



Integration of SeqWare within Galaxy

Zhibin Lu, Morgan Taschuk, Brian
O'Connor and B.F. Francis Ouellette

2012 Galaxy Community Conference
July 27, 2012

Outlines

- Galaxy implementation at OICR
- Introduction of SeqWare
- Integration of SeqWare within Galaxy
- Future development

OICR

- Ontario Institute for Cancer Research
- Launched by the Government of Ontario in December 2005
- An innovative translational research organization dedicated to research on the prevention, early detection, diagnosis and treatment of cancer.
- About 150 researchers in the Genomics and in the Informatics and Bio-computing platforms
- Projects includes ICGC, Bioinformatics.ca, Bioinformatics link directory, ...
- Galaxy fits well in such a large institute

OICR Compute Resources

- 5500 cores
- 185 nodes with 16 GB RAM
- 221 nodes with 24 GB RAM
- 32 nodes with 96 GB RAM
- 5 nodes with 256 GB RAM
- 2.5PB of online storage
- 1Gb, 10Gb and fibre connectivity
- SGE

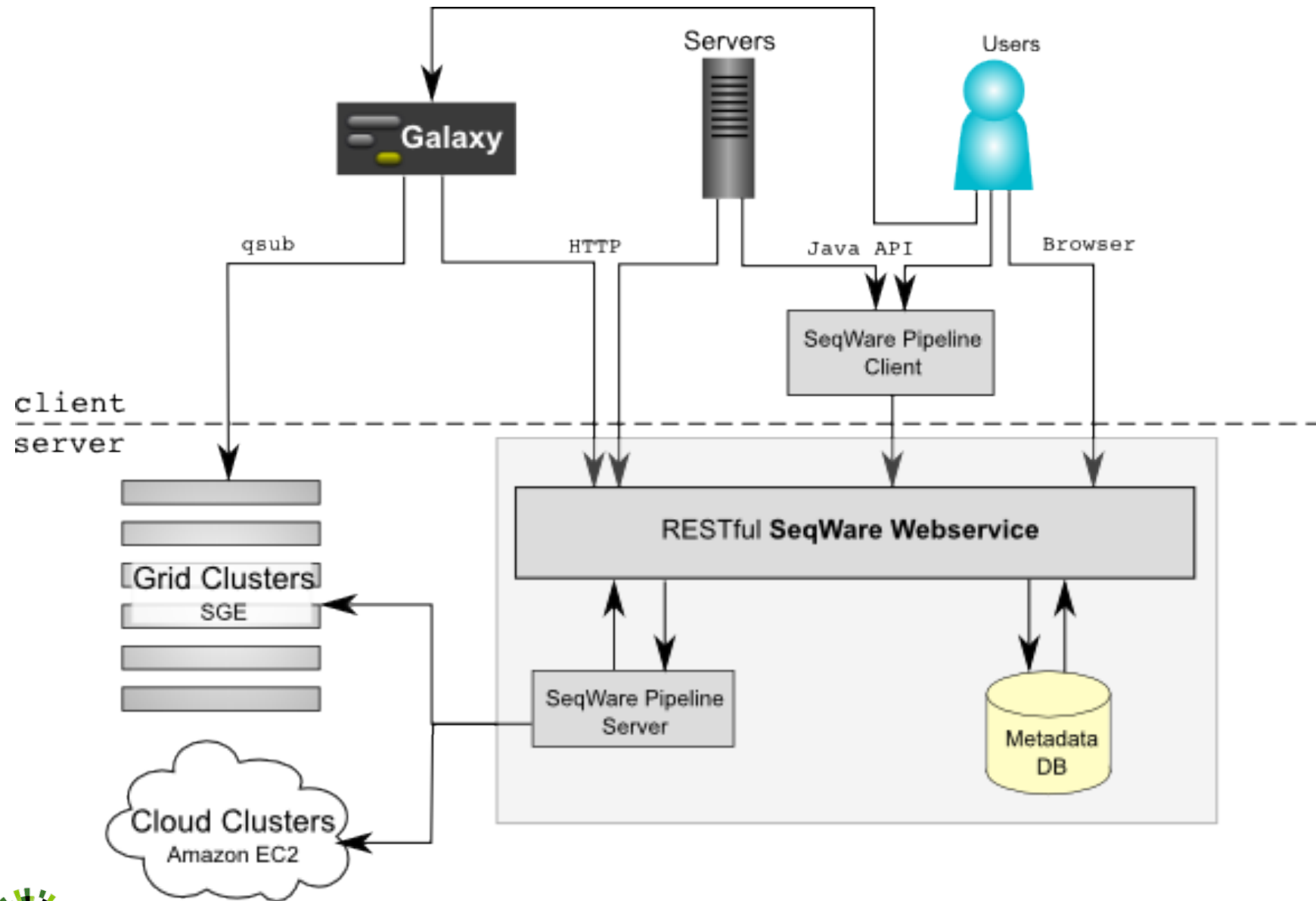
Galaxy Servers at OICR

- Standalone server:
 - 4-cpu
 - 8GB memory
 - Nginx
 - mysql
 - 3 webapp and 1 job runner
- On the cluster:
 - 2-cpu
 - 2GB memory
 - Apache
 - PostgreSQL
 - LDAP authentication
 - Submit jobs to OICR cluster

SeqWare

- Supports massively parallel sequencing analysis
- Developed at UCLA, UNC, OICR
- <http://seqware.sourceforge.net>
- Components:
 - MetaDB
 - Portal
 - Pipeline (Pegasus, Condor, Globus Toolkit)
 - Query Engine
- WebService

