



4  
Wm

Workflow4metabolomics



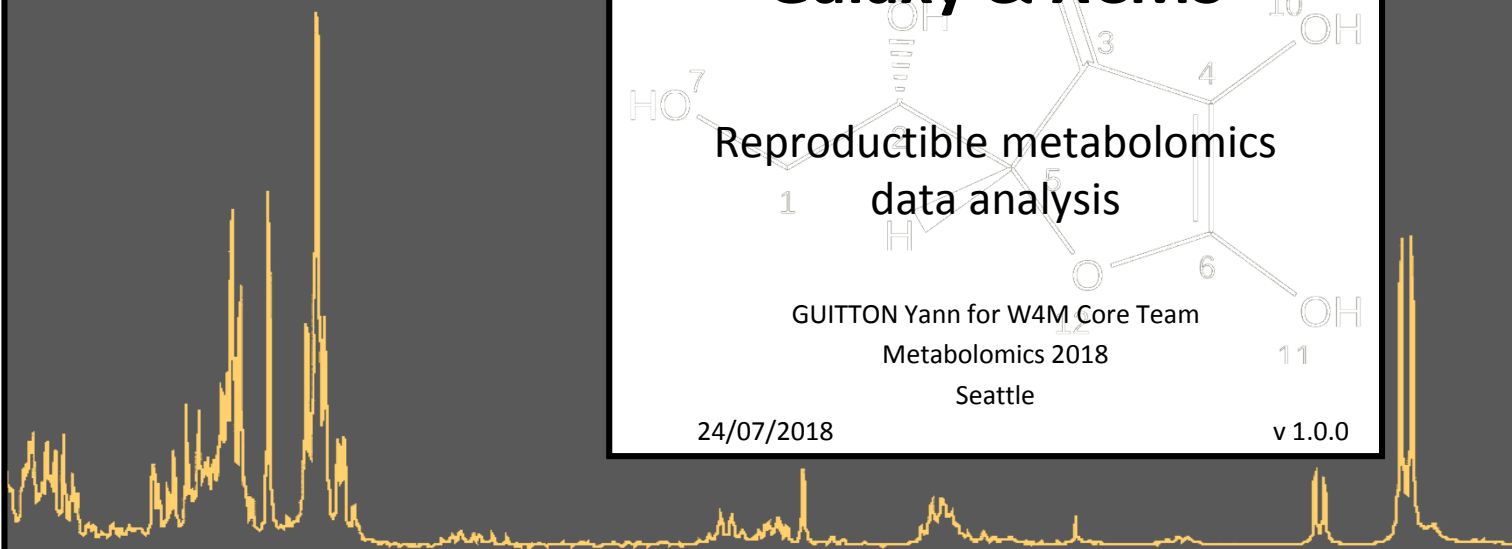
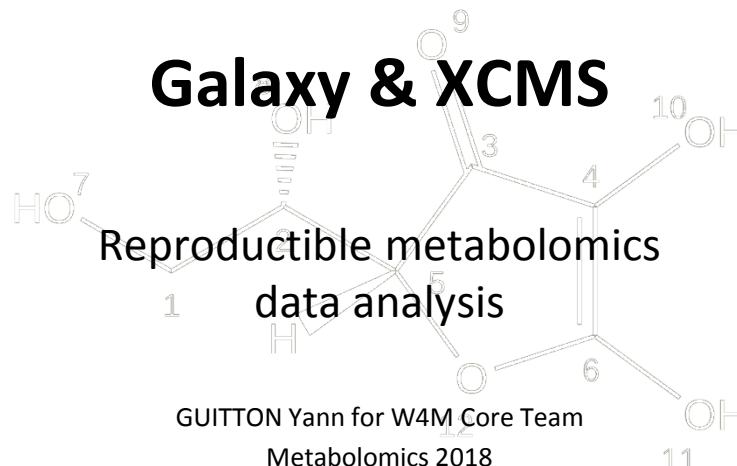
## Galaxy & XCMS

Reproducible metabolomics  
data analysis

GUITTON Yann for W4M Core Team  
Metabolomics 2018  
Seattle

24/07/2018

v 1.0.0



# Presentation



Food Chemical Safety

MS platform (17 instruments)

- Metabolomics
- Bioinformatics (W4M)

