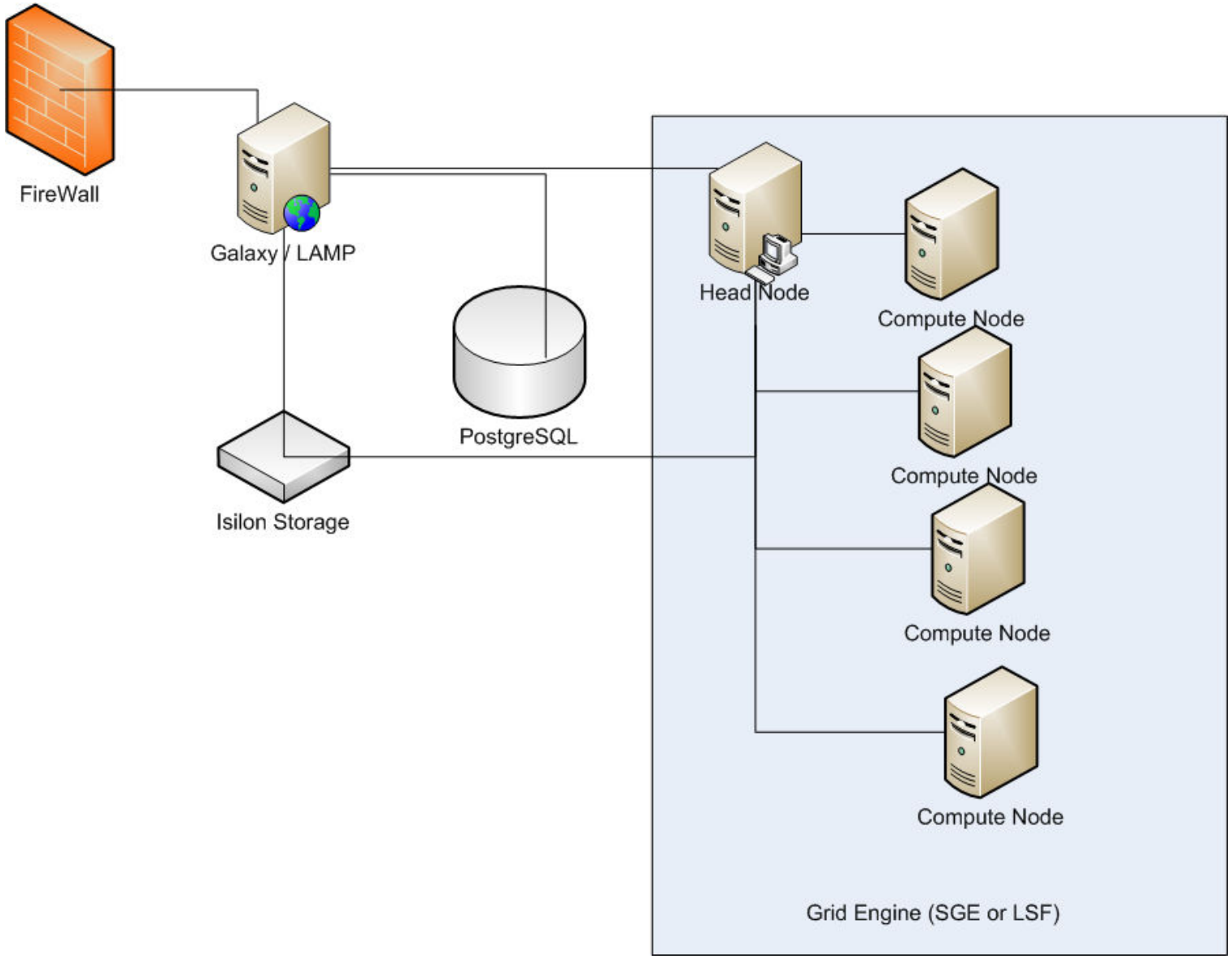


illumina

Significant Challenges

- Big Files
  - Dilemma of linking files and keeping control or uploading and passing control to Galaxy
  - Scanning entire file, computing meta data is useful but costly for compressed files
  - Splitting vs. indexing into large files
- Parallelization within a file
  - No easy solution for handling a large input set without creating many small files
- Temp files consume vast amounts of storage space, disk space monitoring
- Handling tools that produce a variable number of files
  - Scatter - gather/ Collections / Demux model / Variable number of files
- File system caching (file system as communication mechanism)
  - Temp directory may not be there when the job runs (NTFS V3 vs V4 issue?)
  - OS write level caching means that output files may not be completely written before they are used by the next process
- Too much manual intervention required to handle transient conditions
- Organization of results
- Complex interface for multiple roles

