Galaxy & XCMS

Reproductible metabolomics data analysis

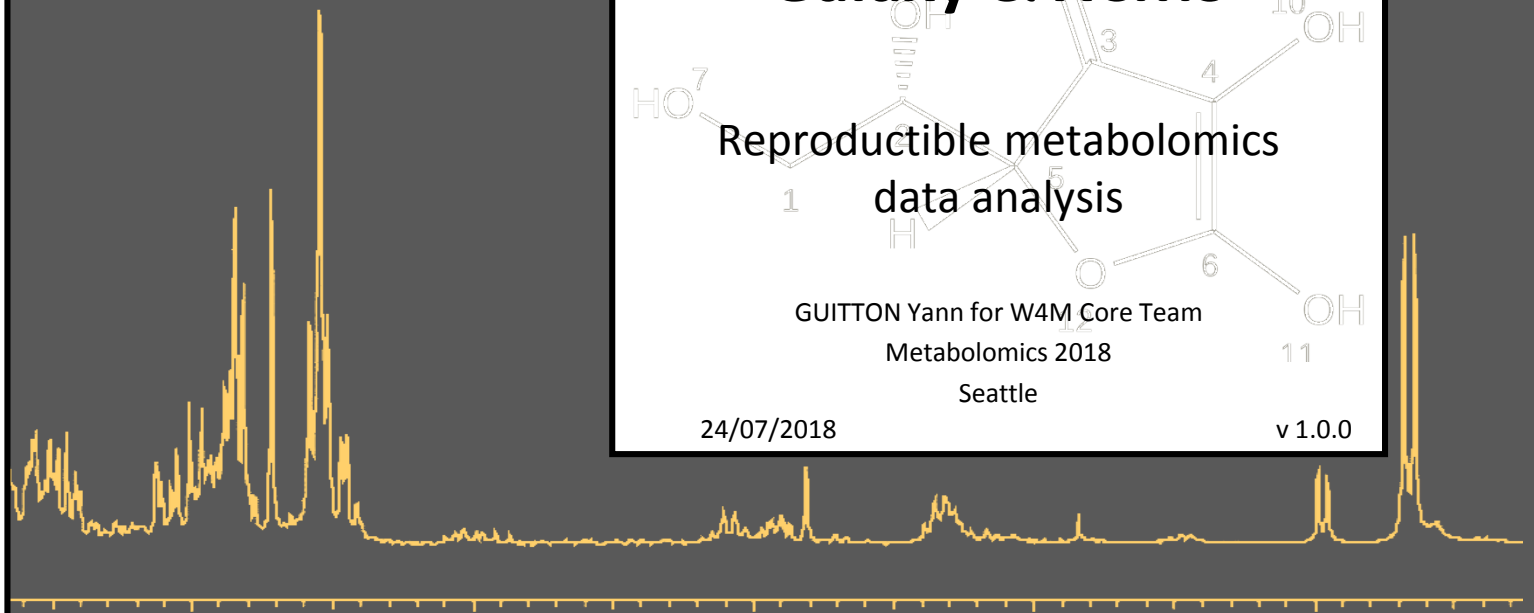
GUITTON Yann for W4M Core Team

Metabolomics 2018

Seattle

24/07/2018

v 1.0.0



Presentation

LABERCA

Food Chemical Safety

MS platform (17 instruments)

- Metabolomics
- Bioinformatics (W4M)

W4M a user story born from a simple questions

XCMS looks perfect BUT please, HELP
what's a script is…?

Sophie Goulitquer –
Chemist on a MS platform

W4M a user story born from a simple questions

Bio-informatics will save my life,
GREAT !

W4M a user story born from a simple questions

Galaxy

…a solution for ME?

Introduction / Galaxy

- ***« Galaxy is a scientific workflow, data integration, and data and analysis persistence and publishing platform that aims to make computational biology accessible to research scientists that do not have computer programming experience. »***
- **http://galaxyproject.org/**

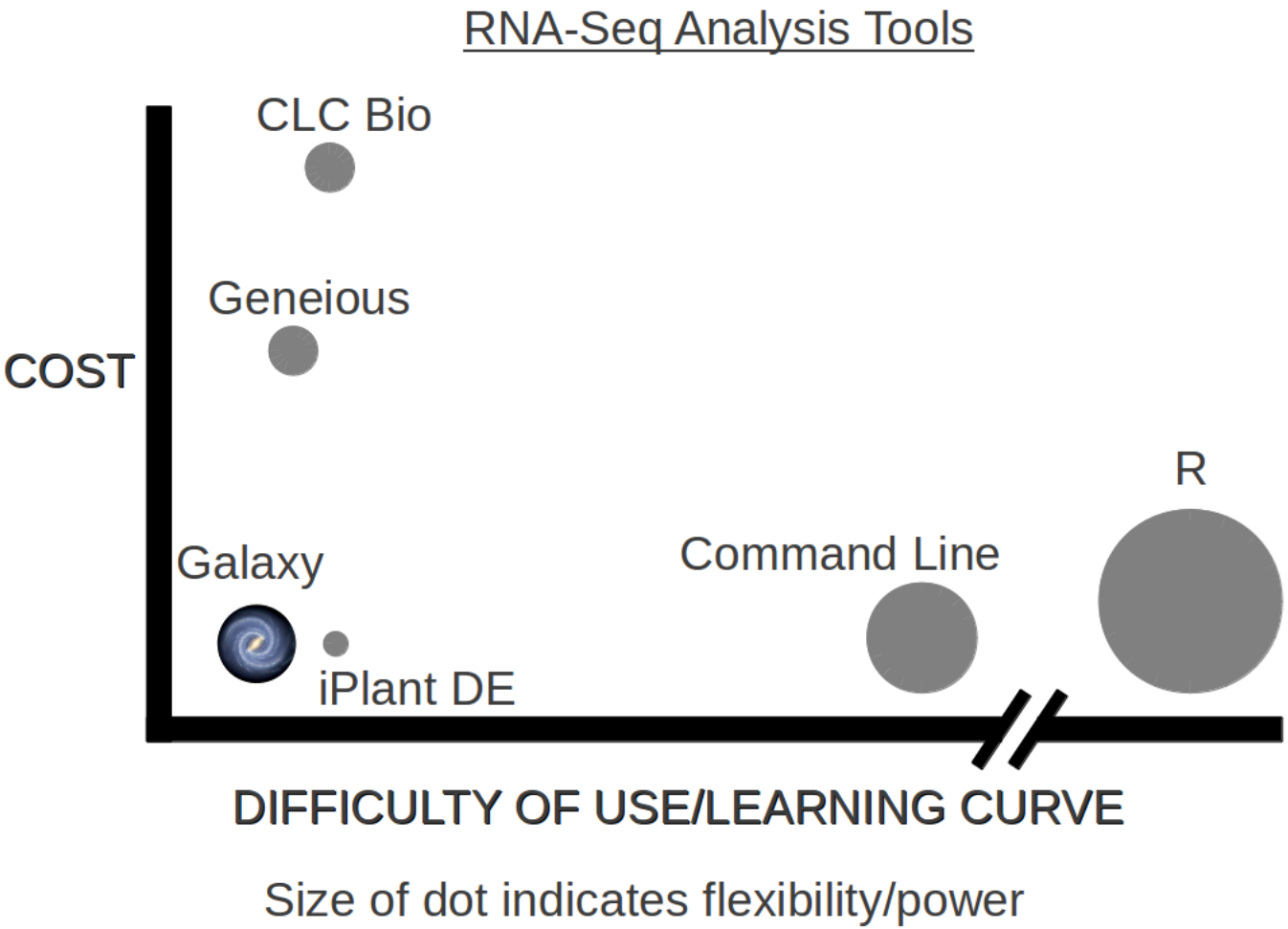
- **Applications suite for 'OMIC' data analysis.**

