SEVENTH FRAMEWORK PROGRAMME

HEALS

Health and Environment-wide Associations based on Large population Surveys

FP7-ENV-2013- 603946

http://www.heals-eu.eu/

7.2 Predictive biomarkers

appropriate for environment-wide

association health assessments

WP 7 Novel bioinformatics for predictive biomarker discovery

Version 1.0

Lead beneficiary: AUTH

Date: 31/5/2017

Nature: Report

Dissemination level: Public

HEALS FP7-ENV-2013-603946	D7.2 - Predictive biomarkers appropriate for EWAS		
	WP7: Novel bioinformatics for predictive biomarker discovery	**Security**:	
	Author(s): Denis A. Sarigiannis	**Version**:	2/111

Document Information

Grant Agreement Number	ENV-603946	**Acronym**	HEALS
Full title	Health and Environment-wide Associations based on Large population Surveys		
Project URL	http://www.heals-eu.eu/		
EU Project Officer	Tuomo Karjalainen,- Tuomo.KARJALAINEN@ec.europa.eu		

Deliverable	**Number**	7.2	**Title**	Predictive biomarkers appropriate for environment-wide association health assessments
Work Package	**Number**	7	**Title**	Novel bioinformatics for predictive biomarker discovery

Delivery date	**Contractual**	M30	**Actual**	30/04/2017
Status	Draft ☐		Final X	
Nature	Demonstrator ☐ Report X Prototype ☐ Other ☐			
Dissemination level	Confidential ☐ Public X			

Author (Partners)	Denis A. Sarigiannis, S. Karakitsios, I. Frydas, N. Papaioannou (AUTH) R. Stierum, E. van Someren (TNO)			
Responsible Author	Denis A. Sarigiannis		**Email**	denis@eng.auth.gr
	Partner	AUTH	**Phone**	+30-2310-994562

Document History

Name	Date	Version	Desciption

TABLE OF CONTENTS

From DOW

D7.2) Report on predictive biomarkers appropriate for environment-wide association health assessments:

This is a synthesis report on the most appropriate biomarkers to use with predictive power in support of environment-wide associations with health effects. [month 30]

Task 7.2 Predictive data mining – data models design and analysis (AUTH, TNO, CERETOX, UPD)

For predictive data mining we will develop a set of models to perform inference on the available (combination of) multidisciplinary datasets. Several techniques can be used for that purpose, ranging from typical approaches based on decision trees or k-nearest neighbors to more sophisticated ones that employ artificial neural networks (ANNs), support vector machines (SVMs) or Bayesian networks (BNs). We will implement and test all of the aforementioned approaches, since our goal is not only to perform mere classification but also to study and unravel the feature attributes concealed in the exposome data.

To estimate the reliability of the proposed predictive models we will employ the k-fold cross-validation schema and measure standard performance indices like sensitivity, specificity and accuracy. ROC analysis will also be carried out to test the robustness of each one of the models. Based on the obtained results we will consider the design of a mixture of models where different approaches will be used to model the diverse data in our dataset with a view to increasing the predictive capacity of the biomarkers identified in WP4 and WP5. This will not only increase the prediction performance in the subsequent environment-wide association health surveys but also incorporate more efficiently the associations and patterns derived from the previous task.

Additionally, the classification models will be further analyzed in terms of visualizing their output and interpreting their deduction mechanism. Visualization mainly refers to mapping the underlying decision hyperplanes of the models as well as depicting the differences of the individual performance indices. Interpretation will be a relatively direct process in the cases of decision trees and k-nearest neighbors while in the models based on ANNs, SVMs or BNs deduction is more complex and harder to interpret. The above analysis will result in better tuning the architecture of the proposed models and optimize their predictions.

Task 7.3 Model integration – biomarkers identification and prediction validation (AUTH, TNO, UPD)

This task will provide the methodological tools for integration of multiple omics biomarkers into a mechanistic description of toxicity pathway interactions, in relation to external/internal exposure. This will be achieved by developing biological pathway models for the endpoints identified in Stream 5 using the bioinformatics approaches described above. In particular, all the findings generated from descriptive and predictive data mining, in the form of clusters, patterns, associations and classifications, will be assessed and incorporated into a meta-modeling framework. By post-processing the meta-model as well as the prediction models proposed above, multivariate decision profiles could be determined specifically tailored for revealing the diagnostic biomarkers. The prediction accuracy of

these biomarkers will be tested against an independent test set, other than that used for training on the previous two tasks.

Similar to the idea of mixture models, introduced in Task 7.2, is that of biomarkers fusion, which lies in the same principles of multivariate analysis. When the problem under study is characterized by high dimensionality or complexity, as is the case of exposome data, it is advantageous to consider as many parameters as possible in order to gain more insight, instead of focusing on a couple of them. Moreover critical aspects about the biomarkers interoperability can be revealed that can lead to even better diagnostic procedures. Biomarkers fusion can be realized efficiently through an inference system that is based on fuzzy logic. Since no prior knowledge exists about the normal and pathological levels of the derived biomarkers, fuzzy logic rule sets are considered as a constructive approach to design robust clinical decision support systems.

1 Summary

A key novelty in HEALS is the integrated use of advanced computational tools supporting environmental and biological data analyses for comprehensive data interpretation. These tools include Physiology-Based BioKinetic models (PBBK), novel bioinformatics strategies for biomarker prediction and advanced multivariate statistics for associating the links between exposure to environmental stressors and health status and investigating causality.

2 Introduction - General considerations

The exposome (Wild, 2005) represents the totality of exposures from conception onwards, simultaneously identifying, characterizing and quantifying the exogenous and endogenous exposures and modifiable risk factors that predispose to and predict diseases throughout a person's life span. Exposome came as a complement to the human genome; although decoding of human genome (Schmutz et al., 2004) increased our understanding of the underlying causes of disease, genome explains only a percentage of population burden. Thus, it is evident that environmental factors are equally or eventually more important and what is actually critical is the interaction of environmental factors with the biological systems. Towards a better understanding of the causal links among genome, environment and disease, unraveling the exposome implies that both environmental exposures and genetic variation are reliably measured simultaneously.

HEALS (Health and Environment-wide Associations based on Large population Surveys) brings together a comprehensive array of novel technologies, data analysis and modeling tools that support the efficient design and execution of large-scale exposome studies. The HEALS approach brings together and organizes environmental, socio-economic, exposure, biomarker and health effect data; in addition, it includes all the procedures and computational sequences necessary for applying advanced bioinformatics coupling advanced data mining, biological and exposure modeling so as to ensure that environmental exposure-health associations are studied comprehensively. The overall approach will be verified in a series of population studies across Europe, tackling various levels of environmental exposure, age windows and gender differentiation of exposure, and socio-economic and genetic variability. The main objective of HEALS is the refinement of an integrated methodology and the application of the corresponding analytical and computational tools for performing environment-wide association studies in support of EU-wide environment and health assessments. For the first time, HEALS will try to reverse the paradigm of "nature versus nurture" and adopt one defined by complex and dynamic interactions between DNA sequence, epigenetic DNA modifications, gene expression and environmental factors that all combine to influence disease phenotypes. HEALS will start from analysis of data collected in on-going epidemiological EU studies involving mother/infant pairs, children, or adults including the elderly to evidence relevant environmental exposure/health outcome associations. These associations will aid in designing pilot surveys using an integrated approach, where the selection of biomarkers of exposure, effects and individual susceptibility results in integrated risk assessment.

The overall methodological concept of HEALS and the different arrays involved is graphically illustrated in Figure 1. This includes a wide array of state of the art technologies across all major disciplines of the environmental exposure, biochemistry, molecular biology, toxicology, bioinformatics and epidemiology arena.

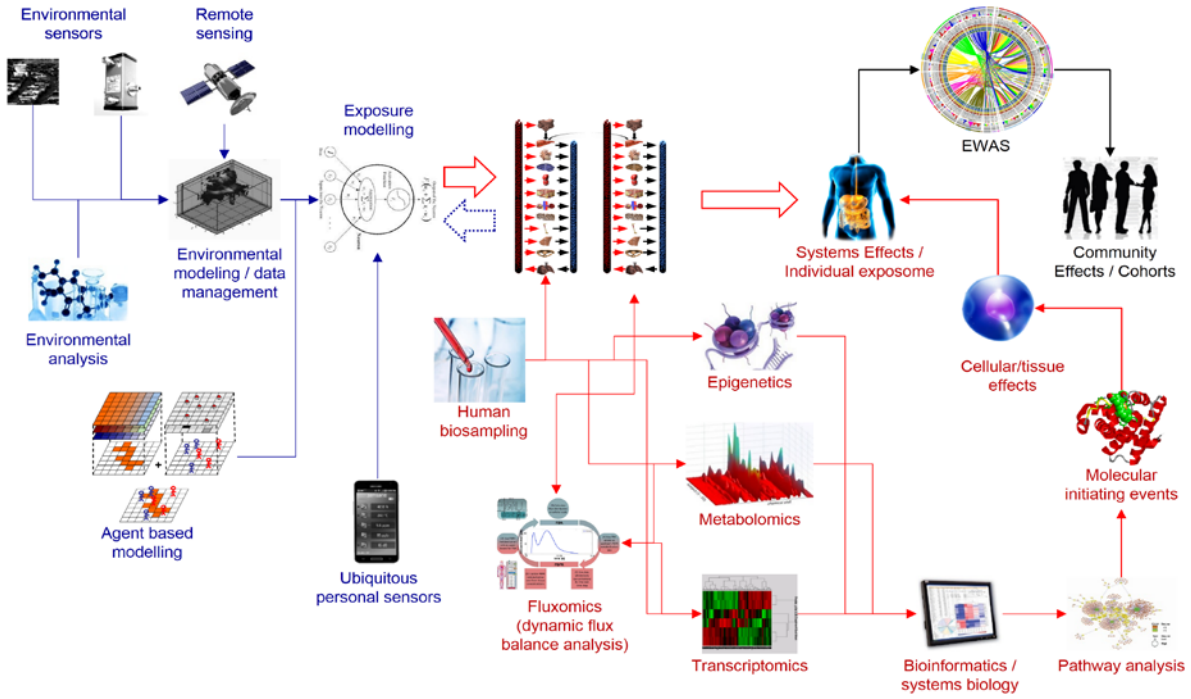


Figure 1: Overall HEALS conceptual methodological framework

Another key aspect of HEALS is the development of innovative **bioinformatics** strategies for biomarker prediction. The bioinformatics tools currently available for biomarker detection and analysis range from statistical approaches to data mining. The latter is the process of discovering valuable information from large amounts of data in the form of associations, patterns, changes, or significant structures. Data mining can be *descriptive* or *predictive*. The distinction between descriptive and predictive data mining models is in many cases unclear mainly because the same tools can be used either way. However, when prediction is under consideration, it can be reached through *classification* or *regression*. The available techniques for both tasks are numerous, with decision trees, neural networks and support vectors more widely used. An additional issue to be taken under consideration related to exposome studies is “big data”. This is the result of the huge amount of information at various levels of biological organization reflected as genome, transcriptome, proteome, metabolome accompanied by other population relevant information (exposure, medical status). The name itself suggests huge amount of data, which, however, represents only one aspect. In general, big data has four important features, so called four V’s (Li and Chen, 2014): volume of data, velocity of processing the data, variability of data sources, and veracity of the data quality.

3 Predictive biomarkers for environment-wide association health assessments

3.1 Allergies and asthma

3.1.1 Transcriptomics

Whole-transcriptome RNA sequencing was performed on nasal airway brushings from 10 control subjects and 10 asthmatic subjects, which were compared with established bronchial and small-airway transcriptomes (Poole et al., 2014). Targeted RNA sequencing nasal expression analysis was used to profile 105 genes in 50 asthmatic subjects and 50 control subjects for differential expression and clustering analyses. High-depth whole-transcriptome sequencing was used to comprehensively determine the degree to which the nasal airway serves as a biologic proxy for the bronchial airway. A novel targeted RNA sequencing (RNA-seq) technology was used, to profile gene expression of candidate airway biomarkers in a larger group of well-characterized children with asthma and healthy control subjects. These data were used to determine the relationship between the nasal transcriptome and subphenotypes of asthma. 90.2% overlap in expressed genes and strong correlation in gene expression ($\rho = .87$) between the nasal and bronchial transcriptomes was found. Previously observed asthmatic bronchial differential expression was strongly correlated with asthmatic nasal differential expression.

3.1.2 Proteomics

3.1.3 Metabolomics

Specific metabolic signatures have also been used for identifying pathologic conditions. To better understand the metabolic phenotypes of asthma, a plasma metabolic signature associated with allergic asthma in ovalbumin (OVA)-sensitized mice by using UPLC-Q-TOF/MS was investigated using sixteen metabolites, characterized as potential pathological biomarkers related to asthma (Yu et al., 2016). The identified potential biomarkers were involved in 6 metabolic pathways and achieved the most entire metabolome contributing to the formation of allergic asthma. Purine metabolism was the most prominently influenced in OVA-induced asthma mice according to the metabolic pathway analysis (MetPA), suggesting that significantly changes in inflammatory responses in the pathophysiologic process of asthma. The metabolites of purine metabolism, especially uric acid (P12) and inosine (P13), may denote their potential as targeted biomarkers related to experimental asthma. The decreased plasma uric acid (P12) suggested that inflammation responses of allergic asthma inhibited the activity of xanthine oxidase in purine metabolism, and manifested the severity of asthma exacerbation. The increased level of inosine (P13) suggests that inflammatory cells induce adenosine triphosphate (ATP) breakdown, resulting in excessive expression of adenosine deaminase (ADA) in the formation of allergic asthma. These findings provided a novel perspective on the metabolites signatures related to allergic asthma, which provided us with new insights into the pathogenesis of asthma, and the discovery of targets for clinical diagnosis and treatment. Chinese patients with mild persistent asthma using GC-MS coupled with a series of multivariate statistical analyses were investigated and clear intergroup separations existed between the asthmatic patients and control subjects (Chang et al., 2015). A list of differential metabolites and several top altered metabolic pathways were identified. The levels of succinate (an intermediate in tricarboxylic acid cycle) and inosine were highly upregulated in the asthmatic patients, suggesting a greater effort to breathe during exacerbation and hypoxic stress due to asthma. Other differential metabolites, such as 3,4-

dihydroxybenzoic acid and phenylalanine, were also identified. Furthermore, the differential metabolites possessed higher values of area under the ROC curve (AUC), suggesting an excellent clinical ability for the prediction of asthma. An integrated approach of combination of LC-GC metabolomics was used to analyze BALF from experimental asthma (Ho et al., 2013). The author investigated the effects of dexamethasone on metabolic profiles of BALF in the murine model of asthma. The findings revealed substantial alterations in energy, lipid, and sterol metabolism in BALF from an experimental murine model of asthma, including potentially important metabolites for phenotyping asthma. Urine samples of 41 atopic asthmatic children (further subdivided in sub-groups according to the symptoms) and 12 age-matched controls were analyzed. Untargeted metabolic profiles were collected by LC-MS, and studied by multivariate analysis (Mattarucchi et al., 2012). The group of the asthmatics was differentiated by a model that proved to be uncorrelated with the chronic assumption of controller drugs on the part of the patients. The distinct sub-groups were also appropriately modeled. NMR-based metabolomics indicates that obese asthmatic (OA) patients are characterized by a respiratory metabolic fingerprint fully different from patients independently affected by asthma or obesity. Such phenotypic difference strongly suggests unique pathophysiological pathways involved in the pathogenesis of asthma in adult obese subjects (Maniscalco et al., 2016). Furthermore, the OA metabotype could define a strategy for patient stratification based on unbiased biomarkers, with important diagnostic and therapeutic implications.

3.1.4 Biochemical markers

Asthma is highly induced by environmental factors such as ambient air particles. Xia et al. (2015) identified components of the Notch pathway, most notably Jagged 1 (Jag1), as targets of PM induction in human monocytes and murine dendritic cells. PM, especially ultrafine particles, upregulated TH cytokine levels, IgE production, and allergic airway inflammation in mice in a Jag1- and Notch-dependent manner, especially in the context of the proasthmatic IL-4 receptor allele Il4raR576. PM-induced Jag1 expression was mediated by the aryl hydrocarbon receptor (AhR), which bound to and activated AhR response elements in the Jag1 promoter. Pharmacologic antagonism of AhR or its lineage-specific deletion in CD11c+ cells abrogated the augmentation of airway inflammation by PM. In conclusion PM activates an AhR-Jag1-Notch cascade to promote allergic airway inflammation in concert with proasthmatic alleles. In another study, the effect of Diesel exhaust particulates (DEPs) on house dust mite (HDM)–specific memory responses was determined by using an asthma model (Brandt et al., 2015). Data from children enrolled in the Cincinnati Childhood Allergy and Air Pollution Study birth cohort were analyzed to determine the effect of DEP exposure on asthma outcomes. DEP coexposure with HDM resulted in persistent TH2/TH17 CD1271 effector/memory cells in the lungs, spleen, and lymph nodes of adult and neonatal mice. After 7 weeks of rest, a single exposure to HDM resulted in airway hyperresponsiveness and increased TH2 cytokine levels in mice that had been previously exposed to both HDM and DEPs versus those exposed to HDM alone. On the basis of these data, Brant et al. (2015) examined whether DEP exposure was similarly associated with increased asthma prevalence in children in the presence or absence of allergen exposure/sensitization in the Cincinnati Childhood Allergy and Air Pollution Study birth cohort. Early-life exposure to high DEP levels was associated with significantly increased asthma prevalence among allergic children but not among non-allergic children. These findings suggested that DEP exposure results in accumulation of allergen-specific TH2/TH17 cells in the lungs, potentiating secondary allergen recall responses and promoting the development of allergic asthma. One major proposed mechanism of air pollutants toxicity involves

the activation of t-helper type 2 (Th2) immune responses which are linked to asthma pathogenesis. Secretion of Th2 inflammatory cytokines (IL-4, IL-5, IL-9, IL-13) may lead to mucus hypersecretion and thickening and contraction of the airway smooth muscle in atopic asthmatics (Larché et al., 2003). The main direct targets of inhaled PM are airway epithelial cells and dendritic cells (DC). Many of the PM components (i.e., polycyclic aromatic hydrocarbons and metals) are redox-active and capable of inducing cellular oxidative stress and injuries including inflammation and cell death. Airway epithelial cells and antigen-presenting dendritic cells (DC) are the major and direct targets of inhaled PM. The epithelial cells can further enhance the DC response to allergen and PM through several immune regulatory cytokines including thymic stromal lymphopoietin (TSLP), IL-33, and IL-25. Among these cytokines TSLP is particularly relevant to the mechanisms by which particulate air pollutants contribute to asthma pathogenesis. Studies have found that TSLP released by PM-exposed human airway epithelial cells could polarize the DC towards a T-helper 2 immune response, which is one of the key immunological mechanisms in asthma pathogenesis. The convergence of regulatory signals generated by PM-induced oxidative stress in DC and the interactions among them may be one of the major mechanisms that are specifically related to the contribution of PM towards asthma pathogenesis (Li and Buglak, 2015). Hinks et al. (2015) performed a comprehensive assessment of TH17 cells, regulatory T cells, mucosal-associated invariant T (MAIT) cells, other T-cell subsets, and granulocyte mediators in asthmatic patients. Sixty patients with mild-to-severe asthma and 24 control subjects underwent detailed clinical assessment and provided induced sputum, endobronchial biopsy, bronchoalveolar lavage, and blood samples. Adaptive and invariant T-cell subsets, cytokines, mast cells, and basophil mediators were analyzed. Significant heterogeneity of T-cell phenotypes was observed, with levels of IL-13–secreting T cells and type 2 cytokines increased at some, but not all, asthma severities. TH17 cells and γδ-17 cells, proposed drivers of neutrophilic inflammation, were not strongly associated with asthma, even in severe neutrophilic forms. MAIT cell frequencies were strikingly reduced in both blood and lung tissue in relation to corticosteroid therapy and vitamin D levels, especially in patients with severe asthma in whom bronchoalveolar lavage regulatory T-cell numbers were also reduced. Bayesian network analysis identified complex relationships between pathobiologic and clinical parameters. Topological data analysis identified 6 novel clusters that are associated with diverse underlying disease mechanisms, with increased mast cell mediator levels in patients with severe asthma both in its atopic (type 2 cytokine–high) and nonatopic forms. The evidence for a role for TH17 cells in patients with severe asthma is limited. Severe asthma is associated with a striking deficiency of MAIT cells and high mast cell mediator levels. This study provides proof of concept for disease mechanistic networks in asthmatic patients with clusters that could inform the development of new therapies.

The systemic cysteine oxidation and its association with inflammatory and clinical features in healthy children and children with difficult-to-treat asthma has been also investigated (Stephenson et al., 2015). It was hypothesized that cysteine oxidation would be associated with increased markers of oxidative stress and inflammation, increased features of asthma severity, decreased clinically defined glucocorticoid responsiveness, and impaired GR function. PBMCs were collected from healthy children (n = 16) and children with asthma (n = 118) aged 6 to 17 years. Children with difficult-to-treat asthma underwent glucocorticoid responsiveness testing with intramuscular triamcinolone. Cysteine, cystine, and inflammatory chemokines and reactive oxygen species generation were quantified, and expression and activity of the GR were assessed. Cysteine oxidation was present in children with difficult-to-treat asthma and accompanied by increased reactive oxygen species generation and increased CCL3 and CXCL1 mRNA expression. Children with the greatest extent of cysteine oxidation had more features of asthma severity, including poorer symptom control, greater medication use, and

less glucocorticoid responsiveness despite inhaled glucocorticoid therapy. Cysteine oxidation also modified the GR protein by decreasing available sulfhydryl groups and decreasing nuclear GR expression and activity. A highly oxidized cysteine redox state promotes a posttranslational modification of the GR that might inhibit its function. Given that cysteine oxidation is prevalent in children with difficult-to-treat asthma, the cysteine redox state might represent a potential therapeutic target for restoration of glucocorticoid responsiveness in this population.

To better understand the inter- and intra-individual variability and seasonal variation of IgE, and high (FcεRI)- and low-affinity (CD23) IgE receptor expression in blood of seasonal allergic rhinitis (SAR) subjects, thirty-two otherwise healthy subjects with a history of SAR to birch pollen and a positive skin prick test to birch pollen were sampled three times out of the pollen season and three times during the pollen season (Carlsson et al., 2015). FcεRI and CD23 expressions were analysed using flow cytometry. Total IgE was analysed using ImmunoCAP(®) and free IgE was analysed with a novel customised research assay using an IgG-FcεRI-chimera protein coupled to ImmunoCAP as capture reagent, ImmunoCAP-specific IgE conjugate and ImmunoCAP IgE calibrators. The performance of the free IgE assay was compared well with the reference ImmunoCAP total IgE assay. The working range of the assay was 0.35-200 kU/l IgE. FcεRI expression on basophils and CD23 expression on B cells showed low intrasubject variability both in and out of the pollen season (<10% CV). There was a small seasonal difference with lower total IgE levels (120 versus 128 kU/l; P = 0.004) and FcεRI expression (283 versus 325 mean fluorescence intensity (MFI); P < 0.001) during the pollen season. IgE, FcεRI expression and CD23 expression fulfilled biomarker and assay requirements of variability, and allergen exposure affected the biomarkers only to a minor degree. The free IgE assay may be used for measurement of free IgE levels in patients after anti-IgE antibody treatment

3.1.5 Susceptibility markers

Genes involved in conferring susceptibility to the development of asthma can be grouped under four headings:

• Genes controlling factors involved in airway development and repair, including remodelling.12 For example, polymorphism of ADAM33, which has been linked with airway wall remodelling, is strongly associated with asthma in diverse populations.

• Genes involved in controlling the responses of the immune system. For example, a polymorphism in the TNF promoter region has been associated with asthma (and its severity).

• Genes controlling bronchial hyperresponsiveness. For example, airway smooth muscle cells producing inadequate levels of C/EBPa, due to a genetic variant, are more susceptible to contractile stimuli than normal cells.

• Genes controlling the production of endogenous anti-oxidants by the airways.18 For example, polymorphic variation of glutathione S-transferase M1, glutathione S-transferase P1 and NAD(P)H: quinine reductase, has been linked with differences in baseline lung function, with airway responsiveness to ozone and with the influence of maternal smoking on asthma.

In addition, a model was based on four SNPs (rs9522789, rs7147228, rs2701423, rs759582) and two metabolites—monoHETE_0863 and sphingosine-1-phosphate (S1P) which could predict asthma control with an AUC of 95% (McGeachie et al., 2015). Integrative ORA identified 17 significantly enriched pathways related to cellular immune response, interferon signaling, and cytokine-related

signaling, for which arachidonic acid, PGE2 and S1P, in addition to six genes (CHN1, PRKCE, GNA12, OASL, OAS1, and IFIT3) appeared to drive the pathway results. Of these predictors, S1P, GNA12, and PRKCE were enriched in the results from integrative and metabolic ORAs.

3.2 Neurodevelopmental and neurodegenerative diseases

3.2.1 Transcriptomics

In terms of transcriptomics signatures, recent studies of genomic variation associated with autism have suggested the existence of extreme heterogeneity. Large-scale transcriptomics should complement these results to identify core molecular pathways underlying autism. Gupta et al. (2014) reported results from a large-scale RNA sequencing effort, utilizing region-matched autism and control brains to identify neuronal and microglial genes robustly dysregulated in autism cortical brain. A gene expression module corresponding to M2-activation states in microglia is negatively correlated with a differentially expressed neuronal module, implicating dysregulated microglial responses in concert with altered neuronal activity-dependent genes in autism brains. These observations provided pathways and candidate genes that highlight the interplay between innate immunity and neuronal activity in the etiology of autism.

3.2.2 Proteomics

3.2.3 Metabolomics

With regard to metabolomics, different metabolite profiles have been identified in children with fetal alcohol spectrum disorders (FASD). The mechanisms underlying FASD are incompletely understood, and biomarkers to identify those at risk are lacking. From a metabolomic analysis of embryoid bodies and neural lineages derived from human embryonic stem (hES) to identify the neural secretome produced in response to ethanol (EtOH) exposure (Palmer et al., 2012). It was found that EtOH treatment induced statistically significant changes to metabolite abundance in human embryoid bodies (180 features), neural progenitors (76 features), and neurons (42 features). There were no shared significant features between different cell types. Fifteen features showed a dose-response to EtOH. Four chemical identities were confirmed: L-thyroxine, 5'-methylthioadenosine, and the tryptophan metabolites, L-kynurenine and indoleacetaldehyde. One feature with a putative annotation of succinyladenosine was significantly increased in both EtOH treatments. As a result, it was found that EtOH exposure induces statistically significant changes to the metabolome profile of human embryoid bodies, neural progenitors, and neurons. Several of these metabolites are normally present in human serum, suggesting their usefulness as potential serum FASD biomarkers. These findings suggest the biochemical pathways that are affected by EtOH in the developing nervous system and delineate mechanisms of alcohol injury during human development.

A variety of possible mechanisms by which neurotoxicants can lead to neurodevelopmental abnormalities is supported by the scientific literature. These mechanisms involve induction of oxidative stress, interfering calcium signaling, effects on neurotransmitter pathways, neuroendocrine effects and epigenetic control (Chen et al., 2011). Especially in the critical stages of nervous system development, the abovementioned effects can impact neuronal growth, differentiation, migration, synaptogenesis, and myelination, leading to an array of neurodevelopmental deficits. Many environmental toxicants, (heavy metals, PBDEs, PCBs, some pesticides etc) posses the ability to

generate ROS and deplete antioxidant capacity. Neuronal cells are especially vulnerable to oxidative stress because of the high amount of ROS generated during normal metabolism and neuronal activity (Milou et al. 2011). In vitro studies have demonstrated that environmental toxicant such as (Pb, Hg, PCBs, PBDEs, …) cause oxidative stress in neuronal cells leading to apoptotic cell death. Evidence is also present for the effect of many subclasses of pollutants to neurotransmitter pathways. Recently, an animal study discovered that repetitive postnatal PCB exposure resulted in increased levels of homovanillic and 5-hydroxyindoleacetic acid, metabolites of DA and 5-HT, in the neostriatum of young adult animals, without changing levels of the transmitters themselves. Moreover, analysis of DA-synapses demonstrated specific effects on a restricted number of specific synaptic proteins, including the presynaptic DAT, the postsynaptic D5 receptor and the PSD-95 scaffolding synapse protein (Dervola et al., 2015). Exposure to PCBs have been found to affect the levels of urinary homovanillic acid, a DA metabolite, in humans (Putschögl et al., 2015).

Autism Spectrum Disorders (ASD) are a group of developmental disorders caused by environmental and genetic factors. Diagnosis is based on behavioral and developmental signs detected before 3 years of age with no reliable biological marker. The potential use of a 2D NMR-based approach to express the global biochemical signature of autistic individuals compared to normal controls was investigated by Mavel et al. (2013). This technique has greater spectral resolution than to 1D H NMR spectroscopy, which is limited by overlapping signals. The urinary metabolic profiles of 30 autistic and 28 matched healthy children were obtained using a 1H–13C NMR-based approach. The data acquired were processed by multivariate orthogonal partial least-squares discriminant analysis (OPLS-DA). Some discriminating metabolites were identified: β-alanine, glycine, taurine and succinate concentrations were significatively higher, and creatine and 3-methylhistidine concentrations were lower in autistic children than in controls. Also, differences in several other metabolites that were unidentified but characterized by a cross peak correlation in 1H–13C HSQC were noted. Statistical models of 1H and 1H–13C analyses were compared and only 2D spectra allowed the characterization of statistically relevant changes [R2Y(cum)¼ 0.78 and Q2(cum)¼ 0.60] in the low abundance metabolites. This method has the potential to contribute to the diagnosis of neurodevelopment disorders but needs to be validated on larger cohorts and on other developmental disorders to define its specificity. Similarly, Wang et al. (2016) performed a metabolomics analysis of serum to identify potential biomarkers for the early diagnosis and clinical evaluation of autism. They analyzed a discovery cohort of patients with autism and participants without autism in the Chinese Han population using ultra-performance liquid chromatography quadrupole time-of-flight tandem mass spectrometry (UPLC/Q-TOF MS/MS) to detect metabolic changes in serum associated with autism. The potential metabolite candidates for biomarkers were individually validated in an additional independent cohort of cases and controls. They built a multiple logistic regression model to evaluate the validated biomarkers, including 73 patients and 63 controls in the discovery cohort and 100 cases and 100 controls in the validation cohort. Metabolomic analysis of serum in the discovery stage identified 17 metabolites, 11 of which were validated in an independent cohort. A multiple logistic regression model built on the 11 validated metabolites fit well in both cohorts. The model consistently showed that autism was associated with 2 particular metabolites: sphingosine 1-phosphate and docosahexaenoic acid. In another study (Emond et al., 2013)., GC-MS urinary metabolic profiles of 26 autistic and 24 healthy children were obtained by liq/liq extraction, and were or were not subjected to an oximation step, and then were subjected to a persilylation step. These metabolic profiles were then processed by multivariate analysis, in particular orthogonal partial least-squares discriminant analysis (OPLS-DA, R2Y(cum) = 0.97, Q2(cum) = 0.88). Discriminating metabolites were identified. The relative concentrations of the succinate and

glycolate were higher for autistic than healthy children, whereas those of hippurate, 3-hydroxyphenylacetate, vanillylhydracrylate, 3-hydroxyhippurate, 4-hydroxyphenyl-2-hydroxyacetate, 1H-indole-3-acetate, phosphate, palmitate, stearate, and 3-methyladipate were lower. Eight other metabolites, which were not identified but characterized by a retention time plus a quantifier and its qualifier ion masses, were found to differ between the two groups. Comparison of statistical models leaded to the conclusion that the combination of data obtained from both derivatization techniques leads to the model best discriminating between autistic and healthy groups of children.

Beyond neurodevelopmental disorders, omics biomarkers have been used for identifying metabolic signatures of neurodegenerative diseases such as Parkinson Disease (PD). PD is a multifactorial disease that is characterized by the progressive loss of dopaminergic neurons of the substantia nigra pars compacta (SN). This progressive loss of dopamine input from the SN to the striatum results in degenerative loss of motor function that manifests in bradykinesia, postural instability, tremor and rigidity (Roede et al., 2014). Epidemiological studies have identified that factors such as living in a rural area, consuming well water, farming, and pesticide exposure may be risk factors for developing PD. These observations have led to the development of an “environmental hypothesis” of PD. This hypothesis states that there are chemicals in the environment that are capable of selectively damaging the dopaminergic neurons of the SN, thus contributing to the development of PD. Two common pesticides, paraquat (PQ) and maneb (MB), have been demonstrated in vivo to preferentially alter the nigrostriatal dopamine system. A combination of the herbicide paraquat (PQ) and fungicide maneb (MB) has been linked to Parkinson's disease. Previous studies show that this involves an additive toxicity with at least two different mechanisms (Roede et al., 2014). However, detailed understanding of mixtures is often difficult to elucidate because of the multiple ways by which toxic agents can interact. In the present study, we used a combination of transcriptomics and metabolomics to investigate mechanisms of toxicity of PQ and MB in a neuroblastoma cell line. Conditions were studied with concentrations of PQ and MB that each individually caused 20% cell death and together caused 50% cell death. Transcriptomic and metabolomic samples were collected at time points prior to significant cell death. Statistical and bioinformatic methods were applied to the resulting 30,869 transcripts and 1358 metabolites. Results showed that MB significantly changed more transcripts and metabolites than PQ, and combined PQ + MB impacted more than MB alone. Transcriptome–metabolome-wide association study (TMWAS) showed that significantly changed transcripts and metabolites mapped to two network substructures, one associating with significant effects of MB and the other included features significantly associated with PQ + MB. The latter contained 4 clusters of genes and associated metabolites, with one containing genes for two cation transporters and a cation transporter regulatory protein also recognized as a pro-apoptotic protein. Other clusters included stress response genes and transporters linked to cytoprotective mechanisms. MB also had a significant network structure linked to cell proliferation. Together, the results show that the toxicologic mechanism of the combined neurotoxicity of PQ and MB involves network level interactions and that TMWAS provides an effective approach to investigate such complex mechanisms. For further elucidation of the mechanisms related to PD, cerebrospinal fluid biomarker studies focused on different disease pathways: oxidative stress, neuroinflammation, lysosomal dysfunction and proteins involved in PD and other neurodegenerative disorders, focusing on four clinical domains: their ability to (1) distinguish PD from healthy subjects and other neurodegenerative disorders as well as their relation to (2) disease duration after initial diagnosis, (3) severity of disease (motor symptoms) and (4) cognitive dysfunction. Oligomeric alpha-synuclein might be helpful in the separation of PD from controls (Andersen et al., 2016). Through metabolomics, changes in purine and tryptophan

metabolism have been discovered in patients with PD. Neurofilament light chain (NfL) has a significant role in distinguishing PD from other neurodegenerative diseases. Several oxidative stress markers are related to disease severity, with the antioxidant urate also having a prognostic value in terms of disease severity. Increased levels of amyloid and tau-proteins correlate with cognitive decline and may have prognostic value for cognitive deficits in PD. Urine metabolic phenotyping has also been associated with the development of PD. A metabolomic study with was performed using gas chromatography - mass spectrometry (GC - MS) and liquid chromatography - mass spectrometry (LC - MS) to characterize the urinary metabolic phenotypes of idiopathic PD patients at three stages (early, middle and advanced) and normal control subjects, with the aim of discovering potential urinary metabolite markers for the diagnosis of idiopathic PD by Luan et al. (2015). Both GC-MS and LC-MS metabolic profiles of idiopathic PD patients differed significantly from those of normal control subjects. 18 differentially expressed metabolites were identified as constituting a unique metabolic marker associated with the progression of idiopathic PD. Related metabolic pathway variations were observed in branched chain amino acid metabolism, glycine derivation, steroid hormone biosynthesis, tryptophan metabolism, and phenylalanine metabolism.

