High level distributed processing pipelines with Galaxy

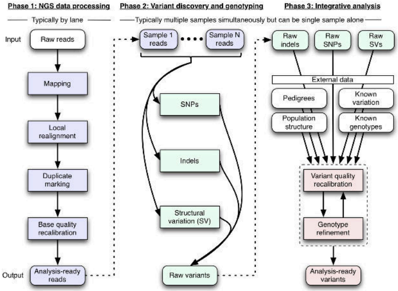
Brad Chapman, Bioinformatics Core at Harvard School of Public Health

Galaxy Community Conference, 27 July 2012

Complex pipelines

Goal: simple interface to complex analyses

- Steps, lots of them
- Intermediate files
- Branching logic: if/else
- Idempotent
- Transactional
- Parallel: by record, by region
- Associated resources
- Experimental metadata



Phase 1: NGS data processing

Typically by lane

Phase 2: Variant discovery and genotyping

Typically multiple samples simultaneously but can be single sample alone

Phase 3: Integrative analysis

Input

Raw reads

Mapping

Local realignment

Duplicate marking

Base quality recalibration

Output

Analysis-ready reads

Sample 1 reads

Sample N reads

SNPs

Indels

Structural variation (SV)

Raw variants

Raw indels

Raw SNPs

Raw SVs

External data

Pedigrees

Known variation

Population structure

Known genotypes

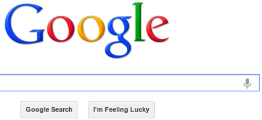
Variant quality recalibration

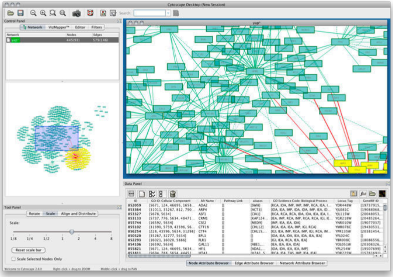
Genotype refinement

Analysis-ready variants

http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v3

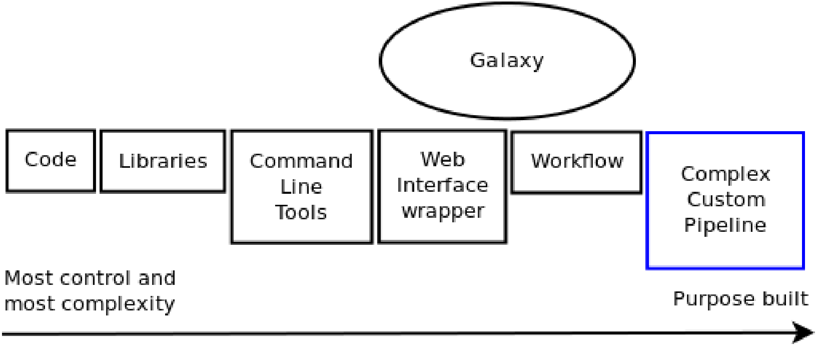
(At least) two types of users





Network	Nodes	Edges
yap*	445(93)	579(146)

