

Introduction à Galaxy

Olivier Inizan

Alban Lermine

Ecole Bioinformatique de Roscoff

14 janvier 2013

Introduction

- « Big data » problem : a small facet of a much bigger challenge
- Meaningful **interpretation** of sequencing data has become particularly important
- Big data interpretation constrains : c.f. « Travailler avec les données NGS »
- Galaxy Project : « **democratization** of biomedical computation so that even the smallest research units with modest budgets are capable of carrying out analyses using appropriate tools in a reproducible fashion »

Democratization

- developing **best practices**
- removing obstacles associated with using heterogenous software on complex high performance computing infrastructure :
accessibility
- promoting the concept of **transparency** and **reproducibility**

Best Practices : emergency !

 APPLICATIONS OF NEXT-GENERATION SEQUENCING — OPINION

Next-generation sequencing
data interpretation: enhancing
reproducibility and accessibility

Anton Nekrutenko and James Taylor

- 1000 Genomes Project : a serie of accepted practices for variant discovery
 - Galaxy P.I survey (Anton Nekrutenko and James Taylor)
 - 2011 : 299 articles that explicitly cite the 1000 genomes project :
 - **10/299** : used tools recommended by the consortium for mapping and variant discovery
 - **4/299** : used the whole workflow
- => The difficulty of reproductibility

Reproductibility : is it so easy ?

- NGS analysis is constant flux
- Not only ONE best practice
- Apply to non-model organisms
- Researchers choose to use more straightforward approaches
- Best practices, accessibility, transparency, reproductibility : the solution with **integrative resources** ?

Integrative resources

- Integrative resources, integrative frameworks : bring together diverse tools under the umbrella of unified interface
- BioExtract, GenePattern, GeneProf, Moby
- Galaxy



Galaxy and « meaningful interpretation »

- a.k.a how Galaxy embrace accessibility, reproductibility and best practices ?
- **Accessibility** : use computational approaches without programming or informatics expertise
- **Reproductibility** : reproduce experimental results
- **Transparency** : analysis can easily be communicated or understood

Accessibility

The screenshot displays the Galaxy web interface. At the top, a dark navigation bar contains the 'Galaxy' logo and menu items: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. A 'Using 0%' indicator is visible in the top right. On the left, a 'Tools' sidebar features a search bar and a list of tool categories such as 'Get Data', 'Send Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Convert Formats', 'FASTA manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Motif Tools', and 'Multiple Alignments'. The main content area is dominated by a large banner for 'Andromeda: A cloud-based Galaxy', which includes logos for NIOO, nbic, SURF SARA, and BiG Grid (the Dutch e-science grid). Below the banner is a 'Live Quickies' section with three video thumbnails: 'Mapping: Single End' (Galactic quickie # 15), 'Uploading Data using FTP' (Galactic quickie # 17), and 'Managing account histories' (Galactic quickie # 19). On the right, a 'History' panel shows 'Unnamed history' (19.6 MB) and a single entry: '1: Galaxy1-[chr4.fastq].fastq' with view, edit, and delete icons. At the bottom of the main content area, a paragraph of text describes Galaxy as an open, web-based platform for data-intensive biomedical research, mentioning its availability on a free public server or a user's own instance, and its affiliation with the Galaxy team at Penn State and Emory University.

Provide a unified, web based interface for bioinformatics analysis

Galaxy Items (1 / 2)

tools

The screenshot displays the Galaxy web interface. On the left is a sidebar titled 'Tools' with a search bar and a list of tool categories: Get Data, Send Data, ENCODE Tools, Lift-Over, Text Manipulation, Convert Formats, FASTA manipulation, Filter and Sort, Join, Subtract and Group, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data, Regional Variation, Multiple regression, Multivariate Analysis, Evolution, Motif Tools, and Multiple Alignments. The main content area features a large banner for 'Andromeda: A cloud-based Galaxy' with logos for nbic, NIOO, SURF, SARA, and BiG Grid. Below the banner is a 'Live Quickies' section with three cards: 'Mapping: Single End', 'Uploading Data using FTP', and 'Managing account histories'. At the bottom, there is introductory text about the Galaxy platform. On the right is a 'History' panel showing 'Unnamed history' (19.6 MB) and a single dataset entry: '1: Galaxy1-[chr4.fastq].fastq'.

