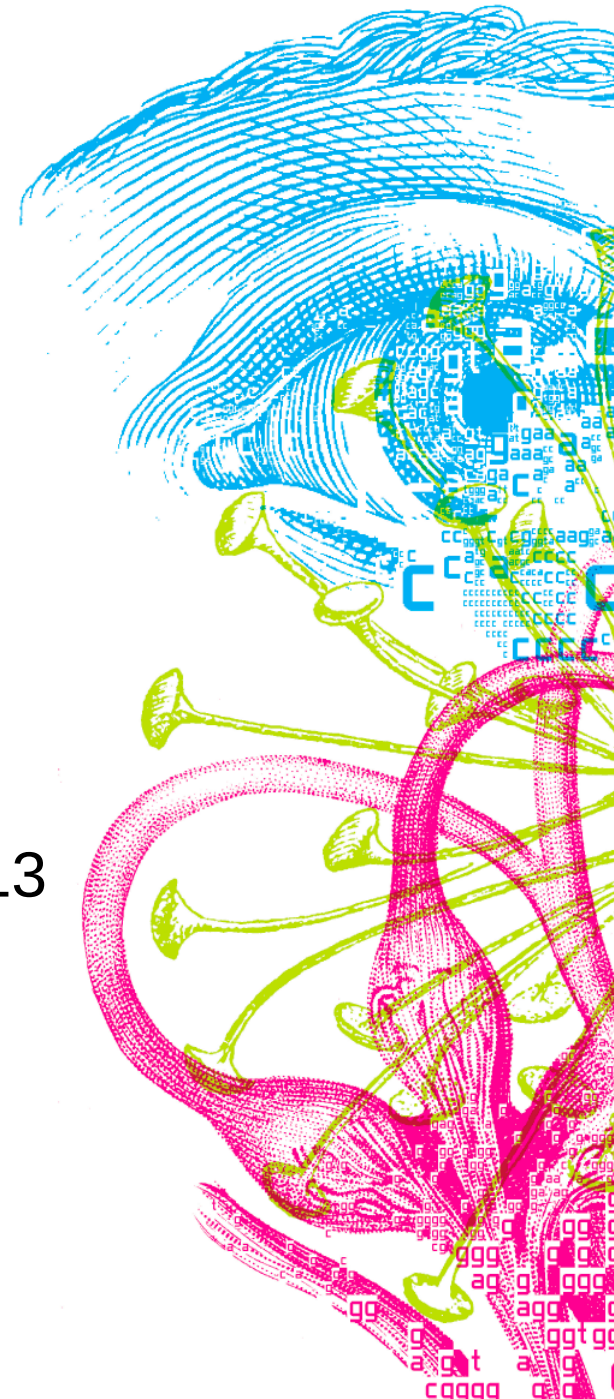




Netherlands  
Bioinformatics  
Centre

# NGS task force & GCC2013 report

<http://wiki.galaxyproject.org/Events/GCC2013>



# Update

- NGS PI meeting on June 18th
  - 2014 onwards, monthly meetings or thematic meetings?
  - SV benchmarking
  - Cross sectors, e.g. PigVD (based on DVD)
  - NGS course portfolio
- Galaxy community conference, June 30th-July 2nd, Oslo
  - RNAseq training, ~60 attendees
  - Amazon performance is disappointing

# GCC 2013

- ~210 attendees
  - NBIC: Leon
  - LUMC: Jeroen, Bowo, Wai Yi
  - EMC: Saskia, Rene
  - NIZO: Judith
  - UvA/NLeSC: Mateusz
  - LU:
  - WUR: Eric, Pieter
- Single track 20 presentations, ~30 posters, and many lightning talks!
  - People are sober thanks to the 5 euro beers.
  - All slides/videos on line  
<http://wiki.galaxyproject.org/Events/GCC2013/Program>

