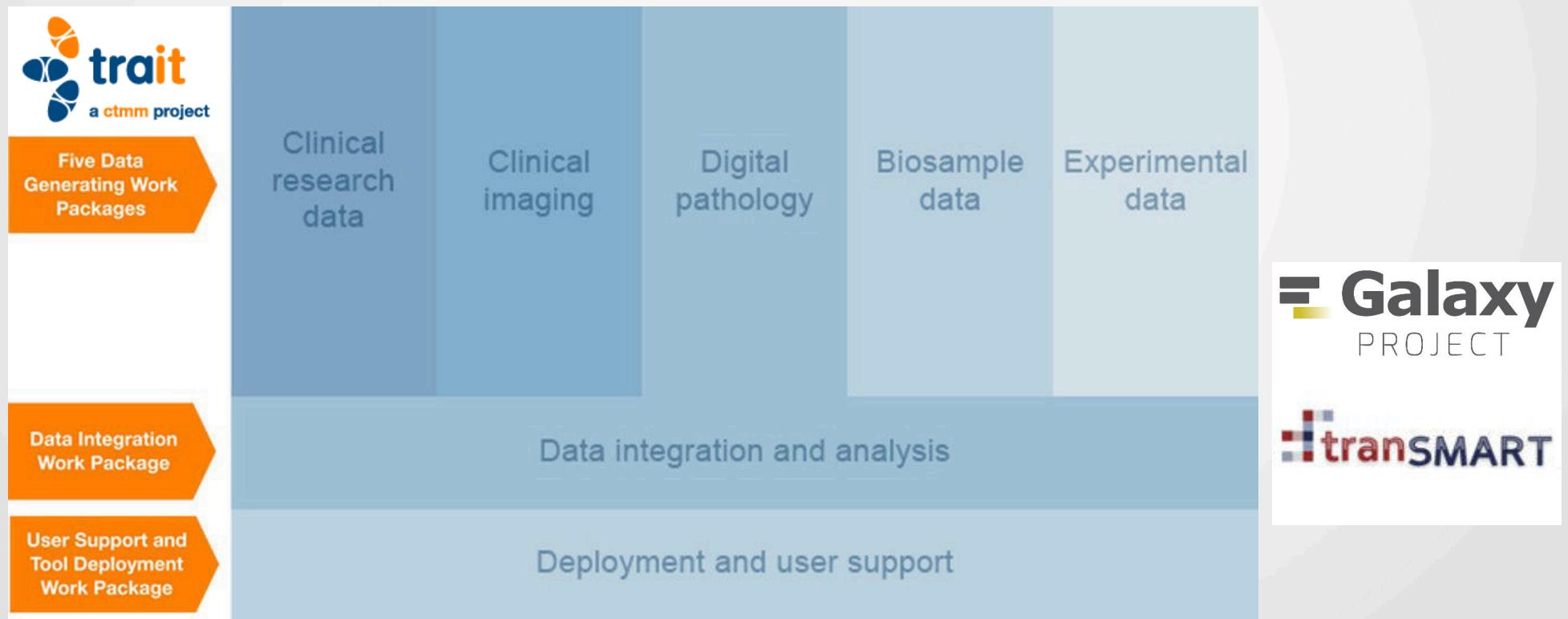


Galaxy as backend for TraIT genotype to phenotype studies

Youri Hoogstrate
GCC2015
7 July 2015, Norwich

TraIT, TranSMART & Galaxy



When you think of Galaxy

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Using 0%

Tools

search tools

Get Data

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

NGS: QC and manipulation

NGS: Mapping

NGS: RNA-seq

NGS: SAMtools

NGS: BAM Tools

NGS: Picard

NGS: VCF Manipulation

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

snpEff

BEDTools

Genome Diversity

EMBOSS

Regional Variation

FASTA manipulation

Evolution

Multiple Alignments

Metagenomic analyses

Motif Tools

NGS TOOLBOX BETA

NGS: Peak Calling

NGS: Variant Analysis

NGS: GATK Tools (beta)

NGS: Picard (beta)

RNA Structure Prediction

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

GCC 2015
Galaxy Community Conference
4-8th July 2015
The Sainsbury Laboratory, Norwich, UK
gcc2015.tsl.ac.uk