With regard to amyotrophic lateral sclerosis (ALS), plasma biomarkers can aid in distinguishing patients with ALS from those with disease mimics were identified (Lawton et al., 2014). In a multi-center study, plasma samples were collected from 172 patients recently diagnosed with ALS, 50 healthy controls, and 73 neurological disease mimics. Samples were analyzed using metabolomics. Using all identified biochemicals detected in > 50% of all samples in the metabolomics analysis, samples were classified as ALS or mimic with 65% sensitivity and 81% specificity by LASSO analysis (AUC of 0.76). A subset panel of 32 candidate biomarkers classified these diagnosis groups with a specificity of 90%/sensitivity 58% (AUC of 0.81). Creatinine was lower in subjects with lower revised ALS Functional Rating Scale (ALSFRS-R) scores. In conclusion, ALS can be distinguished from neurological disease mimics by global biochemical profiling of plasma samples. Our analysis identified ALS versus mimics with relatively high sensitivity. From the study, a subset of 32 metabolites were identify, discriminating patients with ALS with a high specificity. Interestingly, lower creatinine was found to correlate significantly with a lower ALSFRS-R score. Finally, molecules previously reported to be important in disease pathophysiology, such as urate, were also included in our metabolite panel. The evolution of metabolism alteration and its link with disease progression has also been described by Patin et al. (2016). They ran a study focused on (1) the evolution of metabolism disturbance during disease progression through omics approaches and (2) the relation between metabolome profile and clinical evolution. SOD1-G93A (mSOD1) transgenic mice (n = 11) and wild-type (WT) littermates (n = 17) were studied during 20 weeks. Metabolomic profile of muscle and cerebral cortex was analysed at week 20, and plasma samples were assessed at four time points over 20 weeks. The relevant metabolic pathways highlighted by metabolomic analysis were explored by a targeted transcriptomic approach in mice. Plasma metabolomics were also performed in 24 ALS patients and 24 gender- and age-matched controls. Metabolomic analysis of muscle and cerebral cortex enabled an excellent discrimination between mSOD1 and WT mice ($p < 0.001$). In another study (Gray et al., 2015), metabolomic analysis of cerebrospinal fluid (CSF) using proton nuclear magnetic resonance (1H-NMR) spectroscopy for revealing nervous system cellular pathology was used. The 1H-NMR CSF metabolomic signature of ALS was sought in a longitudinal cohort. Six-monthly serial collection was performed in ALS patients across a range of clinical sub-types (n = 41) for up to two years, and in healthy controls at a single time-point (n = 14). A multivariate statistical approach, partial least squares discriminant analysis, was used to determine differences between the NMR spectra from patients and controls. Significantly predictive

models were found using those patients with at least one year's interval between recruitment and the second sample. Glucose, lactate, citric acid and, unexpectedly, ethanol were the discriminating metabolites elevated in ALS. It is concluded that 1H-NMR captured the CSF metabolomic signature associated with derangements in cellular energy utilization connected with ALS, and was most prominent in comparisons using patients with longer disease duration. The specific metabolites identified support the concept of a hypercatabolic state, possibly involving mitochondrial dysfunction specifically. Endogenous ethanol in the CSF may be an unrecognized novel marker of neuronal tissue injury in ALS.

With regard to Alzheimer's disease (AD), an unbiased analysis of steroid-related compounds to identify novel plasma biomarkers using liquid chromatography-atmospheric pressure chemical ionization-mass spectroscopy was carried out by Sato et al. (2012). The analysis revealed that desmosterol was found to be decreased in AD plasma versus controls. To precisely quantify variations in desmosterol, we established an analytical method to measure desmosterol and cholesterol. Using this LC-based method, we discovered that desmosterol and the desmosterol/cholesterol ratio are significantly decreased in AD. Finally, the validation of this assay using 109 clinical samples confirmed the decrease of desmosterol in AD as well as a change in the desmosterol/cholesterol ratio in AD. Difference between mild cognitive impairment and control was also observed. In addition, the decrease of desmosterol was somewhat more significant in females. Receiver operating characteristic (ROC) analysis between controls and AD, using plasma desmosterol shows a score of 0.80, indicating a good discrimination power for this marker in the two reference populations and confirms the potential usefulness of measuring plasma desmosterol levels for diagnosing AD. Further analysis showed a significant correlation of plasma desmosterol with Mini-Mental State Examination scores. Plasma and serum biochemical markers proposed for Alzheimer disease (AD) are based on pathophysiologic processes such as amyloid plaque formation [amyloid β-protein (Aβ), Aβ autoantibodies, platelet amyloid precursor protein (APP) isoforms], inflammation (cytokines), oxidative stress (vitamin E, isoprostanes), lipid metabolism (apolipoprotein E, 24S-hydroxycholesterol), and vascular disease [homocysteine, lipoprotein (a)]. Most proteins or metabolites evaluated in plasma or serum thus far are, at best, biological correlates of AD: levels are statistically different in AD versus controls in some cohorts, but they lack sensitivity or specificity for diagnosis or for tracking response to therapy (Irizarry, 2004). Approaches combining panels of existing biomarkers or surveying the range of proteins in plasma (proteomics) show promise for discovering biomarker profiles that are characteristic of AD, yet distinct from nondemented patients or patients with other forms of dementia. In another study, plasma from 26 AD patients (mean MMSE 21) and 26 cognitively normal controls in a non-targeted approach using multi-dimensional mass spectrometry-based shotgun lipidomics was analysed by Han et al. (2011b) to determine the levels of over 800 molecular species of lipids. These data were then correlated with diagnosis, apolipoprotein E4 genotype and cognitive performance. Plasma levels of species of sphingolipids were significantly altered in AD. Of the 33 sphingomyelin species tested, 8 molecular species, particularly those containing long aliphatic chains such as 22 and 24 carbon atoms, were significantly lower ($p<0.05$) in AD compared to controls. Levels of 2 ceramide species (N16:0 and N21:0) were significantly higher in AD ($p<0.05$) with a similar, but weaker, trend for 5 other species. Ratios of ceramide to sphingomyelin species containing identical fatty acyl chains differed significantly between AD patients and controls. MMSE scores were correlated with altered mass levels of both N20:2 SM and OH-N25:0 ceramides ($p<0.004$) though lipid abnormalities were observed in mild and moderate AD. Within AD subjects, there were also genotype specific differences.

3.2.4 Biochemical markers

3.2.5 Susceptibility markers

Neurodevelopmental disorders with periventricular nodular heterotopia (PNH) are etiologically heterogeneous, and their genetic causes remain in many cases unknown. Missense mutations in NEDD4L mapping to the HECT domain of the encoded E3 ubiquitin ligase lead to PNH associated with toe syndactyly, cleft palate and neurodevelopmental delay (Broix et al., 2016). Cellular and expression data showed sensitivity of PNH-associated mutants to proteasome degradation. Moreover, an in utero electroporation approach showed that PNH-related mutants and excess wild-type NEDD4L affect neurogenesis, neuronal positioning and terminal translocation. Further investigations, including rapamycin-based experiments, found differential deregulation of pathways involved. Excess wild-type NEDD4L leads to disruption of Dab1 and mTORC1 pathways, while PNH-related mutations are associated with deregulation of mTORC1 and AKT activities. Altogether, these data provide insights into the critical role of NEDD4L in the regulation of mTOR pathways and their contributions in cortical development. De novo mutations in CHD8 are strongly associated with autism spectrum disorder, but the basic biology of CHD8 remains poorly understood. Chd8 knockdown during cortical development results in defective neural progenitor proliferation and differentiation that ultimately manifests in abnormal neuronal morphology and behaviors in adult mice (Durak et al., 2016) . Transcriptome analysis revealed that while Chd8 stimulates the transcription of cell cycle genes, it also precludes the induction of neural-specific genes by regulating the expression of PRC2 complex components. Furthermore, knockdown of Chd8 disrupts the expression of key transducers of Wnt signaling, and enhancing Wnt signaling rescues the transcriptional and behavioral deficits caused by Chd8 knockdown. These roles of Chd8 and the dynamics of Chd8 expression during development help negotiate the fine balance between neural progenitor proliferation and differentiation. Together, these observations provide new insights into the neurodevelopmental role of Chd8. Caubit et al. (2016) identified TSHZ3 as the critical region for a syndrome associated with heterozygous deletions at 19q12-q13.11, which includes autism spectrum disorder (ASD). In Tshz3-null mice, differentially expressed genes include layer-specific markers of cerebral cortical projection neurons (CPNs), and the human orthologs of these genes are strongly associated with ASD. Furthermore, mice heterozygous for Tshz3 show functional changes at synapses established by CPNs and exhibit core ASD-like behavioral abnormalities. These findings highlight essential roles for Tshz3 in CPN development and function, whose alterations can account for ASD in the newly defined TSHZ3 deletion syndrome. A genome-wide microRNA (miRNA) expression profiling in post-mortem brains from individuals with ASD and controls and identified miRNAs and co-regulated modules that were perturbed in ASD was performed by Wu et al. (2016). Putative targets of these ASD-affected miRNAs were enriched for genes that have been implicated in ASD risk. A regulatory relationship between several miRNAs and their putative target mRNAs in primary human neural progenitors was confirmed. These include hsa-miR-21-3p, a miRNA of unknown CNS function that is upregulated in ASD and that targets neuronal genes downregulated in ASD, and hsa_can_1002-m, a previously unknown, primate-specific miRNA that is downregulated in ASD and that regulates the epidermal growth factor receptor and fibroblast growth factor receptor signaling pathways involved in neural development and immune function. In a recent application of genome-wide association studies (GWAS) to ASD, Inoue and Inoue (2016) indicated significant associations with the single nucleotide polymorphisms (SNPs) on chromosome 5p14.1, located in a non-coding region between cadherin10 (CDH10) and cadherin9 (CDH9). An in vivo bacterial artificial chromosome (BAC) based enhancer-trapping strategy in mice to scan the gene desert for spatiotemporal cis-regulatory activities was applied. The results showed, that the ASD-associated

interval harbors the cortical area, striatum, and cerebellum specific enhancers for a long non-coding RNA, moesin pseudogene1 antisense (MSNP1AS) during the brain developing stages. Mouse moesin protein levels are not affected by exogenously expressed human antisense RNAs in transgenic brains, demonstrating the difficulty in modeling rather smaller effects of common variants. This in vivo evidence for the spatiotemporal transcription of MSNP1AS however provides a further support to connect this intergenic variant with the ASD susceptibility. With regard to human studies, to identify candidate genes for intellectual disability a meta-analysis on 2,637 de novo mutations, identified from the exomes of 2,104 patient–parent trios was performed (Lelieveld et al., 2016). Statistical analyses identified 10 new candidate ID genes: DLG4, PPM1D, RAC1, SMAD6, SON, SOX5, SYNCRIP, TCF20, TLK2 and TRIP12. In addition, it was showed that these genes are intolerant to nonsynonymous variation and that mutations in these genes are associated with specific clinical ID phenotypes. In addition, mutations in the aspartate/glutamate mitochondrial transporter, SLC25A12, have been associated with ASD (West et al., 2014).

Attention-deficit hyperactivity disorder (ADHD) is a prevalent and highly heritable disorder of childhood with negative lifetime outcomes. Although candidate gene and genome-wide association studies have identified promising common variant signals, these explain only a fraction of the heritability of ADHD. The observation that rare structural variants confer substantial risk to psychiatric disorders suggests that rare variants might explain a portion of the missing heritability for ADHD. A large-scale next-generation targeted sequencing study of ADHD in 152 child and adolescent cases and 188 controls across an a priori set of 117 genes was performed by Hawi et al. (2016). A multi-marker gene-level analysis of rare (<1% frequency) single-nucleotide variants (SNVs) revealed that the gene encoding brain-derived neurotrophic factor (BDNF) was associated with ADHD at Bonferroni corrected levels. Sanger sequencing confirmed the existence of all novel rare BDNF variants. BDNF is a genetic risk factor for ADHD, potentially by virtue of its critical role in neurodevelopment and synaptic plasticity.

For PD, mutations in the PINK1 gene are associated with early onset autosomal recessive parkinsonism (EOP), which is characterized by a phenotypic presentation that, although variable, generally overlaps with that of idiopathic Parkinson Disease (PD) (Rango et al., 2013). Rango et al. detailed clinical and brain metabolomic analysis of a sporadic Italian patient carrying the novel association of compound heterozygous A168P/W437X mutations in PINK1, including brain magnetic resonance spectroscopy (MRS) with assessment of brain mitochondrial function, PET, and SPECT. They also studied the brain metabolomics of a control group of healthy subjects and a group of PD patients; both groups were age-matched and sex-matched., Also, several proteins encoded by PD-related genes are associated with mitochondria including PTEN-induced putative kinase 1 (PINK1), which was first identified as a gene that is upregulated by PTEN. Loss-of-function PINK1 mutations induce mitochondrial dysfunction and, ultimately, neuronal cell death. Mutations in the DJ-1 gene PARK7 are associated with hereditary recessive early onset PD (Andersen et al., 2016). Mutations in the PARK2 gene coding for Parkinson, an enzyme in the ubiquitin-proteasome system, are responsible for autosomal recessive juvenile parkinsonism. Genetic mutations of transthyretin (TTR), a protein transporting thyroxin (T4) and retinol, are related to familial amyloid polyneuropathy, and TTR seems to be related to diseases such as AD, PD and psychiatric disorders. In patients with early PD, higher urate levels conferred a milder clinical and radiographic progression of the disease (Mehta and Adler, 2015). An interesting genetic study went a step further to assess causality. DNA from 808 PD patients was genotyped for 3 SLC2A9 single-nucleotide polymorphisms (SNPs) that identify an allele associated with lower urate concentrations. They found that SNPs in SLC2A9 predicted differences in serum urate and the rate of

progression to a level of disability requiring dopaminergic treatment was faster among those patients carrying the SLC2A9 genotypes associated with lower serum urate.

Regarding ALS, patients with D90A mutation on superoxide dismutase (SOD1) gene conformed the group of subjects more differentiated compared to other mutations and to sporadic ALS cases. In part that different fingerprint was due to decreased levels of primarily amino acids in CSF from carriers of D90A mutation (Ibanez et al., 2015)

3.3 Obesity and childhood diabetes

3.3.1 Transcriptomics

T2D is strongly correlated with other complex disorders such as obesity, insulin resistance, the metabolic syndrome, dyslipidemia, cardiovascular disease, and a variety of diabetic complications (Jenkinson et al., 2016). Transcriptomics data from different tissues including beta-cells, pancreatic islets, arterial tissue, peripheral blood mononuclear cells, liver, and skeletal muscle of 228 samples were integrated with protein-protein interaction data and genome scale metabolic models to unravel the molecular and tissue-specific biomarker signatures of type 2 diabetes mellitus (Calimlioglu et al., 2015). Classifying differentially expressed genes, reconstruction and topological analysis of active protein-protein interaction subnetworks indicated that genomic reprogramming depends on the type of tissue, whereas there are common signatures at different levels. Basal fasting RNA was extracted from adipose tissue biopsies from a subset of 75 unrelated individuals, and gene expression data generated on the Illumina BeadArray platform(Winnier et al., 2015). The number of gene probes with significant expression above baseline was approximately 31,000. A multiple regression analysis of all probes with 15 metabolic traits was performed. Adipose tissue had 3,012 genes significantly associated with the traits of interest (false discovery rate, FDR ≤ 0.05). The significance of gene expression changes was used to select 52 genes with significant (FDR ≤ 10-4) gene expression changes across multiple traits. Gene sets/Pathways analysis identified one gene, alcohol dehydrogenase 1B (ADH1B) that was significantly enriched (P < 10-60) as a prime candidate for involvement in multiple relevant metabolic pathways.

3.3.2 Proteomics

3.3.3 Metabolomics

With regard to diabetes, it has been found that diabetic men showed higher circulating levels of glucose, triglyceride, oxidized low-density lipoprotein (LDL), high-sensitivity C-reactive protein, interleukin (IL)-6, tumor necrosis factor-alpha (TNF-α), homeostasis model assessment-insulin resistance, urinary 8-epi-prostaglandin F_{2a} (8-epi-$PGF_{2\alpha}$) and ba-PWV than nondiabetic man (Ha et al., 2012). In plasma, 19 metabolites including three amino acids, eight acylcarnitines, six lysophosphatidylcholines (lysoPCs), and two lysophosphatidylethanolamines (lysoPEs; C18:2 and C22:6) significantly increased in diabetes men, whereas serine and lysoPE (C18:1) decreased. Decanoyl carnitine, lysoPCs (C14:0, C16:1, C18:1 and C22:6) and lysoPE (C18:1) with variable importance in the projection values >1·0 were major plasma metabolites that distinguished nondiabetic and diabetic men. Decanoyl carnitine positively correlated with oxidized LDL, 8-epi-PGF2α, IL-6, TNF-α and ba-PWV.

ba-PWV correlated positively with lysoPCs C14:0 and C16:1, and negatively with lysoPE C18:1. 8-epi-PGF2α correlated positively with lipoprotein-associated phospholipase A2, ba-PWV and lysoPCs (C14:0 and C16:1). The receiver operating characteristic curve estimation suggested that decanoyl carnitine and lysoPC (C14:0) are the best metabolites for predicting the risk of developing diabetes. Circulating lipid-related intermediate metabolites can be closely associated with inflammation, oxidative stress and arterial stiffness in early diabetes. The human plasma phospholipids in T2DM and DN have been also characterized by Zhu et al. (2011) to identify potential biomarkers of T2DM and DN. Normal phase liquid chromatography coupled with time of flight mass spectrometry (NPLC–TOF/MS) was applied to the plasma phospholipids metabolic profiling of T2DM and DN. As a result, 18 compounds in 7 PL classes with significant regulation in patients compared with healthy controls were regarded as potential biomarkers for T2DM or DN. Among them, 3 DM-specific biomarkers, 8 DN-specific biomarkers and 7 common biomarkers to DM and DN were identified. Ultimately, 2 novel biomarkers, i.e., PI C18:0/22:6 and SM dC18:0/20:2, can be used to discriminate healthy individuals, T2DM cases and DN cases from each other group. NMR-based metabolomic analysis in conjunction with multivariate statistics was applied to examine the urinary metabolic changes in two rodent models of type 2 diabetes mellitus as well as unmedicated human sufferers. The db/db mouse and obese Zucker (fa/fa) rat have autosomal recessive defects in the leptin receptor gene, causing type 2 diabetes (Salek et al., 2007). 1H-NMR spectra of urine were used in conjunction with uni- and multivariate statistics to identify disease-related metabolic changes in these two animal models and human sufferers. Metabolic similarities between the three species examined, including metabolic responses associated with general systemic stress, changes in the TCA cycle, and perturbations in nucleotide metabolism and in methylamine metabolism. All three species demonstrated profound changes in nucleotide metabolism, including that of N-methylnicotinamide and N-methyl-2-pyridone-5-carboxamide, which may provide unique biomarkers for following type 2 diabetes mellitus progression. A total of 19 serum amino acids in T2DM patients and non-diabetics were measured by Drabkova et al. (2015) and ther were 9 amino acids, which were significantly different in these groups ($p<0.05$). Significantly decreased levels of arginine, asparagine, glycine, serine, threonine and significantly increased levels of alanine, isoleucine, leucine, valine in diabetics were found. Significant difference in metabolism of amino acids between diabetics and non-diabetics were observed. The altered levels of amino acids in diabetic patients could be a suitable predictor of diabetes. Using comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry (GC × GC–TOFMS) coupled with pattern recognition methods, diabetic patients and healthy controls could be correctly distinguished based on the metabolic abnormity in plasma (Li et al., 2009). Five potential biomarkers including glucose, 2-hydroxyisobutyric acid, linoleic acid, palmitic acid and phosphate were identified. It was found that elevated free fatty acids were essential pathophysiological factors in diabetes mellitus which reflected either the hyperglycemia or the deregulation of fatty acids metabolism. These potential biomarkers in plasma, e.g. palmitic acid, linoleic acid and 2-hydroxybutyric acid might be helpful in the diagnosis or further study of diabetes mellitus. Similar results have been also obtained from the metabolomics study of Zhao et al. (2010), where normal glucose tolerant (NGT) and IGT subjects were clustered in two distinct groups independent of the investigated metabonome clearly. Pre-diabetes associated alterations in fatty acid-, tryptophan-, uric acid-, bile acid-, and lysophosphatidylcholine-metabolism, as well as the TCA cycle were identified, while Han et al. (2011a) showed that several species of arachidonic acids especially the class of C20 fatty acids might be useful indicators for distinguishing pathological abnormalities among populations in the developments of T2D and obesity. Using 2 highly sensitive metabolomic techniques, distinct serum profile change of a wide range of metabolites from healthy persons to type 2 diabetes mellitus have been reported (Xu et al., 2013). Apart from glucose,

IFG and diabetes mellitus are characterized by abnormalities in amino acid, fatty acids, glycerophospholipids, and sphingomyelin metabolism. These early broad-spectrum metabolic changes emphasize the complex abnormalities present in a disease defined mainly by elevated blood glucose levels. Similarly, Kim et al. (2012) demonstrated that oxidized -LDLM lysoPC (14:0), lycoPC (16:1), high sensitive-C reactive protein (hs-CRP) and interleukin-6 (IL-6) were strongly linked to the events of diabetic complications and higher arterial stiffness in T2D. Also, significant differences on global metabolism were also identified in the a metabonomic study involving 74 patients who were newly diagnosed with type 2 diabetes mellitus (T2DM) and a 48-week single drug treatment of repaglinide, metformin or rosiglitazone (Bao et al., 2009). As compared with the biochemical indices (FPG, 2hPG, HbA1c), T-predicted score showed different impacts on global metabolism by three treatments, respectively. Metabonomic analysis can reveal different treatment effects and provide a novel nonglucose based evaluation strategy for T2DM. A study investigated the differences in plasma metabolomic profiling between overweight/obese and normal-weight men using UPLC−Q-TOF MS (Kim et al., 2010). Overweight/obese (n = 30) and age-matched, normal-weight men (n = 30) were included. Three lyso-phosphatidylcholine (lysoPC) (lysoPC C14:0, lysoPC C18:0, and lysoPC C18:1) were identified as potential plasma markers and confirmed eight known metabolites (valine, leucine, propionyl carnitine, butyryl carnitine, tryptophan, hexanoyl carnitine, and l-carnintine) for overweight/obesity. Differences in plasma concentrations of >350 metabolites in fasted obese T2DM vs. obese non-diabetic African-American women were investigated principal components analysis to identify 158 metabolite components that strongly correlated with fasting HbA1c over a broad range of the latter ($r = -0.631$; $p<0.0001$) were utilized by Fiehn et al. (2010). In addition to many unidentified small molecules, specific metabolites that were increased significantly in T2DM subjects included certain amino acids and their derivatives (i.e., leucine, 2-ketoisocaproate, valine, cystine, histidine), 2-hydroxybutanoate, long-chain fatty acids, and carbohydrate derivatives. Leucine and valine concentrations rose with increasing HbA1c, and significantly correlated with plasma acetylcarnitine concentrations. It is hypothesized that this reflects a close link between abnormalities in glucose homeostasis, amino acid catabolism, and efficiency of fuel combustion in the tricarboxylic acid (TCA) cycle. Metabolomic profiling of obese versus lean humans reveals a branched-chain amino acid (BCAA)-related metabolite signature that is suggestive of increased catabolism of BCAA and correlated with insulin resistance. To test its impact on metabolic homeostasis, we fed rats on high-fat (HF), HF with supplemented BCAA (HF/BCAA) or standard chow (SC) diets (Newgard et al., 2009). Despite having reduced food intake and weight gain equivalent to the SC group, HF/BCAA rats were equally insulin resistant as HF rats. Pair-feeding of HF diet to match the HF/BCAA animals or BCAA addition to SC diet did not cause insulin resistance. Insulin resistance induced by HF/BCAA feeding was accompanied by chronic phosphorylation of mTOR, JNK, and IRS1(ser307), accumulation of multiple acylcarnitines in muscle, and was reversed by the mTOR inhibitor, rapamycin. the Findings showed that in the context of a poor dietary pattern that includes high fat consumption, BCAA contributes to development of obesity-associated insulin resistance.

3.3.4 Biochemical markers

3.3.5 Susceptibility markers

Metabolic changes and the differences in inflammatory markers, oxidative markers and arterial stiffness between early diabetic and nondiabetic subjects remains relatively unstudied (Ha et al., 2012). Compared to non-diabetic men, patients with newly diagnosed type 2 diabetes also showed higher

concentrations of TNF-a, IL-6 and hs-CRP. Higher arterial stiffness assessed by ba-PWV was also identified in patients with diabetes and represents a composite risk factor to identify patients with early atherosclerotic change. Along with the expected glycosuria, changes in the excretion of TCA cycle intermediates, polyols, amines, and amino acids were detected. A profound perturbation in nucleotide metabolism, previously linked with peroxisome proliferation, was also observed and may indicate a metabolic consequence of substrate excess in many tissues, especially the liver (Salek et al., 2007).

In terms of genetic polyrphisms, large-scale sample populations have revealed that there are more than 56 SNPs (Abu Bakar et al., 2015). The SNPs associated with diabetes are mostly found in non-coding regions of the genome that do not encode protein sequences. Regulation of the genomes inherently affects the gene expression and transcription. The SNPs in these metabolic genes are metabolically linked to particular enzymes that expressively lead to the production of specific metabolites of associated genes of interest. Transcription factor 7-like 2 (TCF7L2) polymorphism rs7903146 is identified to be the strongest genetic marker in type 2 diabetes, especially among the Caucasians (Abu Bakar et al., 2015). A SNP in the minor T-allele of rs1260326 in glucokinase (hexokinase 4) regulator (GCKR) might play its role in reducing the risk of type 2 diabetes susceptibility by lowering triglyceride accumulation and dyslipidemia and improving the fasting insulin and glucose level of the subjects. In addition, GCKR is a major pleiotropic risk locus associated with diabetes (Suhre et al., 2011). The melatonin-receptor gene, MTNR1B is linked to the changes in fasting glucose concentrations in T2D (Abu Bakar et al., 2015). Alos, the dysregulation of the phenylalanine hydroxylase gene by the hepatocyte nuclear factor 1a (Hnf1a) in phenylalanine metabolism is observed in T2D. Rare mutations in both KCNJ11 and PPARG are also known to be causal for certain rare monogenic syndromes (neonatal diabetes and lipodystrophies) characterized by severe metabolic disturbance of β-cell function and insulin resistance (Prokopenko et al., 2008). Specific defects in glucagon-like peptide 1-stimulated insulin secretion, glucose-stimulated insulin secretion, insulin exocytosis, insulin granule docking or post-transcriptional processing of insulin have been demonstrated to be associated with different variants, supporting the notion that a range of biological processes are involved in the pathogenesis of type 2 diabetes (Grarup et al., 2014). Novel loci (**Error! Reference source not found.**) such as KCNQ1 and C2CD4A, have been also associated with type 2 diabetes in Japanese individuals (Grarup et al., 2014).

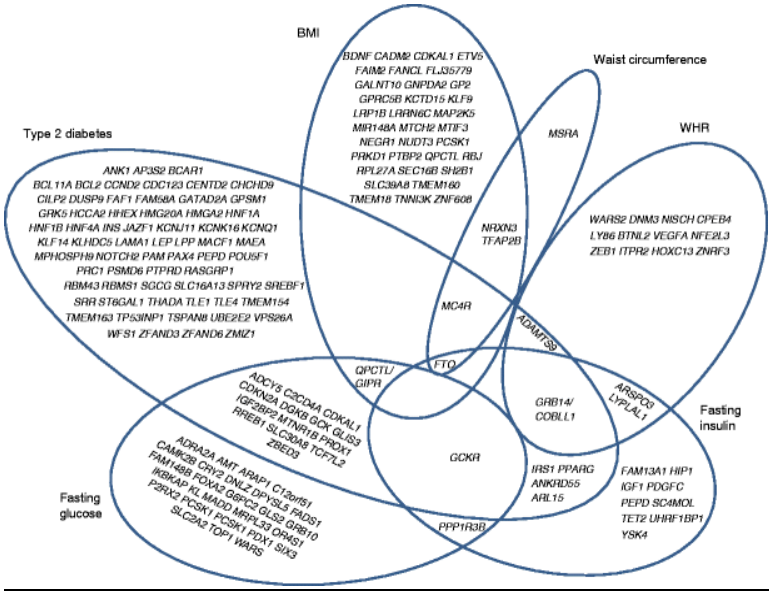


Figure. 1. Venn diagram of intersection between loci associated at genome-wide significance with type 2 diabetes, measures of adiposity and glucose homeostasis. Genome-wide significant associations for six metabolic traits are shown. Gene symbols shown in the plot are by convention the closest gene and not necessarily the functional gene (Grarup et al., 2014).

FTO and MC4R are two identified obesity locus. PCSK1, GP2 and GALNT10 loci in Asian or African populations have been identified for risk variants for obesity (Grarup et al., 2014). Other loci are: MC4R, POMC, LEPR, BDNF, SH2B1, PCSK1 and NTRK2. Five loci are shared both in diabetes and obesity: FTO, MC4R, ADAMTS9, GRB14/COBLL1 and QPCTL/GIPR. GRB14/COBLL1 is an example of an obesity-associated locus with pleiotropic effects on a range of phenotypes related to type 2 diabetes where all associations follow the expected metabolically unhealthy profile. Among genes identified through adipose tissue expression profiling to be regulated by obesity, 16 genes were selected that were encoded in four QTLs for human obesity susceptibility. In a comprehensive allelic association analysis (Jiao et al., 2008), SNPs in PFKFB3 and one haplotype in IRF5 were associated with obesity in cohort 1 with a nominal P-value <0.01. Though no statistically significant association with obesity was observed in a second cohort, PFKFB3 rs1064891 displayed a similar trend of association with obesity in both cohorts and for this SNP 95% confidence intervals of odds ratios from the two cohorts overlapped. Also, a total of 118 polymorphisms in 16 genes were analyzed for association with obesity (Jiao et al., 2008). Single nucleotide polymorphism rs1064891, located in the 3' UTR of the 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 3 (PFKFB3) gene, was nominally associated with obesity in combined analysis of cohorts 1 and 2 (P=0.007) and, in men that were lean or had severe obesity, with BMI (P=or<0.005).

Known obesity quantitative trait loci (QTL) information for the model allowed to further filter genes for increased likelihood of being causal or secondary for obesity. This successfully identified several genes previously linked to obesity (C1qr1, and Np3r) as positional QTL candidate genes elevated specifically in F line adipose tissue. A number of novel obesity candidate genes were also identified (Thbs1, Ppp1r3d, Tmepai, Trp53inp2, Ttc7b, Tuba1a, Fgf13, Fmr) that have inferred roles in fat cell

function (Morton et al., 2011). Quantitative microarray analysis was then applied to the most phenotypically divergent adipose depot after exaggerating F and L strain differences with chronic high fat feeding which revealed a distinct gene expression profile of line, fat depot and diet-responsive inflammatory, angiogenic and metabolic pathways. Selected candidate genes Npr3 and Thbs1, as well as Gys2, a non-QTL gene that otherwise passed our enrichment criteria were characterized, revealing novel functional effects consistent with a contribution to obesity. Known obesity quantitative trait loci (QTL) information for the model allowed to further filter genes for increased likelihood of being causal or secondary for obesity. This successfully identified several genes previously linked to obesity (C1qr1, and Np3r) as positional QTL candidate genes elevated specifically in F line adipose tissue. A number of novel obesity candidate genes were also identified (Thbs1, Ppp1r3d, Tmepai, Trp53inp2, Ttc7b, Tuba1a, Fgf13, Fmr) that have inferred roles in fat cell function (Morton et al., 2011). Quantitative microarray analysis was then applied to the most phenotypically divergent adipose depot after exaggerating F and L strain differences with chronic high fat feeding which revealed a distinct gene expression profile of line, fat depot and diet-responsive inflammatory, angiogenic and metabolic pathways. Selected candidate genes Npr3 and Thbs1, as well as Gys2, a non-QTL gene that otherwise passed our enrichment criteria were characterized, revealing novel functional effects consistent with a contribution to obesity. The linkage of type 2 diabetes mellitus to chromosome 10q1 has been suggestive. 228 microsatellite markers were genotyped in Icelandic individuals with type 2 diabetes and controls throughout a 10.5-Mb interval on 10q. A microsatellite, DG10S478, within intron 3 of the transcription factor 7–like 2 genes (TCF7L2; formerly TCF4) was associated with type 2 diabetes (P = 2.1 times 10-9). This was replicated in a Danish cohort (P = 4.8 times 10-3) and in a US cohort (P = 3.3 times 10-9). Compared with non-carriers, heterozygous and homozygous carriers of the at-risk alleles (38% and 7% of the population, respectively) have relative risks of 1.45 and 2.41. This corresponds to a population attributable risk of 21%. The TCF7L2 gene product is a high mobility group box–containing transcription factor previously implicated in blood glucose homeostasis. It is thought to act through regulation of proglucagon gene expression in enteroendocrine cells via the Wnt signaling pathway2 (Grant et al., 2006).

4 Practical example from HEALS: identifications of biomarkers for neurodevelopmental disorders

4.1 Cohort Studies

Two cohorts were chosen to be analysed within the HEALS framework, PHIME and Repro, in order to adapt and optimize -omics protocols for application on biological matrices for existing and new cohorts, due to the fact that both gene polymorphisms and environmental factors were obtained for a high proportion of study subjects, which will allow the assessment of impact of genetic susceptibility in vulnerability to environmental risk.

An untargeted metabolomics analysis was performed using maternal urine and plasma by both NMR and UPLC-MS/MS platforms for both Repro PL and PHIME cohorts. The REPRO PL cohort provided 149 cord blood samples, with associated subject metadata. The PHIME cohort was comprised of 2 subsets, 165 mothers' samples and 135 children samples. These mothers' and children samples within PHIME were extracted and analysed as two separate metabolomics experiments to allow separate downstream data analysis. The metabolomic analyses performed is described in greater detail in D5.3 Report on best practices for –omics analysis performed on human cohorts.

4.1.1 Data pre-processing

The main goal in data pre-processing or spectral processing is to correctly arrange the huge amount of raw data generated by the MS files into a 2D matrix in which each column represents one sample (object) and each row one ion detected by the MS (variable). Variables are expressed by the combination of three characteristics, two that define the identity of the ion, that is m/z and RT, and the ion intensity measured for each sample. Data pre-processing or spectral processing also aimed at the improvement of signal quality and the reduction of the possible biases present in the raw data, mainly due to batch to batch extraction processes. In most NMR- and MS-based spectra metabolomics studies, the spectral processing includes several steps· baseline correction, filtering, peak detection, peak alignment, normalization, deconvolution (Alonso et al., 2015). There are several open source software available tools providing different methodological options for spectral processing. The most important criteria for the selection of the suitable tool are 1) the analysed biological sample, 2) the analytical technology used, and 3) the tool that has been chosen for further data analysis or pathway analysis, since the generated feature matrix should be in a format compatible with the tool. In most studies, the combination of different metabolomics spectral processing tools seems to be necessary, due to the different features included in each tool. For example, MZmine is recommended for the analysis of human biological samples analysed but LC-MS, but the batch effect correction should be performed using mathematical methods that do not depend on quality control samples (Rusilowicz et al., 2016) (Salerno and Mehrmohamadi, 2017). The following methodology workflows for both plasma and urine samples have been finalised after testing different combinations of tools and evaluate the results, to maximize the well-use of each tool based on the requirements of an exposomic study.

4.1.1.1 LC-MS data pre-processing

Data was acquired using the ThermoFisher Scientific model LTQ Orbitrap Discovery MS, and spectral data processing was performed using the MZMine v.2.21 open-software (Katajamaa et al., 2006). Raw data generated from negative and positive ionization are treated as two different experiments.