SeqWare Architecture



SeqWare Webservice

- REST API

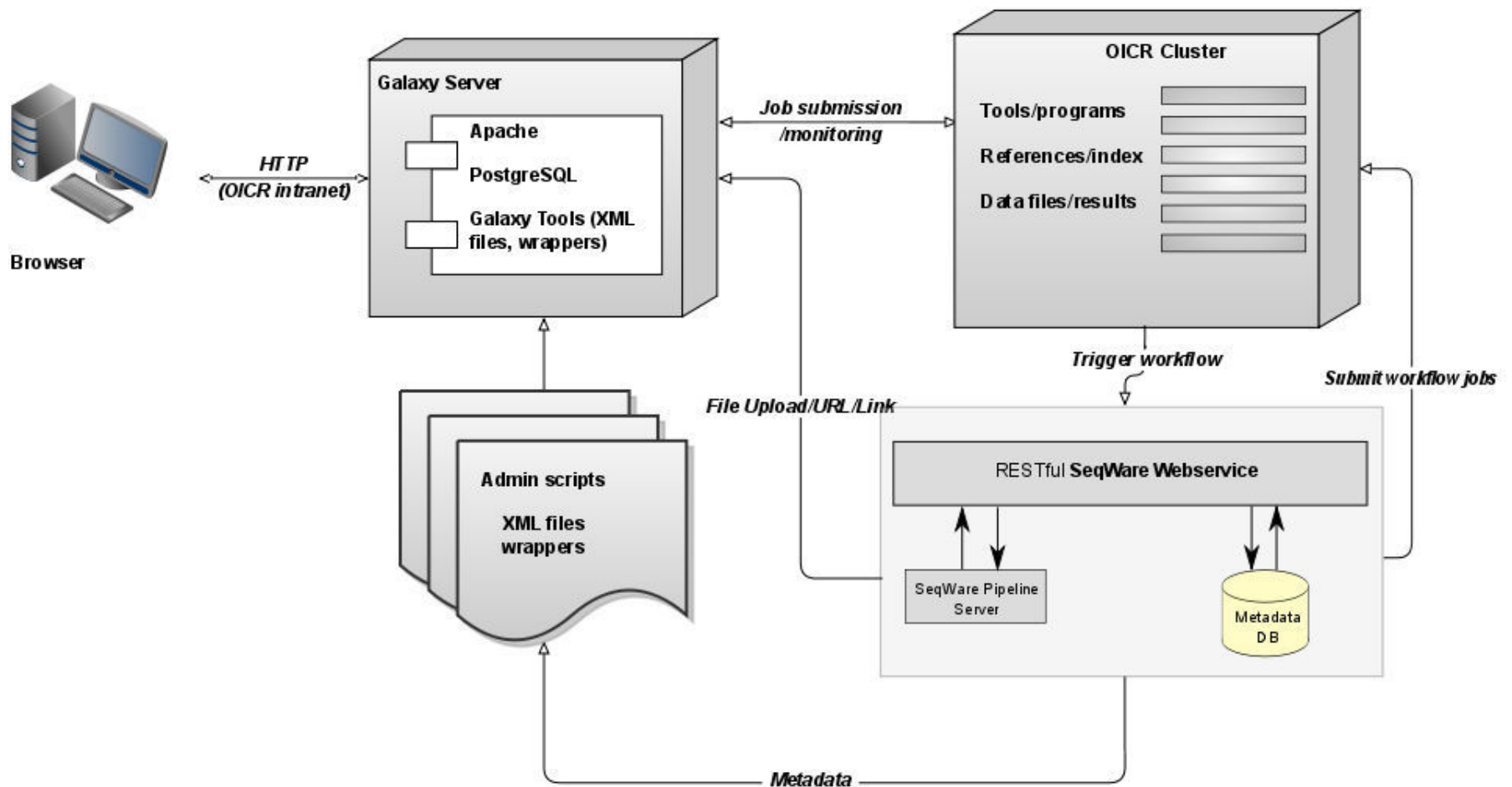
<http://seqware.github.com/seqware/>

- /workflows/
- /workflows/{swAccession}
- /workflows/{swAccession}/run
- /workflowruns/{workflowRunAccession}
- /workflowruns/{workflowRunAccession}/files
- /workflowruns/{workflowRunAccession}/processings