System Architecture to Support Scalability



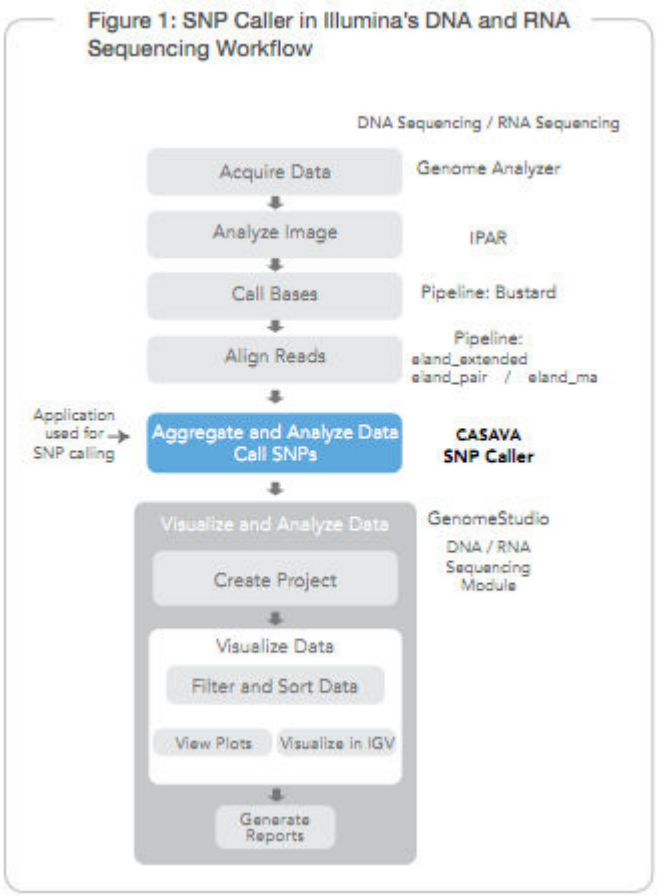
illumina

Ideas and Solutions

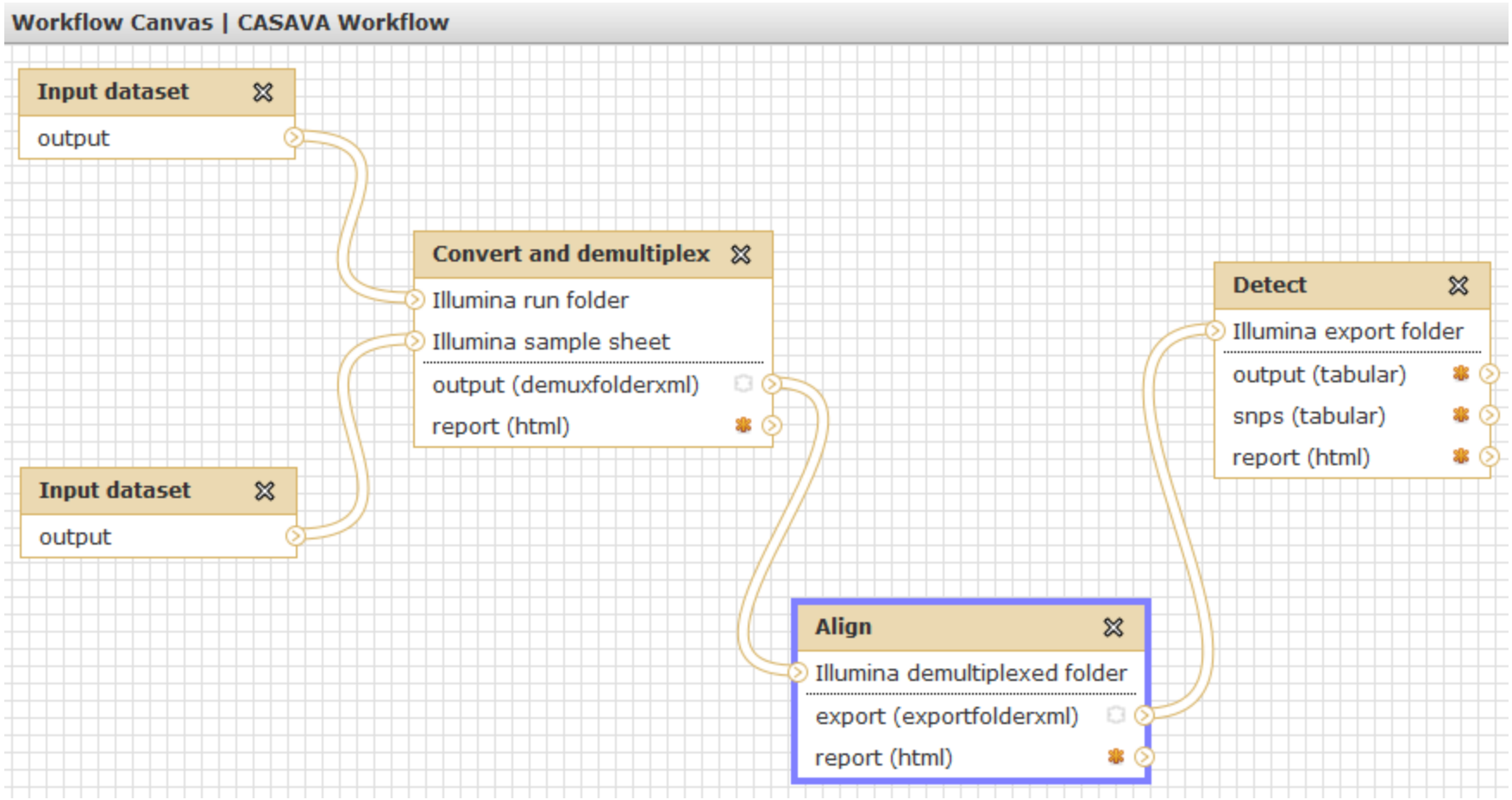
- Within module parallelization may be a useful strategy
- Splitting of FASTQ files (ex. ELAND)
  - Changing tools to operate on a subsets of a file
  - Allow a set (collection) of files as a primitive
  - Parallelization type per input port and use BWA splitting scheme
  - Indexing the gzipped files
- Automatic retry for failed operations to recover from transient events
  - This is useful for write caching issue/ transient events
  - How do you clean up from a failed operation?
- Enhance separation between user interface and function with API to support alternate presentation layers and more automation
  - Allows independent development without a huge investment to support multiple different users requirements
- ???