# Updates from Galaxy team

- UI improvement to handle large # datasets
  - [avoid over-blowing histories](#)
- Integrate toolshed & data manager
  - [Install all tools, dependencies, built-in data via Admin panel](#)
  - [Talks from Greg Von Kuster and Daniel Blankenberg](#)
- Galaxy to become a generic platform, no tools/data associated with the vanilla version.
  - [Several scripts will be created to populate the vanilla Galaxy with standard NGS tools, data.](#)
- Toolshed will be contributed by 3rd party developers and monitored by the IUC (Intergalactic Utilities Commission)
  - [Talk from Dannon Baker](#)

# Shared Interest

- Reproducibility
  - Versioning tools, keep histories, use test, etc.
- Cloud!!!
  - Amazon
  - EMC
  - Other private clouds (in Germany, France, US, Norway, etc), most OpenStack based.
- Professionalization and business model of Galaxy
  - BioTeam SlipStream
- Non-NGS
  - Galaxy-P, cheminformatics, image processing

# Highlight #1 (Tool factory, Ross Lazarus)

Galaxy / @BakerIDI

Analyze Data Workflow Shared Data Visualization Admin Help User

Using 44%

Tools

search tools

Gene Expression  
BakerIDI  
SR Test/Repair/BWA Tools  
Local unreliable SR Quality Tools  
Get Data  
Send Data  
Repeats and Complexity  
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  
NGS: QC and manipulation  
NGS: Assembly  
NGS: Mapping  
NGS: Indel Analysis  
NGS: RNA Analysis  
NGS: SAM Tools  
NGS: GATK Tools (beta)  
NGS: Peak Calling  
NGS: Simulation  
SNP/WGA: Data, Filters  
SNP/WGA: QC, LD, Plots  
SNP/WGA: Statistical Models  
VCF Tools  
NGS: Picard (beta)  
BedTools  
Workflows

Tool Factory (version 0.10)

Select an input file from your history:  
1: activinA\_all\_mm9\_bams2mx.xls

Most scripts will need an input - your script MUST be ready for whatever format you choose

New tool ID and title for outputs:  
activin edgeR paired

This will become the toolshed repository name so please choose thoughtfully to avoid namespace clashes with other tool writers

Create a tar.gz file ready for local toolshed entry:  
No. Just run the script please

Ready to deploy securely!

Create an HTML report with links to all output files and thumbnail links to PDF images:  
Yes, arrange all outputs in an HTML output

Recommended for presenting complex outputs in an accessible manner. Turn off for simple tools so they just create one output

Create a new (default tabular) history output:  
My script writes to a new history output

This is useful if your script creates a single new tabular file you want to appear in the history after the tool executes

Galaxy datatype for your tool's output file:  
Tabular

You may need to edit the xml to extend this list

Select the interpreter for your code. This must be available on the path of the execution host:  
Rscript

Cut and paste the script to be executed here:

```
# edgeR.Rscript  
# updated npv 2011 for R 2.14.0 and edgeR 2.4.0 by ross  
# Performs DGE on a count table containing n replicates of two conditions  
# Parameters  
# 1 - Output Dir  
  
# Original edgeR code by: S.Lunke and A.Kaspi
```

Script must deal with two command line parameters: Path to input tabular file path (or 'None' if none selected) and path to output tabular history file (or 'None').

Execute

⚠ Details and attribution [GTF](#)  
Local Admins ONLY Only users whose IDs found in the local admin\_user configuration setting in universe\_wsgi.ini can run this tool.  
If you find a bug please raise an issue at the bitbucket repository [GTF](#)  
What it does This tool enables a user to paste and submit an arbitrary R/python/perl script to Galaxy.  
Input options This version is limited to simple transformation or reporting requiring only a single input file selected from the history.  
Output options Optional script outputs include one single new history tabular file, or for scripts that create multiple outputs, a new HTML report linking all the files and images created by the script can be automatically generated.  
Tool Generation option Once the script is working with test data, this tool will optionally generate a new Galaxy tool in a gzip file ready to upload to your local toolshed for sharing and installation. Provide a small sample input when you run generate the tool because it will become the input for the generated functional test.  
⚠ Note to system administrators This tool offers NO built in protection against malicious scripts. It should only be installed on private/personal Galaxy instances. Admin\_users will have the power to do anything they want as the Galaxy user if you install this tool.  
⚠ Use on public servers is STRONGLY discouraged for obvious reasons  
The tools generated by this tool will run just as securely as any other normal installed Galaxy tool but like any other new tools, should always be checked carefully before installation. We recommend that you follow the good code hygiene practices associated with safe toolshed.