Dataset

history

2 distributions

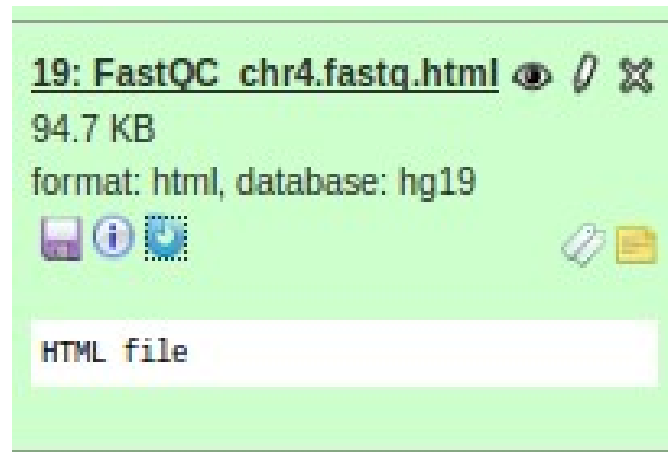
- 2 distributions : central (<https://main.g2.bx.psu.edu/>) and « dist »
- Dist : create your own analysis environment
 - Follow the model Galaxy use for integrating tools
 - A tool = a simple piece of software (cmd line)
 - A developer write a config file (how to run the tool, input and output param)
 - And ... Galaxy works with the tool abstractly : automatic generating web interfaces

Your own analysis env, example

The screenshot displays the Galaxy / ABiMS web interface. At the top, the navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User', along with a 'Using 604.5 MB' indicator. The left sidebar contains a 'Tools' section with a search bar and a list of categories: DATA MANAGEMENT, ECOLE GGB GROUPE 1, ECOLE GGB GROUPE 2, and various analysis tools like 'Mardi 15 RNA-seq' and 'Mercredi 16 RNA-seq de novo'. The main content area features a green welcome message: 'Bienvenue sur le serveur Galaxy de la plateforme ABiMS !' and a notice about a bioinformatics school organized by AVIESAN from January 14-18, 2013. Below this is the AVIESAN logo and logos of partner institutions including cea, CIFS, INRA, Inria, Inserm, Institut Pasteur, and IRD. At the bottom of the main area, it states: 'This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.' The right sidebar shows a 'History' panel with a list of workflow steps, each with an eye icon, a progress indicator, and a delete icon. The steps include: 'imported: TP Initiation' (287.7 MB), '35: Filter on data 34', '34: Compute on data 33', '33: Filter pileup on data 31', '32: MPileup on data 2 and data 28 (log)', '31: MPileup on data 2 and data 28', '30: flagstat on data 28', '29: MarkDups Dupes Marked.html', '28: MarkDups Dupes Marked.bam', '27: flagstat on data 26', '26: SAM-to-BAM on data 2 and data 25: converted BAM', '25: Map with Bowtie for Illumina on data 23 and data 2: mapped reads', and '24: FastQC Filter FASTQ on data 20.html'.

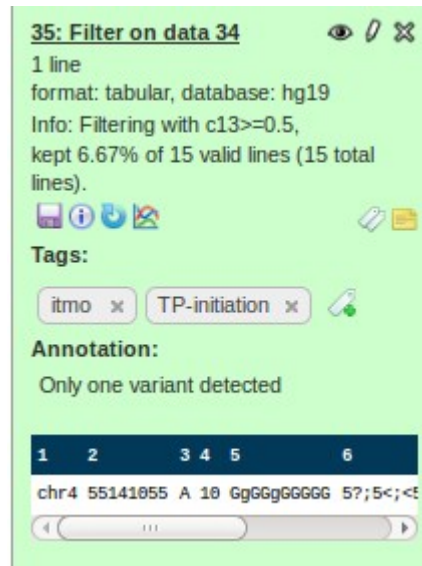
Reproducibility

- Galaxy captures **metadata**
- For each step in an analysis : input dataset, tools used, parameters values and output dataset
- With these metadata users can reproduce the analysis



Reproducibility

- But what about the **intent** of the analysis ?
- Use **annotations** and **tags** (c.f. web practices) to express the intent
- Annotations and tags = user metadata



35: Filter on data 34 👁 0 ✕

1 line
format: tabular, database: hg19
Info: Filtering with c13>=0.5,
kept 6.67% of 15 valid lines (15 total lines).