Tweets

Galaxy Project @galaxyproject Bioinformatics scientist @Rothamsted Closing date: 12 July rothamsted.ac.uk/jobs/1431 #usegalaxy

BF Francis Ouellette @bf0 Our #HTSD15 lecture & lab (w/ Sorana Morrissey) went gr8, thank U @natefoo 4 keeping an eye on things #usegalaxy pic.twitter.com/11LugF2C08

Nate Coraor @natefoo @bjoerngruening merged github.com/galaxyproject/... to make himself @NateCoraor and

History

search datasets

Unnamed history

0 bytes

This history is empty. You can load your own data or get data from an external source

PENNSTATE

JOHNS HOPKINS UNIVERSITY

TACC

iPlant Collaborative

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, and the Department of Biology and at Johns Hopkins University.

This instance of Galaxy is utilizing infrastructure generously provided by the iPlant Collaborative at the Texas Advanced Computing Center, with support from the National Science Foundation.

The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own local Galaxy instance or run Galaxy on the cloud.

Galaxy version 15.05, commit 6423454857bbc4b958ec1966b184cc4133edeb94

When you think of Galaxy

 **Galaxy Project** ⓘ

Galaxy is an open, web-based platform for data intensive biomedical research.

http://galaxyproject.org/ outreach@galaxyproject.org

Filters Find a repository...

galaxy Data intensive biology for everyone. Updated 25 minutes ago

tools-devteam Contains a set of Galaxy Tools mostly written by the Galaxy Team. Updated an hour ago

cloudman Easily create compute clusters on the Cloud. Updated 2 hours ago

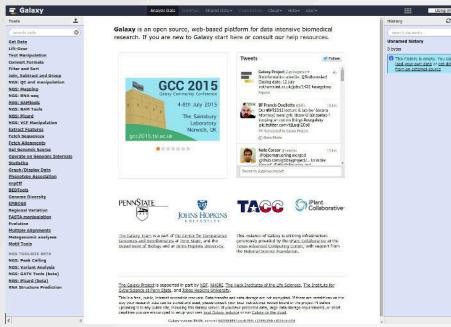
usegalaxy-playbook Ansible Playbook for usegalaxy.org Updated 17 hours ago

ansible-postgresql An Ansible role for managing a PostgreSQL (<http://www.postgresql.org/>) server Updated 18 hours ago

People 8 >



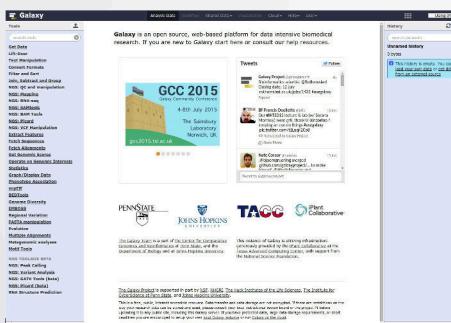
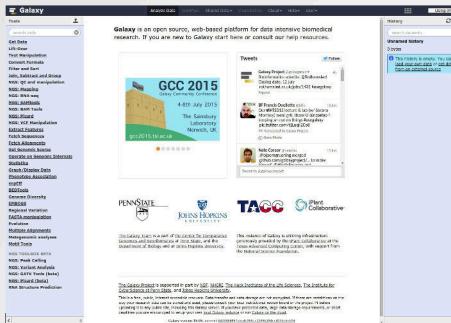
Purpose of Galaxy: using



executing job script:

```
galaxy.jobs.runners.local DEBUG 2015-06-17 14:58:58,947 (34) executing job script: galaxy.jobs DEBUG 2015-06-17 14:58:58,995 (34) Persisting job destination (destina galaxy.jobs.runners.local DEBUG 2015-06-17 14:59:00,771 execution finished: /usr/l galaxy.datatypes.metadata DEBUG 2015-06-17 14:59:00,928 loading metadata from file galaxy.jobs INFO 2015-06-17 14:59:01,086 Collecting job metrics for <galaxy.model.galaxy.jobs DEBUG 2015-06-17 14:59:01,105 job 34 ended (finish() executed in [333. galaxy.datatypes.metadata DEBUG 2015-06-17 14:59:01,105 Cleaning up external metad 127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/histories/5969bf1f7201f12ae/co 7.36 (KHTML, like Gecko) Ubuntu Chromium/43.0.2357.81 Chrome/43.0.2357.81 Safari/5 127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/histories/5969bf1f7201f12ae HT 127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/users/f2db41ef1fa331b3e HTTP/1 ML, like Gecko) Ubuntu Chromium/43.0.2357.81 Chrome/43.0.2357.81 Safari/537.36" 127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/users/f2db41ef1fa331b3e HTTP/1 127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/users/f2db41ef1fa331b3e HTTP/1 like Gecko) Ubuntu Chromium/43.0.2357.81 Chrome/43.0.2357.81 Safari/537.36"
```