ID	GO ID: Cellular Component	Alt Name	Pathway Link	aliases	GO Evidence Code: Biological Process	Locus Tag	GeneRIF ID
852059	[5671, 124, 46695, 1658...	ADA2	[]	[SWI8]	[RCA, IDA, IMP, IMP, IMP, RCA, IEA, I...	YDR448W	[19737915,...
853364	[31011, 35267, 812, 790...	ARP4	[]	[ACT3]	[IDA, IEA, IMP, IDA, IDA, IMP, IEA, ID...	YJL081C	[19088068,...
853327	[5678, 5634]	ASF1	[]	[CIA1]	[RCA, RCA, RCA, IDA, IDA, IEA, IEA, I...	YJL115W	[20048053,...
853133	[5737, 776, 5634, 48471...	CRM1	[]	[KAP124...	[IEA, IMP, RCA, IMP, RCA, RCA, IGI, IE...	YGR218W	[20485264,...
855744	[16592, 5634]	CSE2	[]	[MED9]	[IMP, IEA, IEA, IDA]	YNR010W	[19077037]
855102	[31390, 5739, 43596, 56...	CTF18	[]	[CHL12]	[RCA, IEA, IEA, IMP, IGI, RCA]	YMR078C	[19430531,...
856254	[228, 43596, 5634, 31298]	CTF4	[]	[CHL15...	[IGI, IEA, IMP, RCA, IMP, IGI, RCA, IM...	YPR135W	[20381454,...
850520	[35267, 32777, 5634]	EPL1	[]	[]	[IDA, IEA, IDA, IEA, IDA, IEA, IEA]	YFL024C	[]
852293	[16021, 16020, 5886]	FLR1	[]	[]	[IGI, IEA, RCA, IEA, IEA]	YBR008C	[18086556,...
854106	[16592, 5634]	GAL11	[]	[ABE1,...	[IEA, IEA, IEA, IDA]	YOL051W	[20308326,...
855821	[5671, 124, 46695, 5634...	HFI1	[]	[ADA1,...	[IDA, IDA, IEA, IEA, IMP]	YPL254W	[19507699]
851811	[5694, 788, 5654, 4445]	HTA1	[]	[H2A1, S...	[RCA, IEA, TAS, IMP, IEA, IEA]	YDR225W	[15781691]



Galaxy

Code

Libraries

Command Line Tools

Web Interface wrapper

Workflow

Complex Custom Pipeline

Most control and most complexity

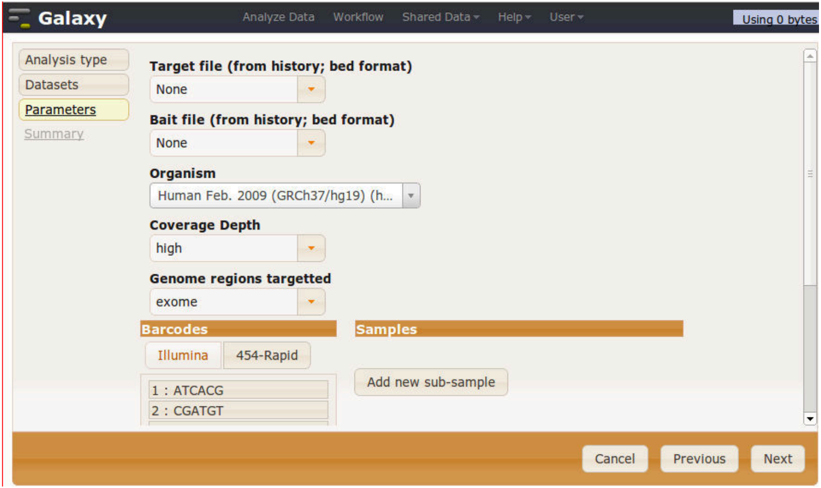
Purpose built

Levels of abstraction

Approaches

- Custom Galaxy interfaces
  - Variant calling pipeline
  - Galaxy views and controllers
- Web tools
  - BioCloudCentral
  - CloudMan
- External data upload
  - ISA-Tab experiments
  - Data Libraries and API
- Interoperable tools
  - Variant analysis platform
  - GenomeSpace

Variant pipeline interface



Target file (from history; bed format): None

Bait file (from history; bed format): None

Organism: Human Feb. 2009 (GRCh37/hg19) (h...

Coverage Depth: high

Genome regions targetted: exome

Barcodes

Illumina | 454-Rapid

1 : ATCACG

2 : CGATGT

Samples

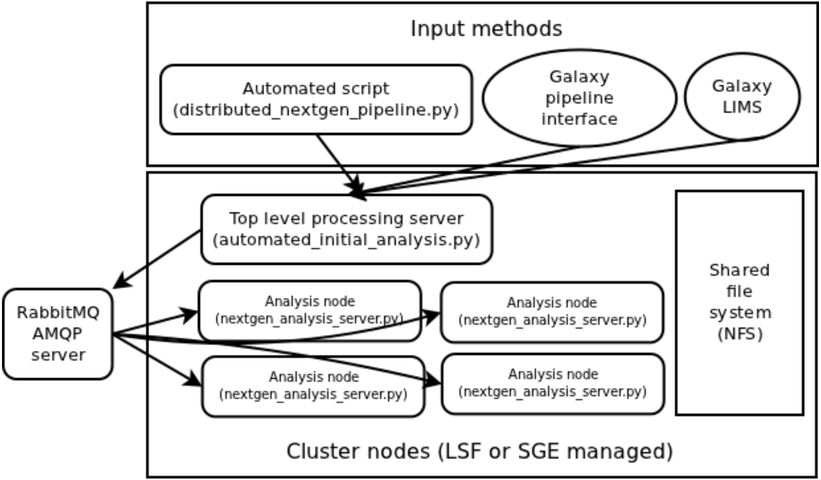
https://bitbucket.org/hbc/galaxy-central-hbc