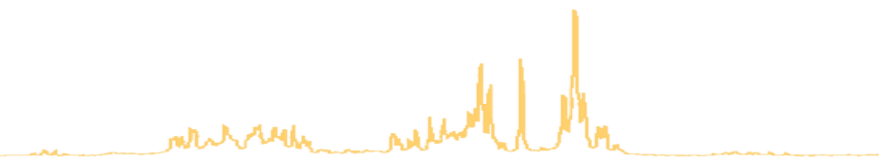


# W4M a user story born from a simple questions

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.**

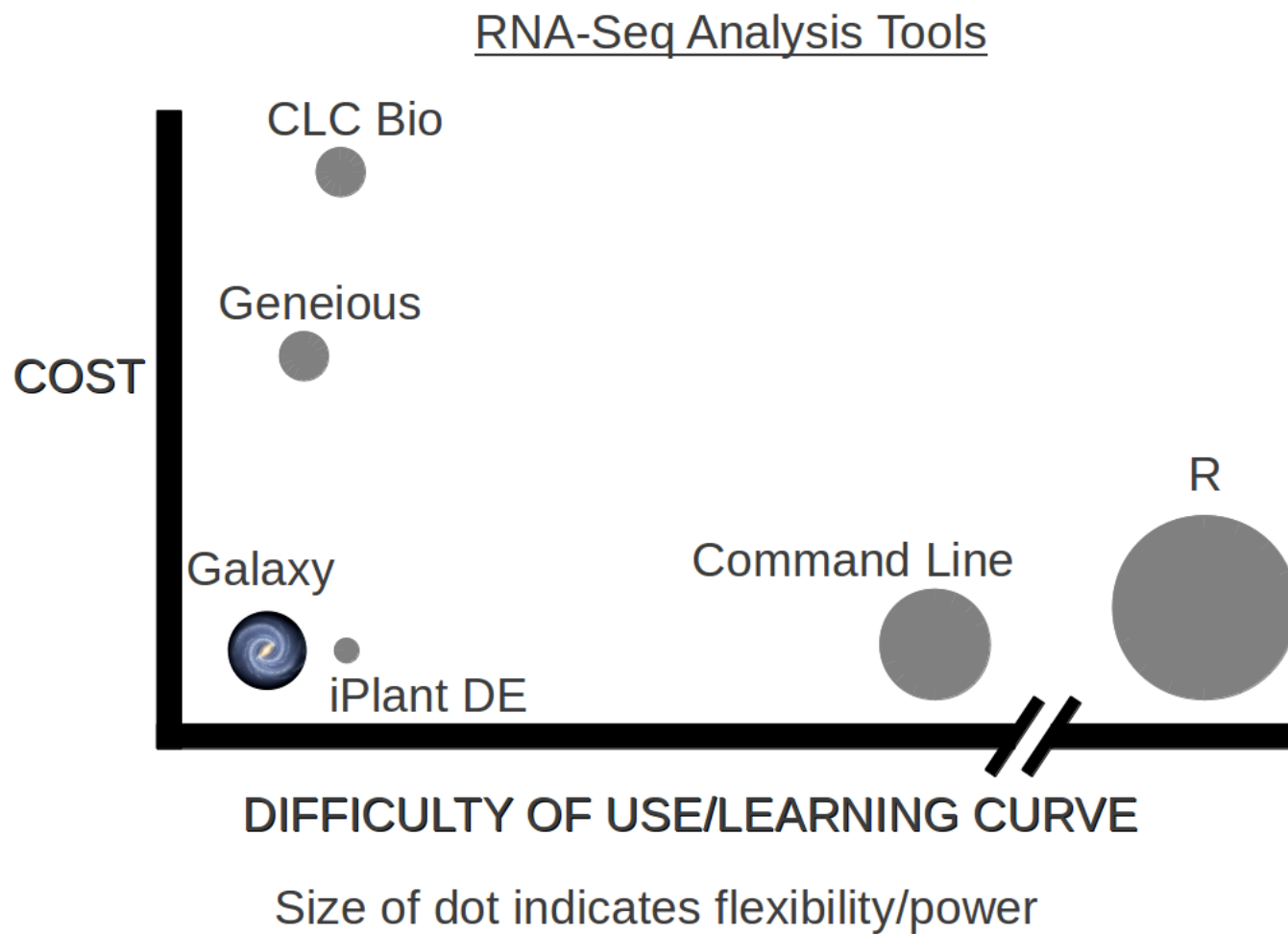
Goecks, J.; Nekrutenko, A.; Taylor, J.; Galaxy Team, T. (2010). "Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences ». *Genome Biology* 11 (8): R86. doi:10.1186/gb-2010-11-8-r86



- 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

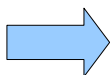


## Why Galaxy ?

- Accessibility
- Reproducibility
- Transparency



## 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
```



```
[login@n0 ~]$ cdprojct
[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