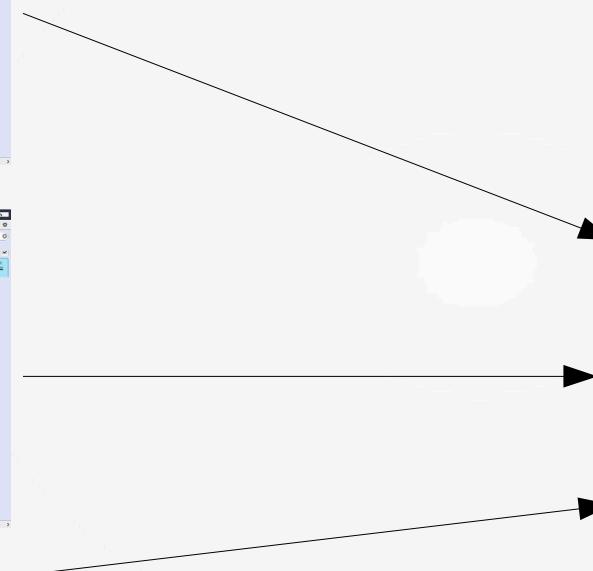
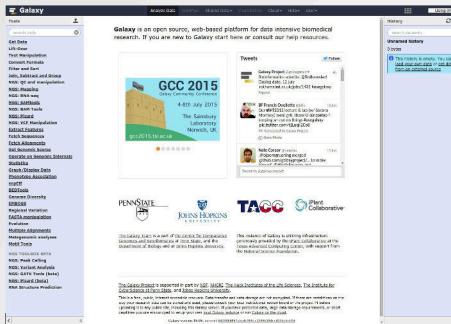
Purpose of Galaxy: using



executing job script:

```
galaxy.jobs.runners.local DEBUG 2015-06-17 14:58:58,947 (34) executing job script:  
galaxy.jobs DEBUG 2015-06-17 14:58:58,995 (34) Persisting job destination (destina  
galaxy.jobs.runners.local DEBUG 2015-06-17 14:59:00,771 execution finished: /usr/l  
galaxy.datatypes.metadata DEBUG 2015-06-17 14:59:00,928 loading metadata from file  
galaxy.jobs INFO 2015-06-17 14:59:01,086 Collecting job metrics for <galaxy.model.  
galaxy.jobs DEBUG 2015-06-17 14:59:01,105 job 34 ended (finish()) executed in (333.  
galaxy.datatypes.metadata DEBUG 2015-06-17 14:59:01,105 Cleaning up external metad  
127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/histories/5969bf1f7201f12ae/co  
7.36 (KHTML, like Gecko) Ubuntu Chromium/43.0.2357.81 Chrome/43.0.2357.81 Safari/5  
127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/histories/5969bf1f7201f12ae HT  
ML, like Gecko) Ubuntu Chromium/43.0.2357.81 Chrome/43.0.2357.81 Safari/537.36"  
127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/uploads/f2db4e1fira331b3e HTTP/1  
like Gecko) Ubuntu Chromium/43.0.2357.81 Chrome/43.0.2357.81 Safari/537.36"
```