History

gregorevic activina results with tool factory script  
133.0 MB

58: activinPairedGSEA.html

53: activinedgeRpaired.html  
43.2 KB  
format: html, database: mm9  
HTML file

52: activinedgeRpaired.tabular

44: activinedgeRpaired.html

43: activinedgeRpaired.tabular

35: pairedSPIA.html

34: pairedSPIA.xls

29: paired\_gsea\_activin\_GSEA.html

27: activinedgeRpaired\_gsea.rnk

22: SPIA 56.html

21: SPIA 56.xls

20: SPIA 14.html

19: SPIA 14.xls

18: SPIA 7.html

17: SPIA 7.xls

16: SPIA 3.html

15: SPIA 3.xls

14: DESeqenesovertime rankings.xls

13: DESeq 56 days DESeq.html

12: DESeq 56 days DESeq.xls

11: DESeq 14

# Highlight #2 (Galaxy tutorial)

[https://genome.edu.au/wiki/Galaxy\\_Tutorials](https://genome.edu.au/wiki/Galaxy_Tutorials)

The screenshot shows a web browser window displaying the Genomics Virtual Laboratory (GVL) website. The browser's address bar shows the URL <https://genome.edu.au/wiki/GVL>. The website header includes the GVL logo and navigation links such as "128.250.103.200 Talk for this IP address" and "Login / create account to edit pages". Below the header, there is a "Page" dropdown menu set to "Discussion" and buttons for "Read", "View source", "View history", "Go", and "Search".

The main content area features a grid of navigation buttons:

- USE**: A button with the text "USE" and the GVL logo.
- LEARN**: A button with the text "LEARN" and the GVL logo.
- GET**: A button with the text "GET" and the GVL logo.
- BROWSE**: A button with the text "BROWSE" and a diagram of a human genome.
- DO**: A button with the text "DO" and the GVL logo.
- Get GVL Data**: A button with the text "Get GVL Data" and a block of DNA sequence: 

```
GTGAGAGGTTCACCTTGGTGAGGGGGTATTG  
GATATTAAGAGGATGATTAGACAGGAGATC  
CTTGGGAGGACATGGGATCCCTTTGATG  
TTGCTGGCCCTTGAGACTCTATTTCCTGG  
CAGGAAAGCATCAACCGTCCACAGGTTAAG  
GGTGGTCCCTCACTGAGACTCTGGTAGTA  
CTAAGAAAGATGATGGCTTGACGGTGAATG  
GGACAGGCTTAAGGAGGGGACCCAGAGGCA  
CTAGACTTGGGGCCGAGACTTGGCCCTAGG  
ATTTCCTTGGCTTTGCTTTTTCAGCCG  
TTGCTGAGGATTAAGGGGACTTCCCTGGGG
```
- HELP**: A button with the text "HELP" and the GVL logo.
- Publications**: A button with the text "Publications" and images of journal covers for "nature" and "Science".
- ABOUT**: A button with the text "ABOUT" and the GVL logo.
- PROJECT UPDATES**: A button with the text "PROJECT UPDATES" and a diagram of a laboratory workflow.