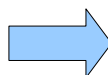
tophat2 panda_v121029 ../input/IllR1-1.fq ../input/IllR1-2.fq
-GTF ../input/panda_v121029.gtf --b2-sensitive -r 100
-num-threads 8

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



**MR. HAPPY**

By Roger Hargreaves



**Galaxy**  
PROJECT

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](mailto:csmith@scripps.edu), Ralf Tautenhahn [rtautenh@gmail.com](mailto:rtautenh@gmail.com), Steffen Neumann [sneumann@ipb-halle.de](mailto:sneumann@ipb-halle.de), Paul Benton [hpaul.benton08@imperial.ac.uk](mailto:hpaul.benton08@imperial.ac.uk) and Christopher Conley [cjconley@ucdavis.edu](mailto: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](mailto: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:

12

Galaxy / ABiMS

Tools

search tools

**Get Data**

ABiMS WORKFLOWS

[Workflow RNA-seq de novo by ABiMS](#)

[Workflow RNA-seq with reference by ABiMS](#)

[Workflow Metabolomic by ABiMS](#)

- [xcms.xcmsSet](#) Filtration and Peak Identification using xcmsSet function from xcms R package to preprocess LC/MS data for relative quantification and statistical analysis
- [xcms.group](#) Group peaks together across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time.
- [xcms.retcov](#) Retention Time Correction using retcov function from xcms R package
- [xcms.fillPeaks](#) Integrate the signal in the region of that peak group not represented and create a new peak
- [xcms.diffreport](#) A report showing the most statistically significant differences in analyte intensities
- [CAMERA.annotateDiffreport](#) Wrapper function for the xcms diffreport and the annotate function. Returns a diffreport within the annotation results.
- [hierarchical clustering](#)
- [PCA](#)

ABiMS TOOLS

Online

- 30-04-13: RNASeq : DESeq is now available for RNASeq expression data with reference (with gtf input).
- 26-04-13: RNASeq : DESeq is now available for denovo RNASeq expression data (without gtf input).
- 26-04-13: RNASeq : sam2counts is now available to count the reads coverage by transcript. It's also a requirement for DESeq denovo.
- 26-04-13: Metabolomic : Workflow Metabolomic by ABiMS, updated to version 20130418

History

XCMS 4 conds  
130.6 MB

1: [mzXML\\_copper\\_stress\\_4cond.ms.zip](#)  
data  
format: ms\_zip, database: 2  
uploaded ms\_zip file

Galaxy

Information  
For any question or request

Galaxy is an open, web-based platform and the Biology and Mathematics and part by NHGRI, NSF, The Huck Institute

✓ **ERGONOMICS**

✓ **PARAMETER COMPLETENESS**

✓ **MODULARITY**

✓ **DATA & WORKFLOW SHARING**

**USERS** →

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



```
[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
```

```
[...]
```

## Galaxy / ABiMS

xcms findChromPeaks (xcmsSet) Chromatographic peak detection (Galaxy Version 3.0.0.0)

Versions

RData file

21: ractporcPOS.raw.RData

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

It contain a xcms3::XCMSnExp object (named xdata) from MSnbase readMSData

Spectra Filters

Extraction method for peaks detection

CentWaveWithPredIsoROIs - performs a two-step centWave-based chromatographic peak detection

See the help section below

Max tolerated ppm m/z deviation in consecutive scans in ppm

8

for the initial ROI definition. (ppm)

Min,Max peak width in seconds

5,50

with the expected approximate peak width in chromatographic space. (peakwidth)

Advanced Options

Signal to Noise ratio cutoff

10

(snthresh)

Prefilter step for for the first analysis step (ROI detection)

3,100

Separate by coma k, I. Mass traces are only retained if they contain at least 'k' peaks with intensity '>= I'. (prefilter)

