A Small Step into Galaxy, a Faster Pace for Metabolomics

Galaxy and the metabolomics analysis Universe

Pierre PERICARD[1], Gildas LE CORGUILLE[1], Urszula CZERWINSKA[1], Marion LANDI[2], Franck GIACOMONI[2], Christophe DUPERIER[2], Jean-François MARTIN[2], Sophie GOULITQUER[1], Estelle PUJOS-GUILLOT[2], and Christophe CARON[1]

[1] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680, Roscoff, France
{pierre.pericard, gildas.lecorguille, ursula.czerwinska, sophie.goulitquer, christophe.caron}@sb-roscoff.fr

[2] PFEM, UMR1019 INRA, Centre Clermont-Ferrand-Theix, 63122, Saint Genes Champanelle, France
landi.marion@gmail.com
{franck.giacomoni, christophe.duperier, jean-francois.martin, estelle.pujos}@clermont.inra.fr

Abstract *Facing the emergence of new technologies in the field of metabolomics, treatment solutions adopted so far (XCMS, R scripts, etc.) clearly show their limits. Bottlenecks affect unified access to core applications as well as computing infrastructure and storage. In the context of collaboration between metabolomics and bioinformatics platforms, we have developed a full pipeline using Galaxy framework for data analysis. This modular and extensible workflow includes existing components (XCMS functions, etc.) but also a whole suite of complementary statistical tools. This implementation is accessible through a web interface, which guarantees the parameters completeness. The advanced features of Galaxy have made possible the integration of components from different sources and of different types. Finally, an extensible environment is offered to the metabolomics community, and enables preconfigured workflows sharing for new users, but also experts in the field.*

Keywords Metabolomics, Galaxy, XCMS, workflow

Un Petit Pas dans Galaxy, et la Métabolomique s'Accélère

Galaxy et l'Univers des analyses métabolomiques

Résumé *Face à l'arrivée de nouvelles technologies dans le domaine de la métabolomique, les solutions de traitements adoptées jusqu'à maintenant (XCMS, scripts R, etc.) montrent clairement des limites. Les verrous concernent aussi bien l'accessibilité unifiée aux applications métiers que les problèmes d'infrastructure de calcul ou de stockage. Dans le cadre d'une collaboration entre les plateformes INRA/PFEM et CNRS/ABiMS-METABOMER, nous avons développé sous Galaxy un pipeline complet d'analyse. Ce workflow modulaire et extensible, inclut des composants existant (fonctions XCMS, etc.) mais aussi toute une suite d'outils statistiques complémentaires. Cette implémentation, accessible au travers d'une interface web, garantie l'exhaustivité des paramètres. Les fonctionnalités avancées de Galaxy ont permis l'intégration de composants provenant de différentes sources et de nature différente. Au final, un premier environnement est proposé à la communauté métabolomique, et permet le partage de workflows préconfigurés à destination d'utilisateurs novices, mais aussi d'experts du domaine.*

Mots-clés Métabolomique, Galaxy, XCMS, workflow

1 Introduction

The development of "omics" sciences give biologists access to the large variety of the life components and allows the study of metabolism reflecting a possible integration of all post genome phenomena. Recent advances in analytical tools such as high performance liquid chromatography coupled with mass spectrometry (LC-MS) and efforts in chemometrics and in biological computing allows imagining new

analysis strategies to study chemical processes involving metabolites. Metabolomics can be described as a global analysis of small molecules of a biological sample, which are produced or modified as a result of stimuli [1]. The production and utilization of metabolites seems to be more directly connected to the phenotype exhibited by an organism than the presence of mRNAs or proteins.

Metabolomic experiments include different steps based on several technical knowledge i) sample preparation, ii) metabolic profiling, iii) extraction and alignment of data, iv) statistical treatment of data set, v) identification of key or discriminatory metabolites vi) their quantification. This approach has the particularity of generating a large amount of massive datasets. Several thousand of features are produced in several hundred of samples. The extraction of ions from acquisition files is a key step in metabolomic studies, as extraction and alignment software packages provide dataset used for further statistical analysis and identification of metabolites. Alignment of multiple samples may produce noise in the list of extracts ions with problem of overfitting model. So, the choice of software settings is crucial and must be consider as one the most important step in the metabolomic data workflow [2].