The main premise in a metabolomics study is that the levels of the detected metabolites reflect the biological status of the system under study. Thus, a quality assurance strategy was implanted in sample preparation and analysis to ensure the quality of the results and to minimize and detect any source of variation (instrumental and experimental drift) unrelated to the biological nature of the samples. Pooled QC samples were employed, since their nature is equal to the problem samples. The following parameters for noise removal, mass detection, deconvolution, data transformation, data reduction etc. were set based on the behavior of QC samples, because QC samples providing an average of all the metabolomes analysed in the study. The presented figures are referring to the analysis of children samples from PHIME cohort in negative mode.

Baseline correction was used to remove low-frequency artifacts and differences between samples that are generated by experimental and instrumental variation. After this, the application of high-frequency filters it was necessary to remove the electronic noise present in the data that is generated by the measurement equipment (Alonso et al., 2015). The use of smoothing, for which the main purpose is to remove high-frequency noise from the chromatograms prior to deconvolution, is highly recommended. The decision of how many mean values will be taken into consideration is crucial and requires testing. The asymmetric baseline corrector was used as the correction method, while the smoothing and the asymmetry were set at 1000 and 0.95 respectively, after optimization. After baseline correction the noise should be minimize at least 1.0E1. For, example in case of the analysis in negative mode of children samples from PHIME cohort the baseline of the chromatogram was at 3.0E5 and transferred at 1.0E4.

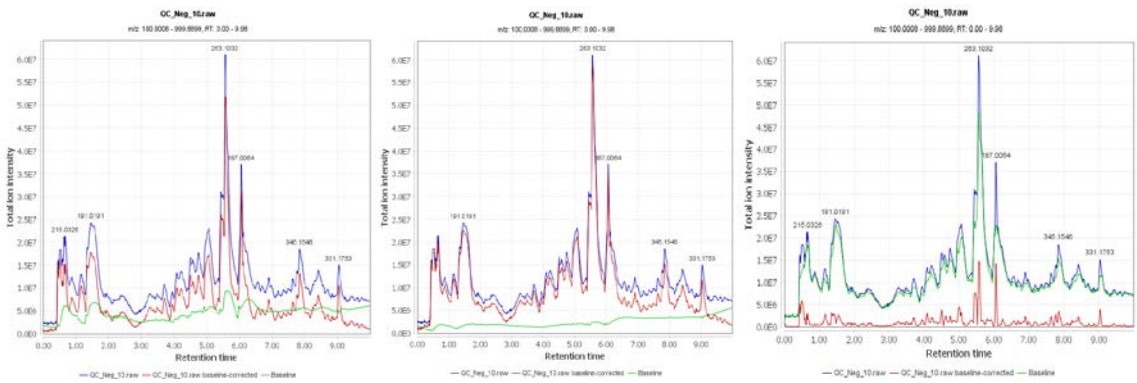
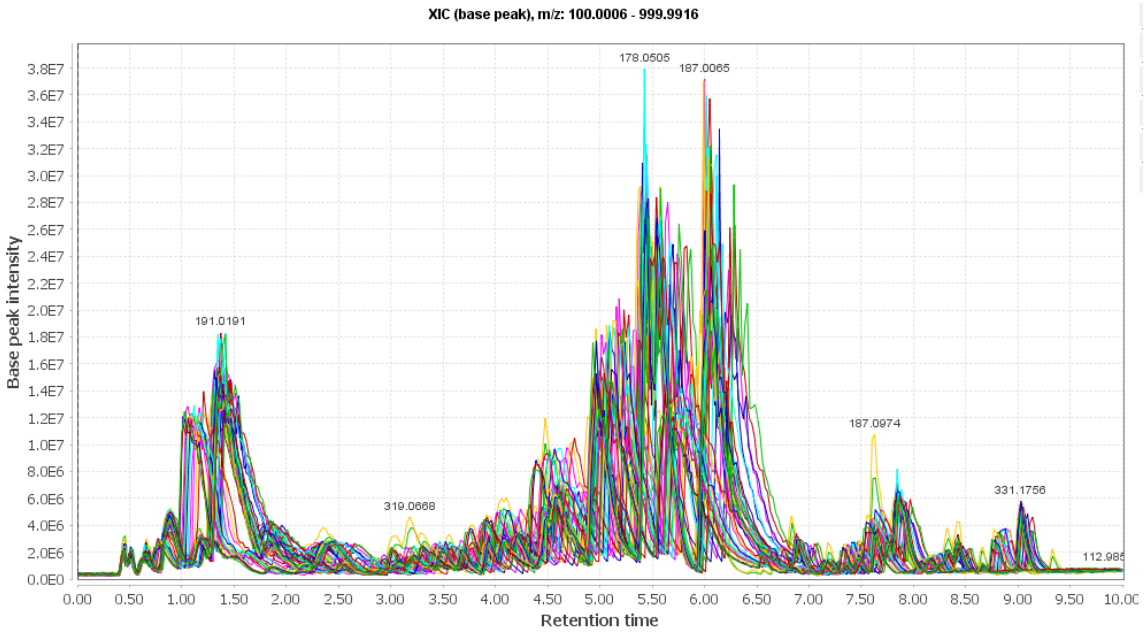


Figure 2. The parameters of smoothing and asymmetry were chosen after three trials. For the first figure from the left the asymmetry was set at 0.0095. For the second figure the asymmetry was set at 0.095 and for the third at 0.95. The noise level of the third figure presents a significant decrease.

A


XIC (base peak), m/z: 100.0006 - 999.9916
Base peak intensity
Retention time
178.0505
187.0065
191.0191
319.0668
187.0974
331.1756
112.985

B

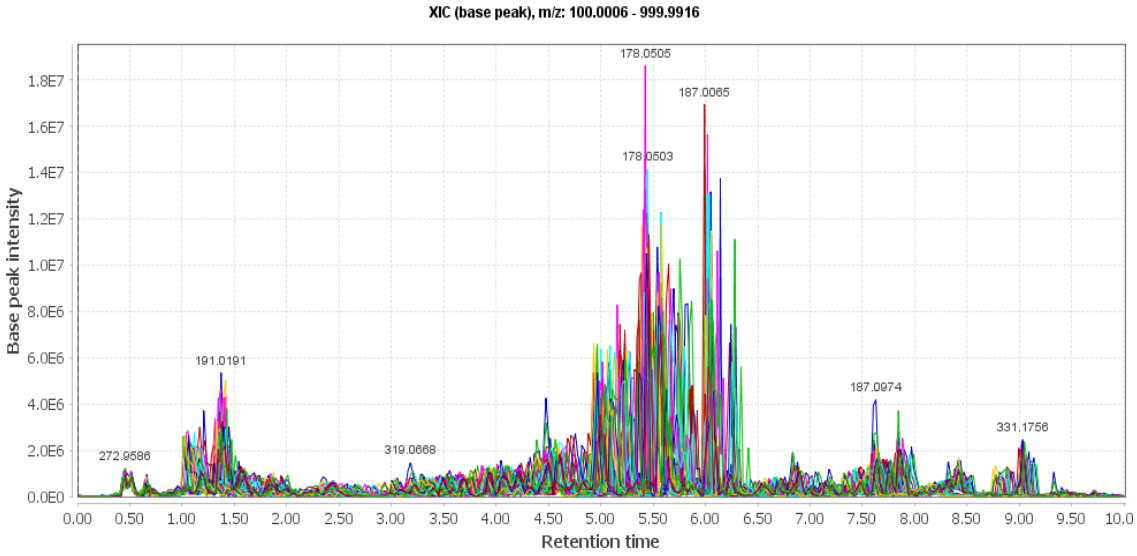


Figure 3. Superimposed TIC of QCs from PHIME mothers' samples. A) Before baseline correction and B) after Baseline correction.

Peak detection algorithms analyze each sample spectrum independently. The different metabolites are identified using one or multiple detection thresholds. These thresholds were applied to different parameters such as the signal-to-noise ratio, the intensity or the area of each peak (Alonso et al., 2015). The Mass detection module generates a list of masses (ions) for each scan in the raw data file. Several algorithms are provided for this step. The choice of the optimal algorithm depends on the raw data characteristics (mass resolution, mass precision, peak shape, noise). For example, if the raw data is already centroid, only Centroid mass detector algorithm can be used, which simply assumes that each signal above given noise level is a detected ion. Note that depending on the instrument, MS data can be recorder in two different modes. The generated files from a Fullscan in Orbitrap, for example, are in mzXML, and these files can be acquired using profile or centroid mode. In centroid mode, each ion is represented as a discrete m/z, intensity pair, while in profile mode the ions are represented by peaks each containing a collection of points. The findings from the different testing results when defining the methodology showed that is preferable to acquire directly in centroid mode, which gives a smaller file and requires less processing time. Moreover, the choice of noise level value depends from the detected peaks that do not represent metabolites and must be excluded from the further analysis. In other words, the choice of noise level value depends on the detected peak of a metabolite with the minimum height.

The second module of peak detection, chromatogram builder, takes the mass lists generated by the mass detection step and builds a chromatogram for each mass that can be detected continuously by the scans. At the end, each chromatogram had more than one peak and that is why the chromatogram deconvolution was a necessary step. In this step, m/z tolerance and ppm must be defined. The m/z tolerance refers to the absolute difference given normally in Da, in amu or u (unified atomic mass unit).

The ppm is the relative tolerance. MZmine calculates the range of tolerance with the maximum of the absolute and relative tolerances.

$$\text{ppm} = \frac{\text{observed mass} - \text{calculated mass}}{\text{calculated mass}} \cdot 10^6$$

For example, when caffeine is used in a typical experiment the following applies:

$$\text{ppm} = \frac{195.0866 - 195.0877}{195.0866} \cdot 10^6$$

$$\text{tolerance} \sim 5\text{ppm}$$

Usually chosen values for TOF are 10 to 15 ppm from 0.003 to 0.004 m/z and for Orbitrap 5 ppm and m/z below 0.0015. In order to define its value, internal standards spectrums, such as caffeine or reserpine, must be used as a reference. This is the reason why reference samples in the beginning of each experiment are used and sometimes runs are repeated during the sequence in non-specific time or at three-time points, begin, middle and at the end of the analysis. Then the observed mass of the reference sample is related to the calculated one to estimate the relative tolerance needed for MZmine as shown above.

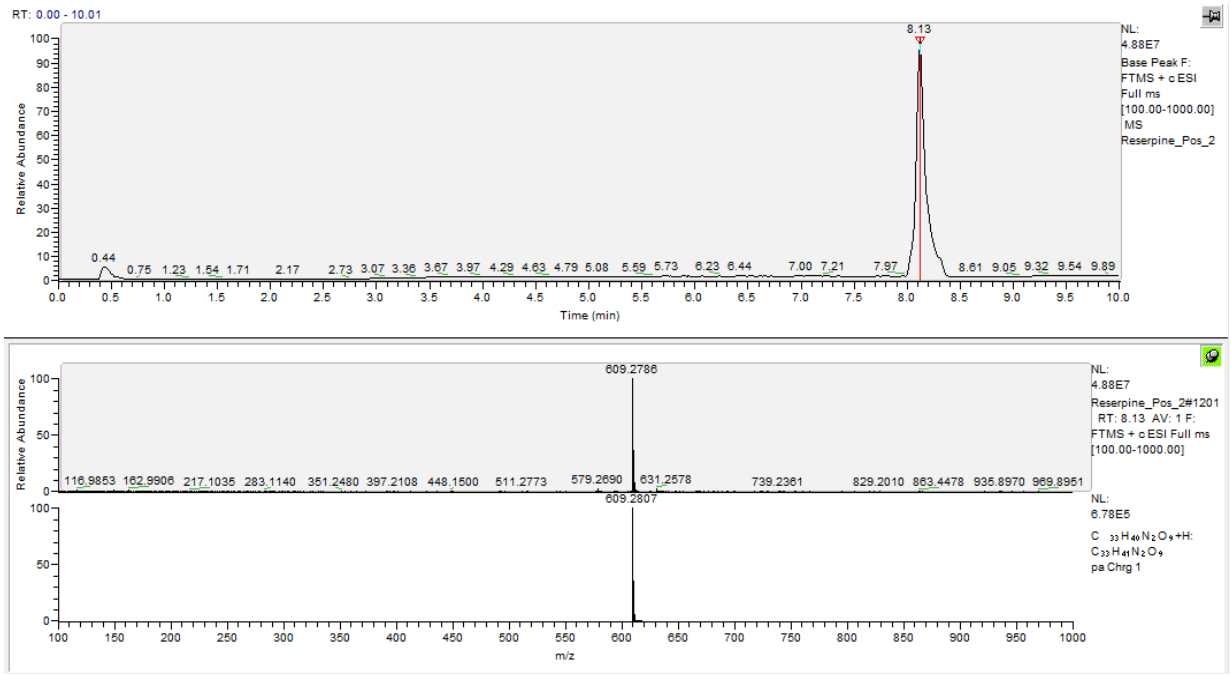


Figure 4. The first figure shows the MS for the reserpine, while the second one presents the calculated MS, according to the bibliography. The Full MS was opened using XCalibur Software, while the same was also used for the calculation of the mass value for reserpine using the isotope simulation mode. The value of most abundant

must be equal to +1 and H has been used as adduct. According to the above, the $mass_{error}$ was 3.447 ppm, and the uncertainty 0.0021 m/z.

Peak overlap is a common problem in MS-based studies. Overlapping peaks are treated as one. To attempt to solve this problem, chromatogram deconvolution methods have been developed (Alonso et al., 2015). One of the most common choice to perform deconvolution is an algorithm called Local Minimum, which also was the best choice for Repro PL and PHIME data sets. The peak with the higher duration time was used to set search in RT range (min) and the minimum relative height. To set the parameters for deconvolution, all the peaks from the generated peak list after chromatogram builder must be selected for the creation of the XIC. For the calculation of the minimum relative height the height of the peak with the higher duration time is divided with the height of the most intense peak. The calculation of the minimum ratio of peak top/edge is based on the peak of reserpine. For the peak duration range (min), the default value was used. Local minimum algorithm was selected due to low noise level. The number of detected peaks of each problem sample has been increased after the deconvolution, as it was expected.

Isotope grouper module attempts to find those peaks in a peak list which form an isotope pattern. The difference between neighboring isotopes is a single neutron or 1.008665 Da, but part of this mass is consumed as a binding energy to other nucleons. This small difference may become significant with high-resolution MS data. The actual mass difference between isotopes depends on the chemical formula of the molecule. Since MZmine does not know the formula at the time of deisotoping, it assumes the default distance of ~ 1.003 Da, with user-defined tolerance the m/z tolerance parameter. For small molecular weight compounds with monotonically decreasing isotope pattern, the most intense isotope should be representative. For high molecular weight peptides, the lowest m/z peptides, the lowest m/z isotope may be representative (Cañaveras, 2015). The values for m/z tolerance had been calculated previously during the chromatogram builder. The estimation of the absolute value of the retention time tolerance was based on the peak with the higher duration time in the chromatogram. The maximum charge was set 2 and the representative isotope was chosen to be the most intense. The monotonic shape was checked.

Peak alignment is one of the main processing steps in metabolomics studies involving multiple samples. The position of the peaks corresponding to the same metabolic feature may be affected by non-linear shifts that are usually introduced by differences in the chemical environment of the samples, like ionic strength, pH, or protein content. In MS-studies peak shifts are observed across the retention time axis (Alonso et al., 2015). The recommended algorithm for peak alignment found to be the Join Aligner. In case, of PHIME children samples the $mass_{error}$ was 3.447 ppm, and the uncertainty 0.0021 m/z, as was calculated during chromatogram builder. The absolute value of the time tolerance was 0.07 min. The weight for RT and m/z was 1 and 0 respectively.

Following alignment, the resulting peak list contained missing peaks as a product of a deficient peak detection or a mistake in the alignment of different peak lists. The fact that one peak is missing after the alignment does not imply that the peak does not exist. In most cases, it is present but was

undetected by the previous algorithms. The algorithm Peak finder was used to fill the "gaps". After gap filling, a file containing the m/z values, retention times, and peak area for each detected peak, was exported in CSV format.

Beside the careful design of the samples acquisition process and the careful cleaning and maintenance of the equipment before a batch analysis, to obtain consistent variables the resulting matrix will be further reduced by the 80% rule. If there are variables with more than 80% missing values (those with ion intensity = 0) will be excluded. The 80% rule should not be applied to all the samples at once, because this would have as a result the loss of important metabolites reflecting to the biological status of the system. The instrument and overall process variability were determined by calculating the median RSD for all the endogenous metabolites, in case of the presented cohorts.

4.1.1.2 NMR data pre-processing

As it has been mentioned before, there are several tools for NMR metaolomics spectral processing, and the choice of the most suitable depends on the following bioinformatics approach. According to the developed methodology in WP17, biological pathways alterations will be investigated using the two modules of GeneSpring platform called Mass Profiler Professional (MPP) and Pathway Architect, developed by *Agilent Technologies. As a result,* in this study, spectral analysis proceeded using MestReNova (Mnova 11.0.3) (http://mestrelab.com), while for the identification of metabolites it was deemed necessary to use in addition ChenoMx (http://www.chenomx.com), since GeneSpring accepts NMR metabolomic profiles that are computed using the Chenomx software suite. Data from urine and plasma samples from both cohorts was analysed using the following workflow.

After loading the spectra of all the samples, the one was placed on the other by using the command superimposed. Next step was to correct the position of the reference peak sample. In this study, the used reference was deuterium oxide (D_2O), due to the used buffer. After the reference correction the alignment of all reference peaks was checked· all reference peaks should be aligned one behind the other. The most widespread algorithms for baseline correction, which was the next step of the data pre-processing, is the Continuous Wavelet Derivative Transform (CWT) and the Smoother Whittaker (Jewison et al., 2014, Carlos Cobas et al., 2006). A measure of algorithm suitability is whether the algorithm creates a problem in the phase of the spectrum. There is the possibility that after baseline correction with the aid of an algorithm some peaks shift from the positive to the negative part of the axis of tension (y-axis) or vice versa. This means that this specific algorithm is not suitable for the analysis. The Smoother Whittaker was chosen for the baseline correction, and after that, spectrum phase was checked and corrected. For phase correction, an automatic algorithm is preferable. A chromatogram resulting from an NMR untargeted metabolic analysis of a sample will contain more than 22000 variables. It is, therefore, necessary to reduce the volume of information to enable the investigator to end up to conclusions after further analysis. This reduction was achieved by grouping these variables (binning or bucketing). In more detail, the user split the x-axis into smaller regions, setting a value for the length of each region. It is common practice to choose length values less than 0.04 ppm. Also, all the peaks should belong to the same region. In each case, for the success of binning, the final image should be identical to that of the original spectrum (Smolinska et al., 2012, Weljie et

al., 2006). Then the spectrum was imported into the program ChenomX NMR Suite 8.2, for peaks identification. The identification of the peaks was based on compounds that exist in the library of the program and requires the definition of pH and concentration of TSP in the buffer that was added to the samples, since these parameters define the position of the peaks (Amiot et al., 2015, Beaudry et al., 2016). Finally, returning to the program MNOVA peak integration of TSP and metabolite, which was identified previously, took place and the peaks of metabolites were identified. In cases where a metabolite was characterized by multiple peaks and/or peaks in different areas of ppm, the area of all these peaks were added to fill the corresponding cell on the sheet of import file to MPP.

4.1.2 Metabolites Annotation

The most crucial step for downstream bioinformatics analysis and one of the biggest challenge is the annotation of metabolites.

In case of LC-MS-based spectra, the mass-to-charge ratio (m/z) value of a molecular ion of interest is searched against metabolite database(s). The metabolites having molecular weights within a specified tolerance to the query m/z value are retrieved from the databases as putative identifications. These putative identifications serve as a foundation for further metabolite verification. It is important to use multiple sources to induce the possibility of missing information. The use of both METLIN (Smith, et al., 2005). and HMDB (Wishart, et al., 2013) databases for metabolite identification is highly recommended. The Human Metabolome Database (HMDB) is a freely available electronic database containing 74,507 metabolite entries including both water-soluble and lipid soluble metabolites. METLIN database, on the other hand, includes masses, chemical formulas and structural detail for over 15,000 endogenous and exogenous metabolites and di- and tri-peptides. The database also facilitates high-volume research with automatic searches using mass lists.

All signal data from both urine and plasma samples from all cohort analyses was assigned using the Human Metabolome DataBases (HMDB) and Metlin. The mass-to-charge ratio (m/z) value of a molecular ion of interest was searched against metabolite database(s). The metabolites with molecular weights within a specified tolerance to the query m/z value, which was defined as 10 ppm, were retrieved from the databases as putative identifications. Cross-referencing across multiple databases was performed when a particular identifier type was missing from a database. The annotation file included the common name, empirical formula and unique CAS (Chemical Abstracts Service) registry number. It is important to note that annotations were purely derived from the accurate mass only and not any other analytical data, so they do not constitute or represent a full identification, and therefore metabolites should still be considered as unknowns. This means the annotations are tentative or a “level 4”, as described in Salek et al. (2013).

The query was limited to endogenous metabolites while performing compound identification against Metlin and/or HMDB database, since this increases the number of the pathways matched to LC-MS data.

Also, sometimes the compounds that were annotated with the CAS ID annotation were synthetic/drug-based compounds. Some metabolites can only be detected in individuals that have used or taken a specific drug (e.g Sevelamer it is a phosphate binding drug used to prevent hyperphosphataemia in patients with chronic renal failure). In case of Repro PL, mothers would not have been chosen as participants if they were suffered from a chronic disease, so drug metabolites were excluded, if these drugs are used to prevent a chronic disease, to minimize the ''false positive'' annotated compounds. In both Repro PL and PHIME samples, the metabolite of morphine, which can only be found in individuals that have used or taken this drug, was detected, and even if could in fact indicate another constituent to the external exposome, the integration into the following bioinformatics analysis should be based on evidence of consumption.

In case of NMR metabolomics, the identification of metabolites was proceeded using the ChenoMx (http://www.chenomx.com), as has been described in section 4.1.1.2 regarding NMR data pre-processing.

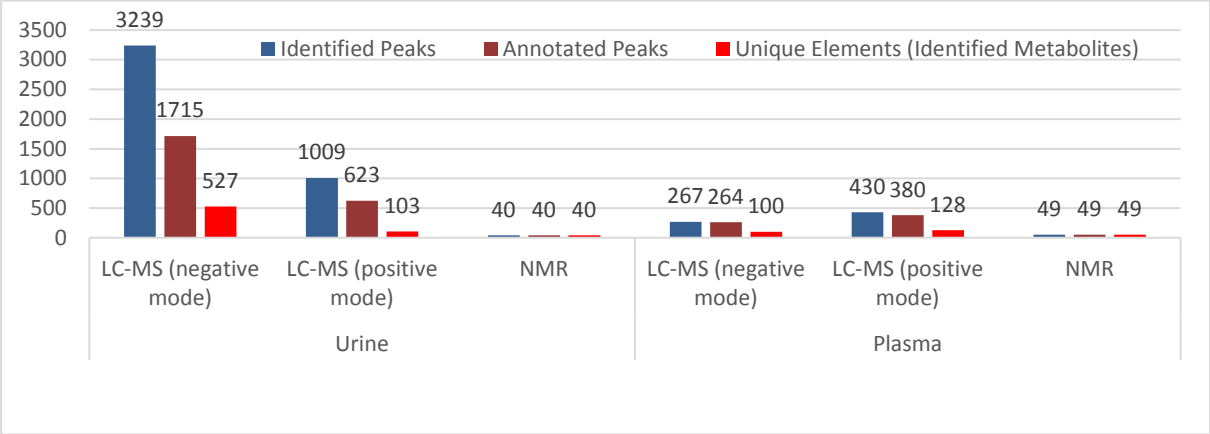


Figure 5. Metabolomics analysis results of REPRO PL cohort study. LC-MS/MS in negative mode analysis of urine samples gave the dataset with the higher number of unique annotated peaks.

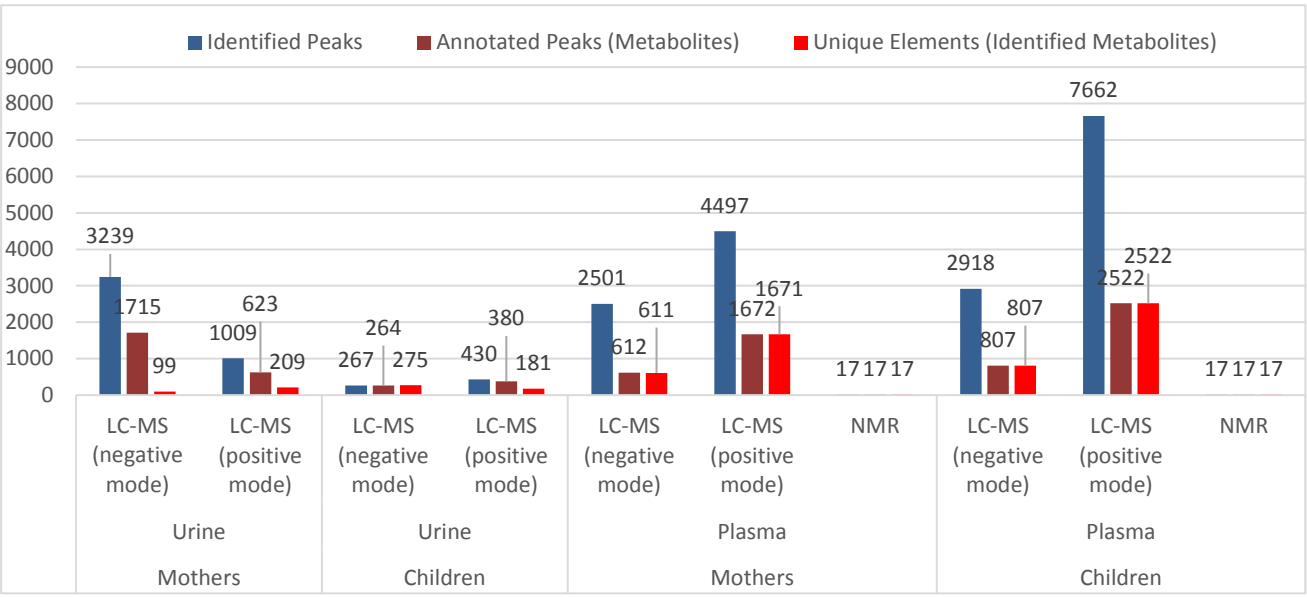


Figure 6. Metabolomics analysis results of PHIME cohort study. In case of the analysis of mothers samples the maximum number of annotated peaks was the result of the analysis in positive mode for both urine and plasma, while for children was the result of the analysis in negative mode for urine, and positive for plasma.

4.1.3 Data processing

A key issue for metabolomics studies is to avoid over-fitting the data. Because of the large number of metabolites and the relatively small sample size, a complex model can over-utilize (over-fit) the data specific information and show very good performance, but in that case the result is useless if it cannot be duplicated using a new set of test data. Proper model evaluation and validation is therefore a necessary step to understand the true performance of a model and the potential biomarkers. So, preprocessing methods, such as filtering, normalization and mean-centering, are crucial to pathway

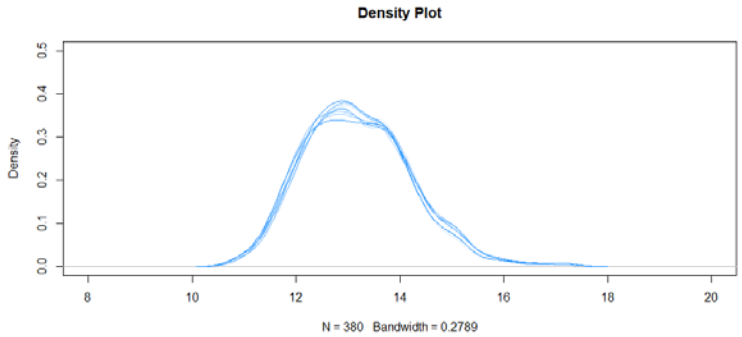
analysis (Bowei Xi et al., 2014). Mass Profiler Professional, a module of GeneSpring GX Agilent's software (Agilent, 2008) software provides the necessary tools for preprocessing steps, statistical analysis and pathway analysis. The resulting files from positive and negative ionization mode are importing in the same experiment.

4.1.3.1 Normalization

One of the critical steps in a metabolomics data analysis is normalizing the samples so that they have common distributions, especially when the distributions are likely being driven by some technical variables. The first step must include the log transformation of the data. Then, distribution of each of different samples must be plotted. This is an Exploratory Analysis, one way to do that is to basically make a plot of the density of each of these samples by writing a loop that loops over each of the other sample. The following step includes the quantum normalization.

In metabolomics research, it is important to reduce systematic error in experimental conditions. To ensure that metabolomics data from different studies are comparable, it is necessary to remove unwanted systematic factors by data normalization. Several normalization methods are used for metabolomic data, but the best method has not yet been identified. Normalization methods adapted from the single-channel microarray literature are commonly used (Jauhiainen et al., 2014). In case of liquid chromatography-mass spectrometry (LC-MS)-based metabolomic data, some of the most used normalization methods are the 1-norm, 2-norm, and quantile normalization. From the aforementioned methods, the quantile normalization is a simple and effective method to reduce non-biological systematic variation from human LC-MS-based metabolomics data, revealing the biological variance (Lee et al., 2012). In these cohorts, the quantile normalization was applied using the pre-process core package, and more specific the normalize.quantiles function. The first step included the log transformation of the data. Then, distribution of each of different samples was plotted. This is an Exploratory Analysis, one way to do that is to basically make a plot of the density of each of these samples by writing a loop that loops over each of the other sample. The following step included the quantum normalization. The density plot for the normalized data of the distribution was used to check the performance of the algorithm. The density plots before and after normalization of the metabolomics results from the untargeted analysis of the urine samples from children within PHIME cohort are given below.

A



B

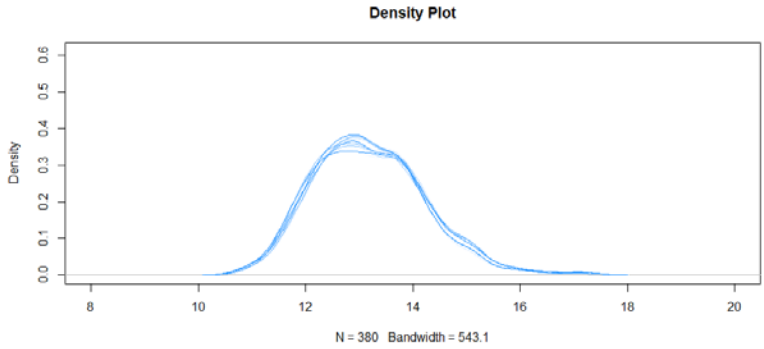
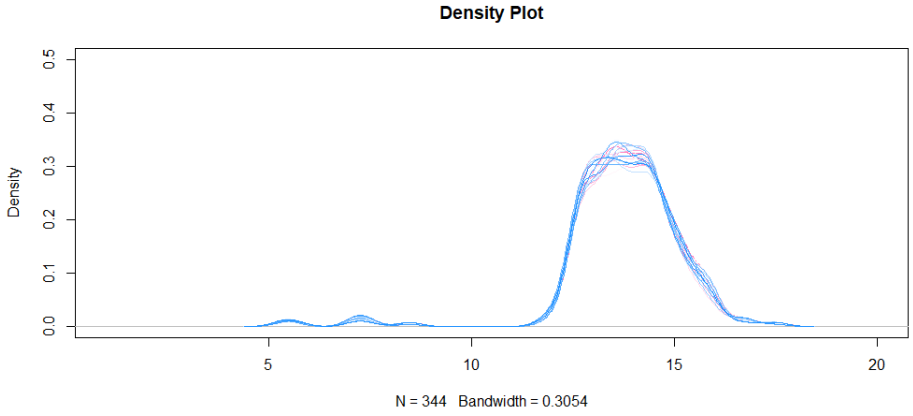


Figure 7. The density plot A illustrates the log transformed raw data from the analysis in the negative ionization mode before normalization, while the density plot B shows the samples distributions after quantile normalization.

A



B

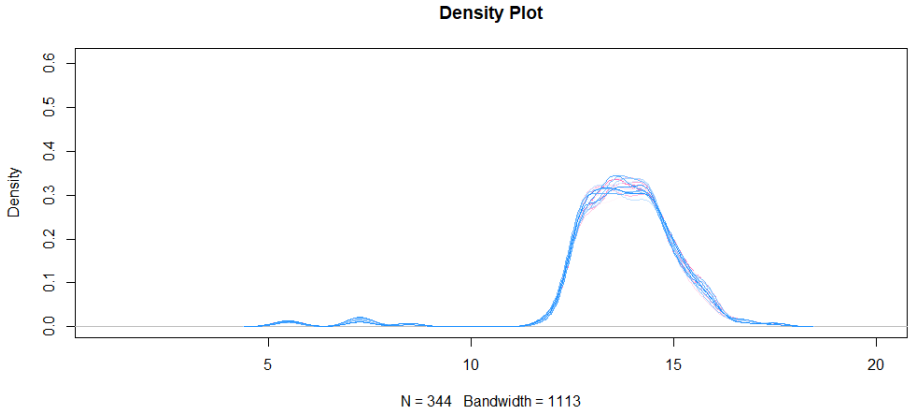


Figure 8. The density plot A illustrates the log transformed raw data from the analysis in the positive ionization mode before normalization, while the density plot B shows the samples distributions after quantile normalization. After quantile normalization of both negative and positive ionization results the bandwidth has dramatically increased, which means that the unwanted systematic variation has been minimized and the samples have almost common distributions.

4.1.3.2 Filtering

Multivariate statistical analysis tends to focus on metabolites with high intensities, but low-concentration metabolites may also play important roles in the biological processes. Filtering is used to change the emphasis from metabolites with high concentrations to those with moderate or small abundances. Variance scaling calculates the standard deviation of each variable (column) and then divides each column by this value (Bowei Xi et al., 2014). In case of the presented analyses the filtering was performed within quality assurance, and was skipped in that stage of the bioinformatics analysis, since the data processing steps of a metabolomics should be invasive, the repetition of a step is not recommended.

4.1.3.3 Alignment of parameters

Unidentified compounds from different samples were aligned or grouped together if their retention times were within the RT window 0.0% + 0.15 min, which were the default parameter. Retention time correction was performed without standards, since internal standards were not added to the QCs or problem samples. The maximum allowed retention time shift was set at 0.5% + 0.5 min, and the mass window at 15.0ppm + 2.0mDa. The total number of the aligned compounds was 344. A baseline to median of all samples was chosen to treat all the compounds equally regardless of their intensity.

4.1.3.4 Multivariate Statistical Analysis

After the metabolomics data are created and organized into matrix format, multivariate analysis methods are performed to identify bio-marker candidates and to examine how well the different groups in the data set are separated using those biomarker candidates through a classification model. These pattern-recognition methods can be classified into two groups: supervised and unsupervised methods. In unsupervised analysis methods, the similarity patterns within the data are identified without taking into account the type or class of the study samples. In supervised methods, the sample labels are used in order to identify those features or features combinations that are more associated with a phenotype of interest. Supervised methods are also the basis for building prediction models.

First, we specify how the samples should be grouped into experiments conditions, so as parameters we choose the 11 phthalates metabolites that have been mentioned in previous section.

4.1.3.4.1 PCA

Unsupervised methods are often applied to summarize the complex metabolomic data. They provide an effective way to detect data patterns that are correlated with experimental and/or biological variables. Principal component analysis (PCA) is the most commonly used unsupervised method in metabolomic studies. PCA is based on the linear transformation of the metabolic features into a set of linearly uncorrelated (i.e., orthogonal) variables known as principal components. This decomposition method maximizes the variance explained by the first component while the subsequent components explain increasingly reduced amounts of variance. At the same time, PCA minimizes the covariance between these components (i.e., they are independent of each other). After applying the PCA method, a set of loading vectors and score vectors are obtained. The loading vectors represent the principal components, and each vector coefficient corresponds to the individual contribution of each variable to the principal component. The score vectors represent the projection of each sample onto the new orthogonal basis. Plotting these sample scores over the first principal components is a convenient way of summarizing the global dataset, since normally these first principal components capture most of the variability in the dataset. PCA is also used in metabolomics studies to assess data quality, since it can identify sample outliers or reveal hidden biases in the study. For example, PCA has been used in several studies to determine the impact of technical variation in the analysis of metabolic profiles (Gika et al., 2008; Winnike et al., 2009; Rasmussen et al., 2011; Yin et al., 2013).