Goecks, J.; Nekrutenko, A.; Taylor, J.; Galaxy Team, T. (2010). "Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences ». Genome Biology 11 (8): R86. doi:10.1186/gb-2010-11-8-r86

Introduction / Galaxy

- Galaxy it's …
  - No need to execute a command line through a terminal
  - Programming or scripting skills are not required
  - Submission of jobs is transparent through a high performance computer cluster
  - Secure histories and data manager
  - A data and protocols sharing system
  - Tool-boxes of several bioinformatics fields
    - NGS
    - Metabolomics
    - Statistics
    - Chemistry
    - Image analysis
    - Etc ...
  - A web-based interface

Introduction / Galaxy



Introduction / Galaxy

Why Galaxy ?

- Accessibility
- Reproductibility
- Transparency

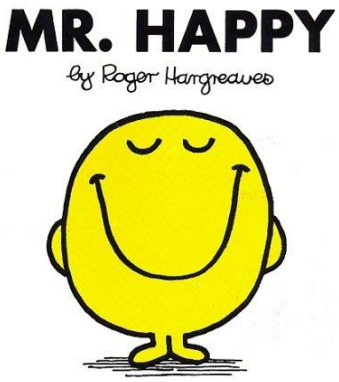
Introduction / Galaxy

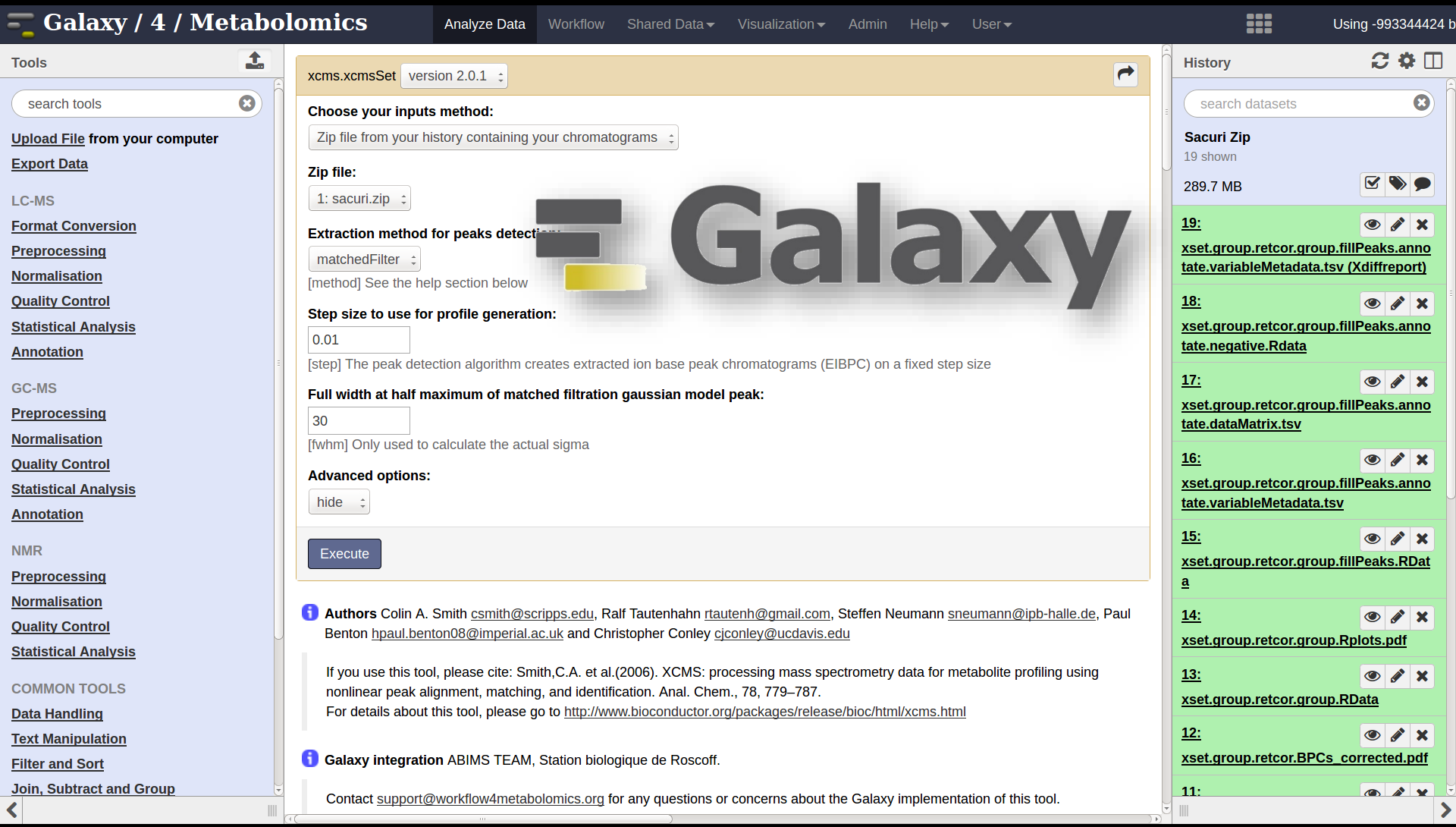
MR. GEEK

```
[login@n0 ~]$ cdprojet
[login@n0 login]$ cd 13-07-29-panda/tmp/mapping
[login@n0 mapping]$ cat tophat.qsub
#!/bin/bash
#$ -S /bin/bash
#$ -M login@sb-roscoff.fr
#$ -m bea
#$ -V
#$ -cwd
#$ -o qsub.out
#$ -e qsub.err

tophat2 panda_v121029 ../input/IllR1-1.fq ../input/IllR1-2.fq
-GTF ../input/panda_v121029.gtf --b2-sensitive -r 100
-num-threads 8

[login@n0 mapping]$ qsub -q long.q -pe thread 8 tophat.qsub
Your job 5338969 ("tophat.qsub") has been submitted
```