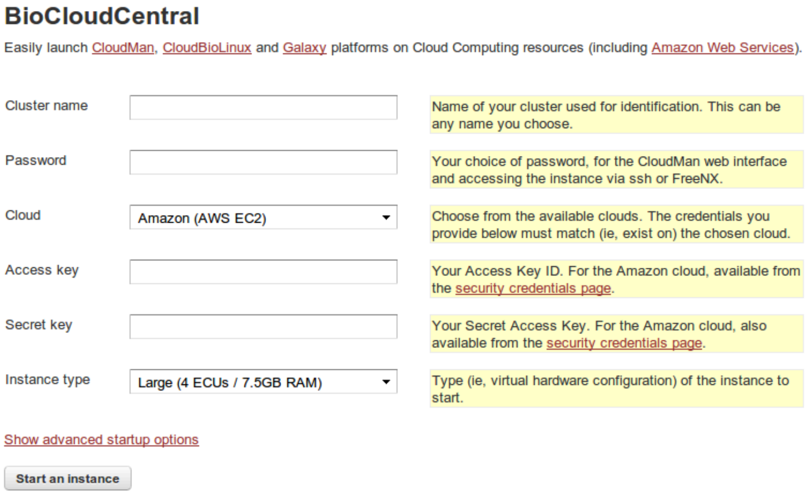
Implementation approach

- Integrated into Galaxy fork
- Custom view (`http://your.galaxy.org/pipeline`)
- Custom controller
- Calls external server for processing
- Results upload to Galaxy Data Libraries

Demo with movies: `http://j.mp/uNXZY6`

Backend variant processing


Input methods
Automated script
(distributed_nextgen_pipeline.py)
Galaxy
pipeline
interface
Galaxy
LIMS
Top level processing server
(automated_initial_analysis.py)
Shared
file
system
(NFS)
RabbitMQ
AMQP
server
Analysis node
(nextgen_analysis_server.py)
Analysis node
(nextgen_analysis_server.py)
Analysis node
(nextgen_analysis_server.py)
Analysis node
(nextgen_analysis_server.py)
Cluster nodes (LSF or SGE managed)



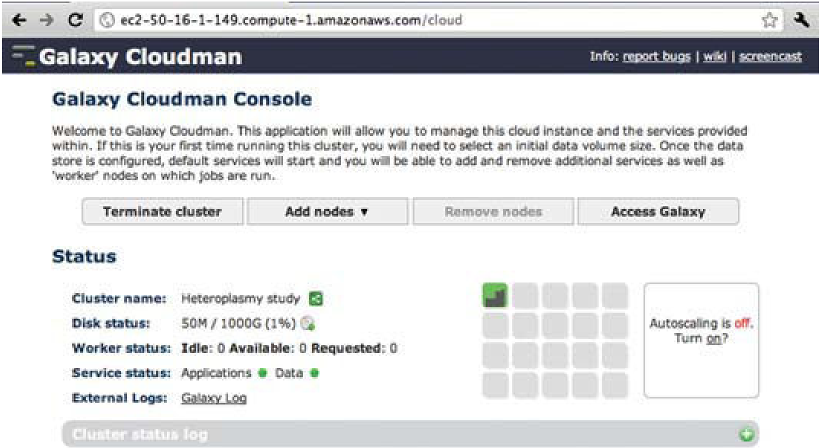
BioCloudCentral

Easily launch CloudMan, CloudBioLinux and Galaxy platforms on Cloud Computing resources (including Amazon Web Services).