It is a common practice to keep the first two or three PCs and examine the score plots. There is no guarantee that the different groups will be well-separated on the PC score plots, since PCA is not designed for classification purposes. However, when the groups are well separated, which happens in many studies, the metabolites that have large loadings in the first two or three PCs can be selected as potential biomarkers (Bowei Xi et al., 2014). We choose 3 principal components, because after some tests we see that the minimum number of principal components that explain our original system is 3.

4.1.4 Metabolic pathway analysis

The next step in the downstream bioinformatics analysis was pathway mapping that revealed the roles that metabolites play in relation to each other and in biological aberrations. For both cohort studies, Repro PL and PHIME, pathway analysis was performed using the GeneSpring Pathway Architect, which can be used to map the results from single experiments onto curated pathways. This module finds relevant pathways associated with the experiment organism from the total number of pathways present in the tool pathways comprising the Biocyc, KEGG and WikiPathways databases based on similar entities (matched entities) between the pathway and the entity list. GeneSpring only reports the number of matched entities for a metabolomics experiment, so no p-values are computed for entities from metabolomics experiments to avoid a misrepresentation of the significance of matching pathways caused by the fact that the technology of a metabolomics experiment is limited to only the measured metabolites with an observable abundance in the experiment and pathways, on the other hand, are likely to contain many other metabolites that may not be present in the technology. This results in a pathway p-value computed with the technology to be higher than a more realistic p-value computed with a comprehensive reference set of global entities. The created pathway lists suggest that the joint use of LC-MS and NMR techniques provide complementary information about the metabolomic fingerprint. The final outcome of the pathway analysis was one list of more than 200 unique pathways for Repro PL and PHIME cohort study.

Using a Venn diagram, it was indicated the amount of overlap and revealed the substantial discrepancies among 3 metabolic pathways dataset of Repro PL. The first dataset was generated by applying pathway analysis to the overall NMR samples. The second dataset consists of the results of the individual pathway analysis of the LCMS samples. Taking into account the metabolites that was detected thought NMR analysis, a third dataset of was generated from the LCMS individual dataset. This dataset was generated under the assumption that at least one metabolite detected by NMR analysis exists to the pathways detected by the LCMS analysis.

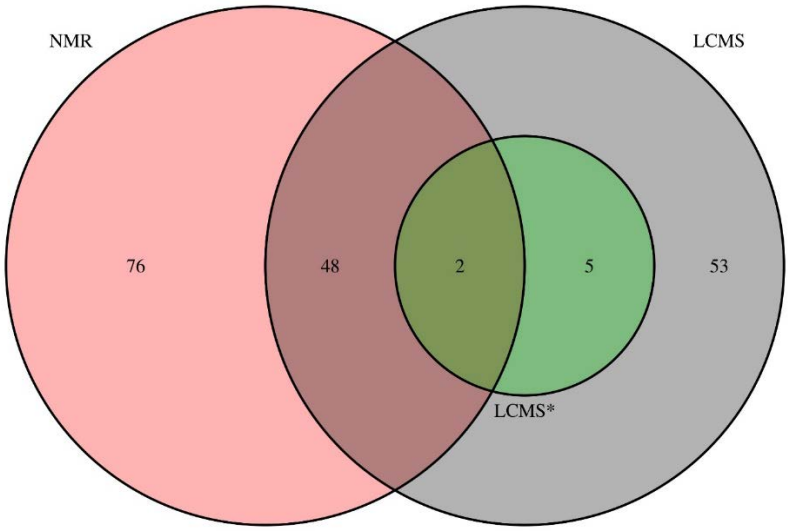


Figure 9. This Venn diagram displays overlaps among three metabolic pathway datasets. NMR dataset consists of 126 pathways while LCMS dataset consists of 108 pathways. NMR and LCMS datasets have 50 common

pathways that were detected on both analysis. LCMS dataset, that was generated based on common detected metabolites, consists of 7 metabolic pathways of which two are common to NMR and LCMS datasets.*

4.1.5 Association of metabolic pathways and child neurodevelopment through EWAS

4.1.5.1 Methodology

The analysis of data was based on EWAS framework. EWAS was introduced by Patel (2010) in order to describe the correlations among multiple variables using the idea "unsupervised learning" (Coates, et al., 2011). This framework provides inferring of hidden structures from unclassified data. Hence, based on this framework, logistic regression was used to associate the results of Bayley test with the metabolic pathways, the environmental factors and the biomonitoring data while adjusting by gender, socio-economic status (Hecker, et al., 1990), mother's height, mother's education and the age of the mother at the child birth. The adjustment took also into account the true conditions of mother's alcohol consumption during, the breastfeeding and the child attendance in a daycare at the first and second year of child life. The form of the logistic regression model was:

$$\begin{aligned}\text{logit}(z) = a + a_{ses}X_{ses} + a_{gender}X_{gender} + a_{mother\_education}X_{mother\_education} + \\ a_{mother\_height}X_{mother\_height} + a_{achohol\_consumption}X_{achohol\_consumption} + \\ a_{breast\_feeding}X_{breast\_feeding} + a_{child\_daycare\_attendance}X_{child\_daycare\_attendance} + \\ a_{mother\_age}X_{mother\_age} + b_{factor}Y_{factor}\end{aligned}$$

where αi is the parameter and Xi is the value of described adjusting factor and b is the parameter and Y is the value of the observed factor. It is noted that using logistic regression was computed the non-parametric correlation coefficient between the adjusting and examined factors.

Additionally, Spearman correlation, that is a non-parametric test, was applied to compute correlations between variables avoiding any distributional assumptions for the variables.

Exposure related variables were arising by applying LC/MS. The skewness of the data was checked and the right skewness were log transformed. Exposures were captured either as continuous or a categorical variable. Furthermore, a z-score transformation was applied scaling by the standard deviation aiming at the comparison of odds ratios from the regressions results.

The False Discovery Rate (FDR) q-value were calculated to associate the factor with Bayley test levels controlling the type I error using the Benjamini-Hochberg step-down approach (Benjamini and Hochberg, 1995). Additionally, applying permutation resampling to Bayley test results was allowed the validation of the FDR results (Mielke Jr and Berry, 2007). It is noted that to choose factors significantly associated with Bayley test, the used significance level was 0.05, which corresponded to a FDR of 10% (Patel, et al., 2010).

The calculations were executed Ubuntu 16.04 server using R . Using the package (Harrell Jr, 2008) was estimated functionalities related to data analysis, variable clustering, utility operations, functions for computing sample size and power, importing and annotating datasets, imputing missing values,

advanced table making. The skewness of the data was checked using the R library "moments" (Komsta and Novomestky, 2015). Moreover, the package 'permute' (Simpson, et al., 2016) was used for the permutations resampling and the package 'RCircos' (Zhang, et al., 2013) for the globe visualization. Finally, the logistic regression and FDR was carried out using the 'X-Wide Association Analyses' package (Chirag Patel's, 2017).

EWAS analysis results of urinary untargeted LC-MS/MS metabolomics analysis results were illustrated using correlation globes and volcano plots, which are given in the ANNEX 1.

4.1.5.2 Interpretation

The key finding of both cohort studies is that maternal exposure to phthalates and metals may affect child neuro development mainly through perturbations in the metabolism of citric acid (TCA cycle), urea cycle, and amino acid metabolism with possible disruption of mitochondrial oxidative phosphorylation. This furthers highlight the importance of this critical window of susceptibility (pregnancy) in later life health.

Urea cycle, inosine-5'-phosphate biosynthesis II, asparagine degradation I, citrulline nitric oxide cycle, and the metabolic pathway of effects of nitric oxide, which were detected due to the presence of glutathione, methionine, cysteine, pyruvate, N-acetylglutamic acid, β-alanine, serine, and arginine, in samples from both ReproPL and PHIME cohort studies, were found to be significant associated to cognitive development according to EWAS analysis. Abnormalities in the aforementioned pathways results in the pathogenesis of oxidative stress stress (Meguid, et al., 2017; Yoshimi, et al., 2016). The detected metabolites could play the role of biomarkers.

Glycolysis/Gluconeogenesis and TCA cycle were ones of the identified perturbed pathways in both ReproPL and PHIME samples. EWAS analysis revealed that metabolic pathways related to glycolysis, are associated with the language development. Metabolomics analysis of ReproPL plasma samples revealed the presence of D-fructose 1,6-bisphosphate, b-Glucose, and acetate, which are intermediates products of glycolysis, while the analysis of urine samples of PHIME revealed the presence of oxoglutaric acid and oxalosuccinic acid, which play a key role in both TCA cycle and glycolysis.

Exposure to endocrine disrupting chemicals (EDCs) is associated with the citric acid cycle and more specific with oxidation by which fatty acids are metabolized to acetyl-CoA, which enters the citric acid cycle, indicating the possibility of disruption of mitochondrial respiration. Oxidation of fatty acids could theoretically influence electron flow by the former pathway, leading to altered ROS levels (Donohoe, et al., 2012; Jones, et al., 2000).

Dysfunction in carnitine metabolism may affect calcium homeostasis, which will be involved in mitochondrial oxidative phosphorylation that leads to neurodevelopmental disorders (Zheng, et al., 2011). Pathway analysis showed that the metabolic pathway of L-carnitine biosynthesis was statistical significant perturbed, which could be involved in mitochondrial oxidative phosphorylation that leads to neurodevelopmental disorders (Zheng, et al., 2011). The metabolites 2-oxoglutarate, succinate, and L-carnitine, which were present in plasma samples from Repro and PHIME children samples, could play the role of biomarkers.

Perturbations on the main pathways of dopamine, serotonin, norepinephrine, and glutamate metabolism could lead to identifying potential biomarkers of neurodevelopment after exposure to

EDCs. Exposure to EDCs disrupt the synthesis, transport, and release of many neurotransmitters which play key roles in modulating behavior, cognition, learning, and memory, according to both the literature search and the pathway analysis of REPRO PL and PHIME samples (Schug, et al., 2015). Metabolites that could be considered as candidate biomarkers, according to metabolomics analysis of samples, are: homovanillic acid, L-tyrosine, L-tryptophan, 5-hydroxy-L-tryptophan, cysteine, 5-hydroxy-L-tryptophan, oxoglutaric acid, L-aspartic acid, glutamate, fumarate, succinate, 5-hydroxyindole acetate, dopamine, and melatonine.

Peroxisome proliferator activated receptors (PPARs) are ligand activated transcription factors with crucial functions in lipid homeostasis, glucose metabolism, anti-inflammatory processes, placental development, and are involved in cognitive functions and neurodegenerative diseases (Zhen, et al., 2007). The following biomarker, that have been detected in urine samples from ReproPL study, arise from PPARalpha effects on tryptophan, corticosterone, fatty acid metabolism and on glucuronidation: Hippurate, 2,8-dihydroxyquinoline-beta-d-glucuronide, 11beta-hydroxy-3,20-dioxopregn-4-en-21-oic acid, 11beta,20-dihydroxy-3-oxopregn-4-en-21-oic acid, nicotinamide, nicotinamide 1-oxide, 1-methylnicotinamide, xanthurenic acid, hexanoylglycine, phenylpropionylglycine, and cinnamoylglycine.

4.2 Identification of metabolites associated with children neurodevelopmental disorders using in vitro assays.

For the mechanistic confirmation of the causative effect between exposure and disease endpoint, in vitro assays assessing complex toxic endpoints relevant to the disease endpoints of the case studies were implemented. For the in vitro cellular model, the human liver-derived HepaRG cell line was chosen for several reasons: i) after differentiation by DMSO, two cell types coexist in the culture (50% hepatocyte-like cells and biliary-type cells). Moreover, bile ducts are present in the culture between hepatocytes, and thus it is one of the best cellular model for human liver. HepaRG are known to express all the xenosensors and xenobiotic metabolizing enzymes which can be implicated in the effects of the pollutants (Antherieu et al, 2010) and ii) these cells can be treated for a long time (several weeks) with low concentrations of the selected pollutants to mimic at best a low, semi-chronic exposure (Josséet al, 2008). More details regarding the cell culture, treatments, preparation of samples, and omics experiments are presented at the D5.4-Database on candidate, in vitro supported, -omics derived biomarkers, for targeted analysis deliverable.

4.2.1 Data Preprocessing

4.2.1.1 Transcriptomics

Raw data was analyzed using GeneSpring GX 14.9 (Agilent Technologies) with the raw data files (.cel) of the Affymetrix Human 2 0-st chip imported into the software. Affymetrix expression technology was created using a custom CDF file obtained from http://brainarray.mbni.med.umich.edu/brainarray/Database/CustomCDF/genomic_curated_CDF.asp. The RAM (Robust Multichip Averaging) algorithm was used to normalize the data where it was subjected to quantile normalization with a mean of all samples taken for baseline transformation.

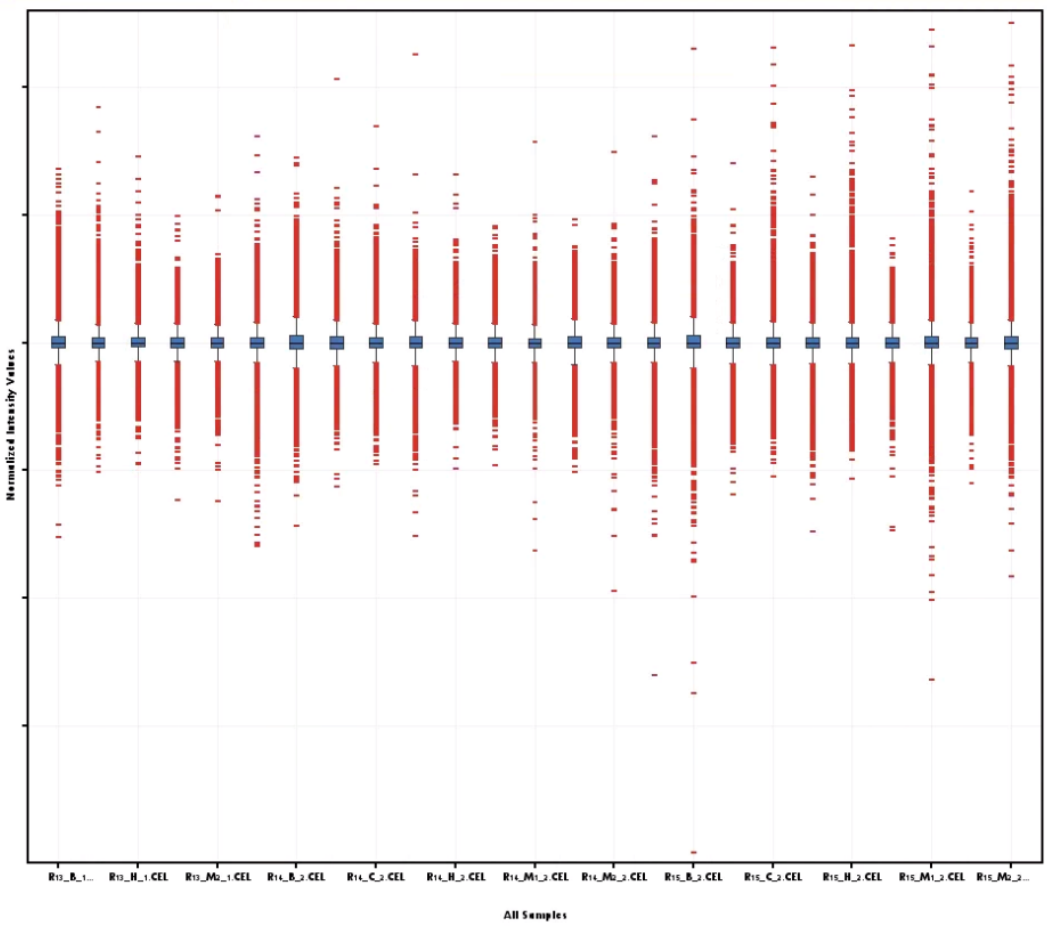


Figure 10. Summary view of the transcriptomics experiment for every sample after the background correction, normalization and probe summarization using the RMA algorithm. According to the Box Whisker diagram the data from all the samples follows normal distribution.

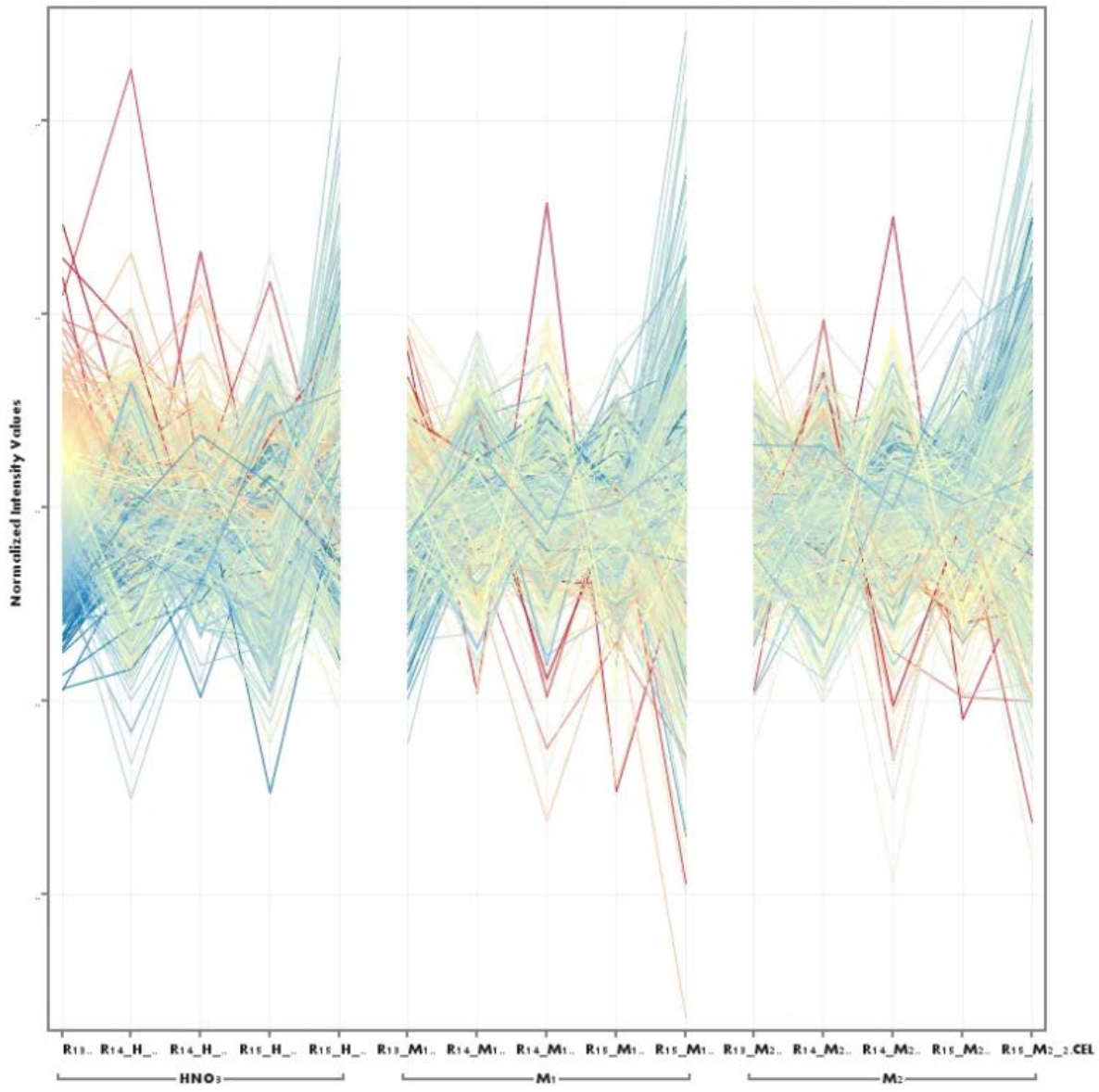


Figure 11. Profile plot of transcriptomics data when the samples have been grouped based on treatment. According to the profile plot samples treated with M1 and M2, and controls, characterized by differences in their genes expression. These differences will be determined during fold change analysis.

4.2.1.2 Proteomics

The data analysis of the proteomics results was based on the same workflow followed for the analysis of transcriptomics data. Data was analyzed using the module of GeneSpring GX 14.9 called Mass Profiler Professional (Agilent Technologies). The data was normalized using the quantile algorithm, after the log transformation. The following figure presents the summary view of the created experiment, using a Box Whisker plot.

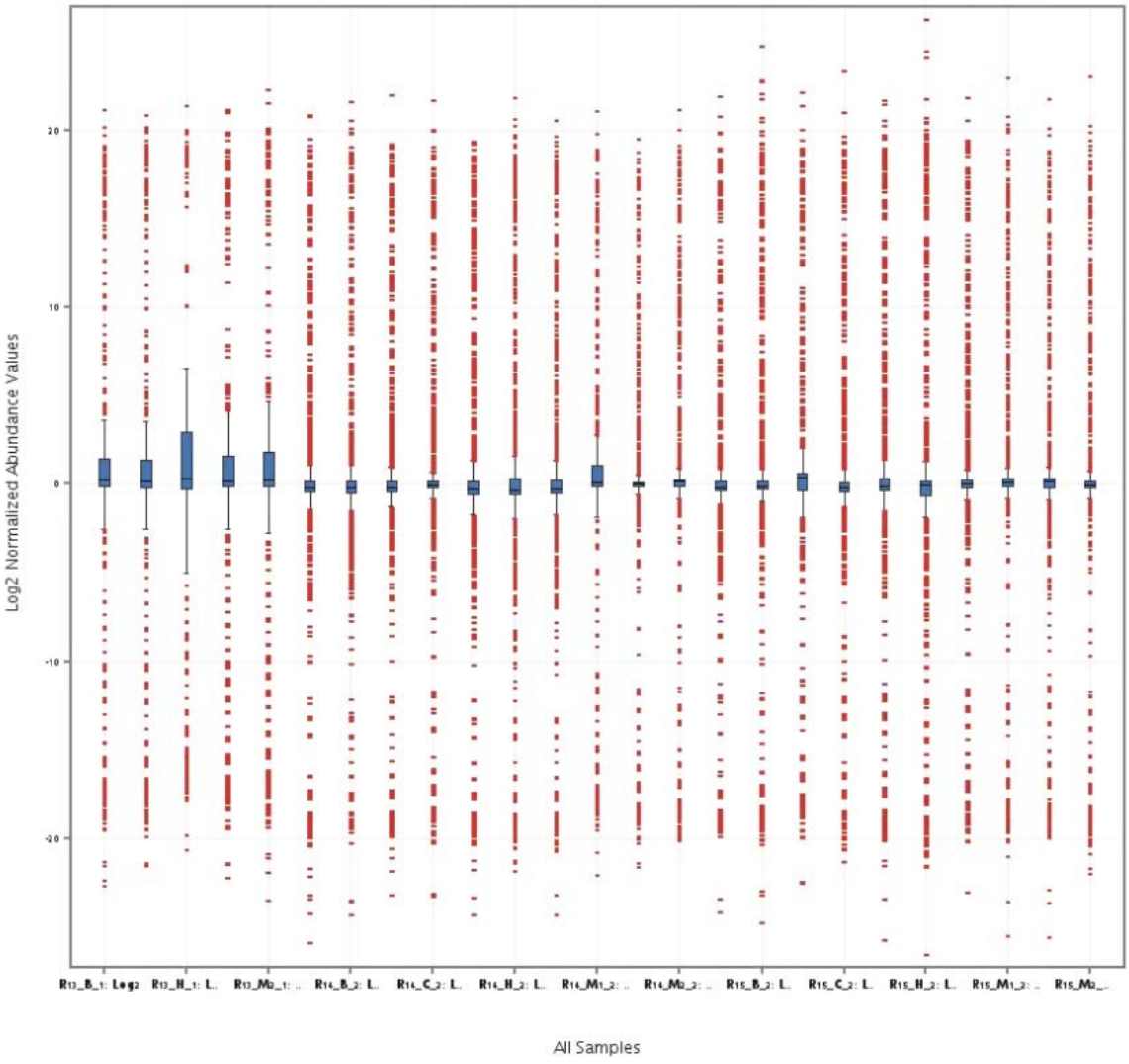


Figure 15. Summary view of the proteomics experiment after the background correction, and normalization.

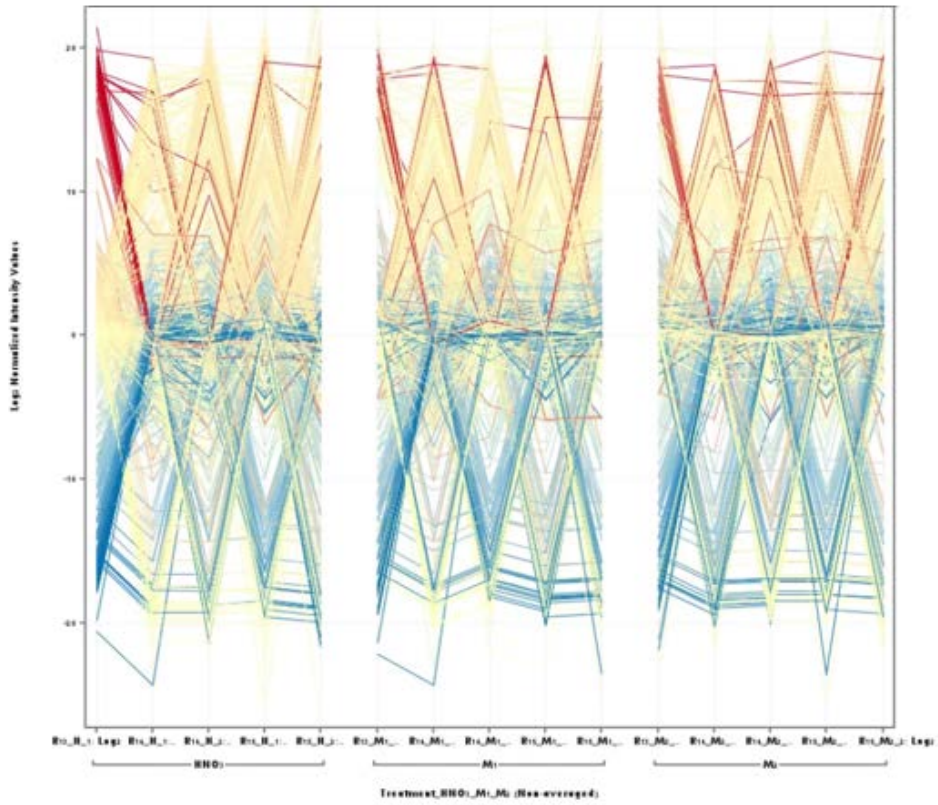


Figure 12. Profile plot of proteomics data when the samples have been grouped based on treatment. According to the profile plot, differences also occur in the level of proteins between samples treated only with solvent, and samples treated with phthalates and metals.

4.2.1.3 Metabolomics

Data was acquired using the ThermoFisher Scientific model LTQ Orbitrap Discovery MS, and for data generated by Non-Agilent Chromatography system (non-Agilent data), spectral data processing was performed using the MZMine v.2.29 open-software (Katajamaa, et al., 2006).

One of the most crucial step of pre-processing is to detect and minimize any sources of variation unrelated to the biological nature of the samples. In this study, this was established by a) the detection of internal standards to the problem samples, b) a careful design of the samples acquisition process, and c) a thorough equipment cleaning and maintenance before a batch analysis. The Quality Control samples (QCs) were generated from samples of cells treated for 1 to 3 weeks, which means that could not be used for quality control or batch effect corrections. Data preprocessing workflow, as well as the used parameters for baseline correction, mass detection, chromatogram builder, deconvolution, deisotoping, alignment and gap-filling are given at the Annex 2. The following figure presents the summary view of the normalized metabolomics data set, using a Box Whisker plot. The data in case of all samples follows normal distribution, since the whiskers have equal length and the median is in the middle of the first and the third quartile. Moreover, there were no outliers.

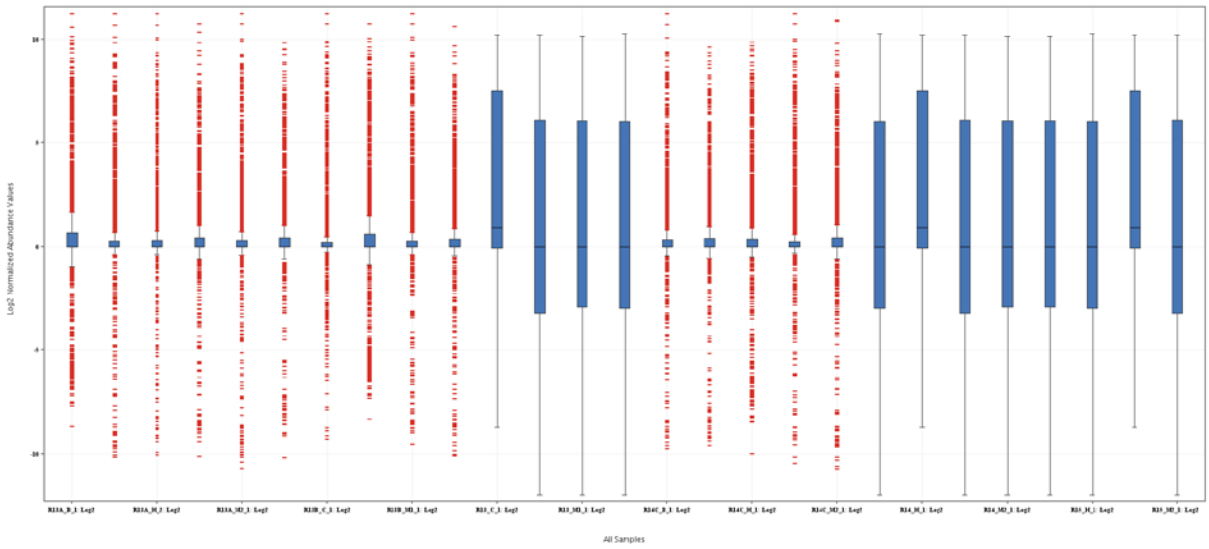


Figure 13. Summary view of the metabolomics experiment after pre-processing. The data in case of all samples follows normal distribution.

4.2.2 Data processing

After the omics data were created and organized into matrix format, multivariate analysis methods were performed to identify bio-marker candidates and to examined how well the different groups in the data set were separated using those biomarker candidates through a classification model. These pattern-recognition methods can be classified into two groups: supervised and unsupervised methods. In unsupervised analysis methods, the similarity patterns within the data are identified without considering the type or class of the study samples. In supervised methods, the sample labels are used to identify those features or features combinations that are more associated with a phenotype of interest. Supervised methods are also the basis for building prediction models.

4.2.2.1 Principal component analysis (PCA)

Unsupervised methods are often applied to summarize the complex omics data. They provide an effective way to detect data patterns that are correlated with experimental and/or biological variables. Principal component analysis (PCA) is the most commonly used unsupervised method. PCA is based on the linear transformation of the features into a set of linearly uncorrelated (i.e., orthogonal) variables known as principal components. This decomposition method maximizes the variance explained by the first component while the subsequent components explain increasingly reduced amounts of variance. At the same time, PCA minimizes the covariance between these components (i.e., they are independent of each other). After applying the PCA method, a set of loading vectors and score vectors are obtained. The loading vectors represent the principal components, and each vector coefficient corresponds to the individual contribution of each variable to the principal component. The score vectors represent the projection of each sample onto the new orthogonal basis. Plotting these sample scores over the first principal components is a convenient way of summarizing the global dataset, since normally these first principal components capture most of the variability in the dataset. PCA is also used in omics studies to assess data quality, since it can identify sample outliers or reveal hidden biases in the study. For example, PCA has been used in several studies to determine the impact of technical variation in the analysis of metabolic profiles (Gika et al., 2008; Winnike et al., 2009; Rasmussen et al.,

2011; Yin et al., 2013). It is a common practice to keep the first two or three PCs and examine the score plots. There is no guarantee that the different groups will be well-separated on the PC score plots, since PCA is not designed for classification purposes. However, when the groups are well separated, which happens in many studies, the genes, proteins or metabolites that have large loadings in the first two or three PCs can be selected as potential biomarkers (Bowei Xi et al., 2014).

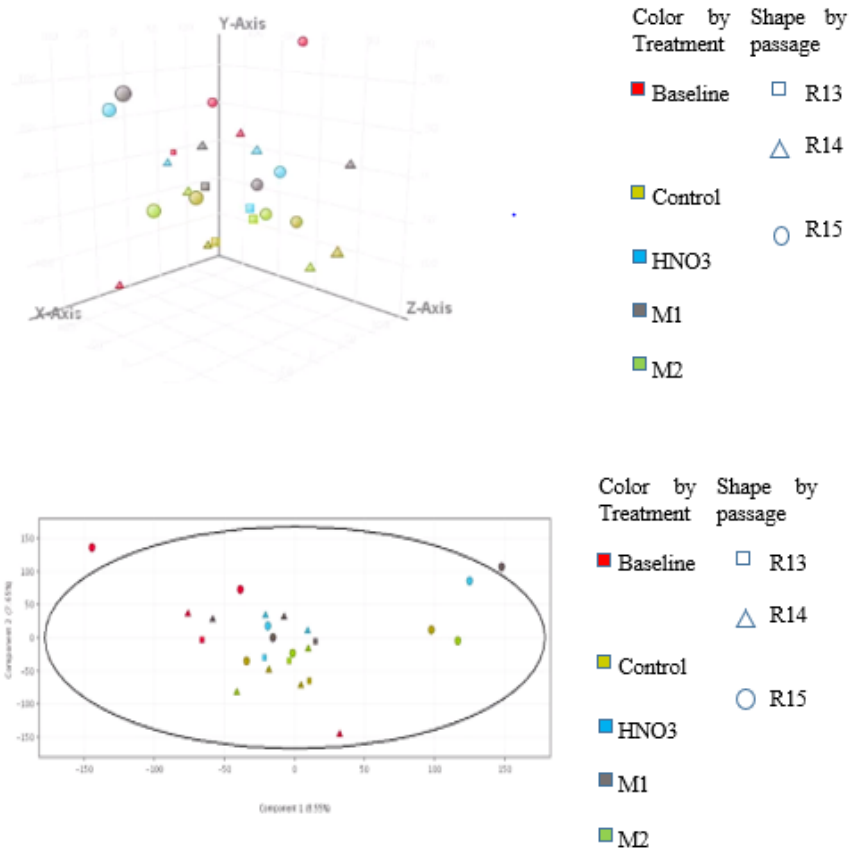


Figure 14. Principal Component Analysis (PCA) performed using normalized trancsriptomics data. Loadings for principal components 1 (PC1) and PC2 are reported in graph (on x, y and z-axes). Moreover, similarity between biological replicates of the same group (baseline, control, M1, and M2) showed a good consistency and

correlation among them.similarity between biological replicates of the same group (baseline, control, M1, and M2) showed a good consistency and correlation among them.