Name of the function to calculate the m/z center of the chromatographic peak

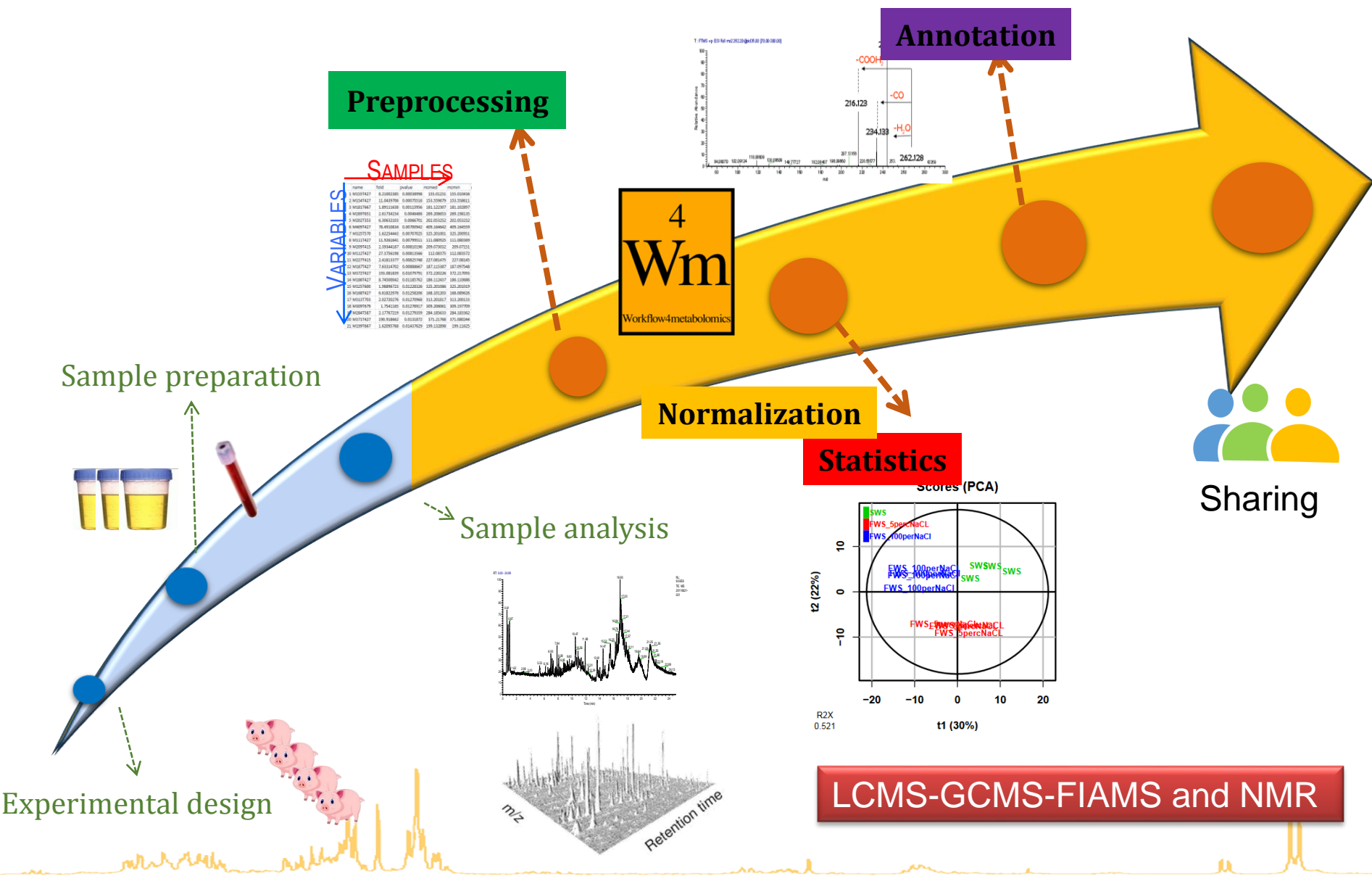
intensity weighted mean of the peak's m/z values

(mzCenterFun)