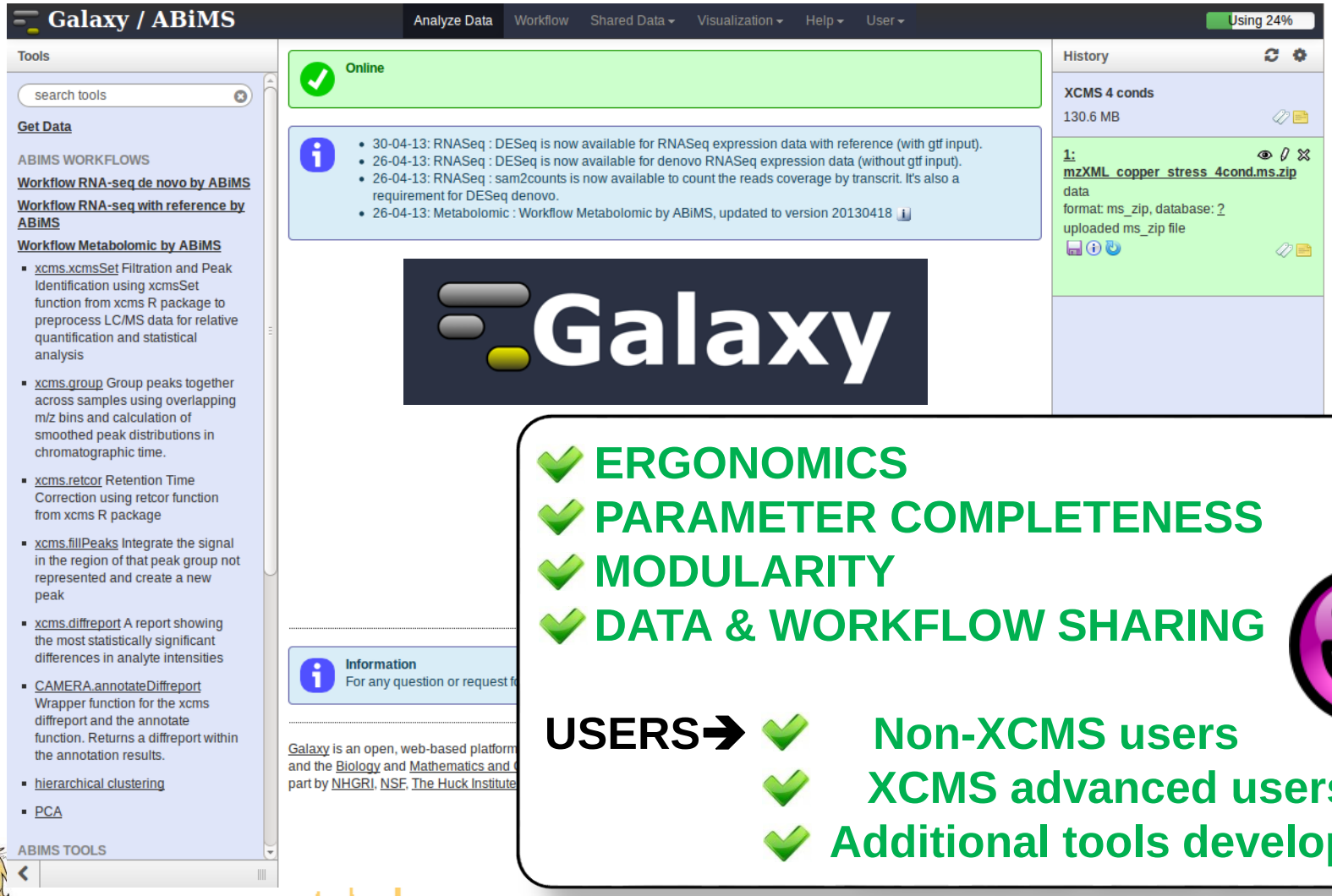
Introduction / Galaxy

```
[login@n0 ~]$ cdprojet
[login@n0 login]$ cd 13-07-29-panda/tmp/mapping
[login@n0 mapping]$ cat tophat.qsub
#!/bin/bash
#$ -S /bin/bash
#$ -M login@sb-roscoff.fr
#$ -m bea
#$ -V
#$ -cwd
#$ -o qsub.out
#$ -e qsub.err

tophat2 panda_v121029 ../input/Ill1R1-1.fq ../input/Ill1R1-2.fq
-GTF ../input/panda_v121029.gtf --b2-sensitive -r 100
-num-threads 8

[login@n0 mapping]$ qsub -q long.q -pe thread 8 tophat.qsub
Your job 5338969 ("tophat.qsub") has been submitted
```




Galaxy / 4 / Metabolomics
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using -993344424 b
Tools
search tools
Upload File from your computer
Export Data
LC-MS
Format Conversion
Preprocessing
Normalisation
Quality Control
Statistical Analysis
Annotation
GC-MS
Preprocessing
Normalisation
Quality Control
Statistical Analysis
Annotation
NMR
Preprocessing
Normalisation
Quality Control
Statistical Analysis
COMMON TOOLS
Data Handling
Text Manipulation
Filter and Sort
Join, Subtract and Group
xcms.xcmsSet
version 2.0.1
Choose your inputs method:
Zip file from your history containing your chromatograms
Zip file:
1: sacuri.zip
Extraction method for peaks detection:
matchedFilter
[method] See the help section below
Step size to use for profile generation:
0.01
[step] The peak detection algorithm creates extracted ion base peak chromatograms (EIBPC) on a fixed step size
Full width at half maximum of matched filtration gaussian model peak:
30
[fwhm] Only used to calculate the actual sigma
Advanced options:
hide
Execute
Authors Colin A. Smith csmith@scripps.edu, Ralf Tautenhahn rtautenh@gmail.com, Steffen Neumann sneumann@ipb-halle.de, Paul Benton hpaul.benton08@imperial.ac.uk and Christopher Conley cjconley@ucdavis.edu
If you use this tool, please cite: Smith,C.A. et al.(2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal. Chem., 78, 779–787.
For details about this tool, please go to http://www.bioconductor.org/packages/release/bioc/html/xcms.html
Galaxy integration ABIMS TEAM, Station biologique de Roscoff.
Contact support@workflow4metabolomics.org for any questions or concerns about the Galaxy implementation of this tool.
History
search datasets
Sacuri Zip
19 shown
289.7 MB
19: xset.group.retcor.group.fillPeaks.annotate.variableMetadata.tsv (Xdiffreport)
18: xset.group.retcor.group.fillPeaks.annotate.negative.Rdata
17: xset.group.retcor.group.fillPeaks.annotate.dataMatrix.tsv
16: xset.group.retcor.group.fillPeaks.annotate.variableMetadata.tsv
15: xset.group.retcor.group.fillPeaks.RData
14: xset.group.retcor.group.Rplots.pdf
13: xset.group.retcor.group.RData
12: xset.group.retcor.BPCs_corrected.pdf
11:

