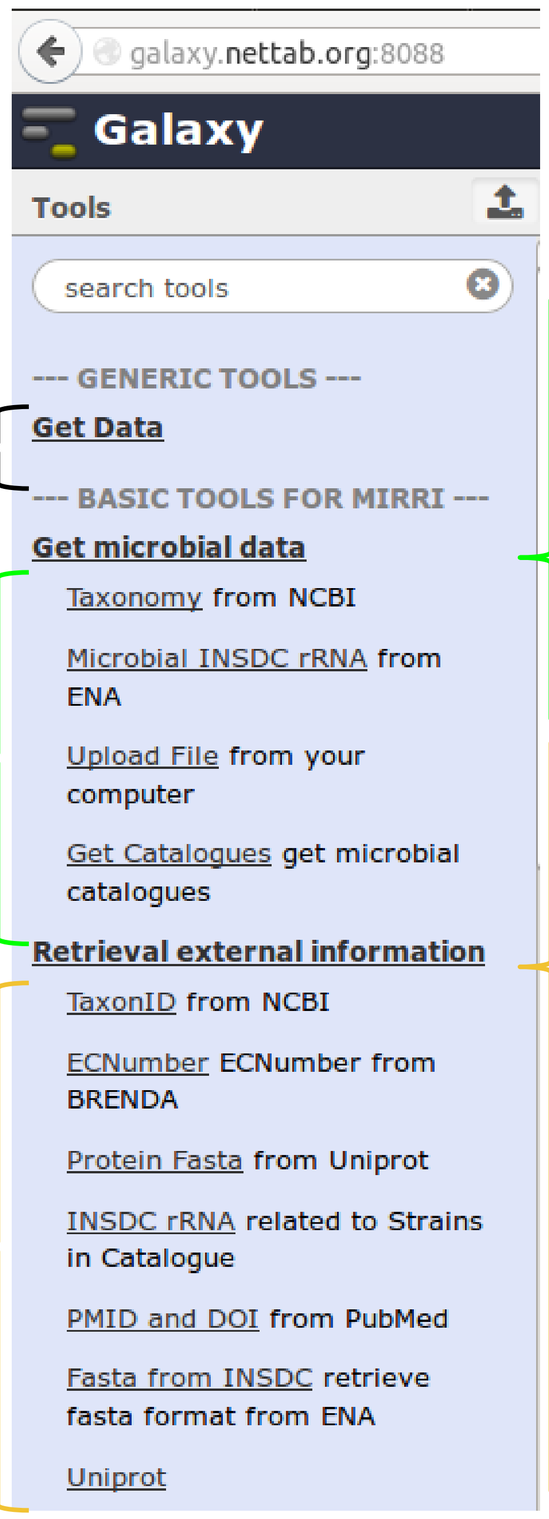
A Galaxy approach to integrate microbial data: the USMI Galaxy demonstrator

D.P. Colobraro, P. Romano
IRCCS AOU San Martino IST, Genova, Italy
{danielepierpaolo.colobraro,paolo.romano}@hsanmartino.it

Galaxy is an open, web-based platform which sorts many tools to retrieve, manage and analyze different kind of information that come from whole life sciences. It provides an intensive start-point to integrate microbiological data and to use them for integrative researches or analyses. A **Galaxy** approach leads to the possibility implementing various tools that allow to researchers and industry communities to perform analyses starting from existing data or to retrieve information about new entries in their microorganism catalogs. **Galaxy**, due to its flexibility and power, can be installed in a local workstation or shared via installation in own web server.



Galaxy' tools: to import and upload various type of data

Generic-developed tools: to import both specific information on microorganisms and catalogues and not specific data.

Specific-developed tools: to gather microorganisms information on:
- *TaxonId*
- *ECNumber*
- *Protein*
- *Sequence of 16S rRNA*
- *PMID and DOI*

Figure 1

***Taxonomy** retrieves all taxonomy information*
***Upload file** is a Galaxy' generic tool*
***Microbial INSDC rRNA** retrieves information by using a Catalogue acronym*
***Get Catalogues** is a 'data_source' tool to import catalogues from external-web storage*
***TaxonID** retrieves taxonomy ID for all Strain in Catalogues*
***ECNumber** gathers information if enzyme name are collected in Catalogue*
***Protein Fasta** retrieves fasta by using protein accession number*
***INSDC rRNA** retrieves INSDCs related to Strains in Catalogue*
***Fasta from INSDC** retrieves fasta by using INSDC*
***PMID and DOI** retrieves Pubmed IDs and Digital Object Identifiers (DOIs) of given bibliographic references*
***Uniprot** retrieves protein accession number related to the given strain*

CABRI, Common access to biological resource and information
Since 2000, CABRI Network Services (http://www.cabri.org/) offer access to 28 catalogues from European Biological Resources Centers (BRCs).
MIRRI Microbial Resource Research Infrastructure, is a pan-European distributed research infrastructure in its preparatory phase which aims to connect all European **mBRCs,** microBiological Resource Centres, with the aim of providing improved and extended services to the research and industry communities. **MIRRI** requires to integrate information on microorganisms with further data that can be found and retrieved from a wide range of biological resources like NCBI, EMBL, BRENDA and UNIPROT.