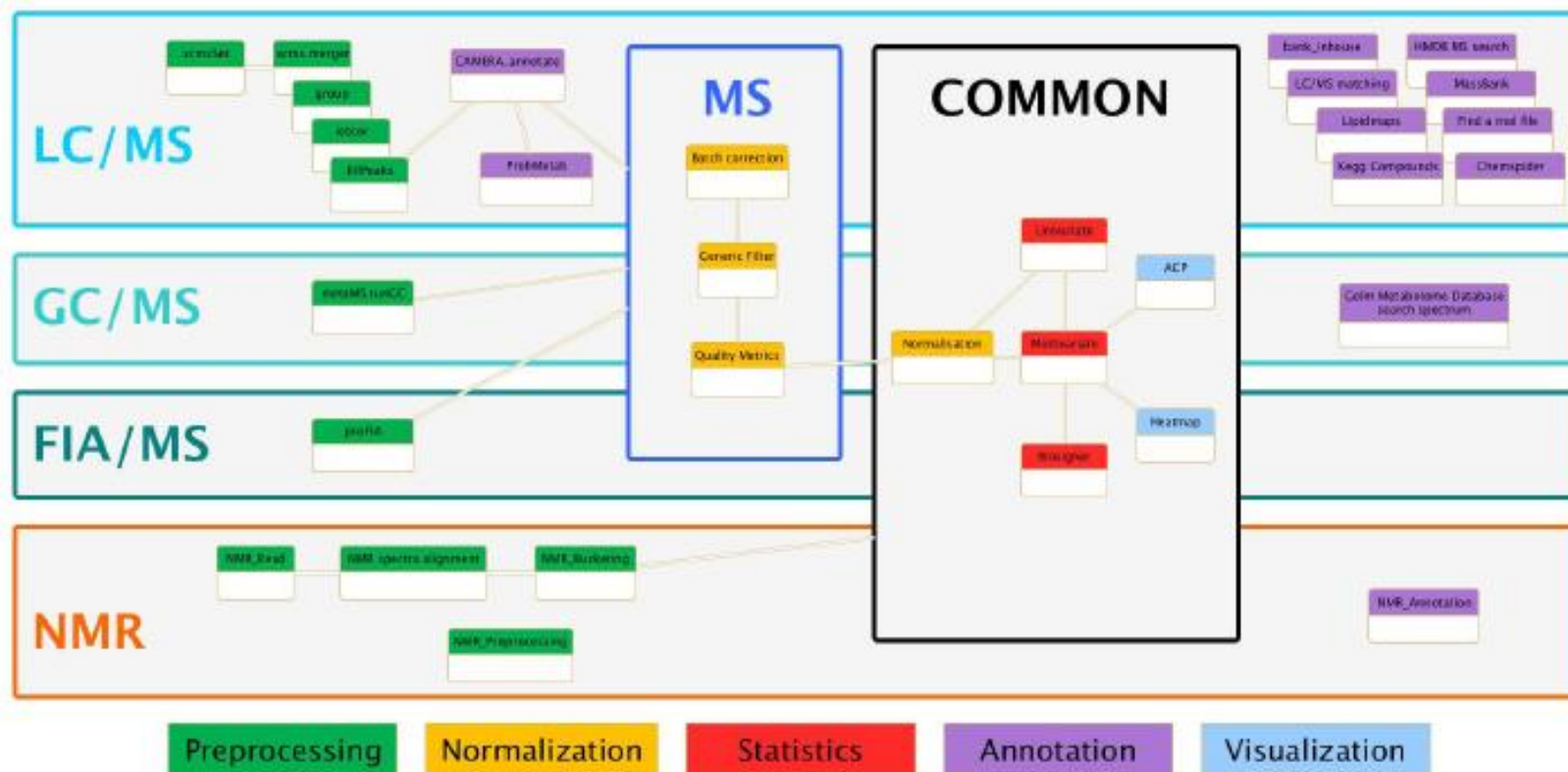
Integration method

peak limits are found through descent on the mexican hat filtered data (more robust, but less exact)

# W4M in the Metabolomics Workflow



- 41 modules and growing



+ NEW Reference your workflows with DOI



# 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.



# Sharing scientific data

## Sharing Data/Histories/Workflows with Galaxy



Using 119.8 GB

**ifb**  
INSTITUT FRANÇAIS  
DE BIOINFORMATIQUE

**METABOHUB**

**CS**

**History**

- HISTORY LISTS**
  - Saved Histories
  - Histories Shared with Me
- CURRENT HISTORY**
  - Create New
  - Copy History
  - Share or Publish**
  - Show Structure
  - Extract Workflow

**13:**  
**xse**  
**df**

**12:**

a Duprier, Marie



# 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 access

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 !

#### Share History with Individual Users

You have not shared this history with any users.

Share with a user

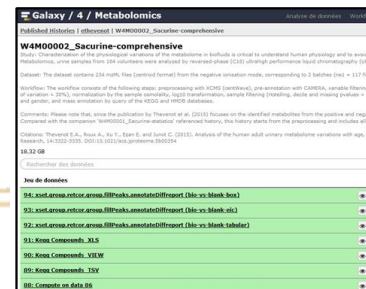
[Back to Histories List](#)