W4M / Galaxy our solution



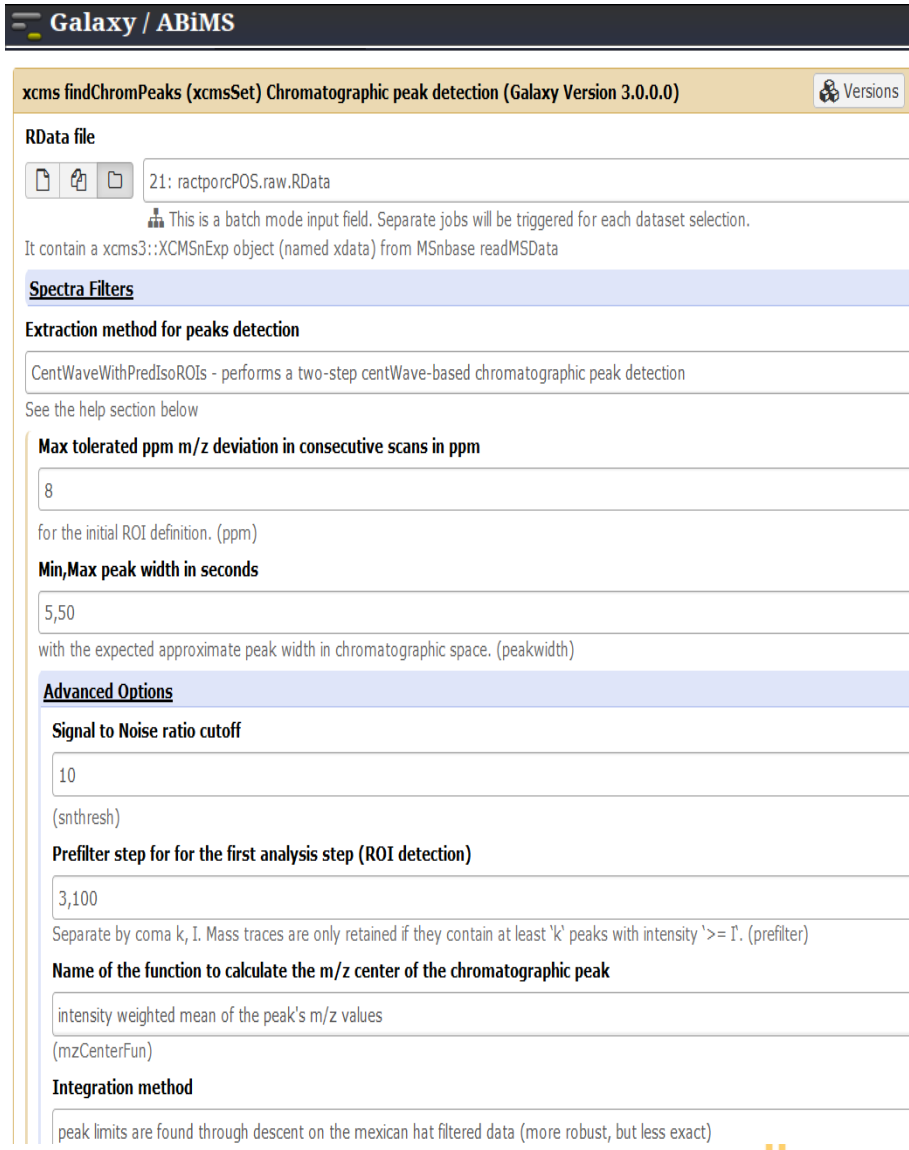
- ERGONOMICS
- PARAMETER COMPLETENESS
- MODULARITY
- DATA & WORKFLOW SHARING

USERS →

- Non-XCMS users
- XCMS advanced users
- Additional tools developers

Introduction / Galaxy

```
[lecorguille@n0 ~]$

object

xcmsSet object
ppm

maxmial tolerated m/z deviation in consecutive scans,
in ppm (parts per million)
peakwidth

Chromatographic peak width, given as range (min,max)
in seconds
snthresh

signal to noise ratio cutoff, definition see below.
prefilter

prefilter=c(k,I). Prefilter step for the first phase.
Mass traces are only retained if they contain at least
k peaks with intensity >= I.

firstBaselineCheck

logical, if TRUE continuous data within ROI is checked
to be above 1st baseline
roiScales

numeric, optional vector of scales for each ROI in
ROI.list to be used for the centWave-wavelet



[...]
```