Cluster name		Name of your cluster used for identification. This can be any name you choose.
Password		Your choice of password, for the CloudMan web interface and accessing the instance via ssh or FreeNX.
Cloud	Amazon (AWS EC2)	Choose from the available clouds. The credentials you provide below must match (ie, exist on) the chosen cloud.
Access key		Your Access Key ID. For the Amazon cloud, available from the security credentials page.
Secret key		Your Secret Access Key. For the Amazon cloud, also available from the security credentials page.
Instance type	Large (4 ECUs / 7.5GB RAM)	Type (ie, virtual hardware configuration) of the instance to start.

http://biocloudcentral.org

CloudMan: automate setup


ec2-50-16-1-149.compute-1.amazonaws.com/cloud
Galaxy Cloudman
Info: report bugs | wiki | screencast

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud instance and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

Status

Cluster name: Heteroplasmy study

Disk status: 50M / 1000G (1%)

Worker status: **Idle:** 0 **Available:** 0 **Requested:** 0

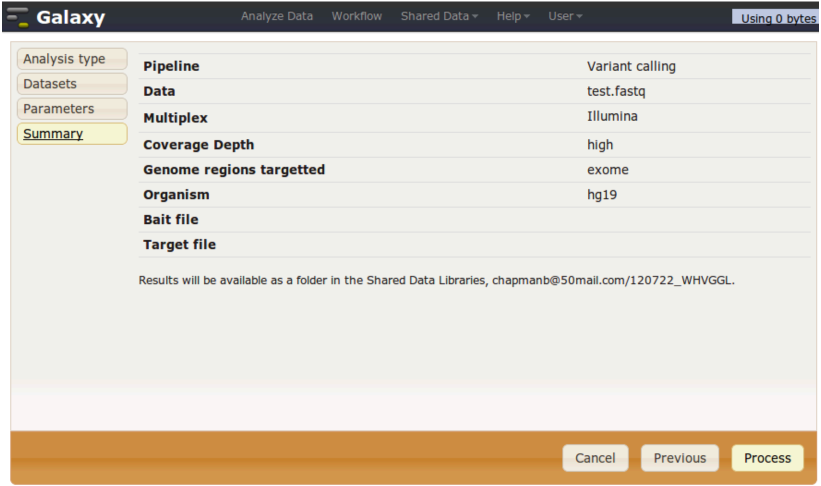
Service status: Applications ● Data ●

External Logs: Galaxy Log

Autoscaling is off. Turn on?

http://usecloudman.org/

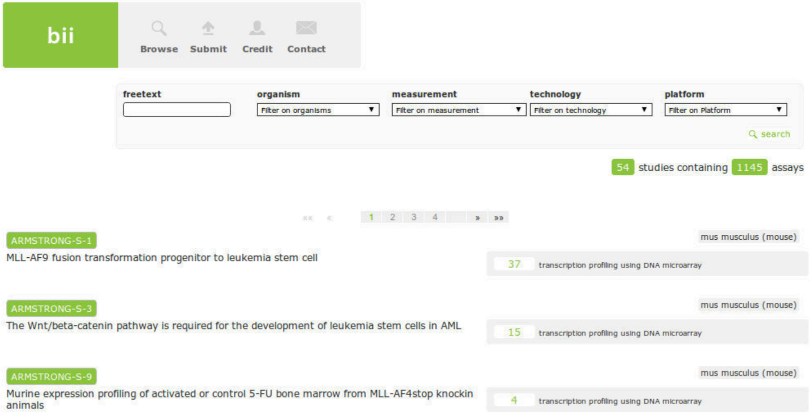
Run analysis



Pipeline	Variant calling
Data	test.fastq
Multiplex	Illumina
Coverage Depth	high
Genome regions targetted	exome
Organism	hg19
Bait file	
Target file	

Results will be available as a folder in the Shared Data Libraries, chapmanb@50mail.com/120722_WHVGGL.

