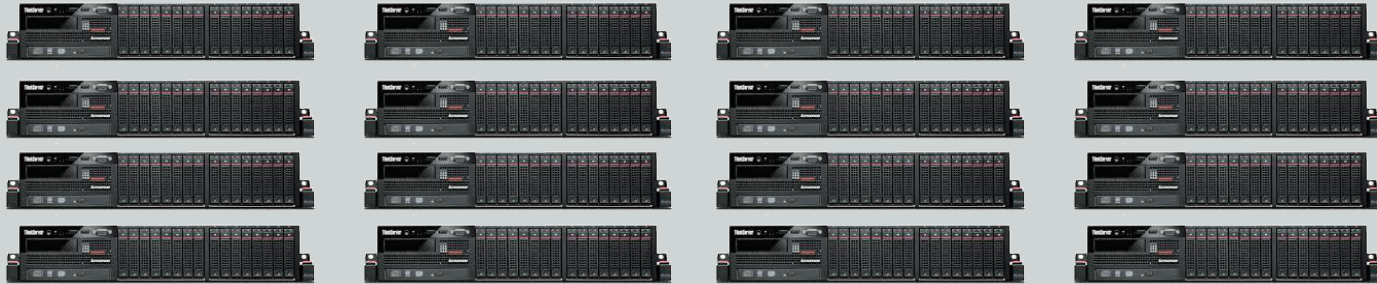

Galaxy Farm

Kyle Ellrott, Dannon Baker

Current Clustering



Jobs



Shared Filesystem and Queue

For High Performance

For large scale genomics using network shared drives doesn't scale

Galaxy needs a concept of data locality

Be able to do workflow end-to-end on remote node, only transfer back the required results

What kind of high performance

Applying standardized pipelines to massive data sets

Example:
Run standard re-alignment pipeline on every TCGA WGS

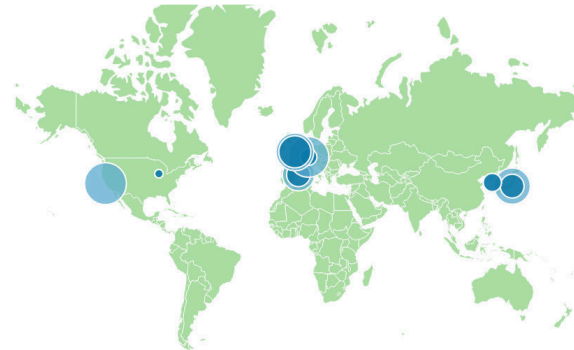
PANCANCER.INFO Sites Uploads

ICGC/TGCA PanCancer Status

The map below shows the number of specimens whose unaligned, lane-level WGS reads have been uploaded to a GNOS repository at one of the PanCancer clouds. Typically there are two specimens per donor, a tumor and a normal, and we expect approximately 3,000 specimens from ICGC and 2,000 from TCGA. It also shows the count of those specimens aligned using the project's BWA-Mem workflow. For more information see our [Wiki Space](#).

Upload & Alignment

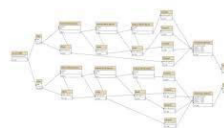
	Barcelona	Chicago	Heidelberg	London	Seoul	Tokyo	Santa Cruz	Total	% of 5000
Aligned	43	offline	13	178	20	42	956	1252	25.04%
Unaligned	200	offline	682	489	20	355	0	1746	34.92%



Farming



Workflow



Things that need to move around

- Work
 - Workflows
 - Tools
 - Data
 - Provenance
-

What do you need

Essentials

- Workflow Request Batches
- UUIDs (on everything)
- Data Referrals (UUID resolution)
- Standardized Job/Workflow/Tool/Provenance Transfer

Fancy Stuff

- Docker
 - Gossip Protocols
 - Peer Authentication
-

What's Done? What's needed?

Done:

- Dataset UUID are now on by default
- Workflow Batches Implemented
- Mesos based launching

Needed:

- Workflow Request Transfer
 - Auto History transfer
-

Going forward?

Get a working prototype running

Do a large scale run

The tools that you need for farming are the same ones you need for federation...

Extra Slides

Workflow Request Batch

In the Galaxy DB, we need to separate the idea of a Workflow Request and a Workflow Instance

Add the concept of Batches, that can be scheduled at the same time (with common scheduling configurations)

UUID resolution

14b228c0-72c5-4c25-aaa5-
da8e5b13eef7



Cancer Genomics Hub

Browser

Cart (0)

Batch search

Help

Accessibility

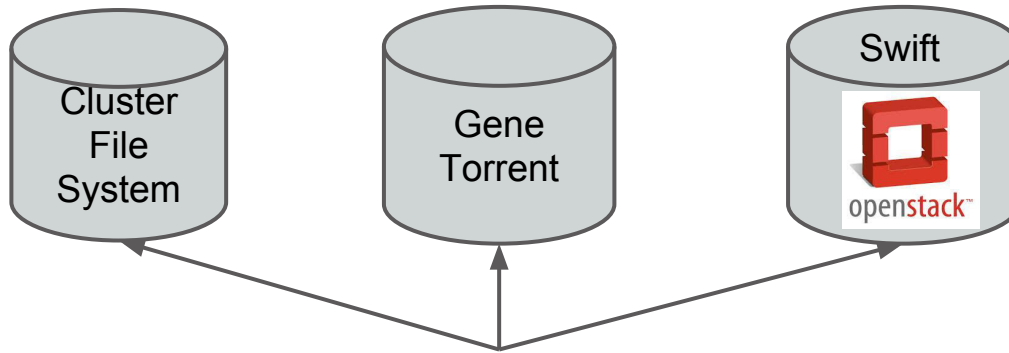
Search

Data Browser

Item details for Analysis Id 14b228c0-72c5-4c25-aaa5-da8e5b13eef7

Study	TCGA
Barcode	TCGA-A6-A567-01A-31D-A28G-10
Disease	COAD
Disease Name	Colon adenocarcinoma
Sample Type	TP
Sample Type Name	01
Analyte Type	DNA
Library Type	WGS
Center	BCM
Center Name	Baylor College of Medicine
Platform	ILLUMINA
Platform Name	Illumina
Assembly	GRCh37-lite
Filename	TCGA-A6-A567-01A-31D-A28G-10_wgs_illumina.bam
Filesize	519.58 GB

Data Referral



14b228c0-72c5-4c25-aaa5-da8e5b13eef7



?



Job Offloading



Work and Data
Transfers