Purpose of Galaxy: using

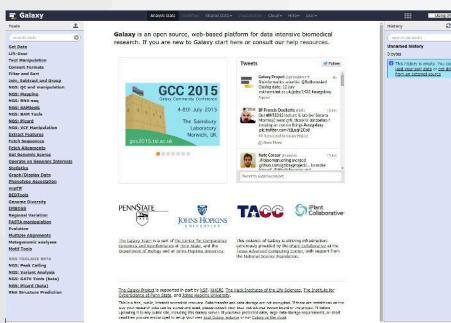
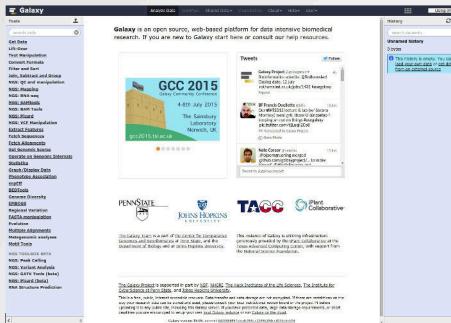


executing job script:

```
galaxy.jobs.runners.local DEBUG 2015-06-17 14:58:58,947 (34) executing job script:  
galaxy.jobs DEBUG 2015-06-17 14:58:58,995 (34) Persisting job destination (destina  
galaxy.jobs.runners.local DEBUG 2015-06-17 14:59:00,771 execution finished: /usr/l  
galaxy.datatypes.metadata DEBUG 2015-06-17 14:59:00,928 loading metadata from file  
galaxy.jobs INFO 2015-06-17 14:59:01,086 Collecting job metrics for <galaxy.model.  
galaxy.jobs DEBUG 2015-06-17 14:59:01,105 job 34 ended (finish()) executed in (333.  
galaxy.datatypes.metadata DEBUG 2015-06-17 14:59:01,105 Cleaning up external metad  
127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/histories/5969bf7201f12ae/co  
7.36 (KHTML, like Gecko) Ubuntu Chromium/43.0.2357.81 Chrome/43.0.2357.81 Safari/5  
127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/histories/5969bf7201f12ae HT  
ML, like Gecko) Ubuntu Chromium/43.0.2357.81 Chrome/43.0.2357.81 Safari/537.36"  
127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/users/f2db41efab331b3e HTTP/1  
like Gecko) Ubuntu Chromium/43.0.2357.81 Chrome/43.0.2357.81 Safari/537.36"
```

```
from bioblend.galaxy.objects import GalaxyInstance  
gi = GalaxyInstance("URL", "API_KEY")  
wf = gi.workflows.list()[0]  
hist = gi.histories.list()[0]  
inputs = hist.get_datasets()[:2]  
input_map = dict(zip(wf.input_labels, inputs))  
params = {"Paste1": {"delimiter": "U"}}  
wf.run(input_map, "wf_output", params=params)
```

Purpose of Galaxy: using



executing job script:

```
galaxy.jobs.runners.local DEBUG 2015-06-17 14:58:58,947 (34) executing job script:  
galaxy.jobs DEBUG 2015-06-17 14:58:58,995 (34) Persisting job destination (destina  
galaxy.jobs.runners.local DEBUG 2015-06-17 14:59:00,771 execution finished: /usr/l  
galaxy.datatypes.metadata DEBUG 2015-06-17 14:59:00,928 loading metadata from file  
galaxy.jobs INFO 2015-06-17 14:59:01,086 Collecting job metrics for <galaxy.model.  
galaxy.jobs DEBUG 2015-06-17 14:59:01,105 job 34 ended (finish()) executed in (333.  
galaxy.datatypes.metadata DEBUG 2015-06-17 14:59:01,105 Cleaning up external metad  
127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/histories/5969bf7201f12ae/co  
7.36 (KHTML, like Gecko) Ubuntu Chromium/43.0.2357.81 Chrome/43.0.2357.81 Safari/5  
127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/histories/5969bf7201f12ae HT  
ML, like Gecko) Ubuntu Chromium/43.0.2357.81 Chrome/43.0.2357.81 Safari/537.36"  
127.0.0.1 - - [17/Jun/2015:14:59:02 +0200] "GET /api/users/f2db41e1fa331b3e HTTP/1  
like Gecko) Ubuntu Chromium/43.0.2357.81 Chrome/43.0.2357.81 Safari/537.36"
```

```
from bioblend.galaxy.objects import GalaxyInstance  
gi = GalaxyInstance("URL", "API_KEY")  
wf = gi.workflows.list()[0]  
hist = gi.histories.list()[0]  
inputs = hist.get_datasets()[:2]  
input_map = dict(zip(wf.input_labels, inputs))  
params = {"Paste1": {"delimiter": "\t"}  
wf.run(input_map, "wf_output", params=params)
```