Experimental metadata: ISA-Tab



54 studies containing 1145 assays

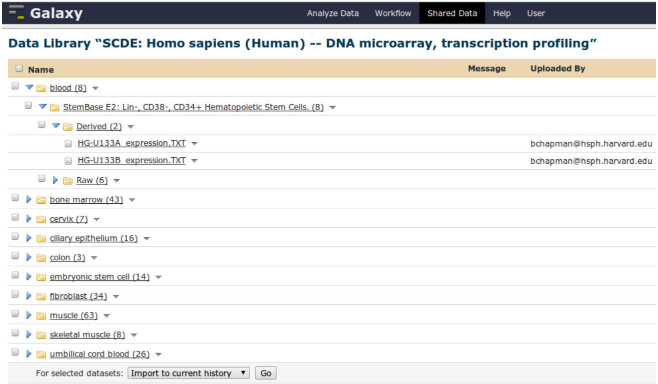
ARMSTRONG-S-1
MLL-AF9 fusion transformation progenitor to leukemia stem cell
mus musculus (mouse)
37 transcription profiling using DNA microarray

ARMSTRONG-S-3
The Wnt/beta-catenin pathway is required for the development of leukemia stem cells in AML
mus musculus (mouse)
15 transcription profiling using DNA microarray

ARMSTRONG-S-9
Murine expression profiling of activated or control 5-FU bone marrow from MLL-AF4stop knockin animals
mus musculus (mouse)
4 transcription profiling using DNA microarray

http://isatab.sourceforge.net/

Data Library integration



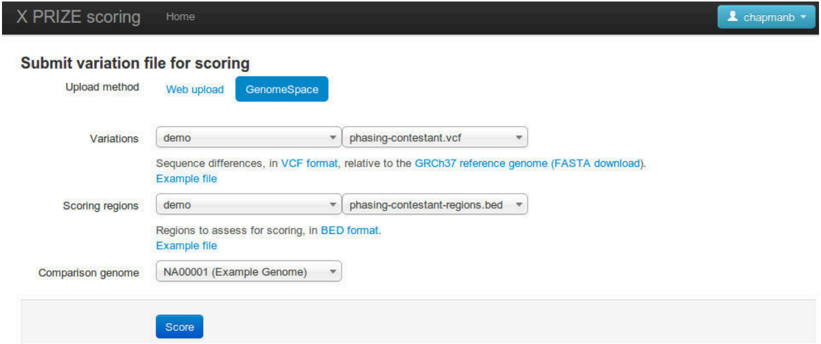
Data Library “SCDE: Homo sapiens (Human) -- DNA microarray, transcription profiling”

☐ Name	Message	Uploaded By
☐ ▼ blood (8) ▾		
☐ ▼ StemBase E2: Lin-, CD38-, CD34+ Hematopoietic Stem Cells. (8) ▾		
☐ ▼ Derived (2) ▾		
☐ HG-U133A_expression.TXT ▾		bchapman@hsph.harvard.edu
☐ HG-U133B_expression.TXT ▾		bchapman@hsph.harvard.edu
☐ ▶ Raw (6) ▾		
☐ ▶ bone marrow (43) ▾		
☐ ▶ cervix (7) ▾		
☐ ▶ ciliary epithelium (16) ▾		
☐ ▶ colon (3) ▾		
☐ ▶ embryonic stem cell (14) ▾		
☐ ▶ fibroblast (34) ▾		
☐ ▶ muscle (63) ▾		
☐ ▶ skeletal muscle (8) ▾		
☐ ▶ umbilical cord blood (26) ▾		

For selected datasets: Import to current history ▾ Go

- Stem Cell Discovery Engine: `http://discovery.hsci.harvard.edu/`
- `https://github.com/hbc/projects/blob/master/scde_deploy/scripts/bii_datasets_to_galaxy.py`

Interoperable tools



Submit variation file for scoring

Upload method: Web upload | GenomeSpace

Variations: demo | phasing-contestant.vcf

Sequence differences, in VCF format, relative to the GRCh37 reference genome (FASTA download).
Example file

Scoring regions: demo | phasing-contestant-regions.bed

Regions to assess for scoring, in BED format.
Example file

Comparison genome: NA00001 (Example Genome)

- https://github.com/chapmanb/bcbio.variation
- http://validationprotocol.org
- http://genomespace.org

Answers and Questions

- Build off core Galaxy features
  - Data Libraries
  - API
  - CloudMan
- Maximize interoperability
- Automate aggressively

Would love to hear your experiences.