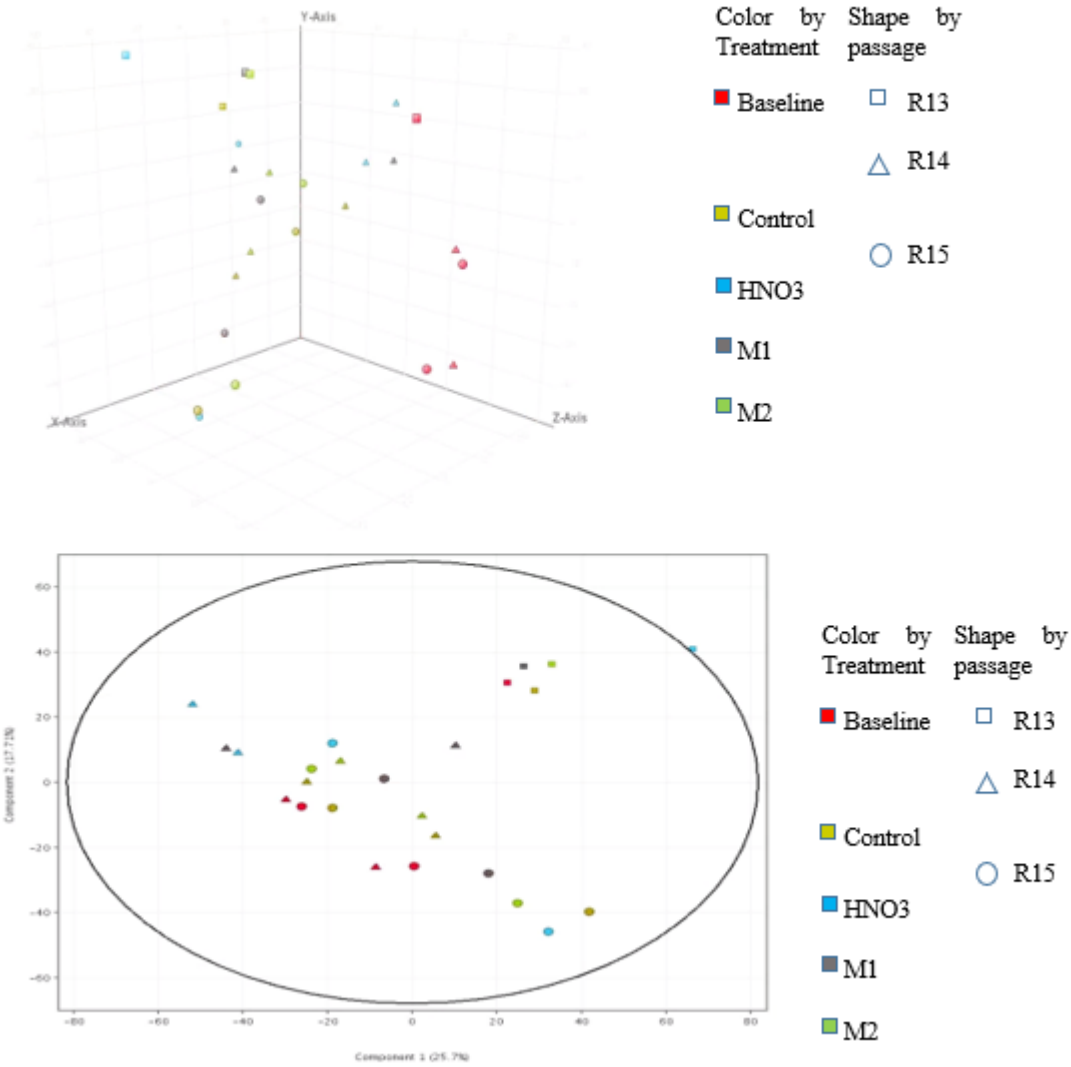


Figure 15. Principal Component Analysis (PCA) performed using normalized trancsriptomics data. Loadings for principal components 1 (PC1) and PC2 are reported in graph (on x, y, and z-axes).

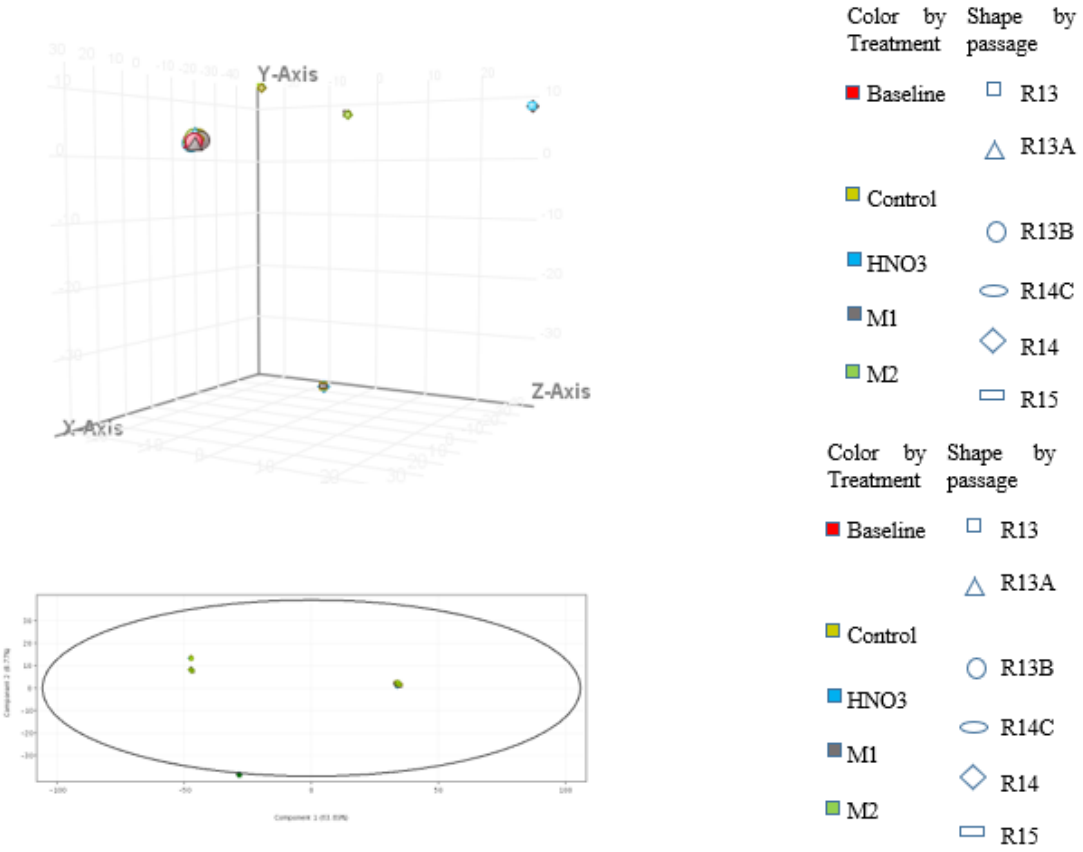


Figure 16. Principal Component Analysis (PCA) performed using normalized metabolomics data. Loadings for principal components 1 (PC1) and PC2 are reported in graph (on x, y, and z-axes). similarity between biological replicates of the same group (baseline, control, M1, and M2) showed a great consistency and correlation among them.

4.2.2.2 Fold Change Analysis

Fold change analysis was used to identify genes with expression ratios or differences between mixture 1 and solvent control, as well between mixture 2 and solvent control, that are outside of a given cut-off or threshold. The ratio between the two conditions was calculated using the equation (Fold change = mixture 1 (or mixture 2)/solvent control). A sensitivity analysis or in other words a testing of a range of threshold conditions was performed as proposed by PAN et al. (Pan et al., 2005) and the results are presented in the Table 1. The same methodology was also followed in case of proteomics, and metabolomics. The resultant list of genes differentially expressed between the two parameters were filtered to include only those that expressed absolute fold change by greater than 1.5-fold. The cut-off criteria were loosened to include more differential expressed genes in the further analysis to find the key metabolic and biological pathways affected by the exposure to phthalates and heavy metals by systems biology analysis using GeneSpring software. The perturbed metabolic pathways for the significant differentially expressed proteins that were outside of a 2.0 cut-off were matched with the ones of the significant differentially expressed genes that were outside of a 2.0 cut-off, which lead to the conclusion that the sources of noise have been successfully removed, resulting in the increased

accuracy of biological variability. The final observed perturbed metabolic pathways were the result of group-specific changes in proteins expression caused by the exposure to the different mixtures of phthalates and heavy metals. The fold-change cut-off criteria was loosened to 1.0, in case of metabolomics experiment, since according to the sensitivity analysis there were no statistically significant differential expressed metabolites comparing the conditions M1 and HNO_3, and M2 and HNO_3.

Table 1. Sensitivity analysis or in other words a testing of a range of threshold conditions for transcriptomics, proteomics, and metabolomics.

Fold Change	Transcriptomics	Proteomics	Metabolomics
2.5	2/48,144	1,174/3,346	604/2419
2.4	2/48,144	1,182/3,346	716/2419
2.3	4/48,144	1,189/3,346	905/2419
2.2	4/48,144	1,201/3,346	948/2419
2.1	5/48,144	1,221/3,346	1073/2419
2.0	7/48,144	1,246/3,346	1123/2419
1.9	8/48,144	1,264/3,346	1190/2419
1.8	10/48,144	1,276/3,346	1233/2419
1.7	16/48,144	1,300/3,346	1269/2419
1.6	28/48,144	1,337/3,346	1315/2419
1.5	57/48,144	1,380/3,346	1503/2419
1.4	100/48,144	1,472/3,346	1532/2419
1.3	249/48,144	1,661/3,346	1722/2419
1.2	815/48,144	2,833/3,346	1816/241
1.1	4,817/48,144	2,892/3,346	1899/2419
1.0	48,144/48,144	3,346/3,346	2419/2419

4.2.2.3 Clustering Analysis

To organize in an efficient way genes/entities and conditions in the dataset, clustering analysis was performed on the set of entities filtered for statistically significant changes. Clustering is a way to visualize differences between samples based on their overall gene (or protein or metabolomics) expression profiles, represented by a heat map depicting the expression of analyzed genes (or proteins or metabolites) in each sample. This format provides a way to cluster genes based on their function, and show how these groups of genes differ between samples, in order to provide a higher level understanding of pathway and functionality differences between samples, while maintaining a more "global" picture of the data (Malarkey et al., 2018). In the analysis of transcriptomic data, many approaches for identifying clusters of genes with similar expression patterns have been used, but in this study three of them were used: Hierarchical, K-means, and Self Organizing Map. For each one of the aforementioned algorithm the Euclidean was defined as the distance metric, while the Ward' s rule was used as the linkage rule in case of the Hierarchical clustering algorithm.

4.2.2.3.1 Hierarchical clustering

Hierarchical clustering, briefly, seeks to pair up data points that are most similar to one another. With the agglomerative (or bottom-up) approach, we begin with N data points forming singleton clusters. For each point, we measure the distance between it and its N−1 neighbors. The pair with the shortest distance between them is taken to form a new cluster. We then look at the distance between the N−2 points remaining and the newly formed cluster, and again pair off the two with shortest distance (either adding to our 2-cluster, or forming another one). This process is repeated until we have a single cluster with N points (regardless of the absolute distance between points) (Zambelli, 2016). A hierarchical tree diagram or dendrogram can be generated to show the linkage points: the clusters are linked at increasing levels of dissimilarity.

A clustering algorithm is required to measure the similarity or difference between entities or conditions. A variety of distance metrics are available to determine the distance between intermediate clusters while computing the similarity scores, but in this study standard sum of squared distance (L2-norm) between two entities or Euclidean was used, since it is the most straightforward and generally accepted way of computing distances between objects in a multidimensional space. Euclidean (and squared Euclidean) distances are usually computed from raw data, not from standardized data. In the case of p variables $X_1, X_2, \dots, X_p$ measured on a sample of n subjects, the observed data for subject i can be denoted by $X_{i1}, X_{i2}, \dots, X_{ip}$, and the observed data for subject j by $X_{j1}, X_{j2}, \dots, X_{jp}$, while the Euclidean distance between these two subjects is given by:

$$dij = \left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \cdots + \left(x_{ip} - x_{jp}\right)^2$$

Another parameter of the hierarchical clustering algorithm is the Linkage rule algorithms, which determine the similarity between clusters for visualization. The chosen method for this study was the Ward's method based on the ANOVA approach, which computes the sum of squared errors around the mean for each cluster. Then, two clusters are joined to minimize the increase in error.

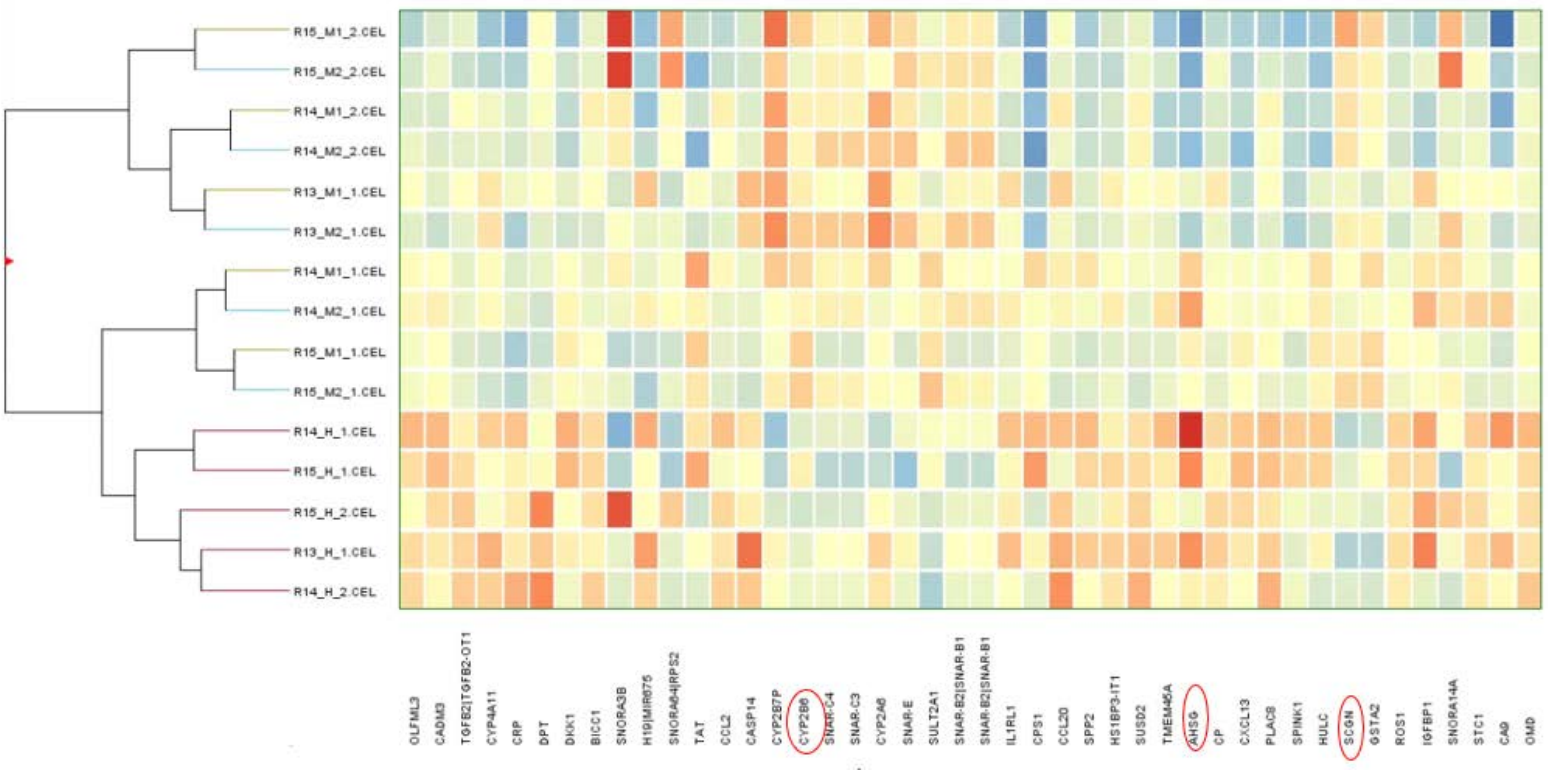


Figure 17. Hierarchical Clustering analysis of significant differential expressed genes. In human hepatocytes, DEHP, via the constitutive androstane receptor, increased the expression of CYP2B6. In the HepaRG cells, both M1 and M2 mixture led to a 1.5-fold increase of its mRNA level. AHSG, the mRNA expression of which is down-regulated in the in vitro study, has been shown to be down-regulated in patients with Alzheimer's disease. In autistic children, SCGN which encodes secretagogin, was found in lower level compared to healthy controls.

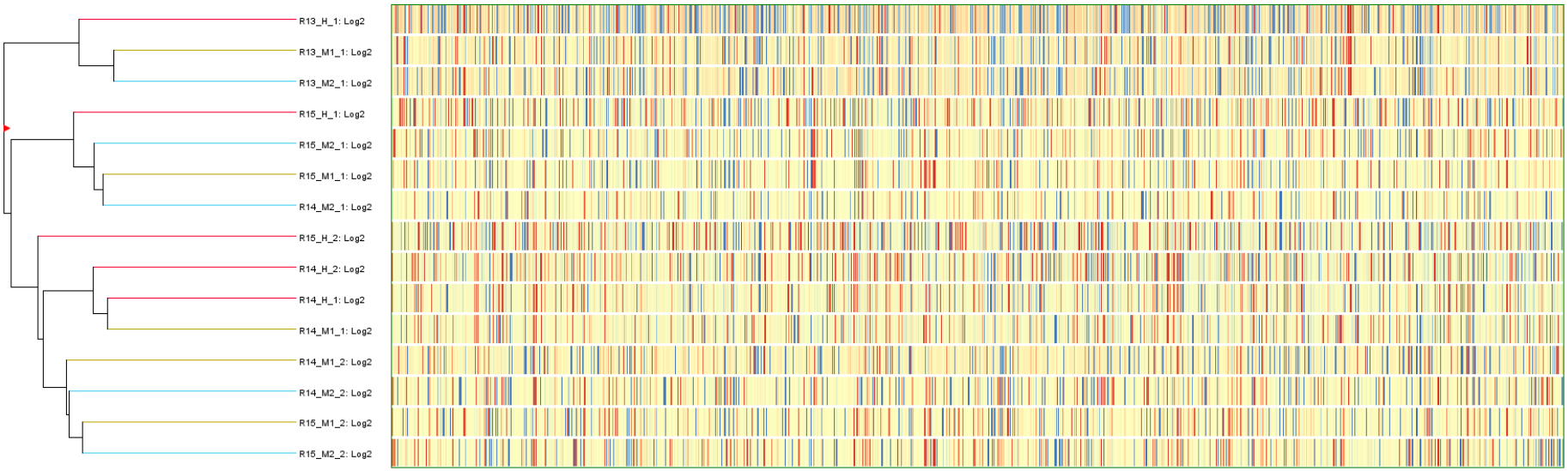


Figure 18. Hierarchical Clustering analysis of significant differential expressed proteins on both entities and conditions. Fold change analysis was used to identify proteins with expression ratios or differences between mixture 1 and solvent control, as well between mixture 2 and solvent control, that are outside of a 2.0 cut-off or threshold.

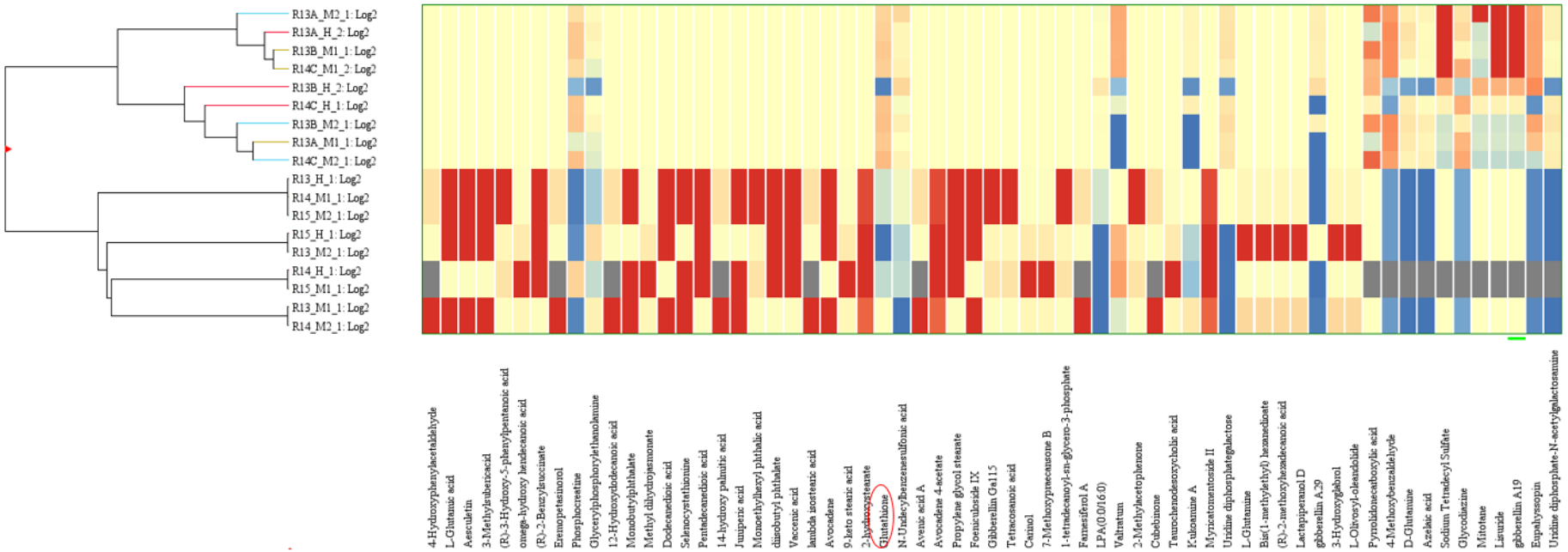


Figure 19. Hierarchical Clustering analysis of significant differential expressed metabolites. The fold-change cut-off criteria was loosened to 1.5, since according to the sensitivity analysis there are no statistical significant metabolites comparing the conditions M1 and HNO3, and M2 and HNO3 when cut-off criteria is greater than 1.5. Glutathione has been also detected in samples from ReproPL and PHIME cohort studies, could play the role of candidate biomarkers for the pathogenesis of neurodevelopment due to oxidative stress.

4.2.2.3.2 K-means

The K-means algorithm is an iterative method that starts with k cluster centers randomly chosen. All observations are then associated to the closest cluster center and new centers are computed as the mean of the observations of a given cluster. The observations are grouped with respect to the new centers iteratively until convergence; that is, no difference occurs in the next iteration (Boccard and Rudaz, 2013). The K-means method is dependent upon the initial set of cluster centroids and different results will usually be found for different initial seeds, which means that the correct number of clusters must be set from the researcher, while in case of HCA the researcher doesn't have to know the correct number of clusters in his data (Gillet and Willett, 2007). Euclidean was used for the measure of the distances between the clusters. The results of K-means analysis are presented in the Annex 3.

4.2.2.3.3 Self-Organizing Map (SOM)

The self-organizing map (SOM) is a clustering technique that illustrates the relationship between groups by arranging them in a two-dimensional map and by dividing entities into groups based on expression patterns. SOMs are useful for visualizing the number of distinct expression patterns in your data and determining which of these patterns are variants of one another. The U-Matrix view is used to display results of the SOM clustering algorithm, which also generates a fixed number of clusters, like the K-Means algorithm. The U-Matrix view displays clusters arranged in a two-dimensional grid such that similar clusters are physically closer to each other in the grid. The grid cells can either be hexagonal or rectangular. Cells in the grid are of two types: nodes and non-nodes. Nodes and non-nodes alternate in the grid. Clusters are associated only with nodes and each node displays the reference vector or the average expression profile of all entities mapped to the node. This average expression profile is plotted in blue. The purpose of non-nodes is to indicate the similarity between neighboring nodes on a gray scale. If a non-node between two nodes is very bright, it indicates that the two nodes are very similar. If the non-node is comparatively darker, then the two nodes are very different. Furthermore, the shade of a node reflects its similarity to its neighboring nodes. Thus, this view displays average cluster profiles as well as how the various clusters are related. The results of self-organizing map (SOM) analysis are presented in the Annex 3.

4.2.3 Multi-omics Pathway Analysis

Differentially expressed genes, proteins and metabolites were co-mapped on available pathways from WikiPathways, BioCyC, and KEGG databases, using the module Pathway Architect of GeneSpring. No p-values were computed for entities from metabolomics and proteomics experiments during a multi-omics experiment to avoid a misrepresentation of the significance of matching pathways caused by the fact that GeneSpring uses the technology (All Entities list) as a reference for p-value computation. The technology of a metabolomics or a proteomics experiment is limited to only the measured metabolites with an observable abundance in the experiment. Pathways on the other hand are likely to contain many other metabolites that may not be present in the technology. This results in a pathway p-value computed with the technology as reference to be higher than a more realistic p-value computed with a comprehensive reference set of global entities. To avoid this apparent increase in significance, GeneSpring only reports the number of matched entities for a metabolomics or proteomics experiment.

According to multi-omics pathway analysis glutathione-mediated detoxification I has been identified across all the three multiple layers. Moreover, the metabolic pathways of folate metabolism and urea has been identified across transcriptomics and proteomics experiments. Imbalance between the cellular ROS and the ability of the cell to detoxify them, lead to oxidative stress. In case that the brain

is highly exposed to increased oxidative stress due to the presence of excitatory amino acids whose metabolism ends with the production of ROS causing neuronal damage. According to literature, abnormalities in the citric acid cycle, urea cycle, and amino acid metabolism play a key role in the pathogenesis of oxidative stress, which may be an effect of exposure to phthalates and metals. Co-exposure to phthalates and heavy metals may result in the imbalance between the cellular ROS and the ability of the cell to detoxify them (Meguid, et al., 2017; Yoshimi, et al., 2016).

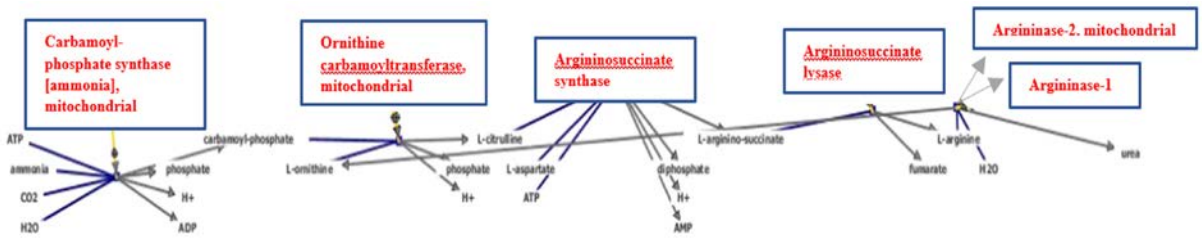


Figure 20. The urea cycle has been co-mapped in case of data integration using transcriptomics and proteomics, due to the statistically significantly difference in the expression of arginase-1, arginase-2 (mitochondrial), arginosuccinate synthase, carbonyl-phosphate synthase (mitochondrial), ornithine carbamoyl transferase (mitochondrial) and arginosuccinate lyase. The identification of urea pathway is of particular interest since has been also identified in case of ReproPL cohort study using untargeted metabolomics NMR plasma analysis.

Multi-omics pathway analysis using data derived from untargeted metabolomics and proteomics analysis revealed dysregulation in purine and pyrimidine synthesis and metabolism. More specific, the following metabolic pathways have been perturbated due to the co-exposure to phthtalates and heavy metals: purine nucleotides de novo biosynthesis II/ pyrimidine ribonucleotides de novo biosynthesis/ pyrimidine ribonucleotides interconversion/ uridine-5'-phosphate biosynthesis. In recent years, a substantial body of evidence has emerged demonstrating that purine and pyrimidine synthesis and metabolism play major roles in controlling embryonic and fetal development and organogenesis. ATP is involved in the development of synaptic transmission and contributes to the establishment of functional neuronal networks in the developing brain. The purinergic control of neurodevelopment is not limited to prenatal life, but is maintained in postnatal life, when it plays fundamental roles in controlling oligodendrocyte maturation from precursors and their terminal differentiation to fully myelinating cells. Based on the above-mentioned and other literature evidence, it is now increasingly clear that any defect altering the tight regulation of purinergic transmission and of purine and pyrimidine metabolism during pre- and post-natal brain development may translate into functional deficits, which could be at the basis of severe pathologies characterized by mental retardation or other disturbances. This can occur either at the level of the recruitment and/or signaling of specific nucleotide or nucleoside receptors or through genetic alterations in key steps of the purine salvage pathway (Fumagalli et al., 2017) It has been found that the exposure to phthalates results in a series of metabolites of purine catabolism, such as xanthine, allantoin and urea. The observed increased levels of these metabolites, together with the elevation of uridine, suggest an activation of nucleic acid degradation (Xia et al., 2011).

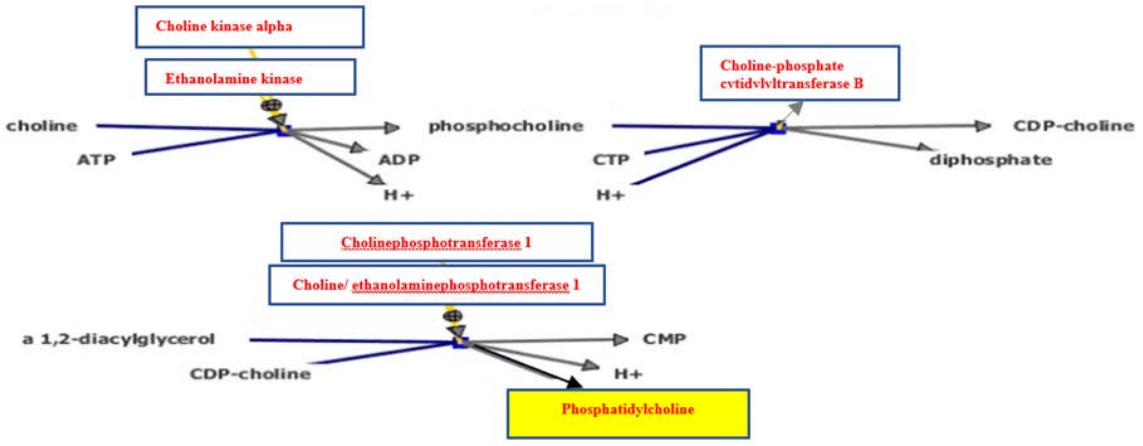


Figure 21. Co-mapping of proteomics and metabolomics data revealed that their common drivers are responsible for the homeostasis of metabolic pathways related to choline, phosphatidylcholine, phospholipases and triacylglycerol metabolism, due to altaretions in the expression levels of phosphatidylcholine, and 1,2-diacyl-sn-glycerol-3-phosphate.

The interpretation of pathway analysis results is described in greater detail in D5.4 Database of candidate, in vitro supported, -omics derived biomarkers, for targeted analysis.

5 Conclusions

The main objective of this deliverable is to provide the methodological tools for integrating multiple biomarkers into a mechanistic description of biological pathways relevant to environment-wide health association studies. The determination of predictive biomarkers based on heterogeneous datasets, resulted from human biomonitoring, omics and PBBK modeling, requires first of all the pre-processing of the large amount of data produced, the discovery of specific data patterns and/or clusters, the creation of data models based on training sets and, finally, the evaluation of the models with regard to their validity and prediction capacity on the basis of the test data. Several approaches were implemented to build the model that best describes our exposome data will be employed to the EXHES and other exposome studies.

A workflow for processing and integrating multi-OMICS data was presented using the example of corresponding metabolomics data sets obtained from the participants in the Repro PL and PHIME cohorts, and from the in vitro assays. Different methods were used to discriminate low-intensity peaks from background noises, including baseline correction, by smoothing signal intensity over the course of multiple scans with a moving window of specified size. The second important presented task was to perform a comparative analysis of matching the same type of metabolites from different samples. A “level 4” annotation was achieved using the Human Metabolome DataBases (HMDB) and Metlin. For the reduction in the data dimensionality Principal Component Analysis (PCA) was implemented because the selection of individual features was not desired, data transformation is often employed. Outlier detection and normal distribution test were also implemented. Well-established clustering methods, such as hierarchical, K-means, and SOP, were used to discover data patterns. Known molecular interactions of the differentially expressed genes, proteins and metabolites were retrieved from Biocyc, KEGG and WikiPathways databases.

The developed comprehensive data processing approach provides the methodological tools for integration of multiple omics biomarkers into a mechanistic description of toxicity pathway interactions, in relation to external/internal exposure.

6 References

ABU BAKAR, M. H., SARMIDI, M. R., CHENG, K. K., ALI KHAN, A., SUAN, C. L., ZAMAN HURI, H. & YAAKOB, H. 2015. Metabolomics - the complementary field in systems biology: a review on obesity and type 2 diabetes. *Mol Biosyst,* 11**,** 1742-74.

AGILENT 2008. GeneSpring GX.

ALONSO, A., MARSAL, S. & JULIA, A. 2015. Analytical methods in untargeted metabolomics: state of the art in 2015. *Front Bioeng Biotechnol,* 3**,** 23.

AMIOT, A., DONA, A. C., WIJEYESEKERA, A., TOURNIGAND, C., BAUMGAERTNER, I., LEBALEUR, Y., SOBHANI, I. & HOLMES, E. 2015. (1)H NMR Spectroscopy of Fecal Extracts Enables Detection of Advanced Colorectal Neoplasia. *J Proteome Res,* 14**,** 3871-81.

ANDERSEN, A. D., BINZER, M., STENAGER, E. & GRAMSBERGEN, J. B. 2016. Cerebrospinal fluid biomarkers for Parkinson's disease - a systematic review. *Acta Neurol Scand*.

BAO, Y., ZHAO, T., WANG, X., QIU, Y., SU, M., JIA, W. & JIA, W. 2009. Metabonomic variations in the drug-treated type 2 diabetes mellitus patients and healthy volunteers. *J Proteome Res,* 8**,** 1623-30.

BEAUDRY, P., CAMPBELL, M., DANG, N. H., WEN, J., BLOTE, K. & WELJIE, A. M. 2016. A Pilot Study on the Utility of Serum Metabolomics in Neuroblastoma Patients and Xenograft Models. *Pediatr Blood Cancer,* 63**,** 214-20.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological),* 57**,** 289-300.

BOCCARD, J. & RUDAZ, S. 2013. Chapter 27 - Mass Spectrometry Metabolomic Data Handling for Biomarker Discovery A2 - Issaq, Haleem J. *In:* VEENSTRA, T. D. (ed.) *Proteomic and Metabolomic Approaches to Biomarker Discovery.* Boston: Academic Press.

BRANDT, E. B., BIAGINI MYERS, J. M., ACCIANI, T. H., RYAN, P. H., SIVAPRASAD, U., RUFF, B., LEMASTERS, G. K., BERNSTEIN, D. I., LOCKEY, J. E., LECRAS, T. D. & KHURANA HERSHEY, G. K. 2015. Exposure to allergen and diesel exhaust particles potentiates secondary allergen-specific memory responses, promoting asthma susceptibility. *Journal of Allergy and Clinical Immunology,* 136**,** 295-303.e7.

BROIX, L., JAGLINE, H., L IVANOVA, E., SCHMUCKER, S., DROUOT, N., CLAYTON-SMITH, J., PAGNAMENTA, A. T., METCALFE, K. A., ISIDOR, B., LOUVIER, U. W., PODURI, A., TAYLOR, J. C., TILLY, P., POIRIER, K., SAILLOUR, Y., LEBRUN, N., STEMMELEN, T., RUDOLF, G., MURACA, G., SAINTPIERRE, B., ELMORJANI, A., DECIPHERING DEVELOPMENTAL DISORDERS, S., MOISE, M., WEIRAUCH, N. B., GUERRINI, R., BOLAND, A., OLASO, R., MASSON, C., TRIPATHY, R., KEAYS, D., BELDJORD, C., NGUYEN, L., GODIN, J., KINI, U., NISCHKE, P., DELEUZE, J.-F., BAHI-BUISSON, N., SUMARA, I., HINCKELMANN, M.-V. & CHELLY, J. 2016. Mutations in the HECT domain of NEDD4L lead to AKT-mTOR pathway deregulation and cause periventricular nodular heterotopia. *Nat Genet,* advance online publication.

CALIMLIOGLU, B., KARAGOZ, K., SEVIMOGLU, T., KILIC, E., GOV, E. & ARGA, K. Y. 2015. Tissue-Specific Molecular Biomarker Signatures of Type 2 Diabetes: An Integrative Analysis of Transcriptomics and Protein-Protein Interaction Data. *Omics,* 19**,** 563-73.