Benefit of the Integration

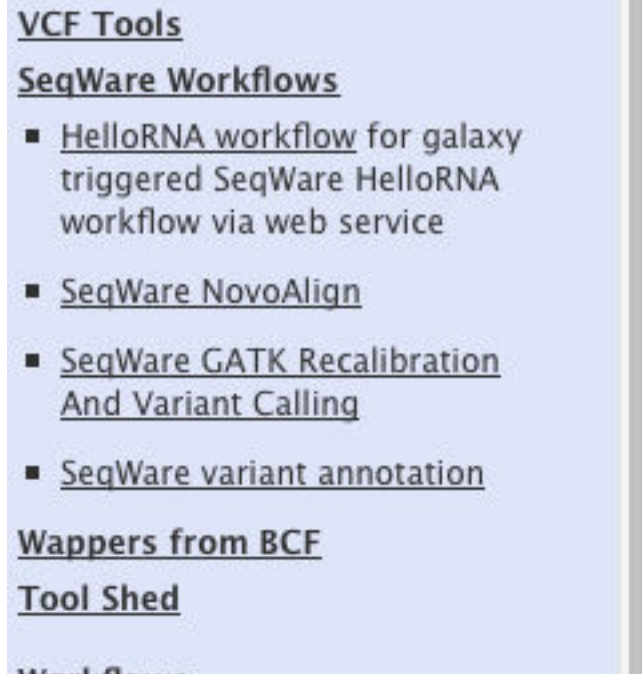
- Take advantage of mature workflows from SeqProd group for NGS analysis
- Easy downstream analysis for SeqProd group
- Reproducible: same dataset, same result
- Version of tools and references
- HPC cluster issues
 - Re-launch jobs when they fail
 - Memory and queue management