Over the past decade, a lot of algorithms and tools were proposed by the scientist community (XCMS, MzMine [4], etc…) and by commercial companies. These applications are standalone software programs or command line scripts, in freeware or commercial versions, with really specific functionalities and often limited performance. Since 2009, some web analysis applications like XCMS Online [5] and MetaboAnalyst [6] are been implemented with a particular effort on ergonomics of tools compared to others described solutions. As an alternative solution, we propose a tool box, easy to grow, functionality by functionality, module by module, from multi-lab sources, accessing by simple web interface and adapted to be used by biologists, MS-experts or statisticians or “bioinformaticians”. Furthermore, this project fits into a methods adaptability context and up-scaling of labs in silico analysis ability.

The web open-source analysis platform Galaxy (http://galaxyproject.org/) was chosen to integrate a suite of current applications in metabolomic community. Galaxy is one of the most useful systems compared to others workflow engines and seems to be adapted to deal with metabolomic data and analysis: no known data size limitations [10], possibilities to automate pipelines, and to ensure reproducibility. Because of its web interface, the system allows true cross platform availability and runs analysis chain by scientists without programming experience.

A metabolomic version of Galaxy including the XCMS suite and statistical modules (Hierarchical clustering, PCA, Normalization) was implemented with the objective to establish a proof of concept based on the new platform adaptability, the interfacing of the original XCMS modularity, and validating ergonomics in comparison to other versions of this framework (R package and XCMS Online versions). This project aims to be the first step of a more ambitious project consisting in the building of complete analysis workflows with extraction, identification and biological interpretation modules.

2 Materials and Methods

2.1 XCMS / CAMERA

XCMS [3] is a free R package which allows extraction and quantification of ions obtained by liquid chromatography coupled with mass spectrometry. XCMS operates on raw acquisition data files converted into NetCDF, mzXML or mzData format. These formats can be produced by all kind of mass spectrometer softwares. XCMS aims to provide a peak list in four steps corresponding to four R functions. Firstly, ions are extracted from each sample independently. Using a Gaussian model, peaks are filtered and integrated. In a second step, called alignment or grouping, individual peaks are matched across all the samples. The third step is an optional correction of retention time drift using a non-linear LOWES regression. Then, if one observed a significant correction of retention time it is necessary to run the second step again. The last step also optional is a gap filling in order to replace missing data by baseline noise. The modularity of these four steps is an attractive feature of XCMS. When the peak list is defined, the diffreport function performs univariate statistics between the different studied conditions and proposes ions annotation based on METLIN database. Finally, the CAMERA package can be used to identify adduct or fragment ions generated in the mass spectrometer ionization source which are redundant informations.

2.2 Galaxy

Our metabolomics tools were integrated to the open, web-based platform: Galaxy [7,8,9]. Initially developed for the computational biomedical research community, its application spectrum has grown increasingly wider. Therefore, more and more bioinformatics platforms, in France and around the globe, have adopted it. Galaxy provides an interface for tools and workflows, and is designed to be: accessible – users without programming experience can easily specify parameters and run tools and workflows; reproducible – Galaxy captures information so that any user can repeat and understand a complete computational analysis; transparent – users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis. Galaxy very active community also insures up-to-date software releases and efficient support for both end-users and tools developers.

2.3 Infrastructure

The Galaxy platform dedicated to metabolomics data analysis was integrated to the standard components of ABiMS (Analysis and Bioinformatics for Marine Science) platform computing infrastructure. Galaxy was deployed in a virtual environment based on VMWARE. Optimization efforts, such as connections pool or web services decoupling, allow a good level of scalability. Computing resources connection uses the standard DRMAA API, and is completed with a dedicated connector (tool runner) in order to make available adequate resources both in terms of high computing performance and memory amount (128GB – 1TB RAM). Finally, a shared and secure storage space completes this layer, essential to smoothly working treatments.