TranSMART

The screenshot shows the TranSMART software interface. At the top, there is a toolbar with "Active Filters and or" (with two filter icons), a "Filter" button, and a "Clear" button. Below the toolbar is a "Navigate Terms" sidebar with a tree view of categories:

- Cell-line (19)
 - Characteristics (19)
 - Cell line constructs (3)
 - Cell line name (16)
 - Cell type (19)
 - Disease (19)
 - Gender (18)
 - Organism (19)
 - Origin tissue (19)
 - Race (17)
 - Age (14)
 - Molecular profiling (19)
 - High-throughput molecular profiling (19)
 - Copy number aberrations (DNA) (14)
 - Expression (miRNA) (2)
 - Expression (mRNA) (18)
 - Expression (protein) (8)
 - Non-highthroughput molecular profiling (13)
 - Copy number aberrations (DNA) (4)
 - Expression (mRNA) (1)

TranSMART

- Clinical data
- Interpreted data
 - VCF files
 - Expression tables
 - Copy number aberration

No intermediate files

- No BAM files
- No raw FASTQ files

The right man for the right job



- + Clinical data
 - + Interactive cohort selection
- + Drag 'n Drop cohorts
- Analysis suite
- Large experimental data



- + API
- + Many bioinformatics apps (>3300)
- + Scaling resources for tools
- Link clinical- to experimental data
- Drag 'n Drop cohorts

The right man for the right job



- + Clinical data
 - + Interactive cohort selection
- + Drag 'n Drop cohorts
- Analysis suite
- Large experimental data



- + API
 - + Many bioinformatics apps
 - + Scaling resources for tools
- Link clinical- to experimental data
- Drag 'n Drop cohorts

The right man for the right job



- + Clinical data
 - + Interactive cohort selection
- + Drag 'n Drop cohorts
- Analysis suite
- Large experimental data



- + API
 - + Many bioinformatics apps
 - + Scaling resources for tools
- Link clinical- to experimental data
- Drag 'n Drop cohorts



Example: PoC: RNA-seq

The screenshot shows a software interface for bioinformatics analysis, specifically for RNA-seq data. On the left, there is a hierarchical search tree under 'Search Terms' and 'Navigate Terms'. The tree includes categories like Clinical Trials, Private Studies (with RNASEQ_GOUD and Biomarker Data), and Public Studies (with TCGAOV and Biomarker Data). A specific node under 'RNASEQ_GOUD\Biomarker Data' is highlighted in yellow. The main workspace on the right displays an 'Analysis' section titled 'Group Test for RNASeq'. It shows a 'Cohorts' section with 'Subset 1: (Private Studies)\RNASEQ_GOUD\' selected. Below this is an 'Input Parameters' table:

RNASeq	Group	Analysis type
...\\RNASeq\\	...\\no-Fusion\\	<input checked="" type="radio"/> two group unpaired
	...\\TMPRSS-ERG\\	<input type="radio"/> multi-group

At the bottom, there is an 'Intermediate Result - Job Name: admin-RNASeqgroupTest-100541' section which is currently empty. Navigation controls at the bottom indicate 'Page 1 of 234'.

- Drag 'n Drop

Example: PoC: RNA-seq

The screenshot shows a software interface for bioinformatics analysis. On the left, there is a hierarchical search tree under 'Search Terms' and 'Navigate Terms'. The tree includes categories like Clinical Trials, Private Studies (selected), RNASEQ_GOUD (27), Biomarker Data (27), Fusion Gene (27) (selected), no-Fusion (16), TMPRSS-ERG (11), Illumina (27), Prostate (27), RNASEQ (27), Biopsy (27), Derivatives (27), Diagnosis (27), Follow-up (27), General data dynamic (27), Laboratory dynamic (27), Patient Information (27), Pca general dynamic (27), Prostatectomy (27), Sample Data (27), Samples dynamic (17), Treatment (27), RNASEQ_PROSTATE (27), and Public Studies (TCGAOV (573), Biomarker Data (80), Chrom (80), GPL4091 (80), Ovary (80)).

The main workspace on the right has tabs for Generate Summary Statistics, Summary, Clear, Save, Comparison, Advanced Workflow, Results/Analysis, Grid View, Data Export, Export Jobs, Analysis Jobs, and Genome. The 'Analysis' tab is selected.