W4M in the Metabolomics Workflow

Annotation

Preprocessing

Normalization

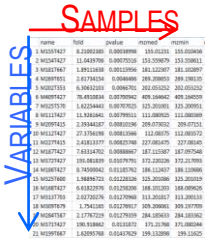
Statistics

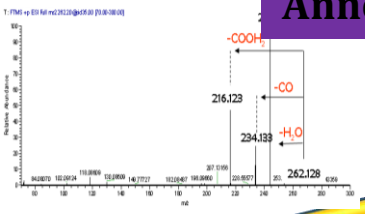
Sharing

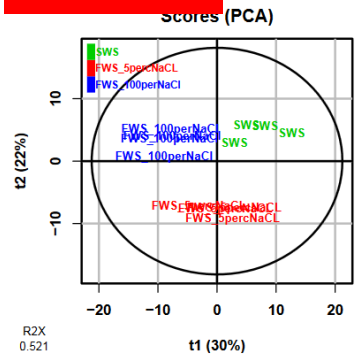
Sample preparation

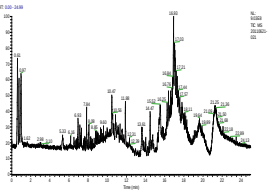
Sample analysis

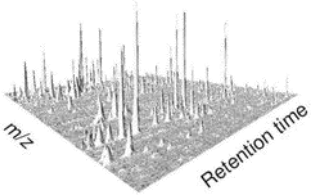
Experimental design







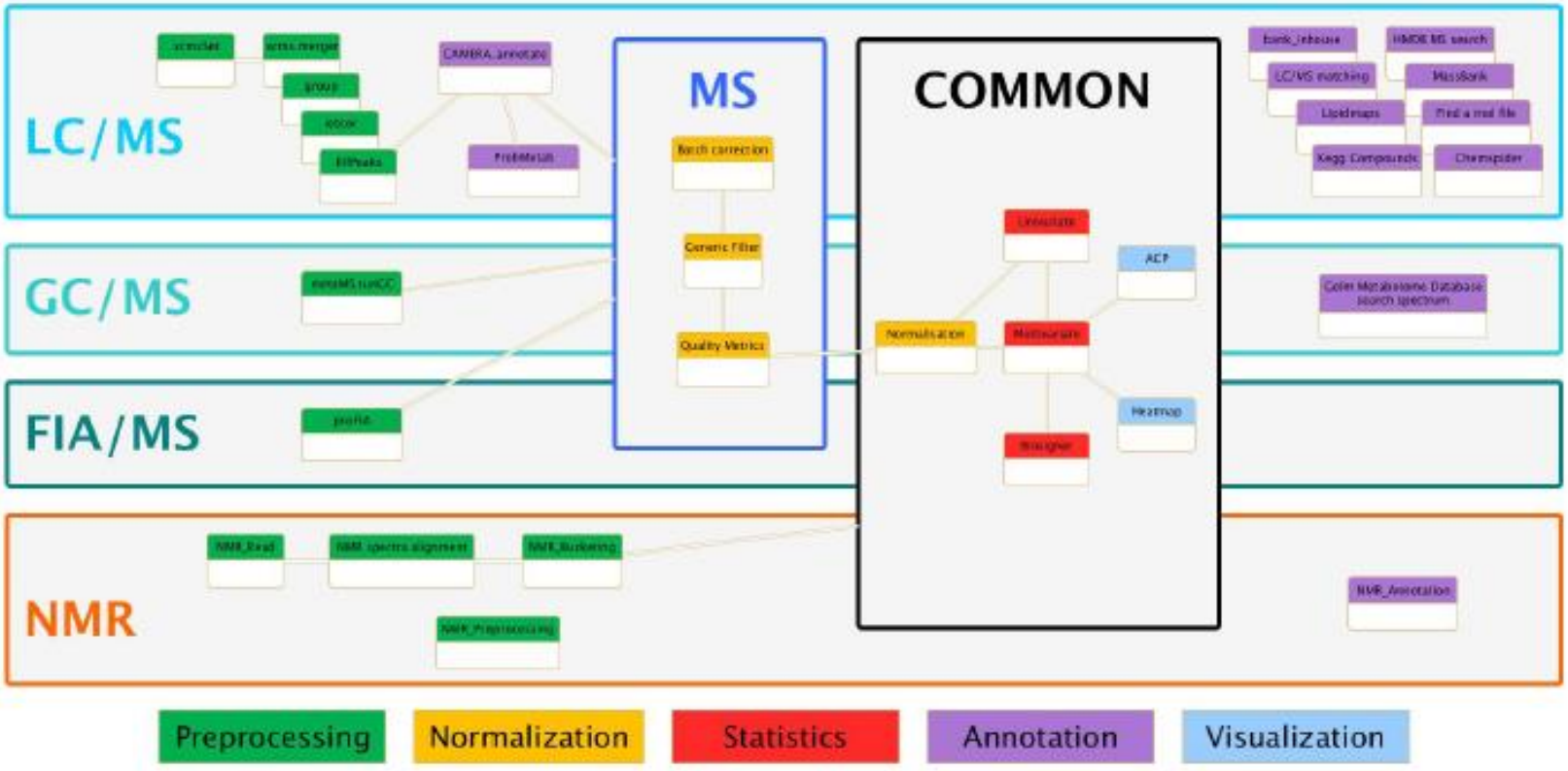




LCMS-GCMS-FIAMS and NMR

W4M tools

- 41 modules and growing



+ NEW Reference your workflows with DOI

DataCite

Sharing scientific data

A prerequisite for repeatability is for the research artifacts that back up the published results to be shared.

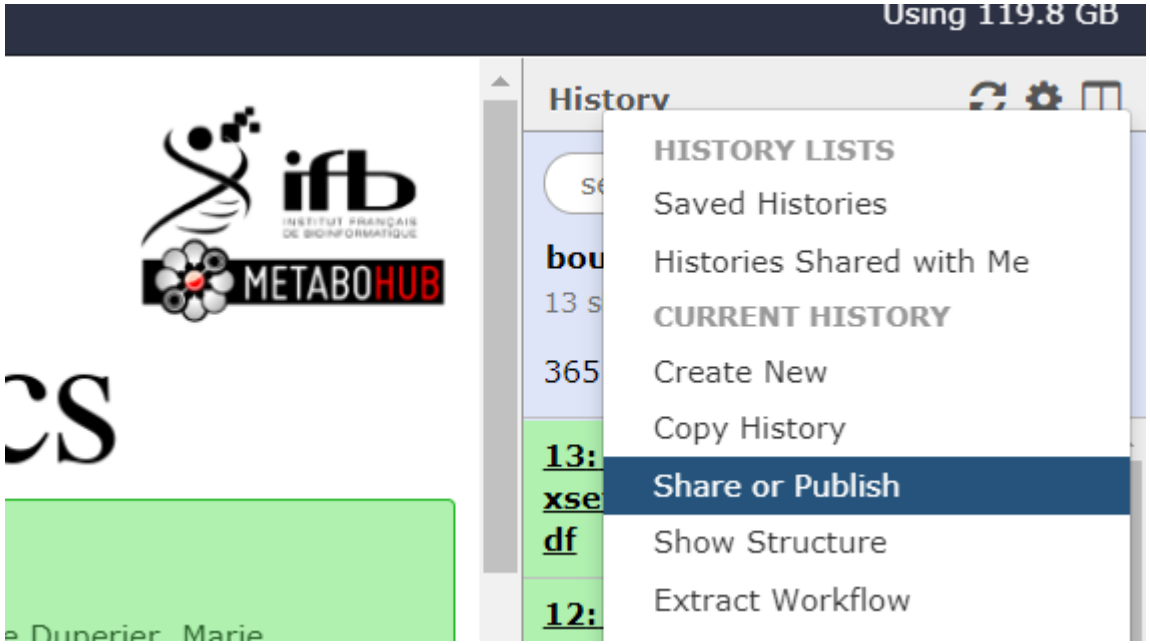
Sharing for repeatability is essential to ensure that **colleagues** and **reviewers** can evaluate our results based on accurate and complete evidence.

Data FAIRport

Find, Access, Interoperate & Re-use Data

Sharing scientific data

Sharing Data/Histories/Workflows with Galaxy



Sharing scientific data

Sharing Data/Histories/Workflows with Galaxy

Share or Publish History

Make History Accessible via Link and Publish It

This history is currently restricted so that only you and the users listed below can acces

Make History Accessible via Link

Generates a web link that you can share with other people so that they can view and in

Make History Accessible and Publish

Makes the history accessible via link (see above) and publishes the history to Galaxy's I

Share History with Individual Users

You have not shared this history with any users.