2.4 Inputs

The XCMS package can read full-scan LC/MS data from NetCDF, mzXML, and mzData files. A single experiment can generate from tens to several hundreds files, which have to be organized in a specific arborescence so that the distinct conditions can be identified by XCMS. To easily handle such inputs, we added to our Galaxy instance a new proprietary datatype based on a zip file but with a specific extension (.ms.zip). To prevent Galaxy to unzip these files we also patched the “Get Data” tool and added a sniffer specific to our new datatype. These modifications allow .ms.zip files to be uploaded and automatically assigned the correct datatype with no manual intervention from the user. To allow communication between our pipeline tools, we also added a second datatype based on RData files, which saves the information from one tool and can be used as the input of the next tool.

2.5 Galaxy Tool XML Definition Files

The first step to integrate a tool in Galaxy consists in creating its XML definition file (cf. Figure 1), which let Galaxy know the execution details of the new tool. Therefore, we wrote nine definition files (one for each step of our pipeline) according to the good development practices guidelines from the IFB (Institut Français de Bioinformatique) Galaxy working group. In order to provide to XCMS users an ergonomic interface, the large majority of the tools parameters were described, typed, and discriminated between main and advanced parameters. We also leveraged the Galaxy conditional system to dynamically display parameters depending on other parameters. Every tool and parameter was documented based on the official documentation and our implementation specificities.

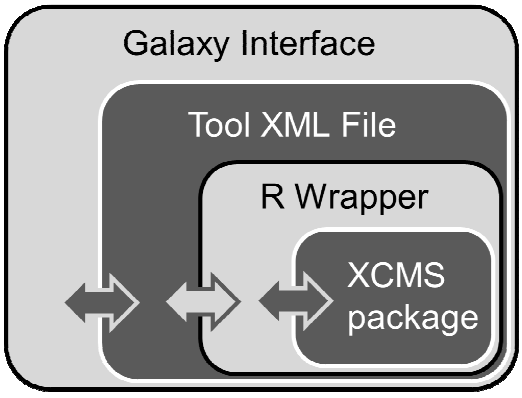

Figure 1. Galaxy multiple layers organization

2.6 R Wrapper and XCMS Functions

Galaxy XML wrappers cannot directly run R functions and/or multiple commands in a single run. Therefore, to run R functions we need to develop wrappers in R (cf. Figure 1). We chose to write only one script/wrapper by package (in this case XCMS/CAMERA) and keep it as inclusive and transparent as possible. As a consequence, adding a parameter or function (eg. following a XCMS update) only require some changes in the XML definition file. This script allows libraries loading, argument passing between the XML definition files and the R functions, and data preprocessing.

2.7 Post Extraction Normalization and Statistical Analyses

After peaklist definition, the pipeline continues with normalization and a series of univariate and multivariate statistical analyses, available using R functions. During LC-MS analyses, especially with large number of samples, the analysis is often interrupted for mass spectrometer maintenance or calibration. Moreover, it is also observed a decrease of intensity due to ionization source clogging during a batch analysis. This leads to analytical intensity drifts. It is possible to correct this evolution of intensity by normalization based on linear regression [11]. In order to select metabolites with significant intensity difference among the different experimental conditions, an analysis of variance can be carried out with Benjamini Hochberg p-values correction (BH). Defining a p-value threshold, significant ions are selected for further unsupervised multivariate analyses. Hierarchical Ascendant Clustering using hclust R function (using different options of distance and aggregation methods) can be used to aggregate ions depending on their abundance in samples. A result file is produced in order to use Treeview [12] software for interactive clustering of sample and metabolites and for heatmap construction. Principal Component Analysis (PCA) is also available using FactoMineR (http://factominer.free.fr/) R package.