📄 ⓘ ↻ 🗑 📌 📄

Tags:

itmo ✕ TP-initiation ✕ 📌

Annotation:
Only one variant detected

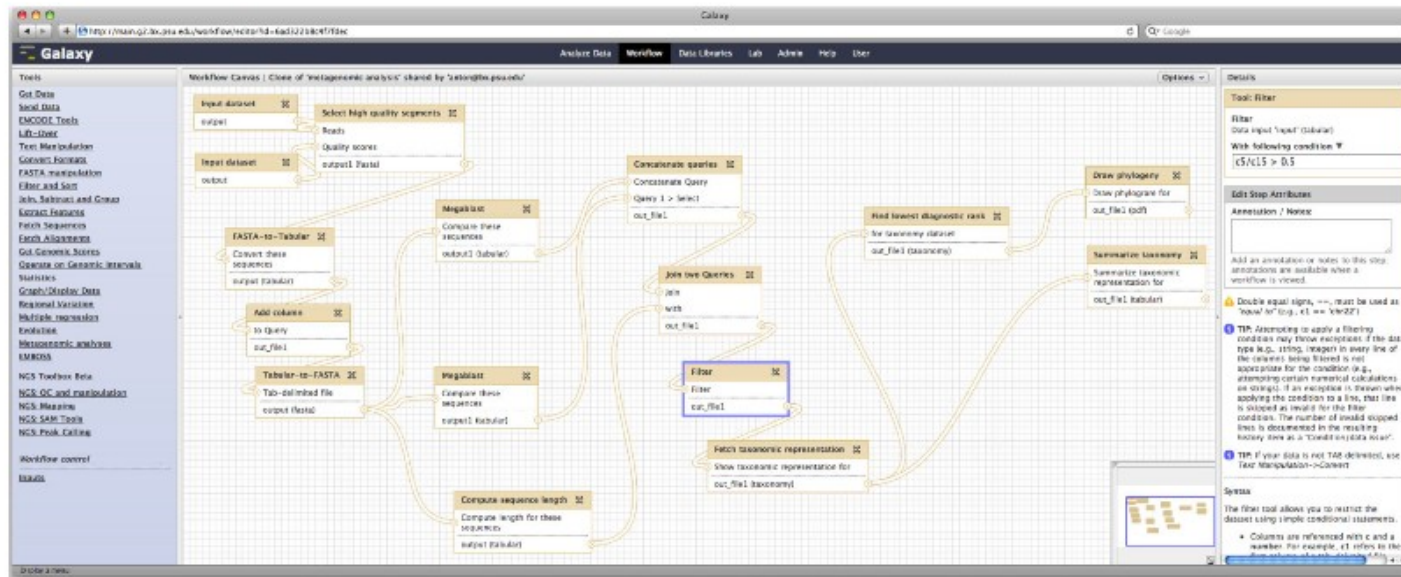
1	2	3	4	5	6
chr4	55141055	A	10	GgGGgGGGGG	5?;5<;<E

◀ ⋮ ▶

Galaxy Items (2/2)

- And ... if I want to reproduce the whole analysis ?
- Galaxy use **workflows**
- Create workflows from scratch, or create from history of your analysis

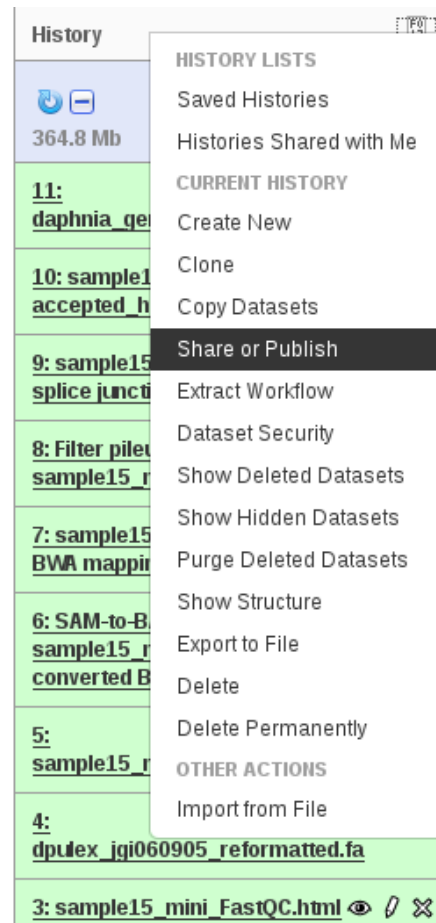
Workflow (example)



Transparency

- Transparency : enable user to share and communicate their experimental results and output
- **3** elements for Galaxy transparency
- **1** : Galaxy **sharing model** = sharing a Galaxy item* : dataset, histories, visualisation and workflows
- **2** : search shared item from **Galaxy Web Based framework**

Sharing model : example



Search shared item : example

Galaxy Analyze Data Workflow **Shared Data** Visualization

Published Histories

[Close Advanced Search](#)

name:

annotation:

owner:

community tags:

- Data Libraries
- Published Histories**
- Published Workflows
- Published Visualizations
- Published Pages