SeqWare and Galaxy Integration



SeqWare Workflow Implemented

- HelloWorld
- Novoalign
- GATK (Exomes, WholeGenome)
- VariantAnnotation (zipped, un-zipped VCF, GFF3 input)



The screenshot displays a web interface with a light blue background. It features several sections:
1. VCF Tools: A section header.
2. SeqWare Workflows: A section header followed by a bulleted list of four items:
 - HelloRNA workflow for galaxy triggered SeqWare HelloRNA workflow via web service
 - SeqWare NovoAlign
 - SeqWare GATK Recalibration And Variant Calling
 - SeqWare variant annotation
3. Wappers from BCF: A section header.
4. Tool Shed: A section header.
The text is in a dark grey font, and the list items are preceded by small square bullet points.

SeqWare NovoAlign Workflow

SeqWare NovoAlign (version 0.1)

Is this library mate-paired?:

Paired-end ↕

Single-end
Paired-end

1: PCSI0022C.val.1.fastq ↕

Fastq input first read file

FASTQ file:

2: PCSI0022C.val.2.fastq ↕

Fastq input second read file

Select a reference genome:

hg19 ↕

NovoAlign genome reference index

ColorSpace:

Color Space

Execute

An alignment workflow using Novoalign to hg19 references prepared by OICR.

SeqWare GATK workflow

SeqWare GATK Recalibration And Variant Calling (version 0.1)

BAM file:

8: SeqWare NovoAlign..apped reads

BAM input

More BAM files

More BAM files 1

Additional BAM file:

8: SeqWare NovoAlign..apped reads

Remove More BAM files 1

Add new More BAM files

Select a reference genome:

Human (Homo sapiens): hg19

reference genomes

output_post_realign_recal_bams:

Output Merged, Realigned, Mark Duplicates, Recalibrated BAM

exomes or whole genome?:

Whole Genome

Exom

Whole Genome

EXECUTE

This workflow is designed to take one or more BAM files, merge them, break it down by chromosome, perform realignment, duplicate flag, and perform variant calling for small indels and SNVs. The result is a VCF file for SNVs and indels that has been filtered. This workflow is designed to use GATK version 1.3.16 which was released on 20111116. The 'Exom' option is identical to the 'Whole Genome' but has an exome-specific quality filter.

SeqWare VariantAnnotation Workflow

SeqWare variant annotation (version 0.1)

Input file type:

GFF
GFF
VCF

GFF3 file

Execute

A SNV and small indel variant annotation tool that will annotate a VCF file input with refGene mutation consequence, cytoBand, transcription factor binding sites from the transfac Matrix Database (v 7.0), snoRNA/miRNA, TargetScan miRNA target site predictions, segmental duplications, phastCons conserved elements, conserved functional RNA via EvoFold, Database of Genomic Variants (structural variations), OMIM, published GWAS results on human disease associations, 1000 Genomes Nov 2010 (SNVs and indels) and May 2011 (SNVs) releases, dbSNP 132, SIFT scores, PolyPhen2, MutationTaster, PhyloP conservation score, LRT, GERP++ scores, and variants from 5400 NHLBI exomes. These tracks were downloaded from the Annovar project (<http://www.openbioinformatics.org/annovar>) on 20111231. This workflow also includes custom tracks and annotates VCF files with the Human Gene Mutation Database (HGMD, public dataset, see <http://www.hgmd.org>), the Human Mitochondrial Genome Database (mtDB, updated March 2007), HmtDB, and MitoTool. This workflow uses the 2011Nov20 version of Annovar, see the docs directory for HTML files downloaded from their site that describe their annotation sources.

Future Development

- More bundled software workflows: RNASeq, Structure Variants, ...
- Different versions of SeqWare workflow bundles
- SeqWare results as data source
- SeqWare adopts Galaxy XML syntax
- Integrating SeqWare MetaDB

Questions and Comments