3 Results and Discussion

A total of nine XCMS R functions and statistical analysis scripts were implemented in Galaxy as a metabolomics analysis pipeline (cf. Figure 2). Each tool can be run independently or as part of a complete workflow. We also successfully implemented the parallelization of the supported XCMS functions. The “Rmpi” R library was installed on the ABiMS cluster and all the related Galaxy wrappers were configured to use the XCMS multi-core options. This configuration allows to greatly speed up some of XCMS most time-consuming functions (xcmsSet, by ex). MPI runs also allow a smoother cluster usage while saving time during CPUs reservation (vs. a usual multi-thread run).

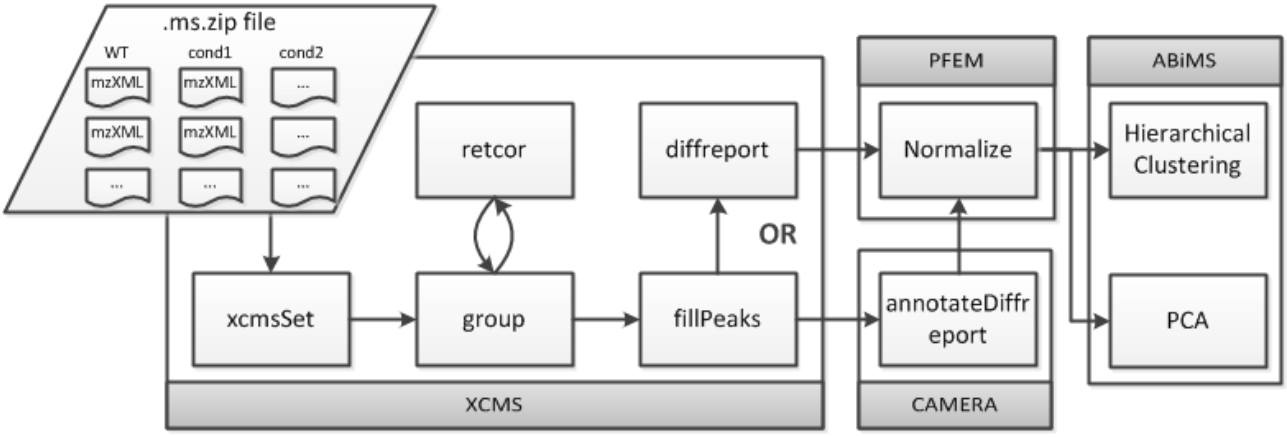

Figure 2. Organization of our metabolomic analysis workflow components.

In order to assess the different tools and verify that they all leads to the same peaklist, a real set of samples using Netcdf files format was used to compare the extraction results using XCMS command line (XCL), XCMS Online (XO) and XCMS implemented in Galaxy (XG). As expected almost the same list of ions was detected: 911 ions for XCL, 912 for XG and 962 for XO with an average coefficient of variation of peak integration of 0.0014%. These results confirm that we obtained very close results with these different techniques.

XCMS for R command line is the reference software. It has a great potential when associated with existing R packages especially for statistical analysis. Though programming makes possible defining the workflow and integrating in-house scripts for normalization and annotation of ions with public databases, it highly impacts ergonomics. Mass spectrometer analysts are generally not R users and so there is a need of interface to easily use the XCMS package. XCMS Online and Galaxy are both very attractive alternatives to XCMS for R as they are both based on a web interface. XCMS Online is dedicated to non XCMS users by providing pre-configured workflows with parameters for each kind of mass spectrometers. However, some experts can be limited by the lack of parameter completeness. Instead, our implementation of XCMS in Galaxy was designed to display as much parameters as the R functions, although non-XCMS users have access to pre-configured workflows shared by the community.

