WHALE: A PACKAGE COMBINING FINGERPRINTING AND TARGETED METABOLITE PROFILING TO IMPROVE THE EXTRACTION OF METABOLIC INFORMATION IN NMR SPECTRA

J. Gomez[1,2*], N. Ramírez[1,2], M. Vinaixa[1,2], Miguel A. Rodríguez[1,2], Reza M. Salek[3], X. Correig[1,2] and Nicolau Cañellas[1,2]

[1]Metabolomics Platform, IISPV, Universitat Rovira i Virgili, Campus Sescelades, Carretera de Valls, s/n, 43007 Tarragona, Catalonia

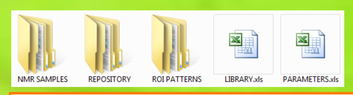
[2]CIBERDEM, Spanish Biomedical Research Centre in Diabetes and Associated Metabolic Disorders, Bonanova 69, 6th floor, 08017 Barcelona, Catalonia

[3]European Bioinformatics Institute (EMBL-EBI) European Molecular Biology Laboratory, Welcome Trust Gemome Campus Hinxton, Cambridge CB10 1SD, UK

[@] josep.gomez@urv.cat

INTRODUCTION

Nuclear Magnetic Resonance (NMR) spectroscopy has been established as one of the most popular tools for high-throughput characterization of metabolites present in complex biological mixtures. It is a non-destructive, highly reproducible and versatile technique, and allows the user to "interrogate" the same sample in different ways by selecting different pulses and acquisition parameters in order to obtain complementary information [1]. An NMR spectrum is composed by resonances of a huge number of metabolites where each metabolite may contribute to the NMR spectra with many individual signals. Depending on the final goal, different approaches can be applied to analyze the samples in a dataset. An untargeted analysis of the sample fingerprint can be enough to discriminate groups in a study, or very useful for finding 'hot spots' in an exploratory analysis. Otherwise, a targeted analysis of the metabolite profile using regions of interest (ROIs) can be performed in order to find molecular patterns of the samples under study [2]. We present a package that allows the user to use both approaches and to include unknown signals and 2D acquisitions for an exhaustive analysis of the NMR data.



•The package needs 5 inputs to perform an analysis:

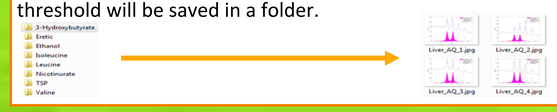
- **NMR SAMPLES**: A folder with the NMR samples dataset.
- **REPOSTORY**: A folder with the reference compounds.
- **ROI PATTERNS**: A folder with the quantification parameters for each ROI.
- **LIBRARY**: A file containing the names of the signals to be quantified.
- **PARAMETERS**: A file containing experiment parameters such as alignment, supression and normalization options.

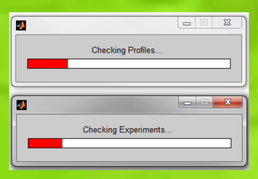
•The output contains information such as the signal area, fitting error and position of each signal.

	A	B	C	D	E	F	G	H	I
1	Exp/Met	3-Hydroxybut	Ethanol	Isoleucine	Leucine	Nicotinurate	Valine	TSP	Eretic
2	Liver_AQ_1	10.3747043	8.04774883	9.98424331	29.5256155	1.82332876	18.8871052	314.855837	29.4433508
3	Liver_AQ_2	9.83060668	18.8363578	9.00500017	25.5773821	2.27558072	17.3778827	317.009357	29.3365742
4	Liver_AQ_3	11.5053245	4.83702491	8.12053033	25.2074507	3.02607506	16.6795667	307.800668	29.3473453

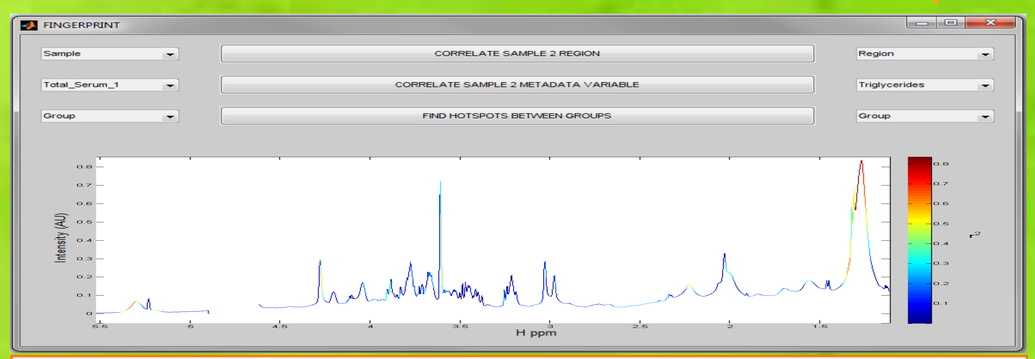
Areas (AU) / Fitting errors (%) / Positions (ppm) / Signal area ratio (%) / ROI baseline area (AU)

•And all the quantifications that have not passed a quality threshold will be saved in a folder.