The 'Analysis' section shows 'Analysis: Group Test for RNASeq' and 'Cohorts: Subset 1: (Private Studies)\RNASEQ_GOUD\}'. Below this is an 'Input Parameters' panel for 'RNASeq' analysis. It includes fields for 'Group' (containing '...\\RNASEQ\\', '...\\no-Fusion\\', and '...\\TMPRSS-ERG\\') and 'Analysis type' (with 'two group unpaired' selected). There is also a 'multi-group' option.

At the bottom, there is an 'Intermediate Result - Job Name: admin-RNASeqgroupTest-100541' section and a page navigation bar indicating 'Page 1 of 234'.

- Drag 'n Drop
 - Creates expression- & design matrix

Example: PoC: RNA-seq

The screenshot shows the Galaxy Project interface. On the left is a navigation tree with categories like Clinical Trials, Private Studies, and Public Studies. Under Private Studies, there's a folder 'RNASEQ_GOUD' containing 'Biomarker Data' with sub-folders 'Fusion Gene (27)', 'Illumina (27)', and 'RNASEQ (27)'. The 'RNASEQ (27)' folder is highlighted. On the right, the main workspace shows an 'Analysis' section titled 'Group Test for RNASEq'. It displays 'Cohorts: Subset 1: (Private Studies)\RNASEQ_GOUD\'. Below this is an 'Input Parameters' panel for 'RNASEq'. It has a 'Group' section with '...\\RNASEq\\' and a 'Analysis type' section where 'two group unpaired' is selected (indicated by a yellow box). A line points from this 'two group unpaired' option to the Galaxy logo. At the bottom, there's an 'Intermediate Result - Job Name: admin-RNASEqgroupTest-100541' and a page navigation bar.

- Drag 'n Drop
 - Creates expression- & design matrix
 - Sends job to Galaxy



Example: PoC: RNA-seq

The screenshot shows the Galaxy Project interface. On the left is a navigation tree with categories like Clinical Trials, Private Studies, and Public Studies. Under Private Studies, there's a folder 'RNASEQ_GOUD' containing 'Biomarker Data' with sub-folders 'Fusion Gene (27)', 'Illumina (27)', and 'RNASEQ (27)'. Under 'Fusion Gene (27)', there are 'abc no-Fusion (16)' and 'abc TMPRSS-ERG (11)'. The main workspace shows an 'Analysis' panel with 'Group Test for RNASEq' selected. It displays 'Cohorts: Subset 1: (Private Studies)\RNASEQ_GOUD\'. Below this is an 'Input Parameters' section for 'RNASEQ' with a 'Group' dropdown set to '...\\RNASEQ\\' and an 'Analysis type' dropdown where 'two group unpaired' is selected. A note at the bottom says 'Intermediate Result - Job Name: admin-RNASEQgroupTest-100541'. At the bottom of the interface is a page navigation bar.

- Drag 'n Drop
 - Creates expression- & design matrix
 - Sends job to Galaxy



- Runs [edger with design matrix](#)

Example: PoC: RNA-seq

The screenshot shows a software interface for bioinformatics analysis. On the left, there is a navigation tree with categories like Clinical Trials, Private Studies, and Public Studies. Under Private Studies, there is a folder 'RNASEQ_GOUD (27)' which is expanded to show 'Fusion Gene (27)', 'Illumina (27)', and 'RNASEq (27)'. Under Illumina, there are sub-folders for 'Prostate (27)' and 'RNASEq (27)'. Other collapsed categories include Biopsy (27), Derivatives (27), Diagnosis (27), Follow-up (27), General data dynamic (27), Laboratory dynamic (27), Patient Information (27), Pca general dynamic (27), Prostatectomy (27), Sample Data (27), Samples dynamic (17), Treatment (27), and RNASEQ_PROSTATE (27). The main window has tabs for 'Generate Summary Statistics', 'Summary', 'Clear', and 'Save'. It also has buttons for 'Comparison', 'Advanced Workflow', 'Results/Analysis', 'Grid View', 'Data Export', 'Export Jobs', 'Analysis Jobs', and 'Genome'. The 'Analysis' tab is selected, showing 'Analysis: Group Test for RNASeq'. The 'Cohorts' section indicates 'Subset 1: (Private Studies)\RNASEQ_GOUD\'. Below this is an 'Input Parameters' section for 'RNASEq' with fields for 'Group' (containing '...\\no-Fusion\\' and '...\\TMPRSS-ERG\\') and 'Analysis type' (with 'two group unpaired' selected). A table titled 'Intermediate Result - Job Name: admin-RNASeqgroupTest-100541' lists genes, logFC, logCPM, PValue, and FDR. The top row of the table is highlighted in yellow. The table contains 10 rows of data.