	R	**XCMS Online**	**XCMS on Galaxy**
Ergonomics	--	++	+
Parameter Completeness	++	-	+
Target	R and XCMS advanced users	Non-XCMS users XCMS advanced users -	Non-XCMS users XCMS advanced users + Additional tools developers
Data & workflow sharing	-	-	++
Modularity	+	-	++

Table 1. XCMS implementation comparison

Modularity features of Galaxy, which allows integrating multi languages applications (Python, Perl, R, Matlab, Bash, C, etc.) from multi repositories (toolshed or others), was used in this project to build components and workflows in two geographical places and in two teams of our community: ABIMS and PFEM platforms. Galaxy plays a central role in merging software engineering methods and sharing private tools (cf. Table 1).

4 Conclusion and Perspectives

To face the predicted fast and large request production of metabolic fingerprints, scientist communities get organized in data representation standards development and in primary database models implementation. In this context, the ability of biological computing actors, proposing in-silico strategies and analysis tools, is

the key factor of the future of metabolomics like a high-speed science, as well as the necessary advances in mass-spectrometry technologies.

We successfully implemented and made available a full bioinformatics pipeline for metabolomics analysis. This pipeline can be used through the ABiMS Galaxy web interface, which provides an easy way to test, parameter, and keep a history of previous analysis. Workflows can also be designed by combining several tools. Experts can pre-configure these workflows using advanced parameters and then make them available to all users through the Galaxy sharing interface. To illustrate this part, the ABiMS Galaxy instance already proposes two expert and pre-configured workflows for metabolomics analysis with either medium or high resolution mass-spectrometers.

With this national collaboration between two metabolomics platforms PFEM and METABOMER, and the ABiMS bioinformatics team, we went over the initial proof of concept and build a primary powerful and extensible analysis environment. Through first results, we obtained fund from Auvergne and Bretagne district councils (calls for projects LIFEGRID3 and BIOGENOUEST – CORSAIRE dispositive) for the next two years (2014-2015) with concrete application in fields of Nutrition and Marine Environment. The aims of future developments will be to increase metabolomic analysis workflow possibilities and to federate action to other scientist groups involved in Metabolomics through a Galaxy environment.

Finally, we intend to develop our links, related to the national research infrastructures (METABOHUB and EMBRC-France) to build a reliable metabolomic workflow infrastructure supported by the IFB in the future.

Acknowledgements

This work was supported by ANR program ‘Investissements d’Avenir’ (METABOHUB, EMBRC-France), and by Auvergne and Bretagne district councils.

References

[1] J.K. Nicholson, J.C. Lindon and E. Holmes, 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29(11):1181-1189, 1999.

[2] W. Dunn, Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology*, 5:011001, 2008.

[3] C.A. Smith, E.J. Want, G.C. Tong, R. Abagyan, and G. Siuzdak, XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779-787, 2006.

[4] M. Katajamaa and M. Oresic, Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics*, 6:179, 2005.

[5] R. Tautenhahn, G.J. Patti, D. Rinehart and G. Siuzdak, XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Analytical Chemistry*, 84(11):5035-5039, 2012.

[6] J. Xia, N. Psychogios, N. Young and D.S. Wishart, MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res*, 37(Web Server issue):W652-60, 2009.

[7] J. Goecks, A. Nekrutenko, J. Taylor and The Galaxy Team, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol, 11(8):R86, 2010.

[8] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko and J. Taylor, Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, 89:19.10.1-19.10.21, 2010.

[9] B. Giardine, C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W.J. Kent and A. Nekrutenko, Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451-5, 2005.

[10] D. Blankenberg, A. Gordon, G. Von Kuster, N. Coraor, J. Taylor, A. Nekrutenko and Galaxy Team, Manipulation of FASTQ data with Galaxy. *Bioinformatics*, 26(14):1783-1785, 2010.

[11] F.M. Van Der Kloet, I. Bobeldijk, E.R. Verheij, R.H. Jellema, Analytical error reduction using single point

calibration for accurate and precise metabolomic phenotyping. *J Proteome Res*, 8(11):5132-41, 2009.

[12] A.J. Saldanha, Java Treeview--extensible visualization of microarray data. *Bioinformatics*, 20(17):3246-3248, 2004.