Experiences with Galaxy

- Custom simplified GUI
- Workflows/tools
  - DNA Seq, RNA Seq, Methylation, Chip Seq, GT (microarrays)
- Utilities
  - Visualization, BEDtools, VCFTools, Broad GATK, Google charts
- Submissions to Galaxy code base
  - Broad IGV
  - API changes
  - Parallelization
- CASAVA in Galaxy
  - With and without make, qmake



CASAVA workflow in Galaxy



Galaxy - Mozilla Firefox

File Edit View History Bookmarks Tools Help http://ussd-dev-swim01.illumina.com:8085/workflow/editor?id=f2db41e1fa331b3e

Galaxy

Galaxy | Analyze Data | Workflow | Shared Data | Help | User

Tools Options

search tools

- CASAVA 1.8
- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NCBI BLAST+
- NGS: QC and manipulation
- NGS: Mapping
- NGS: Indel Analysis
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: Peak Calling

Workflow Canvas | PhiX workflow Options

Input dataset
- output

Eland standalone
- Fastq file
- export_file (eland)
- genome_size_file (genomesize)

ELAND export to BAM version 2
- ELAND file
- GenomeSize file
- output (bam)

Casava 1.8 snp and indel Caller, using Starling
- per chr bam file
- GenomeSize file
- snp_file (Vcf)
- indel_file (Vcf)

Details

Tool: Eland standalone

Fastq file
Data input 'fastq_file' (fastq)

Reference genome:
PhiX

Edit Step Actions

Rename Dataset
export_file
Create

Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

This tool runs eland alignment workflow.

illumina

Mock-up of a Simplified and Stylized GUI

Sequencing Workflow with IHop

Logout

Galaxy

List of Your Runs

Sample Name	Workflow	Status
FASTQ Files for DNA Resequencing	BWA and GATK	Run
FASTQ Files for mRNA Sequencing	Top Hat and Cufflinks	Run
Variant Call Files	VCF Tools	Run
Methylation Samples	Methyl Seq	Run
Aligned BAM Files	View with IGV	Run
FASTQ Files for Small RNA	Flicker	Run
Sample XYZ Read1 | 2011-04-25 19:31		ok
Sample XYZ read 2 | 2011-04-22 23:31		ok
Sample XYZ Read1 | 2011-04-22 23:31		ok
https://s3.amazonaws.com/IlluminaDataForSwim/Data/Galaxy2-%5BSRR002888_1.filt.fastq%5D.fastqsanger | 2011-04-22 23:19		error
Run Workflow - 2011-04-21 18:18		ok
Run Workflowubuntu@ip-10-78-190-226:/mnt/galaxyTools/galaxy-central/scripts/api$ - 2011-04-21 18:20		ok
Run Workflow - 2011-04-21 19:29		ok

Refresh Table

Powered by Galaxy

HiSeq 2000 HiSeq 1000 Genome Analyzer$_{IIx}$ HiScanSQ MiSeq

A sequencer for every need.

Every budget. Every lab.

illumina®

Illumina's Vision

- Set of recommended workflows
  - Used for common sequencing applications
  - Highly optimized for performance
- Promote availability and easy integration of third party tools
- Ability to process locally or in the Cloud (location agnostic)
- Modular workflows with reduced coupling between components – plug and play
- Data playground – sample data sets and performance numbers
- Allow Galaxy users to create their own end-to-end analysis workflows with the CASAVA tool set
- Help our customers get the most out of Galaxy
- Support the open source community

Conclusion

- Galaxy is an attractive workflow engine candidate
  - engineers tend to focus on risk
- A large number of useful workflows already exist and new ones are rapidly being added
- We have found that adding new workflows is straightforward
- Our usage of CASAVA in Galaxy demonstrates the feasibility of running very large data sets efficiently
- Key challenges to relying on Galaxy for our secondary analysis still exist and will need to be resolved in the short term
- We see great potential in the tool and look forward to working with the Galaxy community to create:
  - Modular workflows
  - Efficient analysis in the Cloud

Contributors

- Galaxy Development team
- Galaxy Community
- Illumina
  - Bioinformatics
    - Semyon Kruglyak
    - Jean Lozach
    - Eric Allen
    - Tobias Wohlfrom
  - Services
    - Brad Sickler
  - Software
    - Francisco Garcia
    - Steve Burgett
    - Mauricio Varea
    - Come Racy
    - John Duddy
  - Marketing
    - Jordan Stockton
    - Dipesh Risal
  - Project Management
    - Scott Kirk