Share with a user

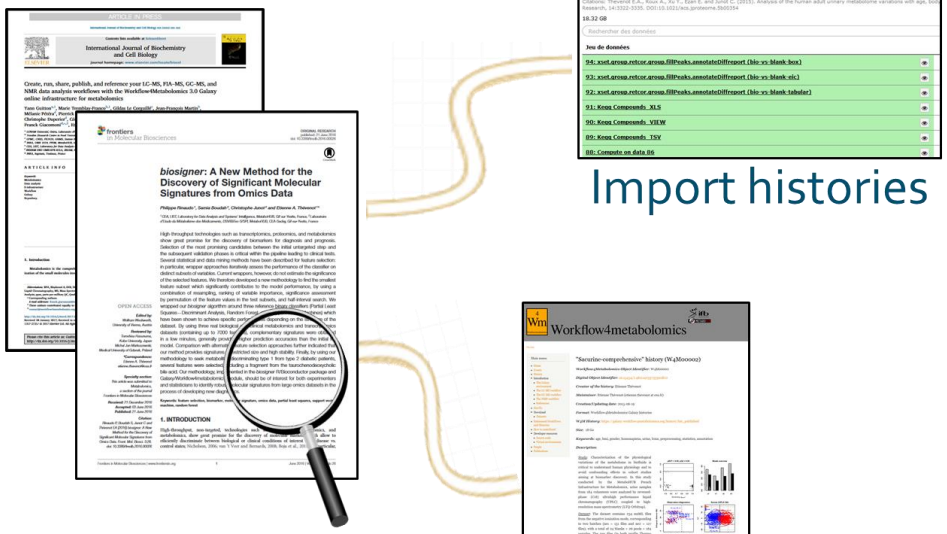
Back to Histories List

Sharing scientific data

Sharing Data/Histories/Workflows with Galaxy

W4M can provide permanent **DOI** usable in papers to reference histories and workflows.

DataCite
FIND, ACCESS, AND REUSE DATA



Import histories

Dedicated webpage

Sharing scientific data

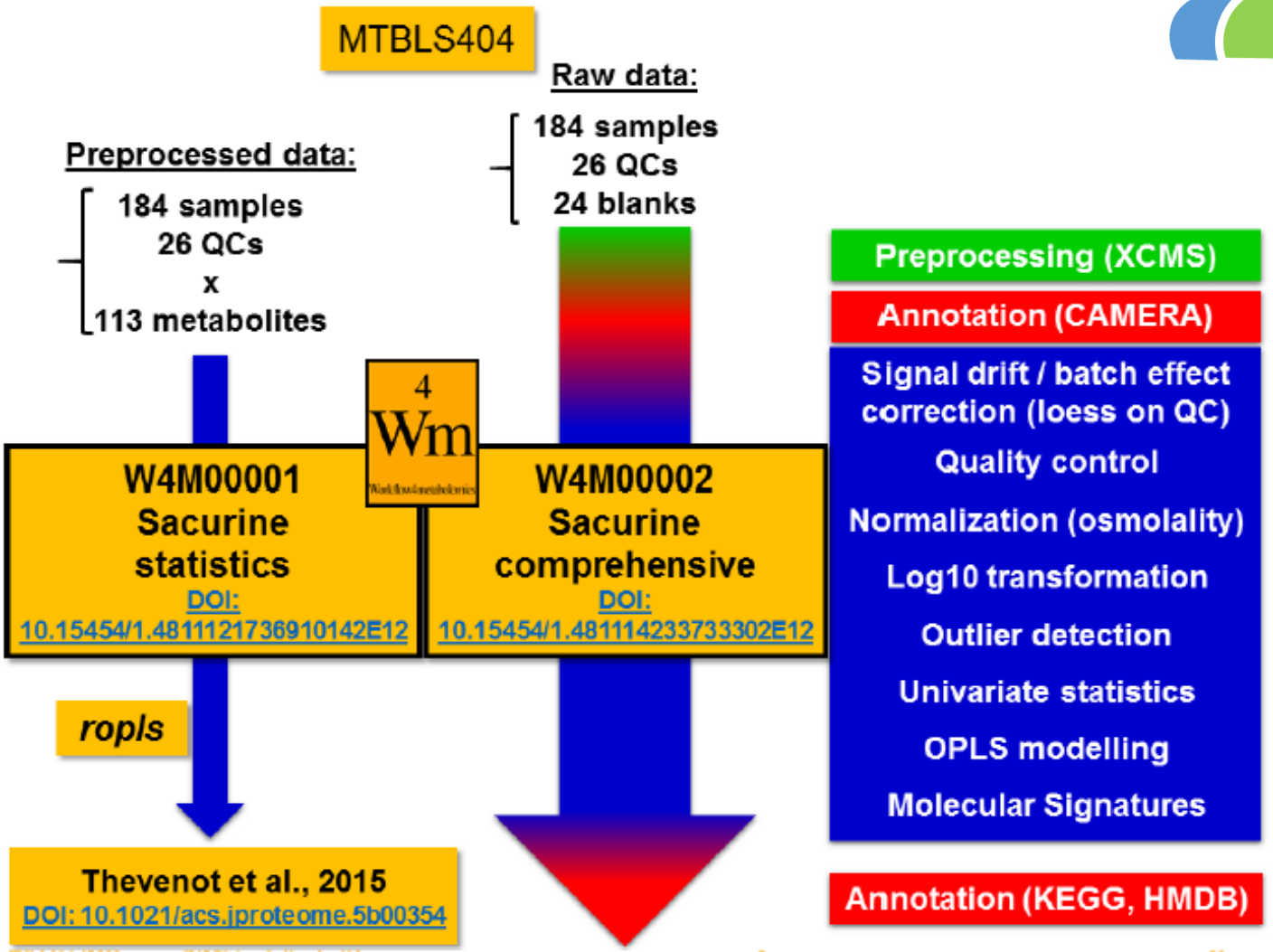
Sharing Data/Histories/Workflows with Galaxy