•This function analyzes all the ROI patterns in all the samples of the dataset automatically. It is very useful in cases where the user has already tested the ROIs and decides to let the software work by itself.

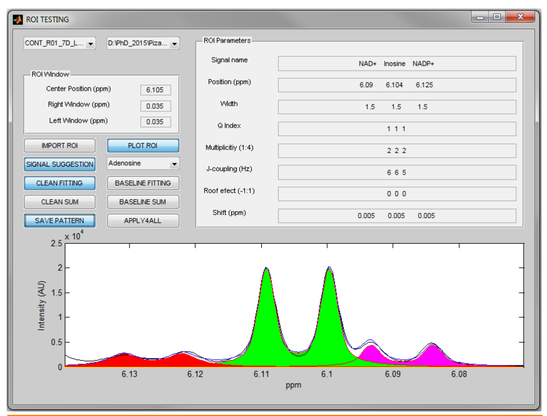


•This panel offers the user 3 options to explore the data in an untargeted way:

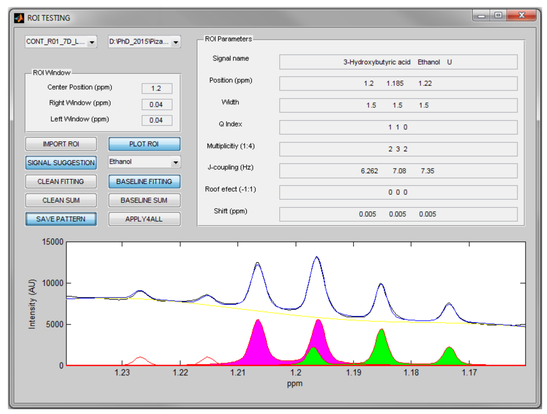
- **CORRELATE SAMPLE 2 REGION**: Generates a colored spectra where the color is showing the correlation coefficient of each point of the spectra with the region selected
- **CORRELATE SAMPLE 2 METADATA VARIABLE**: Generates a colored spectra where the color is showing the correlation coefficient of each point of the spectra with a variable stored in the metadata
- **FIND HOTSPOTS BETWEEN GROUPS**: Generates a colored spectra where the color is showing the level of covariance of each point between groups.

•Here in this figure we show the correlation of one total serum sample with the metadata variable 'Triglycerides'.

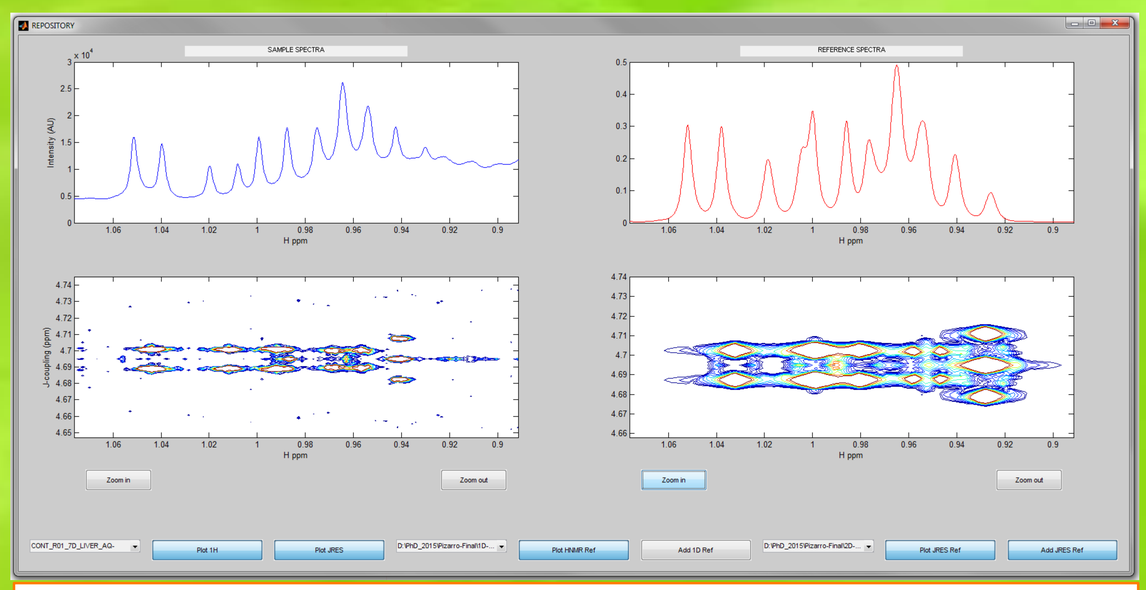
•The **MANUAL PANEL** is the most editable and user-interactive function in the package. It allows the user to adjust signal parameters, visualize spectra and try different quantification modes in order to optimize the analysis. It also suggests signals that may be found in a selected ROI. Our fitting algorithm quality has been already tested in a previous version of the package called Dolphin [3]. Images of each quantification mode performance are shown below.