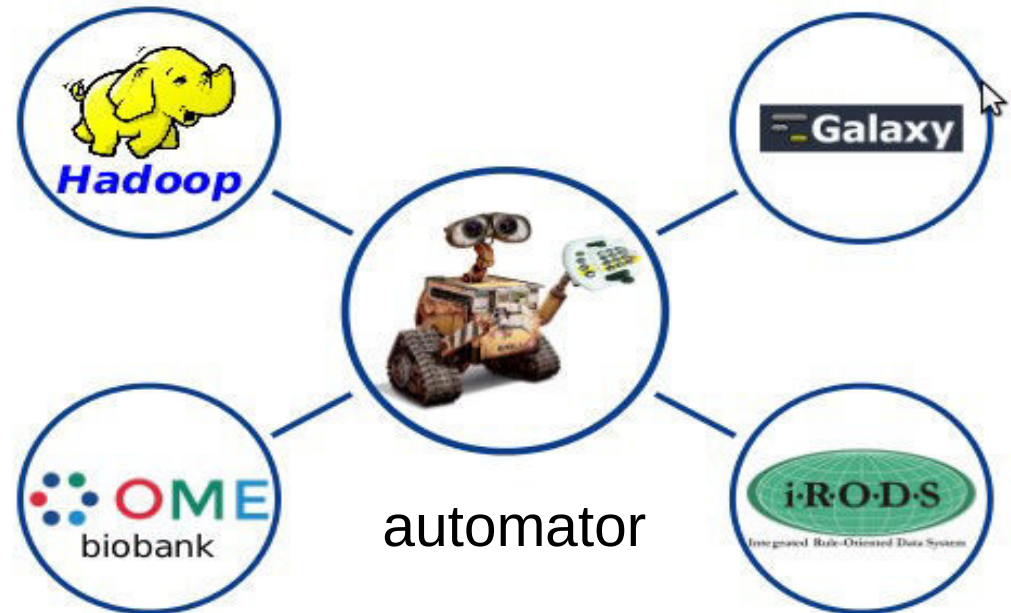
A sidebar on the left contains a "Toolbox" with links: "What links here", "Related changes", "Special pages", "Printable version", and "Permanent link".

At the bottom of the page, a small text line reads: "The Genomics Virtual Laboratory takes the IT out of Bioinformatics. It lets Biologists use a suite of genomics analysis tools that currently often require specialist assistance. GVL".

# Highlight #3 Automated processing and tracking platform

Luca Pireddu, CRS4

- Seal: toolkit for Hadoop-based sequencing data processing
  - demultiplexing, alignment (based on BWA, sorting, etc.)
- Pydoop: Python API for Hadoop
  - A dependency for Seal, but also used for custom tools and scripts
- SeqPig: SQL-like scripting for Hadoop with sequencing-specific functionality



iRODS, distributed file management system, including optimal file transfer support.



# Highlight #4 Genomics Hyperbrowser & Gtrack data type

- Hyperbrowser
  - Geir K Sandve
  - Statistical analysis tool for genomics tracks
  - <http://hyperbrowser.uio.no/test/> Including tutorialss
- Gtrack
  - Sveinung Gundersen
  - A new datatype to harmonize the existing datatypes
  - general purpose, tabular file format for representing data in the form of genomic tracks
  - Several tools and converters available

# Highlight #5 Auditing Galaxy for clinical use

- Sanjay Joshi, from EMC
  - The Clinical Galaxy: A validated platform initiative "We will present an overview of the requirements to move Galaxy into the Clinical realm."

# Highlight #6 BioBlend - automating bioinformatics with Galaxy and CloudMan

- Clare Sloggett
  - <http://bioblend.readthedocs.org/en/latest/>
  - <https://github.com/afgane/bioblend/>

# Public server BOF group discussion

- Security and Billing
  - Galaxy is not designed with security in mind from the ground up.
  - Authentication needs to be more pluggable.
  - Galaxy lacks the reporting feature on cpu hours which is sometimes very useful for funding agency.
- Releases
  - Galaxy lacks a stable release scheme (~twice per year) which makes the life of public Galaxy admin a lot easier.
  - Better Versioning
- Tool shed is currently making things more complicated.
- Dataset profiligation. Can easily end up with 3 copies of most of your datasets, just to get files into Galaxy.
- maintaining a public Galaxy server well requires minimal 0.5 fte