CAÑAVERAS, J. C. G. 2015. *Metabolomics as a tool for the study of drug-induced hepatotoxicity.* DOTORAL, Facultat de CIÉNCIES BIOLÓGIQUES.

CARLOS COBAS, J., BERNSTEIN, M. A., MARTÍN-PASTOR, M. & TAHOCES, P. G. 2006. A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. *Journal of Magnetic Resonance,* 183**,** 145-151.

CARLSSON, M., THORELL, L., SJOLANDER, A. & LARSSON-FARIA, S. 2015. Variability of total and free IgE levels and IgE receptor expression in allergic subjects in and out of pollen season. *Scand J Immunol,* 81**,** 240-8.

CAUBIT, X., GUBELLINI, P., ANDRIEUX, J., ROUBERTOUX, P. L., METWALY, M., JACQ, B., FATMI, A., HAD-AISSOUNI, L., KWAN, K. Y., SALIN, P., CARLIER, M., LIEDEN, A., RUDD, E., SHINAWI, M., VINCENT-DELORME, C., CUISSET, J.-M., LEMAITRE, M.-P., ABDERREHAMANE, F., DUBAN, B., LEMAITRE, J.-F., WOOLF, A. S., BOCKENHAUER, D., SEVERAC, D., DUBOIS, E., ZHU, Y., SESTAN, N., GARRATT, A. N., LE GOFF, L. K. & FASANO, L. 2016. TSHZ3 deletion causes an autism syndrome and defects in cortical projection neurons. *Nat Genet,* advance online publication.

CHANG, C., GUO, Z.-G., HE, B. & YAO, W.-Z. 2015. Metabolic alterations in the sera of Chinese patients with mild persistent asthma: a GC-MS-based metabolomics analysis. *Acta Pharmacologica Sinica,* 36**,** 1356-1366.

CHEN, A., DIETRICH, K. N., HUO, X. & HO, S. M. 2011. Developmental neurotoxicants in e-waste: An emerging health concern. *Environmental Health Perspectives,* 119**,** 431-438.

DERVOLA, K. S. N., JOHANSEN, E. B., WALAAS, S. I. & FONNUM, F. 2015. Gender-dependent and genotype-sensitive monoaminergic changes induced by polychlorinated biphenyl 153 in the rat brain. *NeuroToxicology,* 50**,** 38-45.

DRABKOVA, P., SANDEROVA, J., KOVARIK, J. & KANDAR, R. 2015. An Assay of Selected Serum Amino Acids in Patients with Type 2 Diabetes Mellitus. *Adv Clin Exp Med,* 24**,** 447-51.

DURAK, O., GAO, F., KAESER-WOO, Y. J., RUEDA, R., MARTORELL, A. J., NOTT, A., LIU, C. Y., WATSON, L. A. & TSAI, L.-H. 2016. Chd8 mediates cortical neurogenesis via transcriptional regulation of cell cycle and Wnt signaling. *Nat Neurosci,* advance online publication.

EMOND, P., MAVEL, S., AÏDOUD, N., NADAL-DESBARATS, L., MONTIGNY, F., BONNET-BRILHAULT, F., BARTHÉLÉMY, C., MERTEN, M., SARDA, P., LAUMONNIER, F., VOURC'H, P., BLASCO, H. & ANDRES, C. R. 2013. GC-MS-based urine metabolic profiling of autism spectrum disorders. *Analytical and Bioanalytical Chemistry,* 405**,** 5291-5300.

FIEHN, O., GARVEY, W. T., NEWMAN, J. W., LOK, K. H., HOPPEL, C. L. & ADAMS, S. H. 2010. Plasma Metabolomic Profiles Reflective of Glucose Homeostasis in Non-Diabetic and Type 2 Diabetic Obese African-American Women. *PLoS ONE,* 5**,** e15234.

GILLET, V. J. & WILLETT, P. 2007. 4.08 - Compound Selection Using Measures of Similarity and Dissimilarity A2 - Taylor, John B. *In:* TRIGGLE, D. J. (ed.) *Comprehensive Medicinal Chemistry II.* Oxford: Elsevier.

GRANT, S. F. A., THORLEIFSSON, G., REYNISDOTTIR, I., BENEDIKTSSON, R., MANOLESCU, A., SAINZ, J., HELGASON, A., STEFANSSON, H., EMILSSON, V., HELGADOTTIR, A., STYRKARSDOTTIR, U., MAGNUSSON, K. P., WALTERS, G. B., PALSDOTTIR, E., JONSDOTTIR, T., GUDMUNDSDOTTIR, T., GYLFASON, A., SAEMUNDSDOTTIR, J., WILENSKY, R. L., REILLY, M. P., RADER, D. J., BAGGER, Y., CHRISTIANSEN, C., GUDNASON, V., SIGURDSSON, G., THORSTEINSDOTTIR, U., GULCHER, J. R., KONG, A. & STEFANSSON, K. 2006. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet,* 38**,** 320-323.

GRARUP, N., SANDHOLT, C. H., HANSEN, T. & PEDERSEN, O. 2014. Genetic susceptibility to type 2 diabetes and obesity: from genome-wide association studies to rare variants and beyond. *Diabetologia,* 57**,** 1528-1541.

GRAY, E., LARKIN, J. R., CLARIDGE, T. D. W., TALBOT, K., SIBSON, N. R. & TURNER, M. R. 2015. The longitudinal cerebrospinal fluid metabolomic profile of amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration,* 16**,** 456-463.

GUPTA, S., ELLIS, S. E., ASHAR, F. N., MOES, A., BADER, J. S., ZHAN, J., WEST, A. B. & ARKING, D. E. 2014. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nature Communications,* 5**,** 5748.

HA, C. Y., KIM, J. Y., PAIK, J. K., KIM, O. Y., PAIK, Y.-H., LEE, E. J. & LEE, J. H. 2012. The association of specific metabolites of lipid metabolism with markers of oxidative stress, inflammation and arterial stiffness in men with newly diagnosed type 2 diabetes. *Clinical Endocrinology,* 76**,** 674-682.

HAN, L.-D., XIA, J.-F., LIANG, Q.-L., WANG, Y., WANG, Y.-M., HU, P., LI, P. & LUO, G.-A. 2011a. Plasma esterified and non-esterified fatty acids metabolic profiling using gas chromatography–mass spectrometry and its application in the study of diabetic mellitus and diabetic nephropathy. *Analytica Chimica Acta,* 689**,** 85-91.

HAN, X., ROZEN, S., BOYLE, S. H., HELLEGERS, C., CHENG, H., BURKE, J. R., WELSH-BOHMER, K. A., DORAISWAMY, P. M. & KADDURAH-DAOUK, R. 2011b. Metabolomics in Early Alzheimer's Disease: Identification of Altered Plasma Sphingolipidome Using Shotgun Lipidomics. *PLoS ONE,* 6**,** e21643.

HAWI, Z., CUMMINS, T. D. & TONG, J. 2016. Rare DNA variants in the brain-derived neurotrophic factor gene increase risk for attention-deficit hyperactivity disorder: a next-generation sequencing study.

HINKS, T. S. C., ZHOU, X., STAPLES, K. J., DIMITROV, B. D., MANTA, A., PETROSSIAN, T., LUM, P. Y., SMITH, C. G., WARD, J. A., HOWARTH, P. H., WALLS, A. F., GADOLA, S. D. & DJUKANOVIĆ, R. 2015.

Innate and adaptive T cells in asthmatic patients: Relationship to severity and disease mechanisms. *Journal of Allergy and Clinical Immunology,* 136**,** 323-333.

HO, W. E., XU, Y. J., XU, F., CHENG, C., PEH, H. Y., TANNENBAUM, S. R., WONG, W. S. & ONG, C. N. 2013. Metabolomics reveals altered metabolic pathways in experimental asthma. *Am J Respir Cell Mol Biol,* 48**,** 204-11.

IBANEZ, C., CIFUENTES, A. & SIMO, C. 2015. Recent advances and applications of metabolomics to investigate neurodegenerative diseases. *Int Rev Neurobiol,* 122**,** 95-132.

INOUE, Y. U. & INOUE, T. 2016. Brain enhancer activities at the gene-poor 5p14.1 autism-associated locus. *Scientific Reports,* 6**,** 31227.

IRIZARRY, M. C. 2004. Biomarkers of Alzheimer Disease in Plasma. *NeuroRX,* 1**,** 226-234.

JENKINSON, C. P., GÖRING, H. H. H., ARYA, R., BLANGERO, J., DUGGIRALA, R. & DEFRONZO, R. A. 2016. Transcriptomics in type 2 diabetes: Bridging the gap between genotype and phenotype. *Genomics Data,* 8**,** 25-36.

JEWISON, T., SU, Y., DISFANY, F. M., LIANG, Y., KNOX, C., MACIEJEWSKI, A., POELZER, J., HUYNH, J., ZHOU, Y., ARNDT, D., DJOUMBOU, Y., LIU, Y., DENG, L., GUO, A. C., HAN, B., PON, A., WILSON, M., RAFATNIA, S., LIU, P. & WISHART, D. S. 2014. SMPDB 2.0: big improvements to the Small Molecule Pathway Database. *Nucleic Acids Res,* 42**,** D478-84.

JIAO, H., KAAMAN, M., DUNGNER, E., KERE, J., ARNER, P. & DAHLMAN, I. 2008. Association analysis of positional obesity candidate genes based on integrated data from transcriptomics and linkage analysis. *Int J Obes (Lond),* 32**,** 816-25.

KATAJAMAA, M., MIETTINEN, J. & ORESIC, M. 2006. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics,* 22**,** 634-6.

KIM, C. S., PARK, J. R., YU, S. H., KANG, J. G., RYU, O. H., LEE, S. J., HONG, E. G., KIM, D.-M., YOO, J. M., IHM, S. H., CHOI, M. G. & YOO, H. J. 2012. Impact of Serum Adiponectin Concentration on Progression of Carotid Atherosclerosis in Patients with Type 2 Diabetes Mellitus. *Endocrinol Metab,* 27**,** 31-38.

KIM, J. Y., PARK, J. Y., KIM, O. Y., HAM, B. M., KIM, H.-J., KWON, D. Y., JANG, Y. & LEE, J. H. 2010. Metabolic Profiling of Plasma in Overweight/Obese and Lean Men using Ultra Performance Liquid Chromatography and Q-TOF Mass Spectrometry (UPLC–Q-TOF MS). *Journal of Proteome Research,* 9**,** 4368-4375.

LARCHÉ, M., ROBINSON, D. S. & KAY, A. B. 2003. The role of T lymphocytes in the pathogenesis of asthma. *Journal of Allergy and Clinical Immunology,* 111**,** 450-463.

LAWTON, K. A., BROWN, M. V., ALEXANDER, D., LI, Z., WULFF, J. E., LAWSON, R., JAFFA, M., MILBURN, M. V., RYALS, J. A., BOWSER, R., CUDKOWICZ, M. E. & BERRY, J. D. 2014. Plasma metabolomic biomarker panel to distinguish patients with amyotrophic lateral sclerosis from disease mimics. *Amyotroph Lateral Scler Frontotemporal Degener,* 15**,** 362-70.

LELIEVELD, S. H., REIJNDERS, M. R. F., PFUNDT, R., YNTEMA, H. G., KAMSTEEG, E.-J., DE VRIES, P., DE VRIES, B. B. A., WILLEMSEN, M. H., KLEEFSTRA, T., LOHNER, K., VREEBURG, M., STEVENS, S. J. C., VAN DER BURGT, I., BONGERS, E. M. H. F., STEGMANN, A. P. A., RUMP, P., RINNE, T., NELEN, M. R., VELTMAN, J. A., VISSERS, L. E. L. M., BRUNNER, H. G. & GILISSEN, C. 2016. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat Neurosci,* 19**,** 1194-1196.

LI, N. & BUGLAK, N. 2015. Convergence of air pollutant-induced redox-sensitive signals in the dendritic cells contributes to asthma pathogenesis. *Toxicology Letters,* 237**,** 55-60.

LI, X., XU, Z., LU, X., YANG, X., YIN, P., KONG, H., YU, Y. & XU, G. 2009. Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry for metabonomics: Biomarker discovery for diabetes mellitus. *Analytica Chimica Acta,* 633**,** 257-262.

LI, Y. & CHEN, L. 2014. Big Biological Data: Challenges and Opportunities. *Genomics Proteomics Bioinformatics,* 12**,** 187-189.

LUAN, H., LIU, L.-F., TANG, Z., ZHANG, M., CHUA, K.-K., SONG, J.-X., MOK, V. C. T., LI, M. & CAI, Z. 2015. Comprehensive urinary metabolomic profiling and identification of potential noninvasive marker for idiopathic Parkinson's disease. *Scientific Reports,* 5**,** 13888.

MALARKEY, D. E., HOENERHOFF, M. J. & MARONPOT, R. R. 2018. Chapter 6 - Carcinogenesis: Manifestation and Mechanisms A2 - Wallig, Matthew A. *In:* HASCHEK, W. M., ROUSSEAUX, C. G. & BOLON, B. (eds.) *Fundamentals of Toxicologic Pathology (Third Edition).* Academic Press.

MANISCALCO, M., PARIS, D., MELCK, D. J., D'AMATO, M., ZEDDA, A., SOFIA, M., STELLATO, C. & MOTTA, A. 2016. Coexistence of Obesity and Asthma Determines a Distinct Respiratory Metabolic Phenotype. *J Allergy Clin Immunol*.

MATTARUCCHI, E., BARALDI, E. & GUILLOU, C. 2012. Metabolomics applied to urine samples in childhood asthma; differentiation between asthma phenotypes and identification of relevant metabolites. *Biomedical Chromatography,* 26**,** 89-94.

MAVEL, S., NADAL-DESBARATS, L., BLASCO, H., BONNET-BRILHAULT, F., BARTHÉLÉMY, C., MONTIGNY, F., SARDA, P., LAUMONNIER, F., VOURC′ H, P., ANDRES, C. R. & EMOND, P. 2013. 1H‑13C NMR-based urine metabolic profiling in autism spectrum disorders. *Talanta,* 114**,** 95-102.

MCGEACHIE, M. J., DAHLIN, A., QIU, W., CROTEAU-CHONKA, D. C., SAVAGE, J., WU, A. C., WAN, E. S., SORDILLO, J. E., AL-GARAWI, A., MARTINEZ, F. D., STRUNK, R. C., LEMANSKE, R. F., LIU, A. H., RABY, B. A., WEISS, S., CLISH, C. B. & LASKY-SU, J. A. 2015. The metabolomics of asthma control: a promising link between genetics and disease. *Immunity, Inflammation and Disease,* 3**,** 224-238.

MEHTA, S. H. & ADLER, C. H. 2015. Advances in Biomarker Research in Parkinson's Disease. *Current Neurology and Neuroscience Reports,* 16**,** 7.

MORTON, N. M., NELSON, Y. B., MICHAILIDOU, Z., DI ROLLO, E. M., RAMAGE, L., HADOKE, P. W. F., SECKL, J. R., BUNGER, L., HORVAT, S., KENYON, C. J. & DUNBAR, D. R. 2011. A Stratified Transcriptomics

Analysis of Polygenic Fat and Lean Mouse Adipose Tissues Identifies Novel Candidate Obesity Genes. *PLoS ONE,* 6**,** e23944.

NEWGARD, C. B., AN, J., BAIN, J. R., MUEHLBAUER, M. J., STEVENS, R. D., LIEN, L. F., HAQQ, A. M., SHAH, S. H., ARLOTTO, M., SLENTZ, C. A., ROCHON, J., GALLUP, D., ILKAYEVA, O., WENNER, B. R., YANCY, W. E., EISENSON, H., MUSANTE, G., SURWIT, R., MILLINGTON, D. S., BUTLER, M. D. & SVETKEY, L. P. 2009. A Branched-Chain Amino Acid-Related Metabolic Signature that Differentiates Obese and Lean Humans and Contributes to Insulin Resistance. *Cell metabolism,* 9**,** 311-326.

PALMER, J. A., POENITZSCH, A. M., SMITH, S. M., CONARD, K. R., WEST, P. R. & CEZAR, G. G. 2012. Metabolic Biomarkers of Prenatal Alcohol Exposure in Human Embryonic Stem Cell-derived Neural Lineages. *Alcoholism, clinical and experimental research,* 36**,** 1314-1324.

PAN, K. H., LIH, C. J. & COHEN, S. N. 2005. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc Natl Acad Sci U S A,* 102**,** 8961-5.

PATIN, F., CORCIA, P., VOURC'H, P., NADAL-DESBARATS, L., BARANEK, T., GOOSSENS, J. F., MAROUILLAT, S., DESSEIN, A. F., DESCAT, A., MADJI HOUNOUM, B., BRUNO, C., LEMAN, S., ANDRES, C. R. & BLASCO, H. 2016. Omics to Explore Amyotrophic Lateral Sclerosis Evolution: the Central Role of Arginine and Proline Metabolism. *Mol Neurobiol*.

POOLE, A., URBANEK, C., ENG, C., SCHAGEMAN, J., JACOBSON, S., O'CONNOR, B. P., GALANTER, J. M., GIGNOUX, C. R., ROTH, L. A., KUMAR, R., LUTZ, S., LIU, A. H., FINGERLIN, T. E., SETTERQUIST, R. A., BURCHARD, E. G., RODRIGUEZ-SANTANA, J. & SEIBOLD, M. A. 2014. Dissecting childhood asthma with nasal transcriptomics distinguishes subphenotypes of disease. *Journal of Allergy and Clinical Immunology,* 133**,** 670-678.e12.

PROKOPENKO, I., MCCARTHY, M. I. & LINDGREN, C. M. 2008. Type 2 diabetes: new genes, new understanding. *Trends in Genetics,* 24**,** 613-621.

PUTSCHÖGL, F. M., GAUM, P. M., SCHETTGEN, T., KRAUS, T., GUBE, M. & LANG, J. 2015. Effects of occupational exposure to polychlorinated biphenyls on urinary metabolites of neurotransmitters: A cross-sectional and longitudinal perspective. *International Journal of Hygiene and Environmental Health,* 218**,** 452-460.

RANGO, M., ARIGHI, A., MAROTTA, G., RONCHI, D. & BRESOLIN, N. 2013. PINK1 parkinsonism and Parkinson disease: Distinguishable brain mitochondrial function and metabolomics. *Mitochondrion,* 13**,** 59-61.

ROEDE, J. R., UPPAL, K., PARK, Y., TRAN, V. & JONES, D. P. 2014. Transcriptome–metabolome wide association study (TMWAS) of maneb and paraquat neurotoxicity reveals network level interactions in toxicologic mechanism. *Toxicology reports,* 1**,** 435-444.

RUSILOWICZ, M., DICKINSON, M., CHARLTON, A., O'KEEFE, S. & WILSON, J. 2016. A batch correction method for liquid chromatography–mass spectrometry data that does not depend on quality control samples. *Metabolomics,* 12**,** 56.

SALEK, R. M., MAGUIRE, M. L., BENTLEY, E., RUBTSOV, D. V., HOUGH, T., CHEESEMAN, M., NUNEZ, D., SWEATMAN, B. C., HASELDEN, J. N., COX, R. D., CONNOR, S. C. & GRIFFIN, J. L. 2007. A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. *Physiological Genomics,* 29**,** 99-108.

SALERNO, S., JR. & MEHRMOHAMADI, M. 2017. RRmix: A method for simultaneous batch effect correction and analysis of metabolomics data in the absence of internal standards. 12**,** e0179530.

SATO, Y., SUZUKI, I., NAKAMURA, T., BERNIER, F., AOSHIMA, K. & ODA, Y. 2012. Identification of a new plasma biomarker of Alzheimer's disease using metabolomics technology. *Journal of Lipid Research,* 53**,** 567-576.

SCHMUTZ, J., WHEELER, J., GRIMWOOD, J., DICKSON, M., YANG, J., CAOILE, C., BAJOREK, E., BLACK, S., CHAN, Y. M., DENYS, M., ESCOBAR, J., FLOWERS, D., FOTOPULOS, D., GARCIA, C., GOMEZ, M., GONZALES, E., HAYDU, L., LOPEZ, F., RAMIREZ, L., RETTERER, J., RODRIGUEZ, A., ROGERS, S., SALAZAR, A., TSAI, M. & MYERS, R. M. 2004. Quality assessment of the human genome sequence. *Nature,* 429**,** 365-368.

SMOLINSKA, A., BLANCHET, L., BUYDENS, L. M. C. & WIJMENGA, S. S. 2012. NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica Chimica Acta,* 750**,** 82-97.

STEPHENSON, S. T., BROWN, L. A. S., HELMS, M. N., QU, H., BROWN, S. D., BROWN, M. R. & FITZPATRICK, A. M. 2015. Cysteine oxidation impairs systemic glucocorticoid responsiveness in children with difficult-to-treat asthma. *Journal of Allergy and Clinical Immunology,* 136**,** 454-461.e9.

SUHRE, K., SHIN, S.-Y., PETERSEN, A.-K., MOHNEY, R. P., MEREDITH, D., WÄGELE, B., ALTMAIER, E., CARDIOGRAM, DELOUKAS, P., ERDMANN, J., GRUNDBERG, E., HAMMOND, C. J., DE ANGELIS, M. H., KASTENMÜLLER, G., KÖTTGEN, A., KRONENBERG, F., MANGINO, M., MEISINGER, C., MEITINGER, T., MEWES, H.-W., MILBURN, M. V., PREHN, C., RAFFLER, J., RIED, J. S., RÖMISCH-MARGL, W., SAMANI, N. J., SMALL, K. S., WICHMANN, H. E., ZHAI, G., ILLIG, T., SPECTOR, T. D., ADAMSKI, J., SORANZO, N. & GIEGER, C. 2011. Human metabolic individuality in biomedical and pharmaceutical research. *Nature,* 477**,** 10.1038/nature10354.

WANG, H., LIANG, S., WANG, M., GAO, J., SUN, C., WANG, J., XIA, W., WU, S., SUMNER, S. J., ZHANG, F., SUN, C. & WU, L. 2016. Potential serum biomarkers from a metabolomics study of autism. *J Psychiatry Neurosci,* 41**,** 27-37.

WELJIE, A. M., NEWTON, J., MERCIER, P., CARLSON, E. & SLUPSKY, C. M. 2006. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal Chem,* 78**,** 4430-42.

WEST, P. R., AMARAL, D. G., BAIS, P., SMITH, A. M., EGNASH, L. A., ROSS, M. E., PALMER, J. A., FONTAINE, B. R., CONARD, K. R., CORBETT, B. A., CEZAR, G. G., DONLEY, E. L. & BURRIER, R. E. 2014. Metabolomics as a tool for discovery of biomarkers of autism spectrum disorder in the blood plasma of children. *PLoS One,* 9**,** e112445.

WILD, C. P. 2005. Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers and Prevention,* 14**,** 1847-1850.

WINNIER, D. A., FOURCAUDOT, M., NORTON, L., ABDUL-GHANI, M. A., HU, S. L., FAROOK, V. S., COLETTA, D. K., KUMAR, S., PUPPALA, S., CHITTOOR, G., DYER, T. D., ARYA, R., CARLESS, M., LEHMAN, D. M., CURRAN, J. E., CROMACK, D. T., TRIPATHY, D., BLANGERO, J., DUGGIRALA, R., GÖRING, H. H. H., DEFRONZO, R. A. & JENKINSON, C. P. 2015. Transcriptomic Identification of <italic>ADH1B</italic> as a Novel Candidate Gene for Obesity and Insulin Resistance in Human Adipose Tissue in Mexican Americans from the Veterans Administration Genetic Epidemiology Study (VAGES). *PLoS ONE,* 10**,** e0119941.

WU, Y. E., PARIKSHAK, N. N., BELGARD, T. G. & GESCHWIND, D. H. 2016. Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. *Nat Neurosci,* advance online publication.

XIA, M., VIERA-HUTCHINS, L., GARCIA-LLORET, M., NOVAL RIVAS, M., WISE, P., MCGHEE, S. A., CHATILA, Z. K., DAHER, N., SIOUTAS, C. & CHATILA, T. A. 2015. Vehicular exhaust particles promote allergic airway inflammation through an aryl hydrocarbon receptor–notch signaling cascade. *Journal of Allergy and Clinical Immunology,* 136**,** 441-453.

XU, F., TAVINTHARAN, S., SUM, C. F., WOON, K., LIM, S. C. & ONG, C. N. 2013. Metabolic Signature Shift in Type 2 Diabetes Mellitus Revealed by Mass Spectrometry-based Metabolomics. *The Journal of Clinical Endocrinology & Metabolism,* 98**,** E1060-E1065.

YU, M., CUI, F.-X., JIA, H.-M., ZHOU, C., YANG, Y., ZHANG, H.-W., DING, G. & ZOU, Z.-M. 2016. Aberrant purine metabolism in allergic asthma revealed by plasma metabolomics. *Journal of Pharmaceutical and Biomedical Analysis,* 120**,** 181-189.

ZAMBELLI, A. E. 2016. A data-driven approach to estimating the number of clusters in hierarchical clustering. *F1000Res,* 5.

ZHAO, X., FRITSCHE, J., WANG, J., CHEN, J., RITTIG, K., SCHMITT-KOPPLIN, P., FRITSCHE, A., HÄRING, H.-U., SCHLEICHER, E. D., XU, G. & LEHMANN, R. 2010. Metabonomic fingerprints of fasting plasma and spot urine reveal human pre-diabetic metabolic traits. *Metabolomics,* 6**,** 362-374.

ZHU, C., LIANG, Q.-L., HU, P., WANG, Y.-M. & LUO, G.-A. 2011. Phospholipidomic identification of potential plasma biomarkers associated with type 2 diabetes mellitus and diabetic nephropathy. *Talanta,* 85**,** 1711-1720.

Annex 1

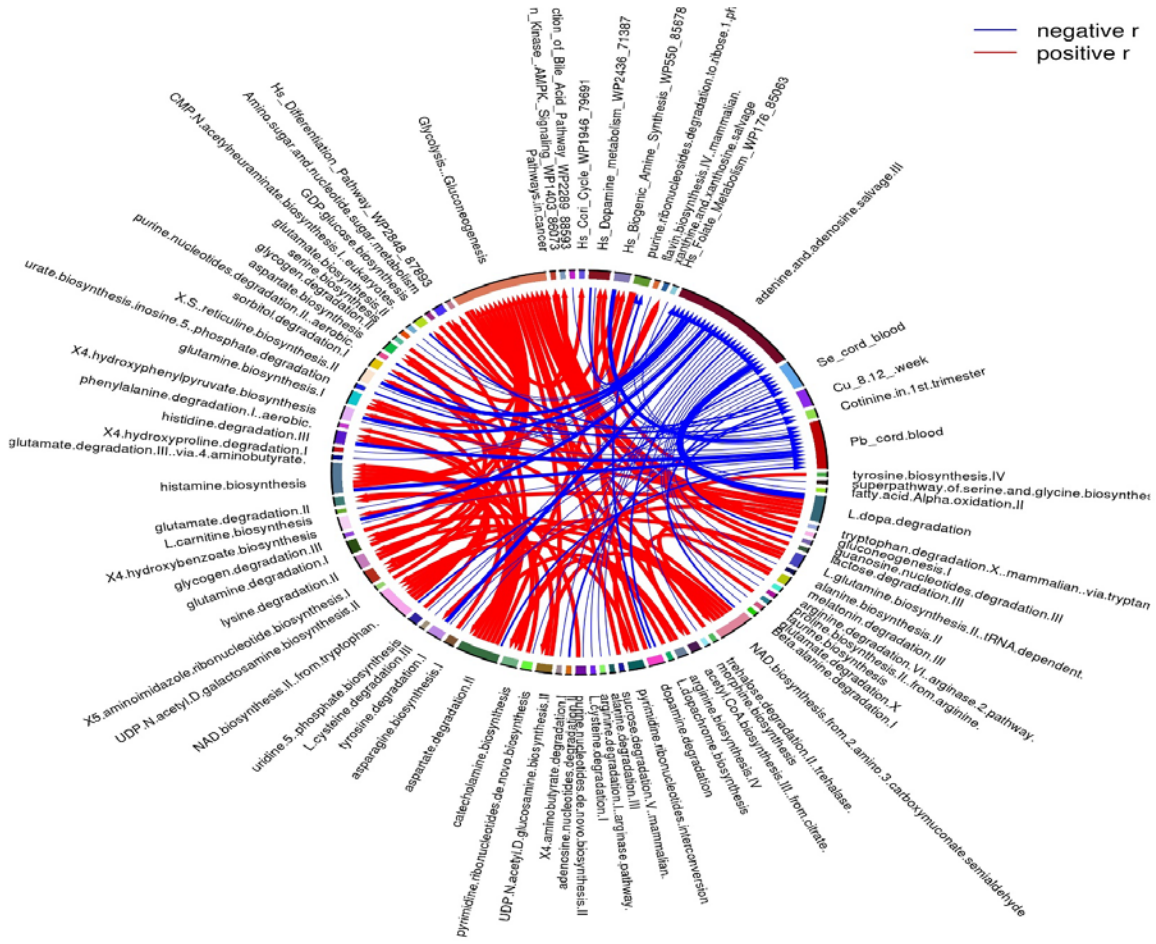


Figure 22. Correlation globe plot of external exposure and metabolic pathway perturbation for the REPRO PL study. The green and the dark yellow connection lines are subsets of the blue and red lines respectively. These subsets have been calculated considering the q value of the EWAS analysis, applying as validation method the false discovery rate on 4 random subsets of the dataset (Benjamini and Hochberg, 1995). Line thickness is proportional to size of the absolute value of correlation coefficient.

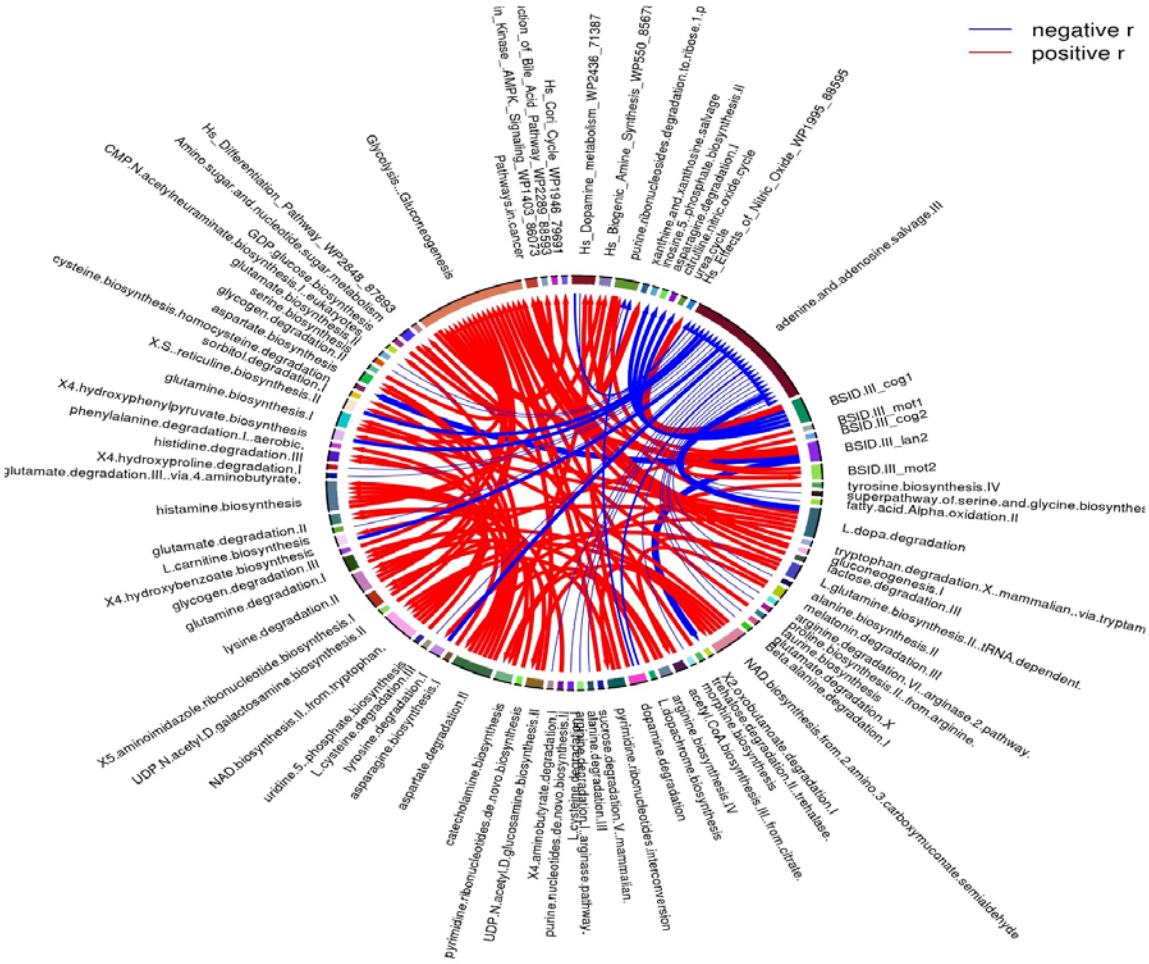


Figure 23. Correlation globe for the neurodevelopmental health outcomes and metabolic pathway perturbations of the REPRO PL study.

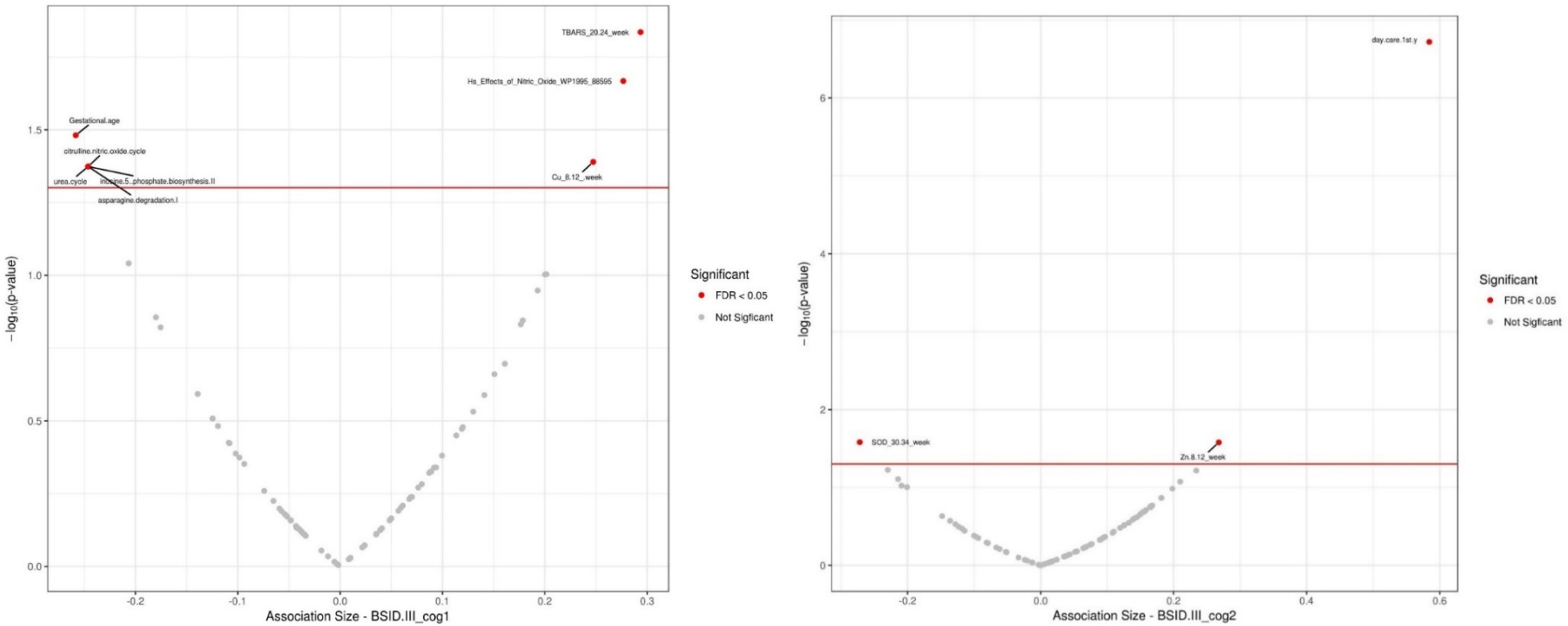


Figure 24. Association of the cognitive development volcano plots of at one year of age, on the left, and at two years of age, on the right, with exposure and modifiers. The cognitive development during the first year is positively associated with the concentration of Cu and TBARS during the second trimester of pregnancy, and negatively associated with the metabolic pathways of urea cycle, asparagine degradation I, inosine-5'-phosphate biosynthesis II, and citrulline nitric oxide cycle. During the second

year of life, the cognitive development is positively associated with the concentration of Zn during the first weeks (8-12) of pregnancy and negatively associated with the abundance of superoxide dismutase (SOD) during the last weeks of pregnancy.

