Introduction à Galaxy

Olivier Inizan
Alban Lermine
Ecole Bioinformatique de Roscoff
18 novembre 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

• 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!

Next-generation sequencing data interpretation: enhancing reproducibility and accessibility

Anton Nekrutenko and James Taylor

- 1000 Genomes Project: a series 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 reproducibility
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 ressources ?
Integrative ressources

- Integratives ressources, integrative frameworks: bring together diverse tools under the umbrella of unified interface
- BioExtract, GenePattern, GeneProf, Mobyle
- 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

Provide a unified, web based interface for bioinformatics analysis
Galaxy Items (1 /2)
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 developper 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
Reproductibility

- 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
Reproductibility

- 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
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

### Published Histories

<table>
<thead>
<tr>
<th>Name</th>
<th>Annotation</th>
<th>Owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive v. Memory for Patient 001D</td>
<td></td>
<td>meganesto</td>
</tr>
<tr>
<td>Dexamethasone</td>
<td></td>
<td>marcpech</td>
</tr>
<tr>
<td>human 22 chr SNPs</td>
<td></td>
<td>mvecala</td>
</tr>
<tr>
<td>ChIPseq example</td>
<td></td>
<td>larandall</td>
</tr>
<tr>
<td>VGN FASTQ</td>
<td></td>
<td>jv5</td>
</tr>
<tr>
<td>Databases</td>
<td></td>
<td>sj320</td>
</tr>
<tr>
<td>Unnamed History</td>
<td></td>
<td>huongie</td>
</tr>
</tbody>
</table>
Transparency

• 3 : Galaxy pages
• Web based document that enable user to communicate their experiment
• 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
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)
References and links

  - Use galaxy: galaxy-central, a free public server
  - Get a galaxy distribution
  - Learn galaxy: tutorials, screencast
  - Get involved: mailing lists and wiki


- Galaxy: a comprehensive approach for supporting accessible, reproducible and transparent computational research in life science. Jeremy Goecks et al. - 2010 – Genome Biology