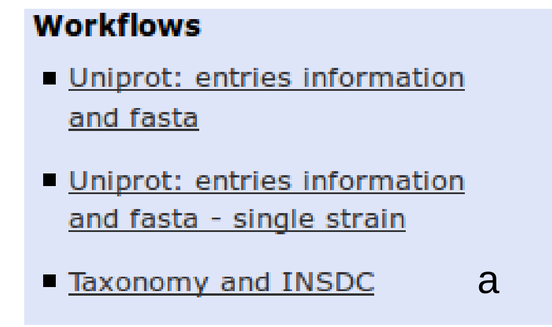
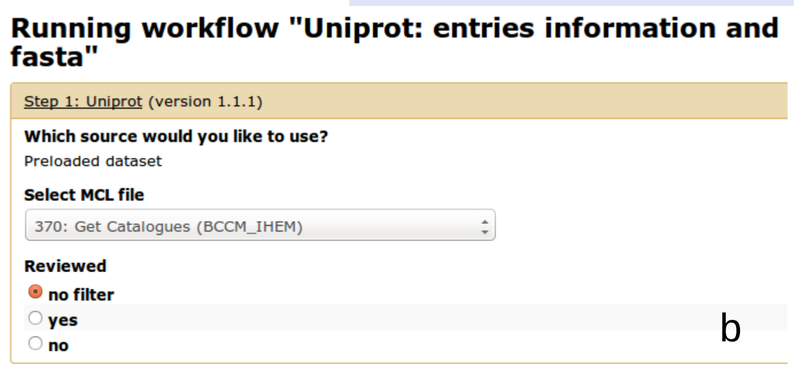
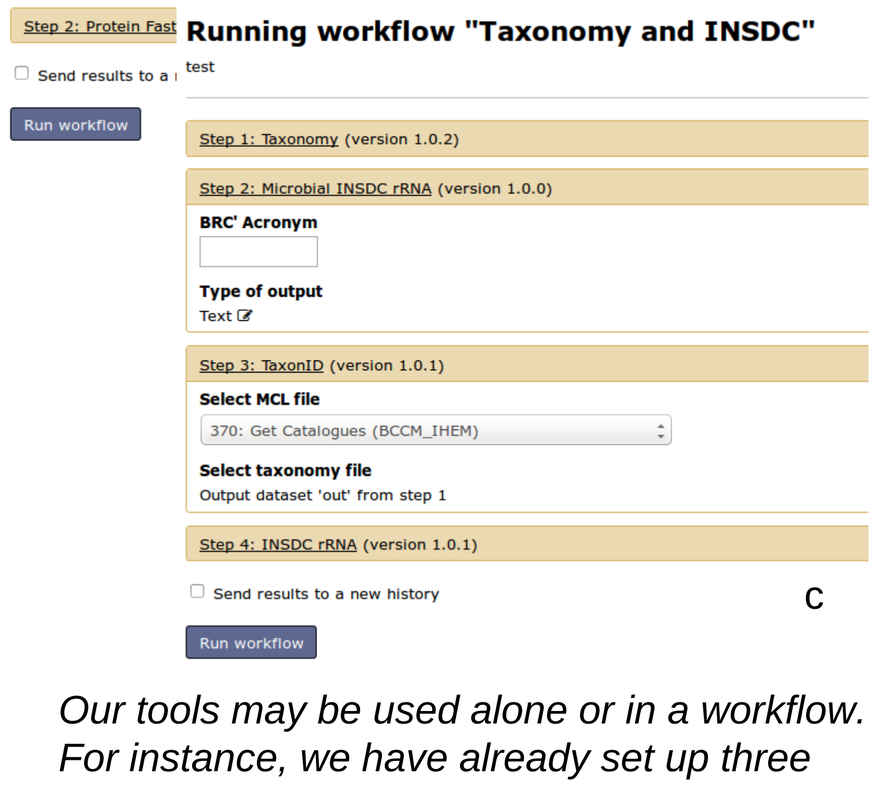


Figure 2

a



b



c

Microbial Catalogues sometimes gather metabolites, genomics, proteomics and taxonomy data, although those information characterize and validate the microbial species that are collected into mBRC. So, our **purposes** are i. to find a method to merge all microbiological source, ii. to offer a clear vision of microorganism data, iii. to aid curators of catalogues to improve annotations, iv. to make available this data to pipelines of analyse and v. to integrate information.

Galaxy version 2014.10.06 are publicly available on-line at http://galaxy.nettab.org:8088. As shown in fig. 1, The developed tools are available in two section, *Get microbial data* and *Retrieval external information*, under the general label 'BASIC TOOLS FOR MIRRI'. Galaxy allows to set up workflows to rerun, record, store and share both specific analyses and import data. As shown in fig. 2a-c, tools may be set in various ways in order to define a own pipeline. Indeed, changing tools as modular elements allows to make up several pipelines.

Our tools may be used alone or in a workflow. For instance, we have already set up three workflows that allow, in the first two cases (b), to retrieve uniprot accession number and fasta format related to both each strain into given Catalogue file and for a single strain.
In the last example (c), for a given microbial catalogue, the workflow returns the taxonomy and every 16S rRNA sequence related to each strain into given catalogue file.

References
- The MIRRI Project: www.mirri.org
- The Galaxy project: http://usegalaxy.org
- Blankenberg D et al., Database, vol. 2011. doi:10.1093/database/bar011

Acknowledgements
The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 312251.