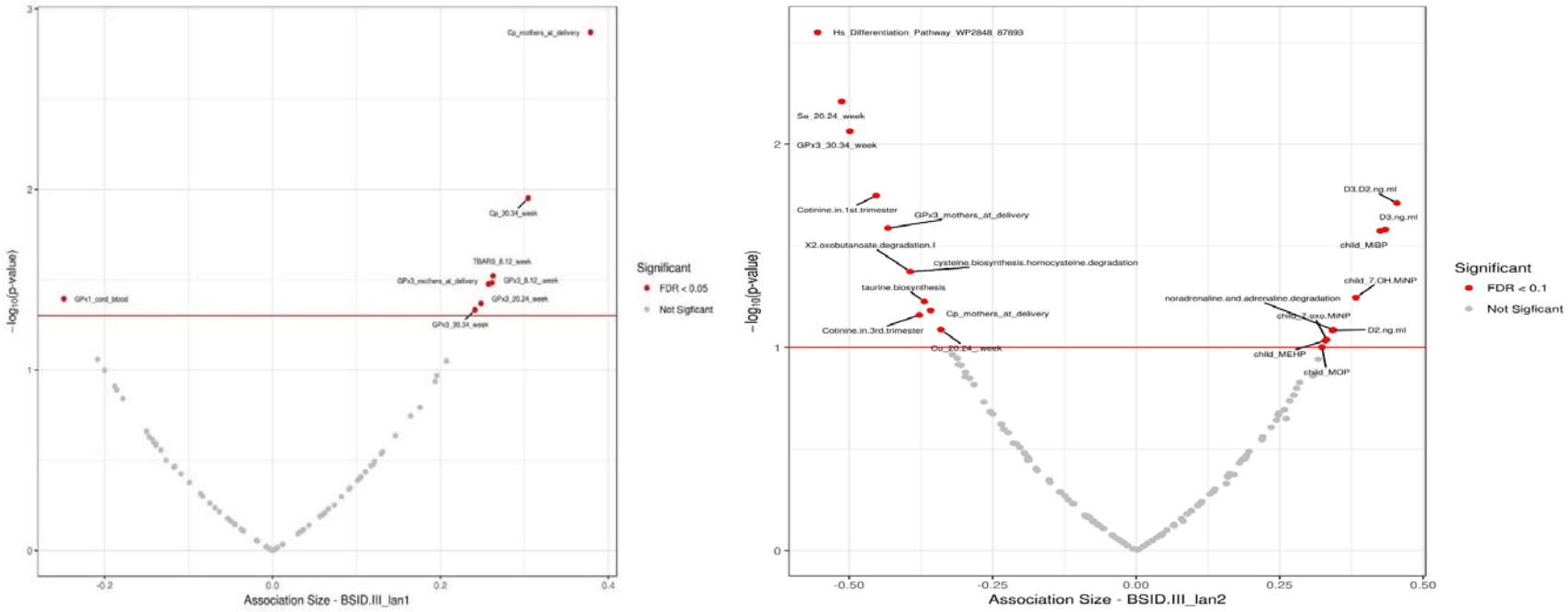


Figure 25. Association of the language development volcano plots of at one year of age, on the left, and at two years of age, on the right, with exposure and modifiers. The language development during the first year is positively associated with the concentration of glutathione peroxide (GPx3) during the pregnancy, the concentration of TBARS during the first trimester of pregnancy, and the concentration of SOD during the last trimester, but is negatively associated with the concentration of glutathione peroxide (GPx1) at the cord blood. During the second year of life, the language development is positively associated with the metabolic pathways of oxobutanoate

degradation I, cysteine biosynthesis/homocysteine degradation, and the glycolysis/gluconeogenesis. Concentration of Zn during the first weeks (8-12) of pregnancy and the concentration of SOD during the last weeks of pregnancy, are positively and negatively associated respectively.

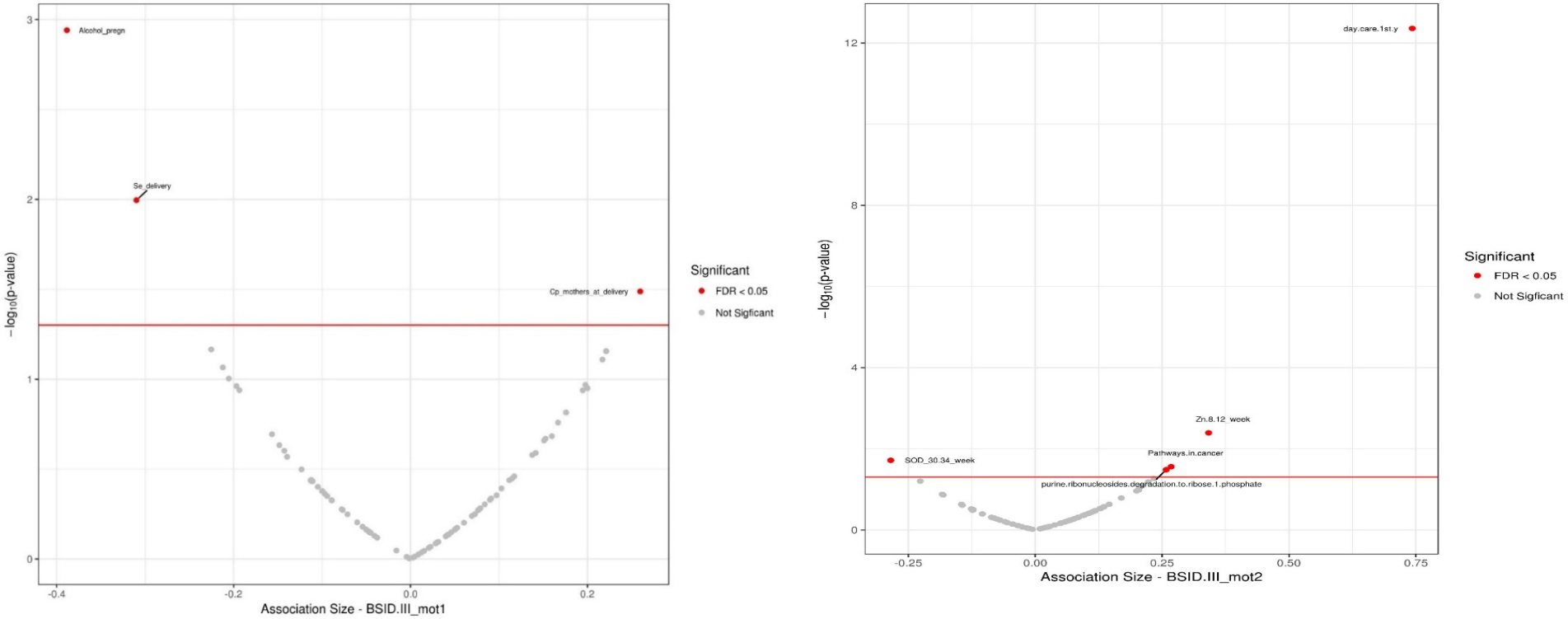


Figure 26. Association of the motor development volcano plots of at one year of age, on the left, and at two years of age, on the right, with exposure and modifiers. Lysine degradation and purine ribonucleotides degradation to ribose 1 phosphate are positively associated with the motor development. The concentration of Se at delivery and SOD during the last trimester are negatively associated with motor development, while the concentration of Zn during the first weeks of pregnancy is positively associated with the motor development during the second year of life.

Annex 2

Data preprocessing for the metabolomics experiment

Analytical Method: LC-MS/MS ESI (-)

Raw Data Import

In this study the number of the imported problem samples was 21, while the number of replicates of the solvent mixtures of I.S was 3. The preview of the superimposed TIC of problem samples and mixtures of I.S respectively, led to the observation of the baseline or the noise level.

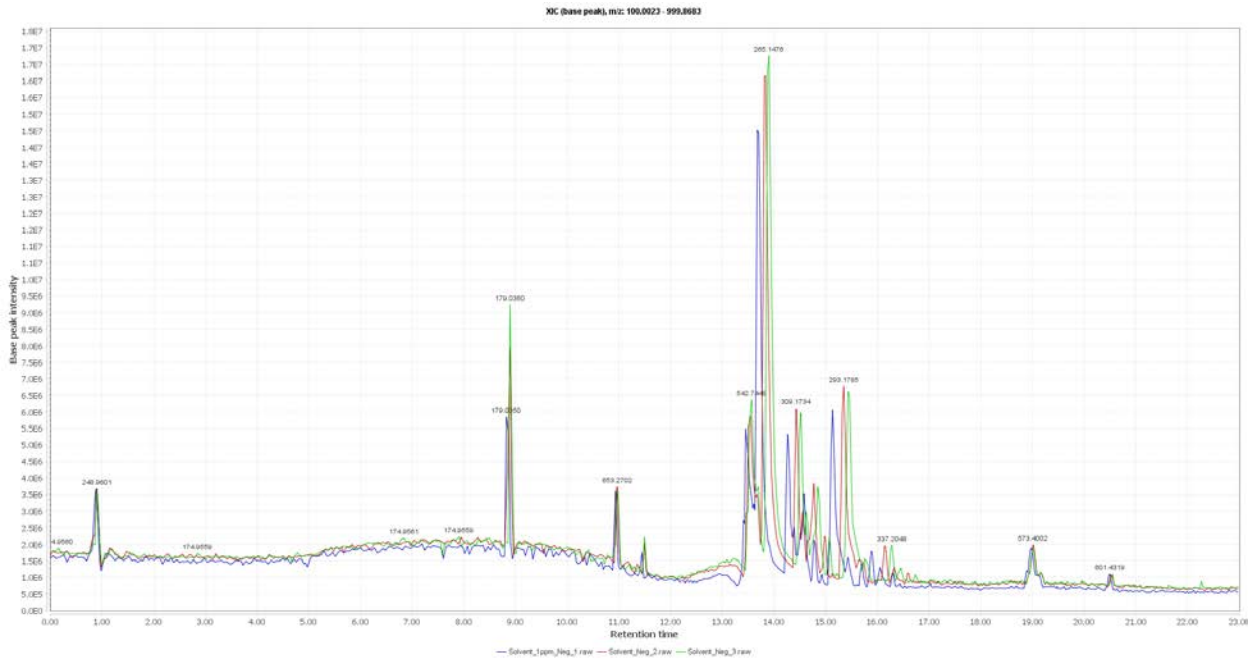


Figure 27. Superimposed TIC of the three mixtures of internal standards that have been analyzed during the experiment in negative mode.

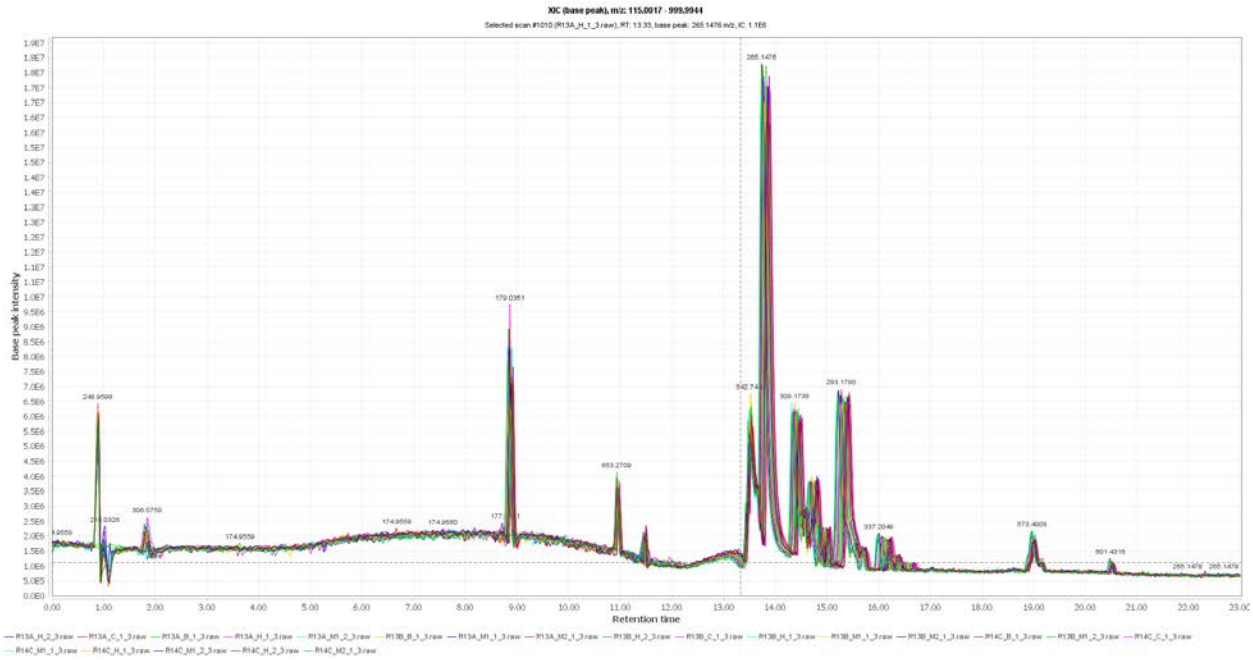


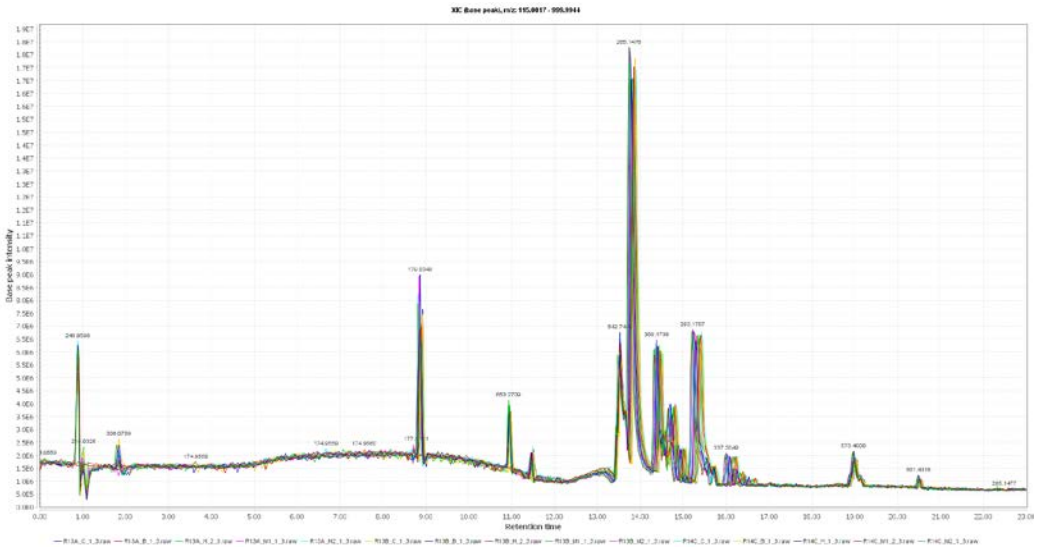
Figure 28. Superimposed TIC of the problem samples of cells after 3 weeks of treatment with two different mixtures of phthalates and metals, that have been analyzed during the experiment in negative mode.

Baseline Correction

The asymmetric baseline corrector was the most suitable method for the correction of the varying baseline of the above chromatograms. The smoothing and the asymmetry were set at 500 and 0.5 respectively. The baseline of the problem samples chromatogram was decreased from 8.1E5 to 1.9E4.

Figure 29. The choice of the most preferable parameters for smoothing and asymmetry was based on trial and error. For the first figure from the left the asymmetry was set at 0.5, for the second at 0.05, while the smoothing was set at 500, and for the third the smoothing was set at 1000 and the asymmetry at 0.5.

A



B

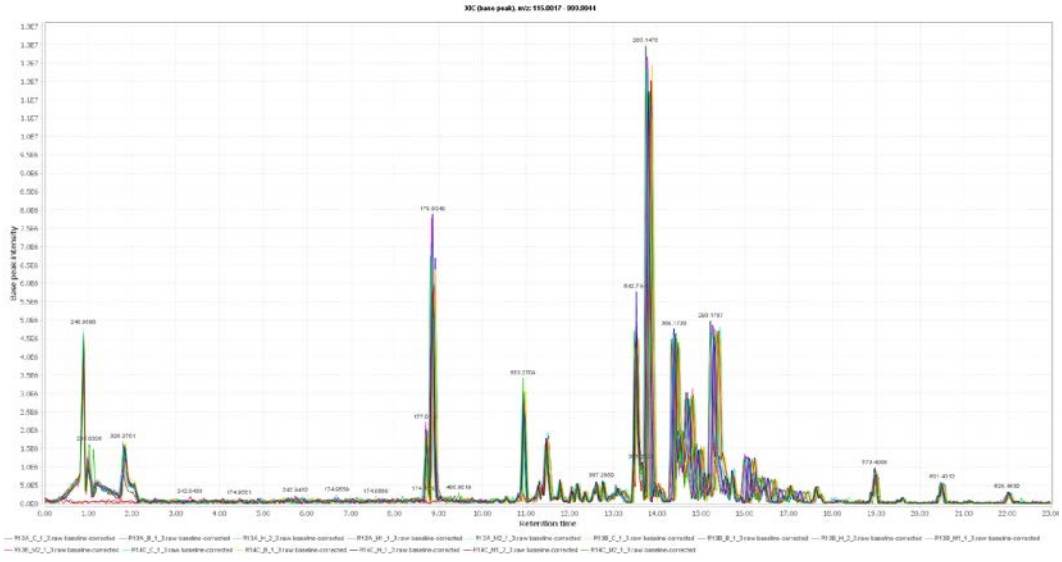


Figure 30. Superimposed TIC of all the problem samples that will be used for further analysis A. before baseline correction and B. after baseline correction.

Mass Detection

For the first step of mass detection, a centroid mass detector was used, since the data was centroided. The noise level after baseline correction was 5.0E3.

Chromatogram Builder

Following the mass detection, the chromatograms were build using the results from the tetrabromobisphenol A for the calculation of m/z tolerance, since the MS of this internal standard was closer to the calculated one.

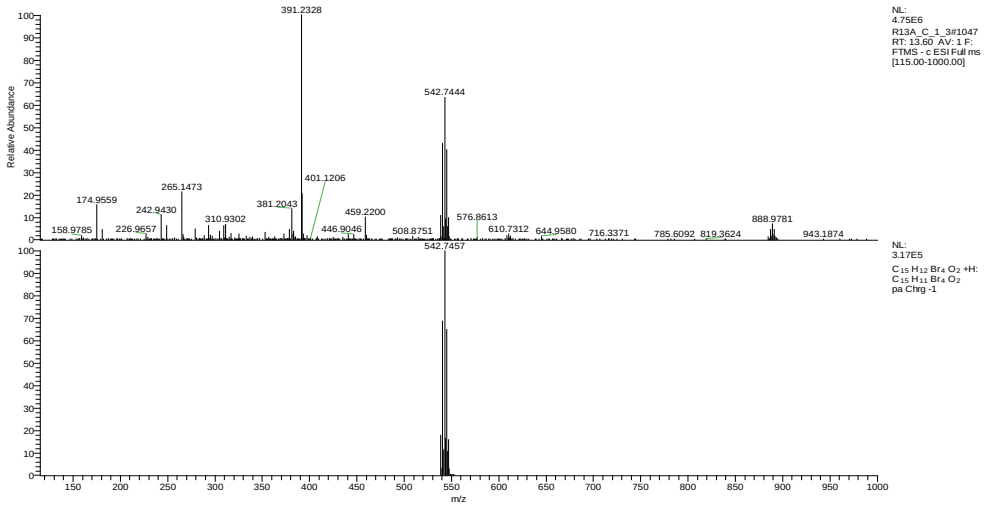


Figure 31. The first figure shows the MS for the tetrabromobisphenol A that has been added to the problem sample R13_C_1_1, while the second presents the calculated MS, according to the bibliography.

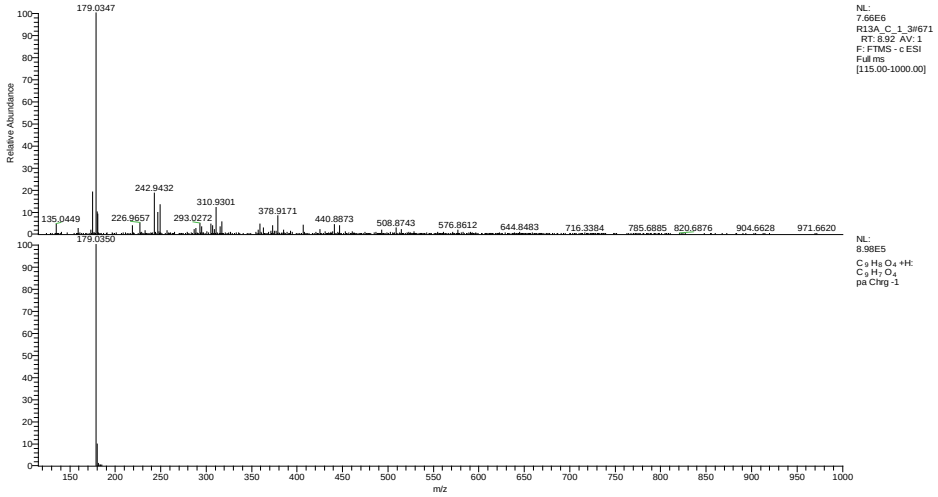


Figure 32. The first figure shows the MS for the caffeic acid that has been added to the problem sample R13_C_1_1, while the second presents the calculated MS, according to the bibliography.

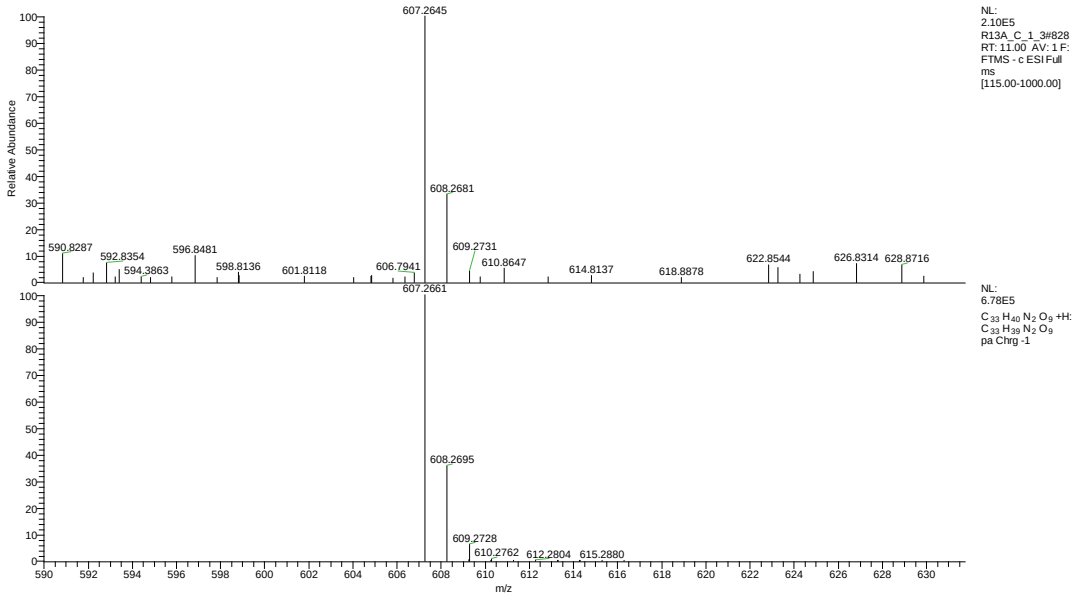


Figure 33. The first figure shows the MS for the reserpine that has been added to the problem sample R13_C_1_1, while the second presents the calculated MS, according to the bibliography.

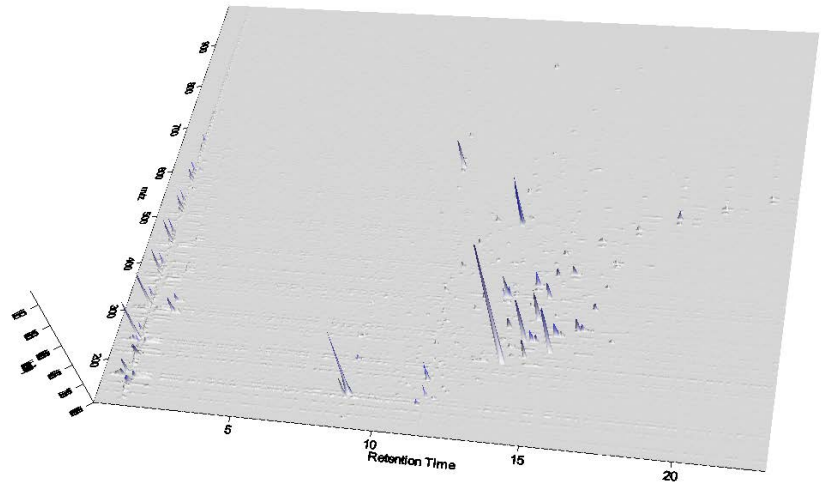


Figure 34. 3D visualizer can be used in order to detect hidden peaks behind another. The presented 3D TIC has been created after the baseline correction and the mass detection for the R13_C_1_1.

.

Table 2. Parameters for the chromatogram builder step during data preprocessing for the problem samples.

Sample	Calculated Mass			Observed Mass			Observed RT		
	Tetrabromobisphenol A	Caffeic acid	Reserpine	Tetrabromobisphenol A	Caffeic acid	Reserpine	Tetrabromobisphenol A	Caffeic acid	Reserpine
R13A_B_1_3	542.7457	179.0305	607.2661						
R13B_B_1_3	542.7457	179.0305	607.2661	542.7446	179.0348	607.2658	13.53	8.86	10.95
R14C_B_1_3	542.7457	179.0305	607.2661	542.7445	179.0347	607.2651	13.57	8.93	10.98
R13A_C_1_3	542.7457	179.0305	607.2661	542.7444	179.0347	607.2645	13.6	8.92	10.98
R13B_C_1_3	542.7457	179.0305	607.2661	542.7448	179.0349	607.2651	13.53	8.92	10.99
R14C_C_1_3	542.7457	179.0305	607.2661	542.7448	179.0348	607.265	13.5	8.82	10.93
R13A_H_2_3	542.7457	179.0305	607.2661	542.7444	179.0347	607.265	13.51	8.91	10.96
R13B_H_2_3	542.7457	179.0305	607.2661						
R14C_H_1_3	542.7457	179.0305	607.2661	542.7445	179.0349	607.2657	13.52	8.84	10.94
R13A_M1_1_3	542.7457	179.0305	607.2661	542.7446	179.0349	607.2656	13.52	8.86	10.95
R13B_M1_1_3	542.7457	179.0305	607.2661	542.7444	179.0349	607.2656	13.55	8.85	10.93
R14C_M1_2_3	542.7457	179.0305	607.2661	542.7446	179.0348	607.2657	13.49	8.89	10.93
R13A_M2_1_3	542.7457	179.0305	607.2661	542.7448	179.0349	607.2657	13.53	8.88	10.96
R13B_M2_1_3	542.7457	179.0305	607.2661	542.7444	179.0348	607.2646	13.46	8.84	10.9
R14C_M2_1_3	542.7457	179.0305	607.2661	542.7449	179.0349	607.2662	13.45	8.82	10.93

Table 3. Parameters for the chromatogram builder step during data preprocessing for the problem samples.

Sample	Mass error in ppm	Uncertainty in m/z	Min time span			Min height
			End	Start	Span	
R13A_B_1_3	2.0267	0.0011	4.35	4.24	0.11	8.60E4
R13B_B_1_3	2.0267	0.0011	4.35	4.24	0.11	8.60E4
R14C_B_1_3	2.2110	0.0012	5.17	5.03	0.14	9.30E4
R13A_C_1_3	2.3952	0.0013	3.59	3.5	0.09	8.80E4
R13B_C_1_3	1.6582	0.0009	4.07	3.96	0.11	1.20E5
R14C_C_1_3	1.6582	0.0009	4.82	4.71	0.11	7.30E4
R13A_H_2_3	2.3952	0.0013	4.61	4.5	0.11	1.00E5
R13B_H_2_3	2.3952	0.0013	4.61	4.5	0.11	1.00E5
R14C_H_1_3	2.2110	0.0012	3.43	3.33	0.1	9.50E4
R13A_M1_1_3	2.0267	0.0011	3.74	3.63	0.11	1.30E5
R13B_M1_1_3	2.3952	0.0013	4.32	4.21	0.11	8.00E4
R14C_M1_2_3	2.0267	0.0011	4.85	4.74	0.11	7.50E4
R13A_M2_1_3	1.6582	0.0009	3.9	3.8	0.1	7.50E4
R13B_M2_1_3	2.3952	0.0013	3.13	3.02	0.11	4.80E4
R14C_M2_1_3	1.4740	0.0008	3.48	3.37	0.11	5.10E4

Deconvolution

For the last step of mass detection, the chromatogram deconvolution, the local minimum algorithm was the most suitable due to the small level of noise. The top/edge parameter was estimated based on the peak of caffeic acid (m/z~179.0305, RT~8.86).

Table 4. Comparison of the number of peaks before and after deconvolution

Sample	Number of detected peaks before deconvolution	Number of detected peaks after deconvolution
R13A_B_1_3	74	102
R13B_B_1_3	151	214
R14C_B_1_3	81	151
R13A_C_1_3	199	287
R13B_C_1_3	106	196
R14C_C_1_3	147	250
R13A_H_2_3	125	192
R13B_H_2_3	79	93
R14C_H_1_3	161	246
R13A_M1_1_3	103	135
R13B_M1_1_3	122	198
R14C_M1_2_3	160	279
R13A_M2_1_3	176	290
R13B_M2_1_3	192	383
R14C_M2_1_3	303	625

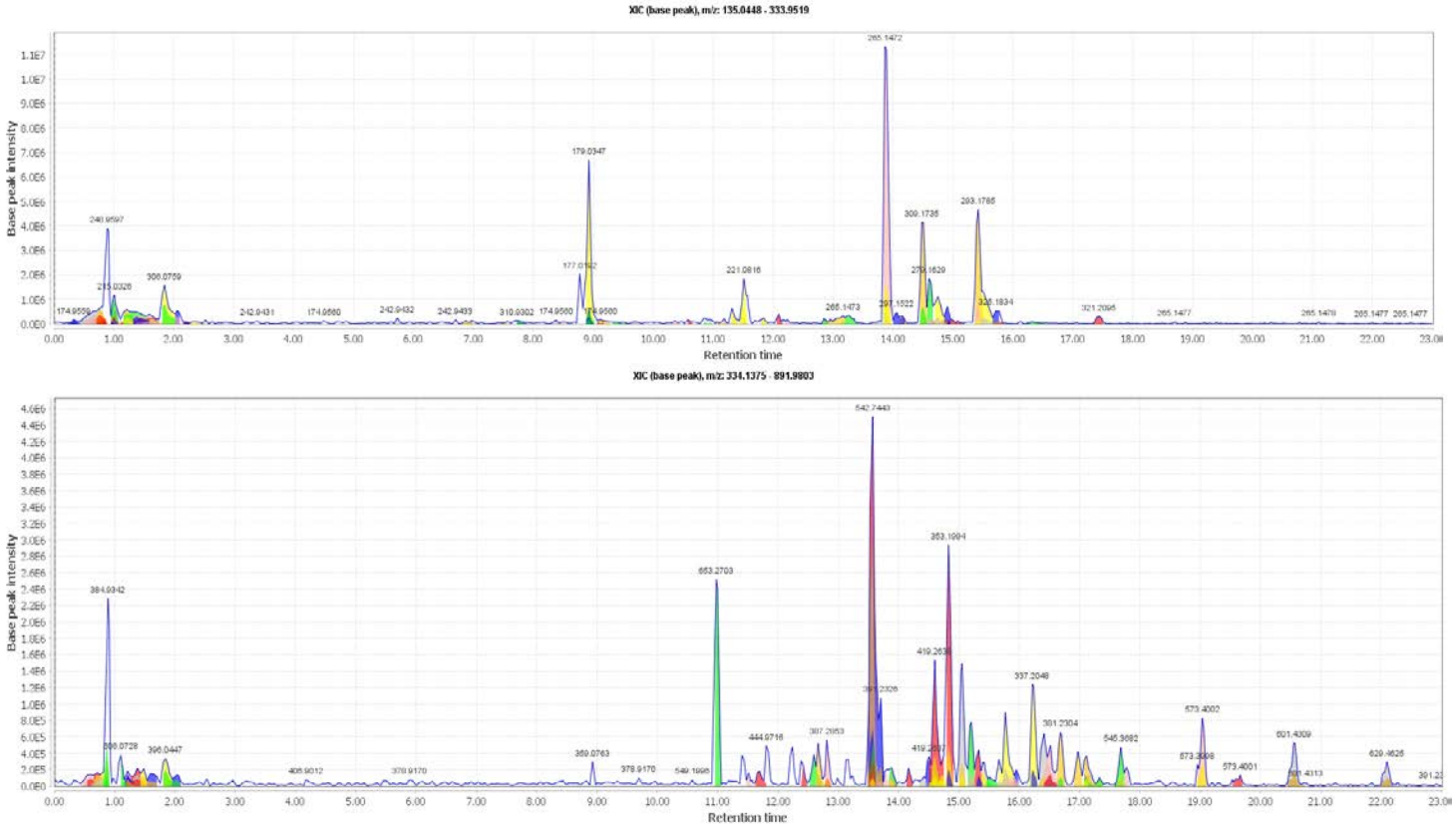


Figure 35. The number of the detected peaks has increased after deconvolution for all the problem samples. The first XIC has been created from the peak list before deconvolution, while the second after deconvolution R13_C_1_1.

Table 5. Parameters for the chromatogram deconvolution step during data preprocessing for the problem samples.