<u>Name</u>	<u>Annotation</u>	<u>Owner</u>
Naive v Memory for Patient 001D		meganesto
Dexamethasone		marpiech
human 22 chr SNPs		mvangala
ChIPseq example		tarandall
VGN FASTQ		jjv5
Databases		sr320
Unnamed history		huongle

Transparency

- **3 : Galaxy pages**
- Web based document that enable user to communicate their experieiment
- A mix of text and graph describing the experiment analysis
- embedded Galaxy items in the page used for the experiment
- Pages and Galaxy sharing model

Page (example)

Published Pages | [nchoisne](#) | [TP_MAPHITS_tutorial](#)













Welcome to MAPHiTS (Mapping Analysis Pipeline for High-Throughput Sequences) tutorial page.

In this page you will learn to use the tools of the MAPHiTS suite.

A little advice before starting : rename your results, choose explicitly filenames.

MAPHiTS is a pipeline developed for SNP discovery after mapping short-reads on a reference genome. This pipeline is currently running with the following public tools "BWA or Bowtie", "Samtools" and "VarScan". The input data files are : a fasta file for the reference genome (Genome.fasta) and 2 fastq files of short-reads sequenced in paired-ends and corresponding to the forward (SR_1.fastq) and the reverse (SR_2.fastq) sequences.

Import "input data" in your current history:

	Galaxy Dataset Genome.fasta	  
	Galaxy Dataset SR_2.fastq	  
	Galaxy Dataset SR_1.fastq	  

Rename your datasets : select "Edit Attributes"

- Genome.fasta
- SR_1.fastq (1250 sequences) => **forward**
- SR_2.fastq (1250 sequences) => **reverse**

Embedded Galaxy item (example)

Import "input data" in your current history:

Galaxy Dataset | Genome.fasta

This dataset is large and only the first megabyte is shown below. | [Show all](#)

```
>C10HBa0111D09 LR276 15142 24441 |Longueur=9300
GAACAAACAACCCCTTTTTGGAGGTGTTGGCGCGTCGTGCAGCTTACACTCAAAGTTAA
AAAGTTGCCTTGCGATGCGGTCATGTTACAAACCTCTCTGCCTTAAATTAAATTCCATAA
CCAAGATTTGGAGGTGCCTCAACGATGCGCAGCCATGTCCCATATTTGGTCGCCTCGTTT
AAAAGTCAAGTTAGACTTAATTAAGAGGTCCAAC TAGTGTAGGGGCGTTTTGAGTACTTG
TGGGATTTATTATAAACGGTTTTGAGTCACTTTAAACCCACTTCACCAATTA AAACAAA
TCCTCAAGTTAAACTCAATATCTTTCCATTCTCTCTCTCTAAAACCTTCATTGGAGATA
TTTGAAGCTCCACGGAAGAAGGTTAATTTTCCAAGGTTTCAATGAAAATTTCGTGTATAG
GTCTTCAATAAGGTATGGTGATTTATCCTTGATTCTTCTATCATTCAAGGATCCAATTC
AAAGGTTTTTCAAAGATCTCAAAAATCCTATTTCGAATTCTAAGTATGGGTTCTTCCAT
TTAAAGGTTTAAATGGATGAATTATGATGTTTTCAATGTTAGTTGATGTTTTTATGATAA
AAAAACTCCATGAACCCATGAGCATCCTAATTCTCTAATTTTGTCTTGTAATTGAGTTT
GATAATTGTGATTGGTTATGGATGGAATTGTATTTAGATTGCTCTATATTGTTGATTCTT
ATTGTTAACCTATCTCTATATATGTAGAATTGAGATTGTAAGGATGAGTTAGTAATCTTG
CTTTTATCGCCCTTTCAATTCGCGCTTTAAGCCCTGCACTTAAGCCCATCTCTCGCCCTTT
```

Galaxy Dataset | SR 2.fastq

Galaxy Dataset | SR 1.fastq

Rename your datasets : select "Edit Attributes"

References and links

- Galaxy Project home page : <http://galaxyproject.org/>
 - Use galaxy : galaxy-central, a free public server
 - Get a galaxy distribution
 - Learn galaxy : tutorials, screencast
 - Get involved : mailing lists and wiki
- Next-generation sequencing and data interpretation : enhancing reproductibility and accessibility. Anton Nekrutenko ; James Taylor – 2012 – Nature Review Genetics.
- Galaxy : a comprehensive approach for supporting accessible, reproductible and transparent computational research in life science. Jeremy Goecks *et al.* - 2010 – Genome Biology