Referenced W4M histories

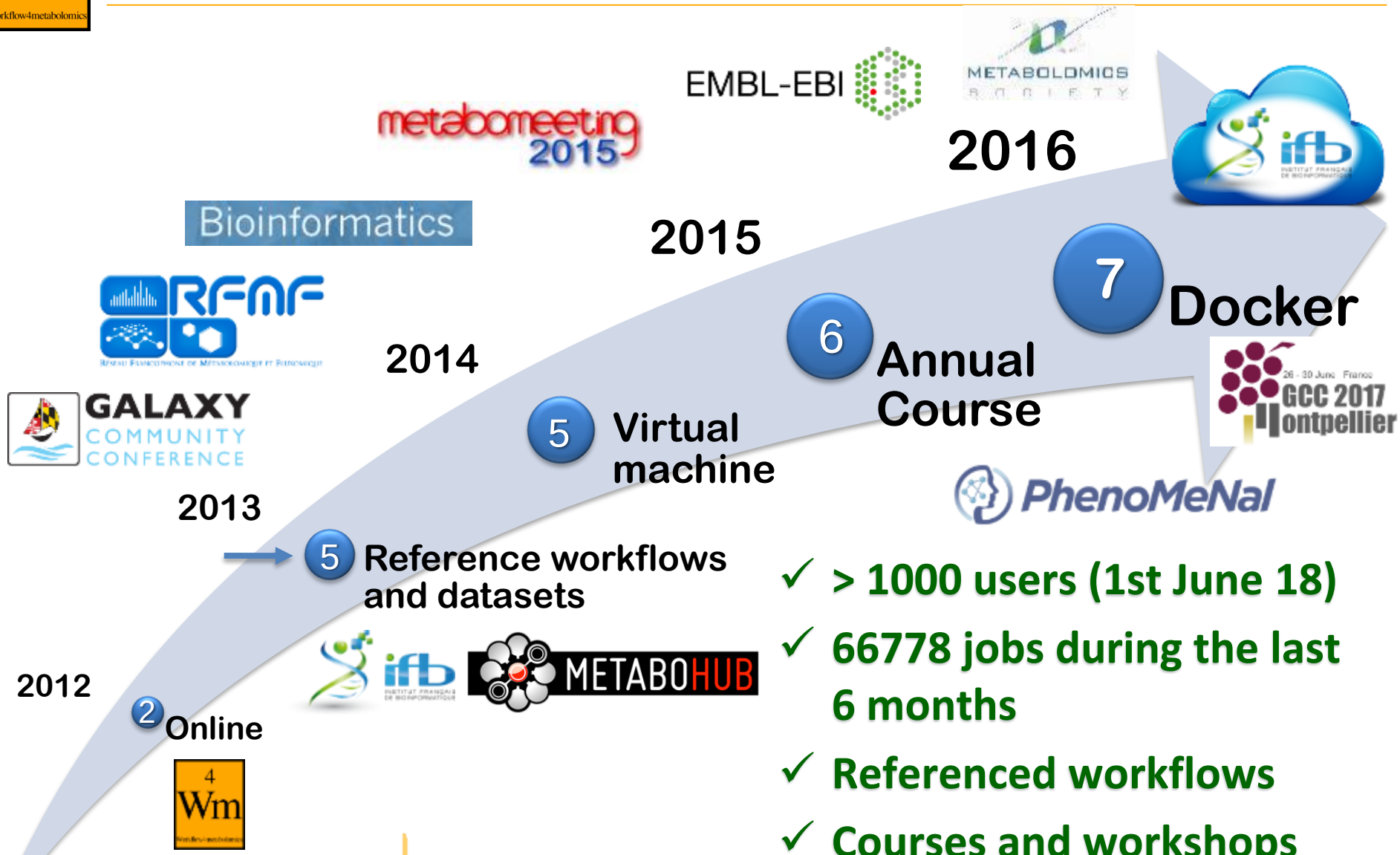
WOI	Name & DOI	Technology	Species	Matrice	Factor	Samples
W4M00001	*"Sacurine-statistics"* *10.15454/1.4811121736910142E12*	LC-MS	*H. sapiens*	Urine	age, BMI, gender	184
W4M00002	*"Sacurine-comprehensive"* *10.15454/1.481114233733302E12*	LC-MS	*H. sapiens*	Urine	age, BMI, gender	184
W4M00003	*"Diaplasma"* *10.15454/1.4811165052113186E12*	LC-MS	*H. sapiens*	Plasma	diabetic type 2	69
W4M00004	*"GCMS Algae"* *10.15454/1.4811272313071519E12*	GC-MS	*E. siliculosus*	Algae	Salinity	12
W4M00005	*"Ractopamine"* *10.15454/1.4811287270056958E12*	LC-MS	*S. scrofa*	Serum	Ractopamine	124
W4M00006	*"BPA-MMusculus"* *10.15454/1.4821558812795176E12*	NMR	*M. musculus*	Brain	BPA	24
W4M00007	*"Coffea leaves"* *10.15454/1.4985472277740251E12*	LCMS	Coffea sp.	Leaves	N/A	169

Sharing scientific data

Sharing Data/Histories/Workflows with Galaxy

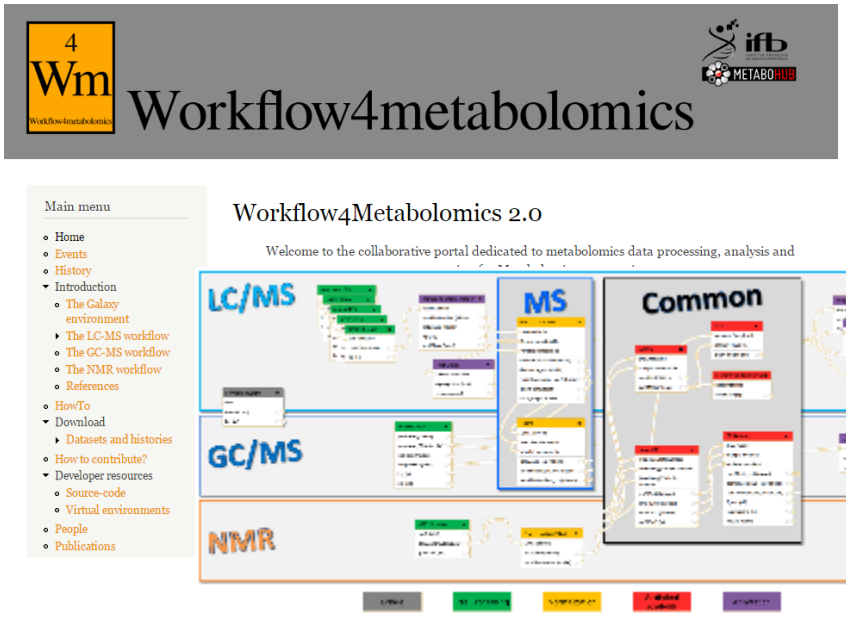


W4M a long term infrastructure



- ✓ > 1000 users (1st June 18)
- ✓ 66778 jobs during the last 6 months
- ✓ Referenced workflows
- ✓ Courses and workshops

W4M Core Team (The Guys Behind the scene)



- **41 modules** for LCMS, GCMS, FIAMS and NMR preprocessing, statistics and annotation
- Advanced functionalities for **workflow management**
- **Help desk**
  - 15 bioinformaticians
  - 7 platforms
- **HPC** environment

Giacomoni et al. (2015). *Bioinformatics*, 31:1493-1495
Guitton et al. (2017). *IJBC*, 93:89-101

Try it

- Just connect to workflow4metabolomics.org (ask for free account)