•The **'CLEAN FITTING'** mode is able to quantify accurately overlapping signals in ROIs where neither baseline nor broad signals are affecting.

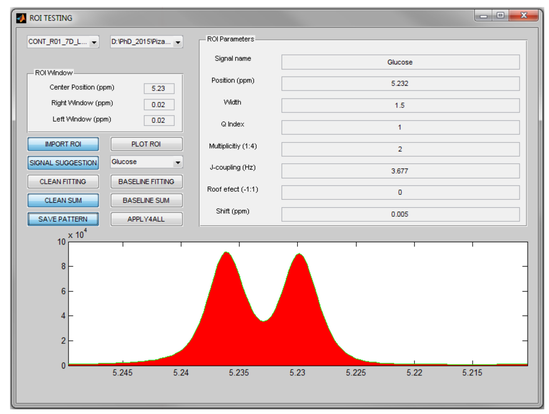


•The **'BASELINE FITTING'** mode allows deconvolving targeted signals in ROIs where baseline or broad signals are affecting the final shape of the region; it will take more computation time but is the optimal solution in those cases.

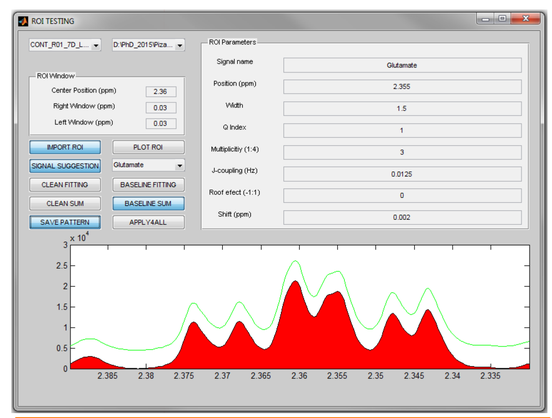


•The reference panel gives the user the chance to visualize reference compounds in **1D NOESY, 2D JRES, 2D COSY** and **2D HSQC** and generate reference mixtures in order to compare them with the target sample.

•Here in this figure we show the Branched-Chain Amino Acids (BCAAs) region in our sample spectra in 1D NOESY (upper-left corner) and 2D JRES (lower-left corner); and the superposition of those BCAAs in 1D NOESY (upper-right corner) and 2D JRES (lower-right corner) using reference spectra stored in the repository.



•The **'CLEAN SUM'** mode is very useful for those ROIs that contain isolated and pure (without any baseline) signals because the computation time is severely reduced while the quantification remains accurate.



•The **'BASELINE SUM'** mode is the best option for those ROIs that contain isolated signals but overlapped with a little and almost constant baseline. The computation time is almost the same than the 'Clean Sum' option.

CONCLUSIONS

•The fingerprint analysis allows the user to find hotspot regions through correlation and covariance functions using all the spectra and the metadata, which is very useful to detect relevant information in an untargeted step.

•The combination of user-interaction and automatic quantification methods makes this tool a good agreement between editability and robustness, avoiding black-box processes and subjectivities sample-sample.

•The reference panel helps the user to identify metabolites using different NMR pulses of reference compounds. It allows the user to visally check

•The output Plots2Check allows the user to visually check the quantifications that have not passed a quality threshold and rapidly detect the variables to correct for improving the results.

REFERENCES

[1] Nicholson JK, et al. (1999) : 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

[2] Gómez, J et al. (2015): Dolphin 1D: Improving Automation of Targeted Metabolomics in Multi-matrix Datasets of 1H-NMR Spectra. 9th International Conference on Practical Applications of Computational Biology and Bioinformatics 59-67

[3] Gómez, J et al. (2014): Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D ^{1}H-NMR data. Analytical and Bioanalytical Chemistry 406:7967-7976