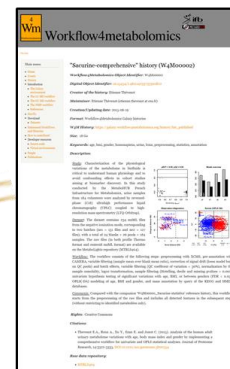
# Sharing scientific data

## Sharing Data/Histories/Workflows with Galaxy

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



Import histories



Dedicated webpage

# Sharing scientific data

Sharing Data/Histories/Workflows with Galaxy



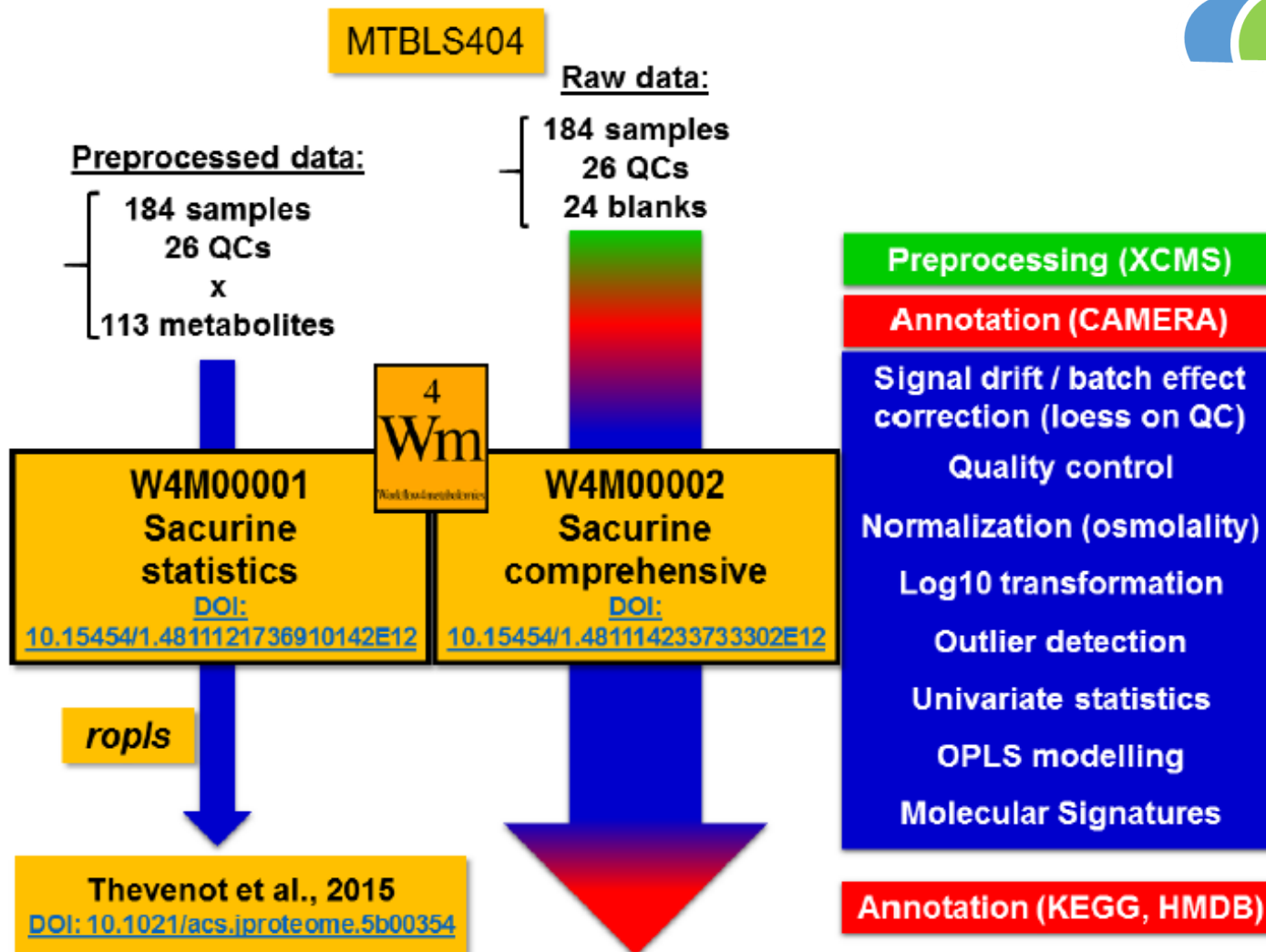