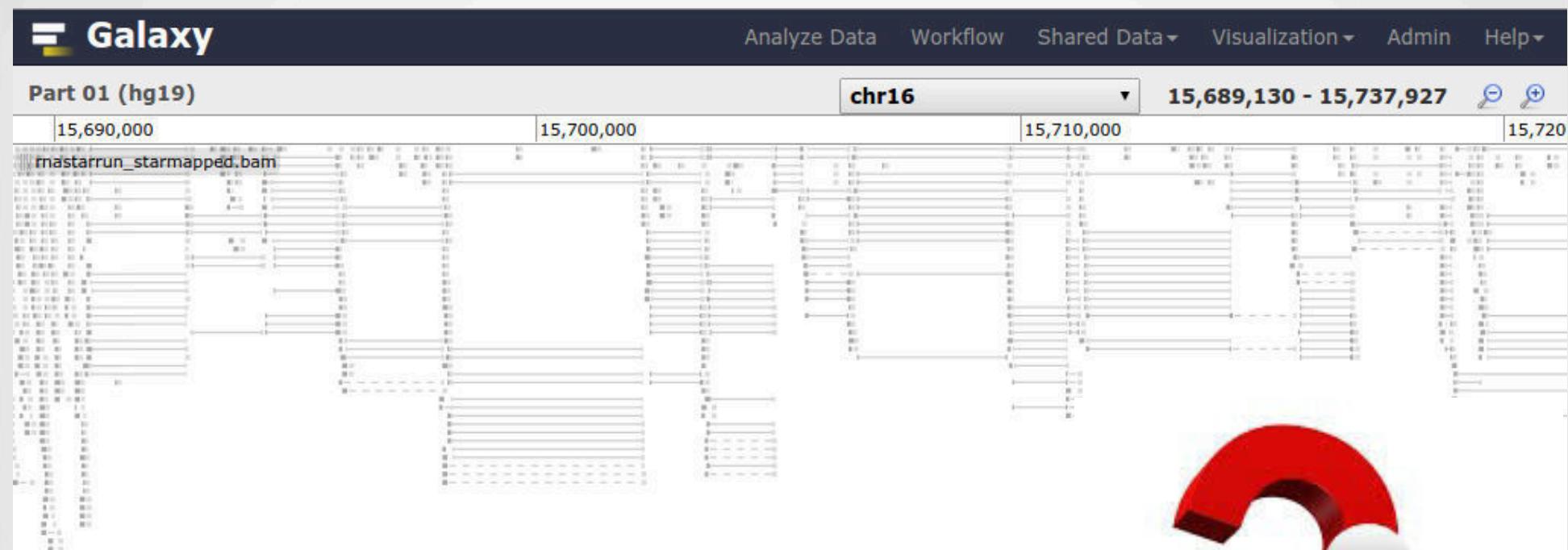
genes	logFC	logCPM	PValue	FDR
ERG	7.10928707922517	8.35628964115133	8.75358681813074e-77	2.04553816766079e-72
Gene9724	-5.78278722779815	7.66920103825944	2.80151278046494e-31	3.27328753269524e-27
Gene1382	7.10050324567181	2.03445035002978	7.89182363409181e-30	6.14720448938191e-26
Gene18489	1.84097063954614	8.65291950387125	8.64713841917363e-29	5.05165826448123e-25
Gene8050	4.96006717564777	6.93655936022684	2.51934189828676e-28	1.1774396295833e-24
Gene14261	-2.2509826167846	7.29595832688266	9.2941741148182e-28	3.61977101191786e-24
Gene2672	2.88488868773622	7.82754661421219	3.53105818705219e-27	1.17876811021479e-23
Gene10322	3.14288894830092	6.86964019840253	1.03031920942848e-25	3.00956241074059e-22