  Generic account available during Metabolomics2018

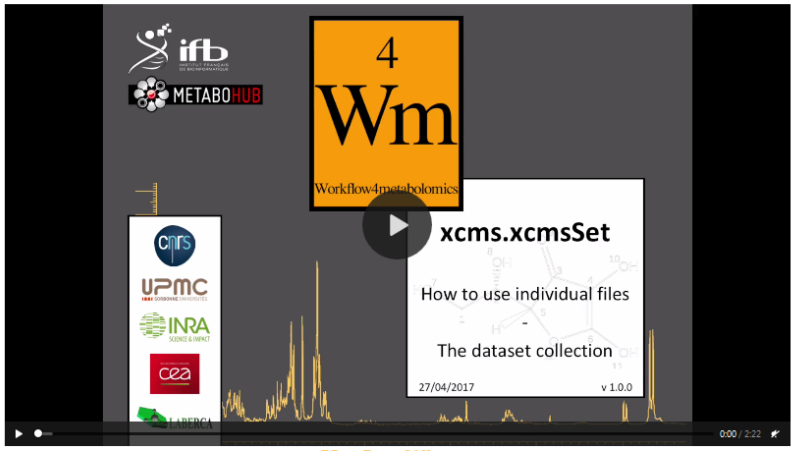
  Account : secole1 or secole2

  Pwd : Inra.0618

HOW TO

CATEGORY	FIELD	Description
Galaxy		Prepare input datasets
Galaxy		Import datasets
Galaxy		Cleanup datasets
Galaxy		Build And Configure A Workflow
Galaxy		Share Histories And Workflows
Preprocessing	LC-MS	Perform Preprocessing using XCMS
Quality Control	Common	Format Data For Postprocessing
Batch correction	LC-MS and GC-MS	Perform Drift And Batch Correction
Statistics	Common	Perform Univariate Analyzes
Statistics	Common	Perform Multivariate Analyzes
Annotation	LC-MS	Perform LCMS Annotations
Annotation	GC-MS	Use NIST

Use single files as input and run xcmsSet in parallel



Tutorials

Main menu

- Home
- Events
- History
- Introduction
  - The Galaxy environment
  - The LC-MS workflow
  - The GC-MS workflow
  - The NMR workflow
  - References
- HowTo
- Download
  - Datasets
- Referenced WorkFlows and Histories
- How to contribute?
- Developer resources
  - Source-code
  - Virtual environments
- People
- Publications

GET: An Open Project

All public W4M tools are distributed:

Dev:

github.com/workflow4metabolomics

Wrappers: on main Galaxy ToolShed

toolshed.g2.bx.psu.edu/groups#/175812cd7caaf439

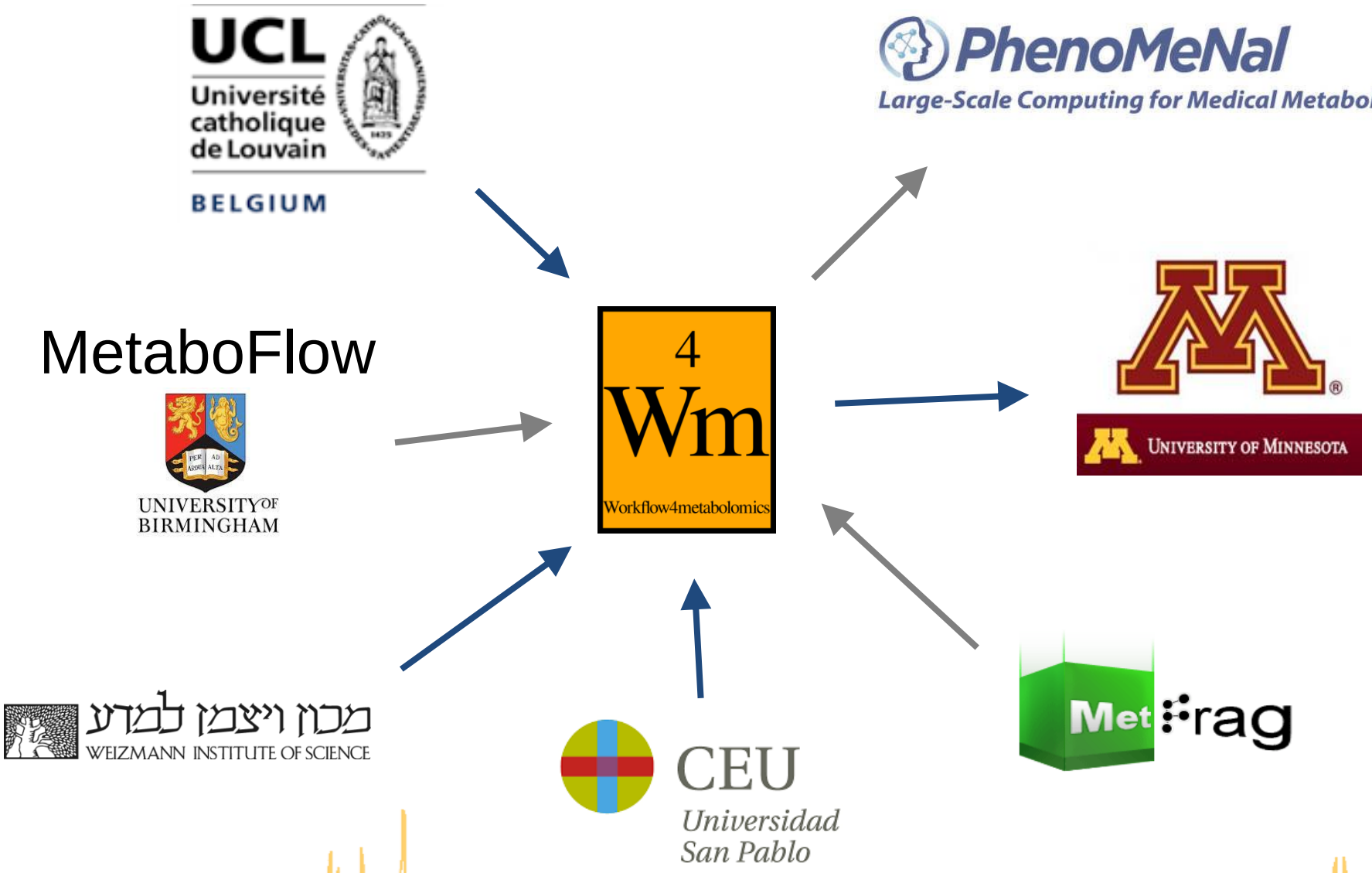
VM: with the most common W4M tools

workflow4metabolomics.org/virtual-environments

Docker/Docker: A Galaxy Flavor

```
docker run -d -p 8080:80 workflow4metabolomics/galaxy-workflow4metabolomics
```

Contribution in progress



PUSH: Contribution

The main W4M instance as a Showcase

Hosting on W4M main instance

Quality standards (IUC Standard)

Advanced mode

- Integration in our reference workflows
- Follow your exchange formats between tools
- Collaboration mode if help is needed

However, the support must be done by the developers themselve

https://github.com/workflow4metabolomics/workflow4metabolomics#how-to-contribute

Tribute to

Christophe Caron [1968-2018]

founding member of W4M

Questions ?

Comme to
Poster 283

References

- References

1 -Giacomoni et al. (2015) 'Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics', *Bioinformatics*, 31: 1493-95.

2- http://workflow4metabolomics.org/referenced_W4M_histories

3- Wilkinson et al. (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.