## Referenced W4M histories

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

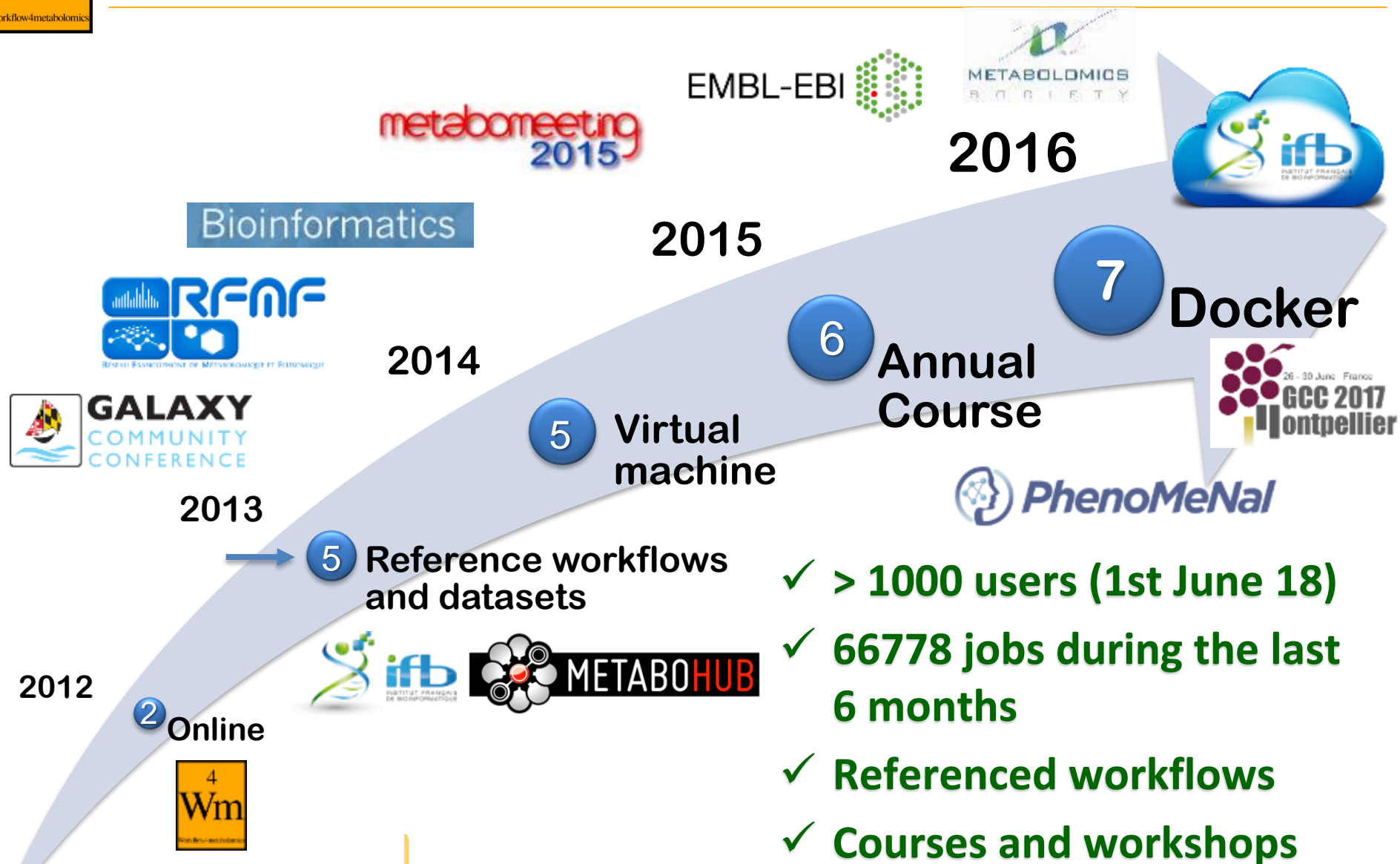


# Sharing scientific data

Sharing Data/Histories/Workflows with Galaxy

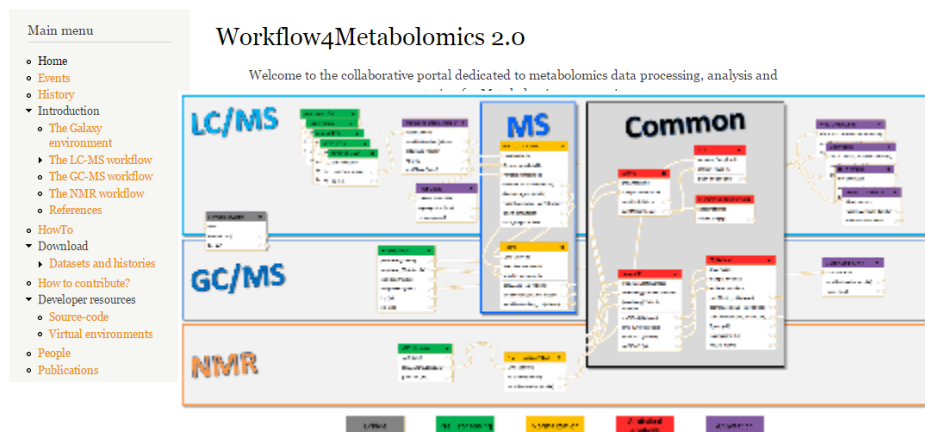


# W4M a long term infrastructure



- ✓ > 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  
 Guillon et al. (2017). *IJBC*, 93:89-101