- Drag 'n Drop
 - Creates expression- & design matrix
 - Sends job to Galaxy



- Runs [edger with design matrix](#)
- Returns statistics & plots

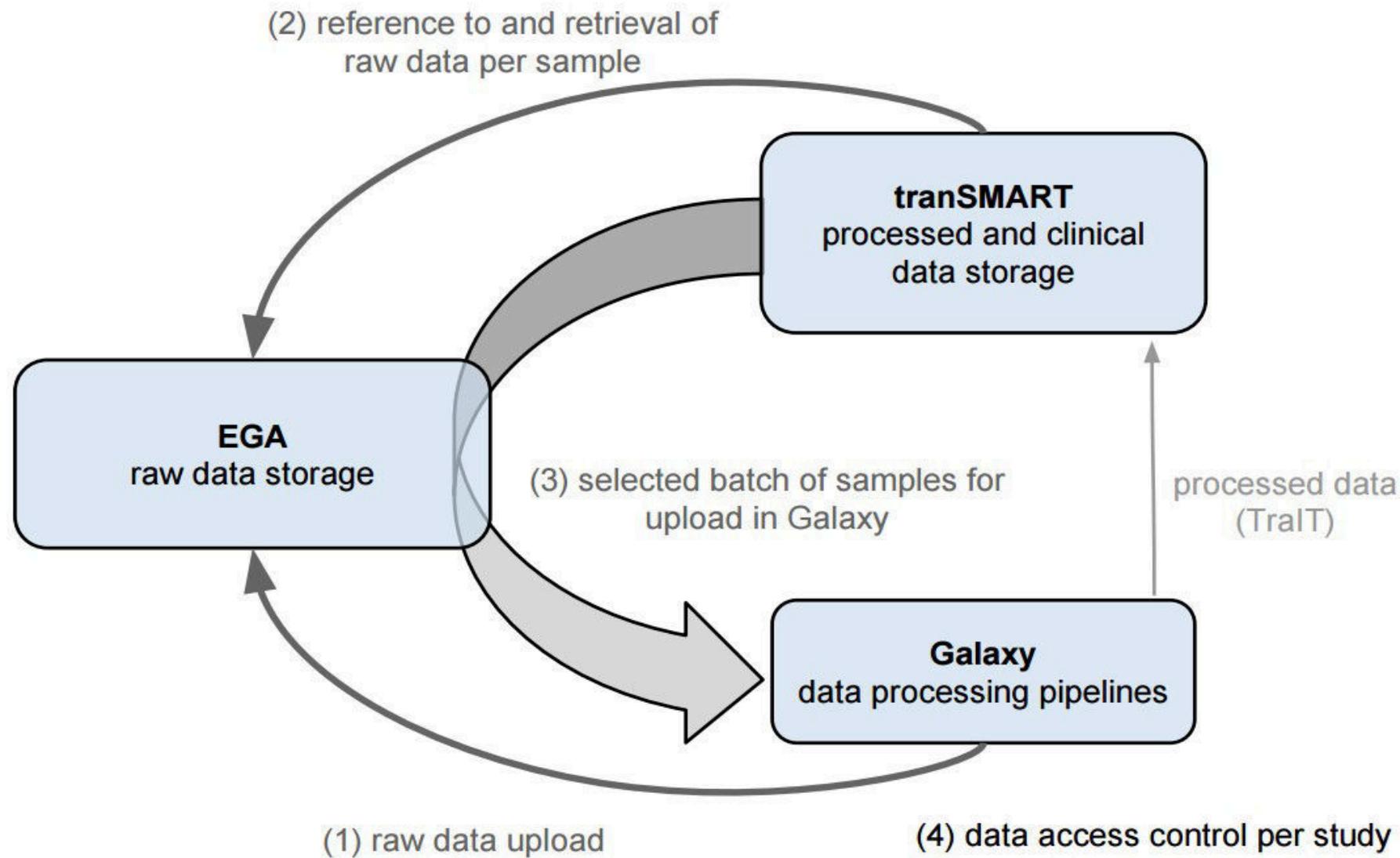
Example: PoC: RNA-seq



- How to get back to the reads?



Export: EGA



Acknowledgements

- Andrew Stubbs
- Freek de Brujin << **TranSMART to Blend4j interface**
- Guido Jenster
- Jan-Willem Boiten
- Jochem Bijlard
- Remond Fijneman
- Ruslan Forostianov << **Internal TranSMART modifications**
- Sanne Abeln
- Wim van der Linden