# Try it

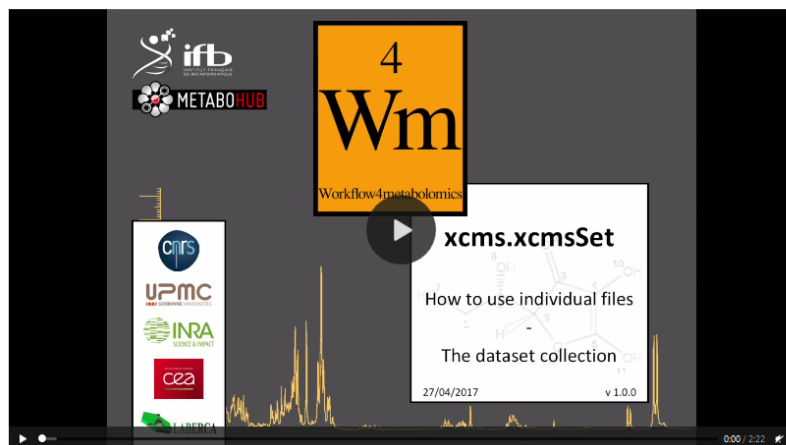
- Just connect to [workflow4metabolomics.org](http://workflow4metabolomics.org) (ask for free account)  
Generic account available during Metabolomics2018

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 Analyses
Statistics	Common	Perform Multivariate Analyses
Annotation	LC-MS	Perform XCMS 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](https://github.com/workflow4metabolomics)



Wrappers: on main Galaxy ToolShed

[toolshed.g2.bx.psu.edu/groups#/175812cd7caaf439](https://toolshed.g2.bx.psu.edu/groups#/175812cd7caaf439)



VM: with the most common W4M tools

[workflow4metabolomics.org/virtual-environments](https://workflow4metabolomics.org/virtual-environments)



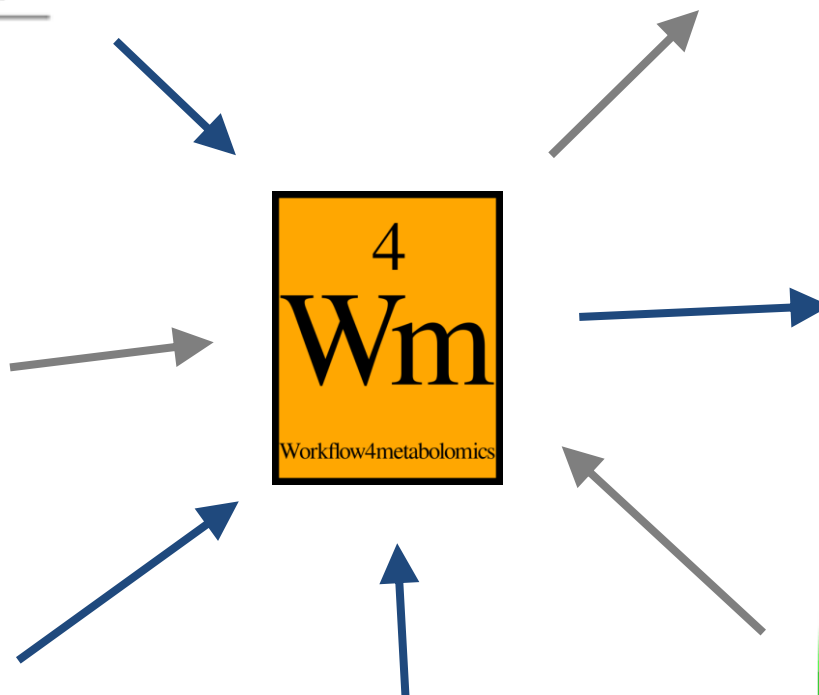
Docker/Docker: A Galaxy Flavor

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

# Contribution in progress



MetaboFlow



# 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 themselves

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

## Christophe Caron [1968-2018] founding member of W<sub>4</sub>M



# Questions ?

---



Comme to  
Poster 283



- 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](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.