Sample	Min Absolute height	Max Absolute Height	Minimum Relative height (%)	Search minimum in RT range (min)					
				End	Start	Span	Top	Edge	top/edge
R13A_C_1_3	8.80E4	1.00E7	0.88	13.93	13.74	0.19	7.20E5	7.90E6	0.0911
R13B_C_1_3	8.80E4	7.90E7	0.11	1.1	0.93	0.17	7.20E5	7.90E6	0.0911
R14C_C_1_3	9.40E4	4.70E6	2.00	17.23	17.02	0.21	7.20E5	7.90E6	0.0911
R13A_H_2_3	8.90E4	1.10E7	0.81	1.1	0.95	0.15	1.40E5	6.70E6	0.0209
R13B_H_2_3	1.20E5	6.40E6	1.88	1.09	0.93	0.16	1.40E5	6.40E6	0.0219
R14C_H_1_3	7.40E4	1.20E7	0.62	1.09	0.93	0.16	6.20E4	7.10E6	0.0087
R13A_M1_1_3	1.00E5	1.10E7	0.91	1.14	0.99	0.15	9.00E4	6.00E6	0.0150
R13B_M1_1_3	1.00E5	1.20E7	0.83	17.19	16.96	0.23	9.00E4	6.00E6	0.0150
R14C_M1_2_3	9.70E4	1.10E7	0.88	12.86	12.69	0.17	1.60E5	5.80E6	0.0276
R13A_M2_1_3	1.40E5	5.80E6	2.41	1.1	0.93	0.17	4.10E5	5.80E6	0.0707
R13B_M2_1_3	8.70E4	1.10E7	0.79	14.9	14.69	0.21	8.50E4	5.30E6	0.0160
R14C_M2_1_3	7.50E4	1.20E7	0.63	1.12	0.95	0.17	4.40E4	5.90E6	0.0075
R13A_C_1_3	7.70E4	1.10E7	0.70	11.1	10.90	0.2	2.30E4	5.50E6	0.0042
R13B_C_1_3	4.90E4	1.20E7	0.41	1.16	0.89	0.27	4.20E5	7.80E6	0.0538
R14C_C_1_3	7.40E3	6.80E6	0.11	2.02	1.86	0.16	1.70E5	6.80E6	0.0250

Deisotoping

Chromatograms were deisotoped using the isotopic peaks grouper algorithm. The values for the m/z tolerance were estimated previous during the chromatogram builder. The estimation of the absolute value of the retention time tolerance was based on the peak with the higher duration time in the chromatogram. The maximum charge has been set at 2 and the representative isotope was chosen to be the most intense. The monotonic shape was checked.

Table 6. Parameters for the chromatogram deisotoping step during data preprocessing for the problem samples.

Sample	Mass error in ppm	Uncertainty in m/z	Min time span		
			End	Start	Span
R13A_B_1_3	2.0267	0.0011	6.39	6.29	0.1
R13B_B_1_3	2.0267	0.0011	1.06	0.95	0.11
R14C_B_1_3	2.2110	0.0012	8.29	8.24	0.05
R13A_C_1_3	2.3952	0.0013	1.07	0.97	0.1
R13B_C_1_3	1.6582	0.0009	1.48	1.43	0.05
R14C_C_1_3	1.6582	0.0009	9.26	9.21	0.05
R13A_H_2_3	2.3952	0.0013	8.56	8.5	0.06
R13B_H_2_3	2.3952	0.0013	15.9	15.83	0.07
R14C_H_1_3	2.2110	0.0012	14.8	14.75	0.05
R13A_M1_1_3	2.0267	0.0011	0.81	0.76	0.05
R13B_M1_1_3	2.3952	0.0013	15.1	15.02	0.08
R14C_M1_2_3	2.0267	0.0011	7.63	7.57	0.06
R13A_M2_1_3	1.6582	0.0009	5.8	5.75	0.05
R13B_M2_1_3	2.3952	0.0013	1.43	1.38	0.05
R14C_M2_1_3	1.4740	0.0008	3.72	3.67	0.05

Alignment

Peak alignment was performed using the Join aligner method (m/z tolerance at 0.0002 (or 1.1171ppm), absolute RT tolerance at 0.11 min, weight for m/z and weight for RT at 0). Caffeic acid was used as reference to evaluate the performance of the chosen alignment algorithm.

Gap-Filling

The peak list was eventually gap-filled with the peak finder module (intensity tolerance at 80%, m/z tolerance at 0.0002 (or 1.1171 ppm), and absolute RT tolerance of 0.44 min). In parallel, RT correction was

performed. Finally, the exported list of detected metabolites was included the row m/z, row retention time, peak m/z, peak RT, peak height, and peak area. The total number of detected metabolites was 1155.

Instrument Variability and Overall Process Variability

From the 1127 detected metabolites 105 were detected only in one out of 15 samples (missing values = 93.33%). To obtain consistent variables these metabolites were deleted, in accordance to the 80% missing values rule. The resulting 2D matrix contains 1022 metabolites.

Instrument variability was determined by calculating the median RSD of RT, peak area and mass accuracy for the I.S of the 4 treatment conditions, while overall process variability was determined by calculating the median RSD of RT, peak area and mass accuracy for all the endogenous metabolites of the 4 treatment conditions.

Table 7. Instrument's and Overall process variability results.

Instrument Variability based on peaks area						
I.S	RSD_B %	RSD_C %	RSD_H %	RSD_M1 %	RSD_M2 %	RSD_ALL SAMPLES %
Caffeic acid	86.72	10.73	87.15	5.49	12.59	42.01
tetrabromobisphenol A	86.65	4.02	86.61	0.70	4.63	41.10
Reserpine	-	-	-	-	-	-
Median	86.68	7.37	86.88	3.10	8.61	41.55
Instrument Variability based on peaks mass						
I.S		RSD_C %	RSD_H %	RSD_M1 %	RSD_M2 %	RSD_ALL SAMPLES %
Caffeic acid	86.60	2.1021E-05	86.60	4.90216E-05	3.72314E-05	40.60
tetrabromobisphenol A	86.60	6.49268E-06	86.60	4.65707E-05	5.40299E-05	40.60
Reserpine	-	-	-	-	-	-
Median	86.60	1.37568E-05	86.60	4.77961E-05	4.56306E-05	40.60
Instrument Variability based on peaks RT						
I.S		RSD_C %	RSD_H %	RSD_M1 %	RSD_M2 %	RSD_ALL SAMPLES %
Caffeic acid	86.60	0.65	86.60	0.22	0.36	40.60
tetrabromobisphenol A	86.60	0.24	86.60	0.13	0.23	40.60
Reserpine	-	-	-	-	-	-
Median	86.60	0.45	86.60	0.17	0.30	40.60
Overall Process Variability (considering all metabolites)						
area	62.89	0.94	86.60	1.73	1.57	43.24
mass	4.39954E-05	2.16229E-05	86.60	3.63352E-05	3.07178E-05	40.60
RT	62.89	0.94	86.60	1.73	1.57	43.24

The reason why the RSD values for control samples were significant higher than the other ones is that none of the caffeic acid and tetrabromobisphenol A were detected in R13A_B_1_3 or R13B_H_2_3. The reason why the peaks of the specific internal standards were not detected in R13A_B_1_3 or R13B_H_2_3 is unknown.

Metabolites Annotation

The internal standards [caffeic acid (m/z: 179.0347, rt:8.93), and tetrabromobisphenol A (m/z: 542.7446 rt:13.53)] were excluded from the further bioinformatics analysis. Reserpine was not included in the final peak list. MetaboSearch Tool v.1.2 was used for metabolites annotation. The used customized databases were the HMDB, Metlin, and Lipid Maps, and the mass tolerance was 5 ppm. The total number of unique identified metabolites was 45/1020.

Analytical Method: LC-MS/MS ESI (+)

Raw Data Import

The imported raw data consistent of 21 problem samples and 3 replicates of solvent mixtures of I.S.

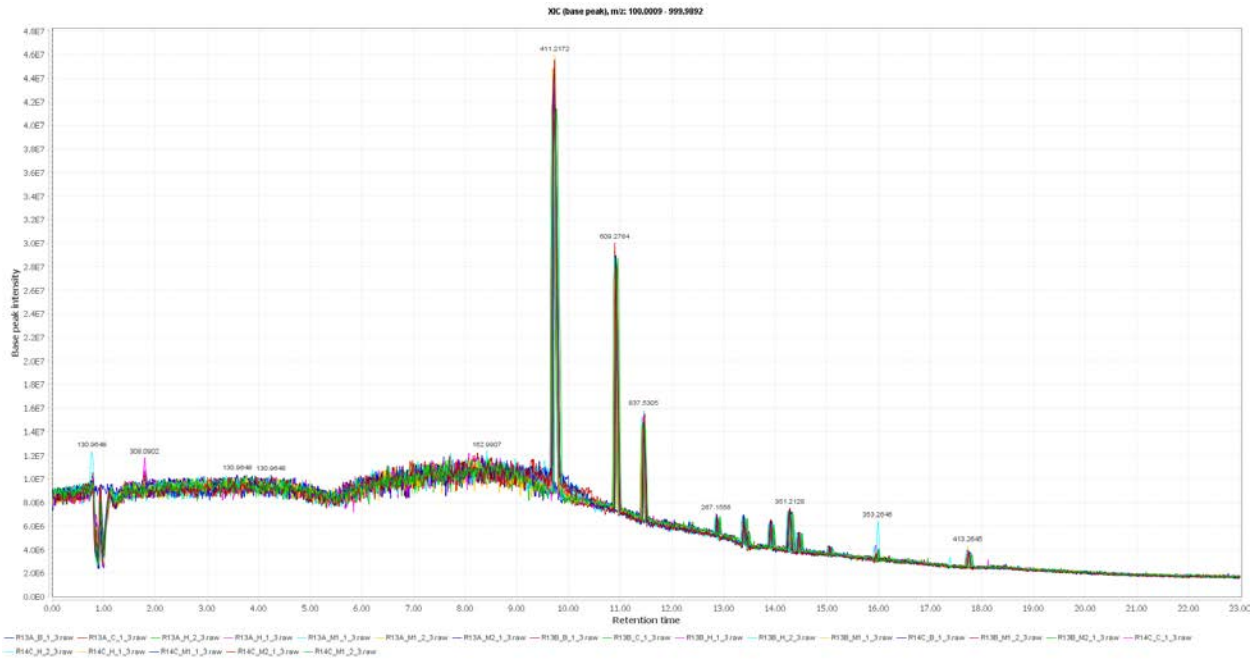


Figure 36. Superimposed TIC of the three mixtures of internal standards that have been analyzed during the experiment in positive mode.

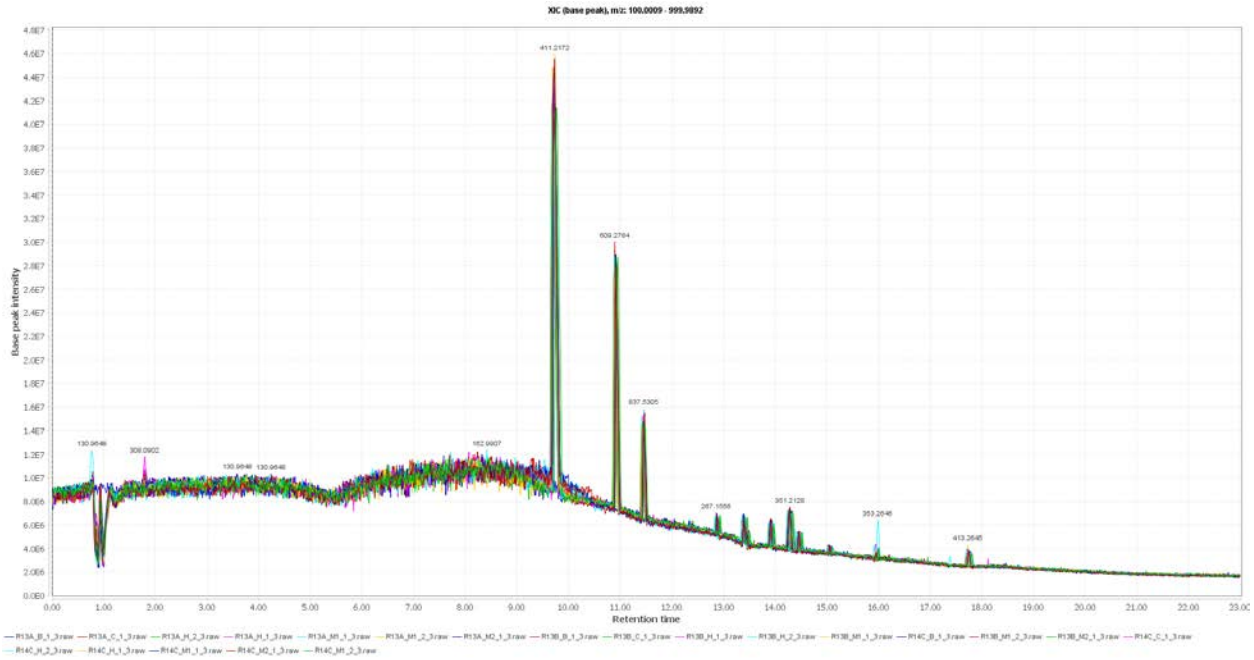


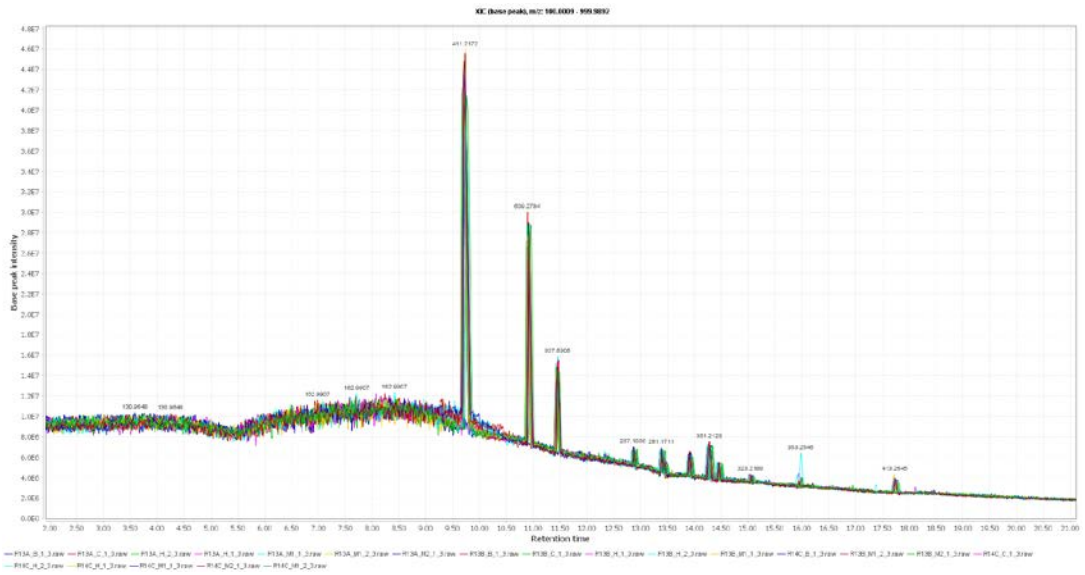
Figure 37. Superimposed TIC of the problem samples of cells after 3 weeks of treatment with two different mixtures of phthalates and metals, that have been analyzed during the experiment in positive mode.

Baseline Correction

The baseline correction was proceeded using the asymmetric baseline corrector and setting the smoothing and the asymmetry parameters at 500 and 0.5 respectively.

The baseline of the above chromatogram was transferred from 9.1E6 to 3.1E5. No significant different intensity values were observed for most of the detected peaks. The samples belonging to the same treatment condition had similar behavior, as it was expected.

A.



B.

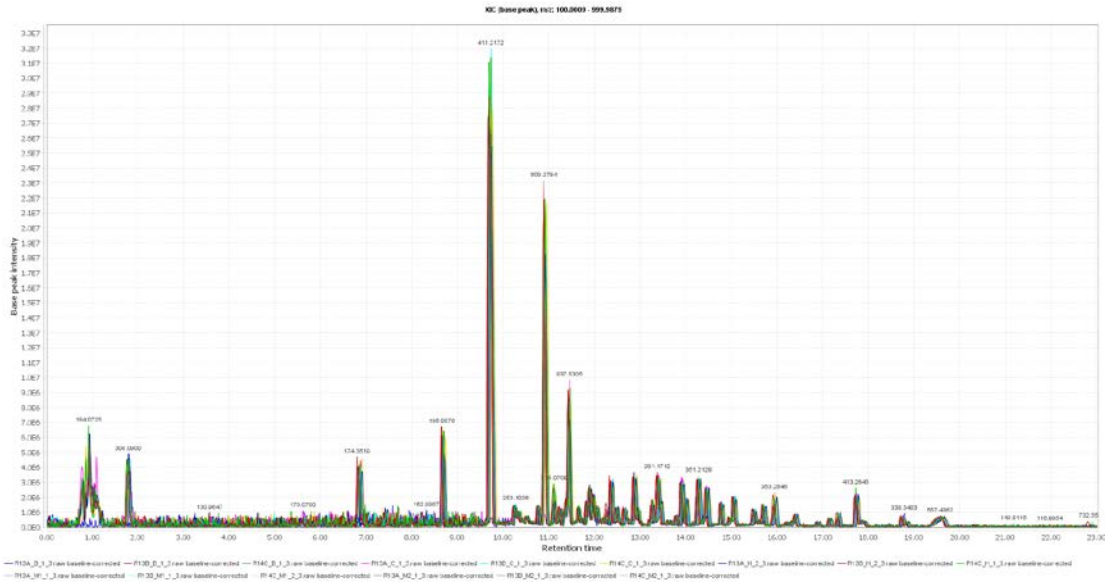


Figure 38. Superimposed TIC of all the problem samples that will be used for further analysis A. before baseline correction and B. after baseline correction.

Mass Detection

The noise was further reduced using the centroid mass detector. The noise level in case of positive mode was estimated at 2.0E3.

Chromatogram Builder

The calculations for the parameters $mass_{error}$(ppm) and uncertainty in m/z was based on the I.S risperidone since it was the most intense in most of the cases (Figure 35).

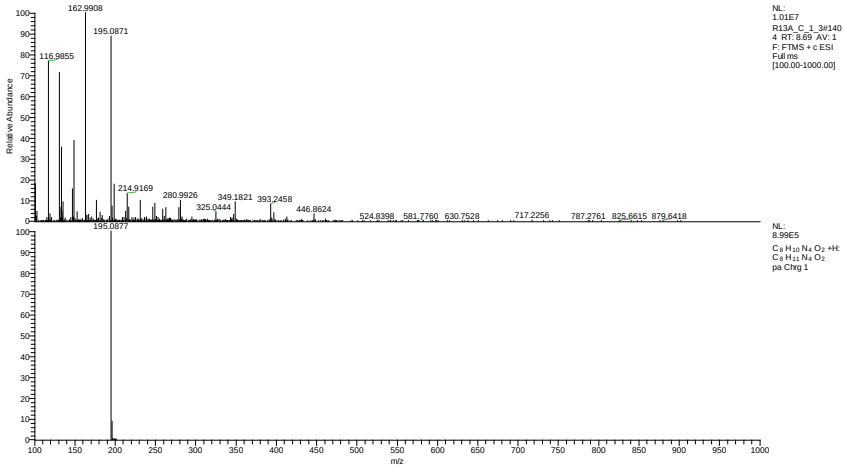


Figure 39. The first figure shows the MS for the caffeine that has been added to the problem sample R13_C_1_1, while the second presents the calculated MS, according to the bibliography.

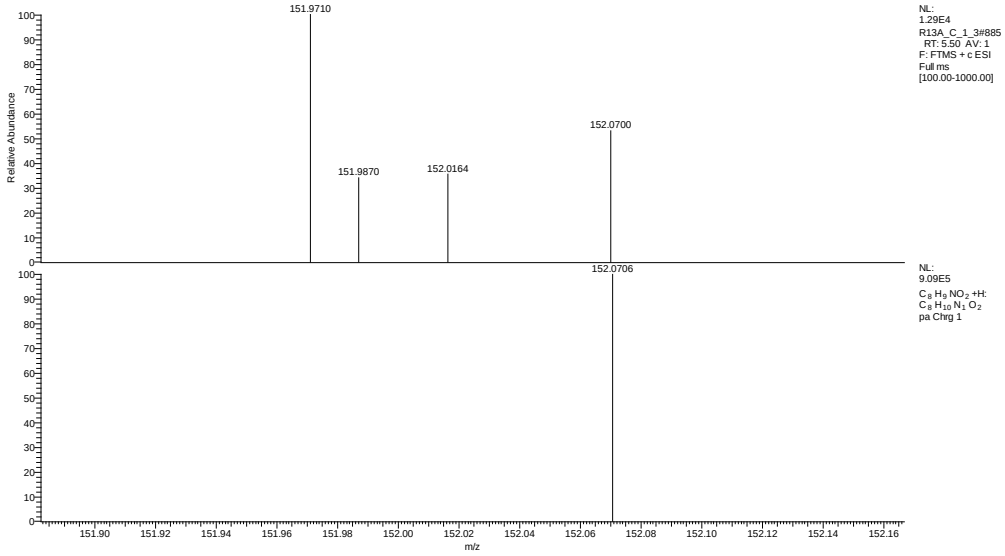


Figure 40. The first figure shows the MS for the paracetamol that has been added to the problem sample R13_C_1_1, while the second presents the calculated MS, according to the bibliography.

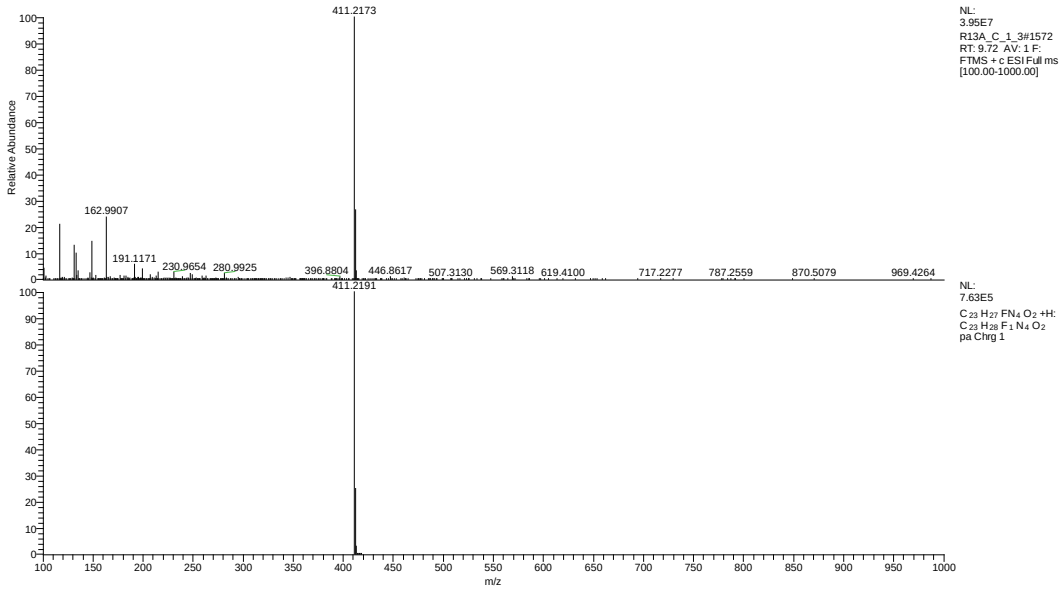


Figure 41. The first figure shows the MS for the risperidone that has been added to the problem sample R13_C_1_1, while the second presents the calculated MS, according to the bibliography.

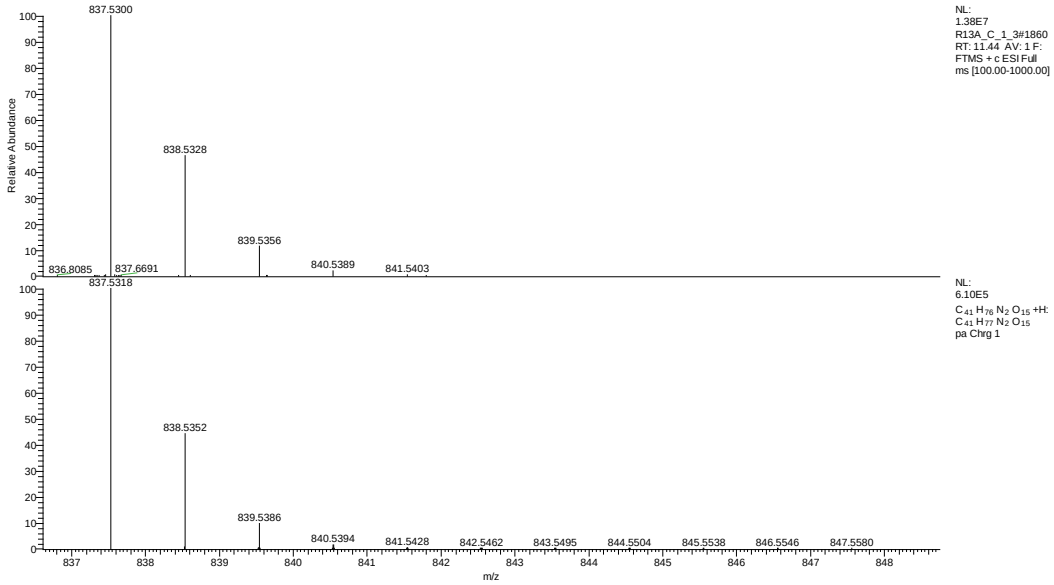


Figure 42. The first figure shows the MS for the roxithromycin that has been added to the problem sample R13_C_1_1, while the second presents the calculated MS, according to the bibliography.

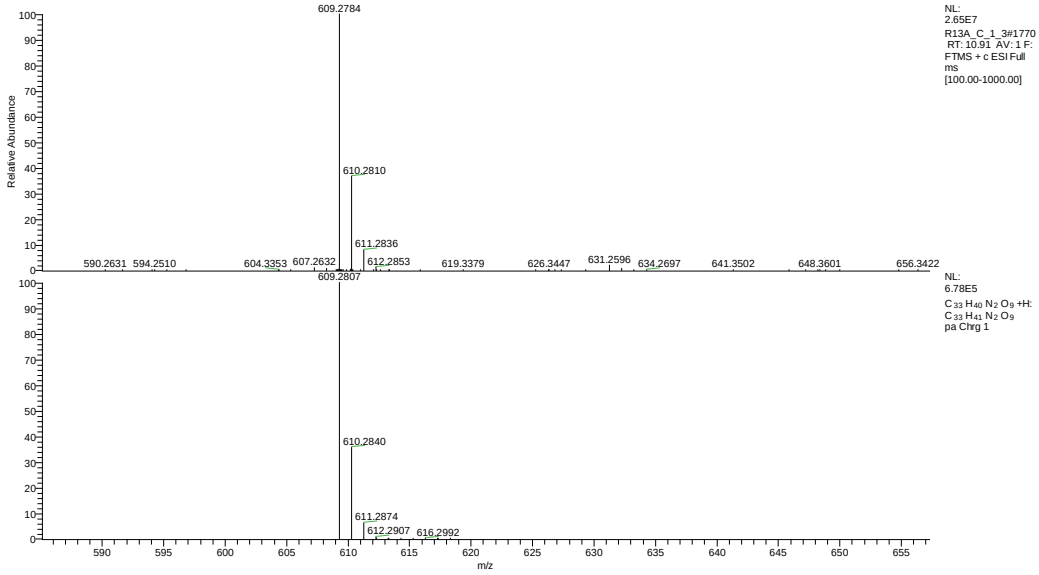


Figure 43. The first figure shows the MS for the reserpine that has been added to the problem sample R13_C_1_1, while the second presents the calculated MS, according to the bibliography.

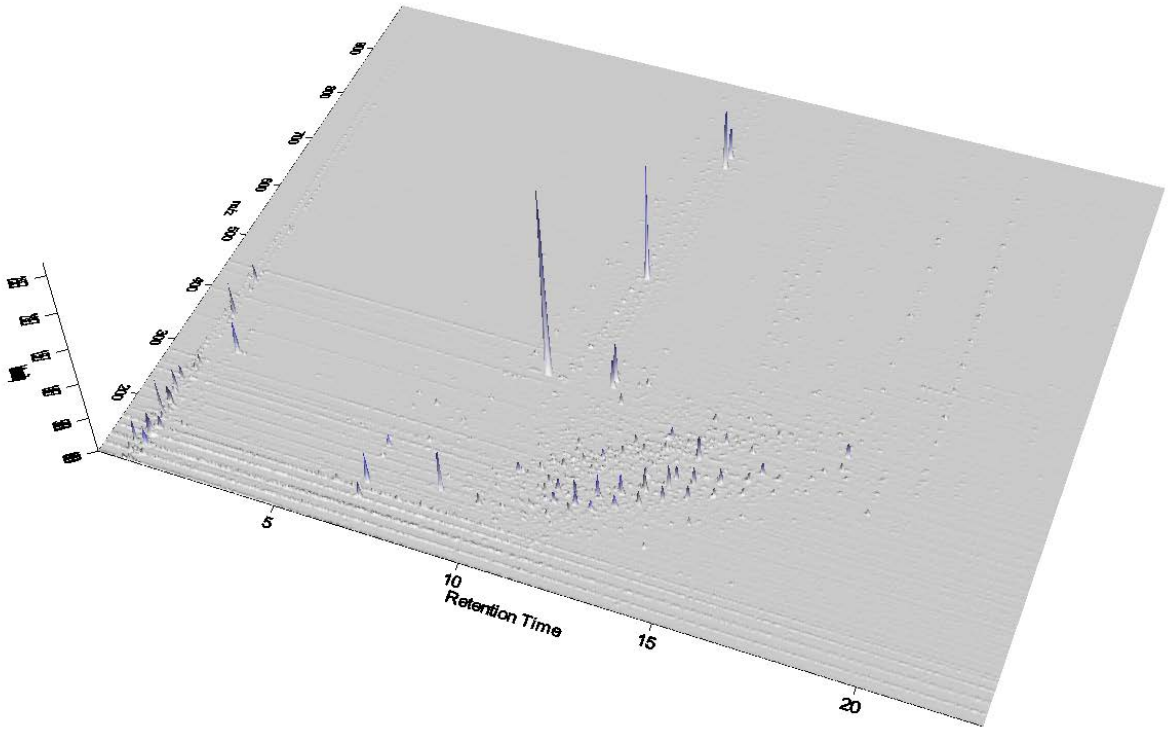


Figure 44. 3D visualizer can be used in order to detect hidden peaks behind another. The presented 3D TIC has been created after the baseline correction and the mass detection for the R13_C_1_1.

Table 8. Calculations of mass error in ppm and uncertainty in m/z for all the internal standards.

Sample	Calculated Mass			Observed Mass			Observed RT		
	Caffeine	Paracetamol	Resperidone	Caffeine	Paracetamol	Resperidone	Caffeine	Paracetamol	Resperidone
R13A_B_1_3	195.0877	152.0706	411.2191	195.087	152.0706	411.2172	7.62	6.27	9.67
R13B_B_1_3	195.0877	152.0706	411.2191	195.087	152.0706	411.2172	7.62	6.27	9.67
R14C_B_1_3	195.0877	152.0706	411.2191	195.0877	152.0701	411.2171	7.32	6.14	9.72
R13A_C_1_3	195.0877	152.0706	411.2191	195.0871	152.0701	411.2173	8.69	5.5	9.72
R13B_C_1_3	195.0877	152.0706	411.2191	195.0877	152.0701	411.2171	8.79	6.47	9.75
R14C_C_1_3	195.0877	152.0706	411.2191	195.087	152.0701	411.2191	7.84	5.89	9.13
R13A_H_2_3	195.0877	152.0706	411.2191	195.087	152.0701	411.2173	7.45	6.8	9.72
R13B_H_2_3	195.0877	152.0706	411.2191	195.0872	152.0703	411.2171	7.65	5.98	9.68
R14C_H_1_3	195.0877	152.0706	411.2191	195.0871	152.0701	411.2176	8.47	5.8	9.69
R13A_M1_1_3	195.0877	152.0706	411.2191	195.0871	152.0698	411.217	7.09	5.93	9.74
R13B_M1_1_3	195.0877	152.0706	411.2191	195.0871	152.0701	411.2172	7.72	6.38	9.73
R14C_M1_2_3	195.0877	152.0706	411.2191	195.0873	152.0701	411.2173	6.43	5.81	9.76
R13A_M2_1_3	195.0877	152.0706	411.2191	195.0871	152.0701	411.217	7.71	6.86	9.73
R13B_M2_1_3	195.0877	152.0706	411.2191	195.0872	152.0704	411.2172	6.42	5.67	9.7
R14C_M2_1_3	195.0877	152.0706	411.2191	195.0871	152.07	411.2173	6.09	5.78	9.73

Sample	Calculated Mass		Observed Mass		Observed RT	
	Roxithromycin	Reserpine	Roxithromycin	Reserpine	Roxithromycin	Reserpine
R13A_B_1_3	837.5318	609.2807	837.53	609.2783	11.43	10.87

Sample	Calculated Mass		Observed Mass		Observed RT	
	Roxithromycin	Reserpine	Roxithromycin	Reserpine	Roxithromycin	Reserpine
R13B_B_1_3	837.5318	609.2807	837.53	609.2783	11.43	10.87
R14C_B_1_3	837.5318	609.2807	837.53	609.2784	11.47	10.92
R13A_C_1_3	837.5318	609.2807	837.53	609.2784	11.44	10.91
R13B_C_1_3	837.5318	609.2807	837.5318	609.278	11.45	10.94
R14C_C_1_3	837.5318	609.2807	837.5398	609.2783	11.44	10.93
R13A_H_2_3	837.5318	609.2807	837.5301	609.2783	11.43	10.9
R13B_H_2_3	837.5318	609.2807	837.53	609.2781	11.41	10.88
R14C_H_1_3	837.5318	609.2807	837.53	609.2782	11.42	10.91
R13A_M1_1_3	837.5318	609.2807	837.5305	609.2781	11.46	10.93
R13B_M1_1_3	837.5318	609.2807	837.53	609.2783	11.46	10.93
R14C_M1_2_3	837.5318	609.2807	837.5298	609.2781	11.48	10.98
R13A_M2_1_3	837.5318	609.2807	837.53	609.2784	11.45	10.91
R13B_M2_1_3	837.5318	609.2807	837.5302	609.2786	11.43	10.9
R14C_M2_1_3	837.5318	609.2807	837.5303	609.2786	11.47	10.9

Table 9. Parameters for the chromatogram builder step during data preprocessing for the problem samples.

Sample	Mass error in ppm	Uncertainty in m/z	Min time span			Min height
			End	Start	Span	
R13A_B_1_3	4.6204	0.0019	22.36	22.33	0.03	1.10E5
R13B_B_1_3	4.6204	0.0019	21.31	21.29	0.02	1.30E5
R14C_B_1_3	4.8636	0.002	22.04	22.01	0.03	1.50E5
R13A_C_1_3	4.3772	0.0018	22.36	22.33	0.03	1.90E5

Sample	Mass error in ppm	Uncertainty in m/z	Min time span			Min height
			End	Start	Span	
R13B_C_1_3	4.8636	0.002	21.19	21.17	0.02	2.50E5
R14C_C_1_3	0	0	20.08	20.11	-0.03	1.50E5
R13A_H_2_3	4.3772	0.0018	21.01	20.99	0.02	1.80E5
R13B_H_2_3	4.8636	0.002	20.99	20.97	0.02	1.60E5
R14C_H_1_3	3.6477	0.0015	22.13	22.11	0.02	1.30E5
R13A_M1_1_3	5.1068	0.0021	21.54	21.52	0.02	1.10E5
R13B_M1_1_3	4.6204	0.0019	21.41	21.38	0.03	1.40E5
R14C_M1_2_3	4.3772	0.0018	22.68	22.66	0.02	1.60E5
R13A_M2_1_3	5.1068	0.0021	21.77	21.74	0.03	1.70E5
R13B_M2_1_3	4.6204	0.0019	21.98	21.96	0.02	2.00E5
R14C_M2_1_3	4.3772	0.0018	20.98	21.01	-0.03	1.50E5

Deconvolution

For chromatogram deconvolution the local minimum algorithm was the most suitable, due to the decreased level of noise after the baseline correction. All the selected parameters are presented in the following tables.

Table 10. Comparison of the number of peaks before and after deconvolution.

Sample	Number of detected peaks before deconvolution	Number of detected peaks after deconvolution
R13A_B_1_3	423	786
R13B_B_1_3	563	1037
R14C_B_1_3	482	741
R13A_C_1_3	414	599
R13B_C_1_3	320	487
R14C_C_1_3	521	850
R13A_H_2_3	420	724
R13B_H_2_3	484	855
R14C_H_1_3	570	1055
R13A_M1_1_3	637	1235
R13B_M1_1_3	503	792
R14C_M1_2_3	452	789
R13A_M2_1_3	418	647
R13B_M2_1_3	384	634
R14C_M2_1_3	509	811

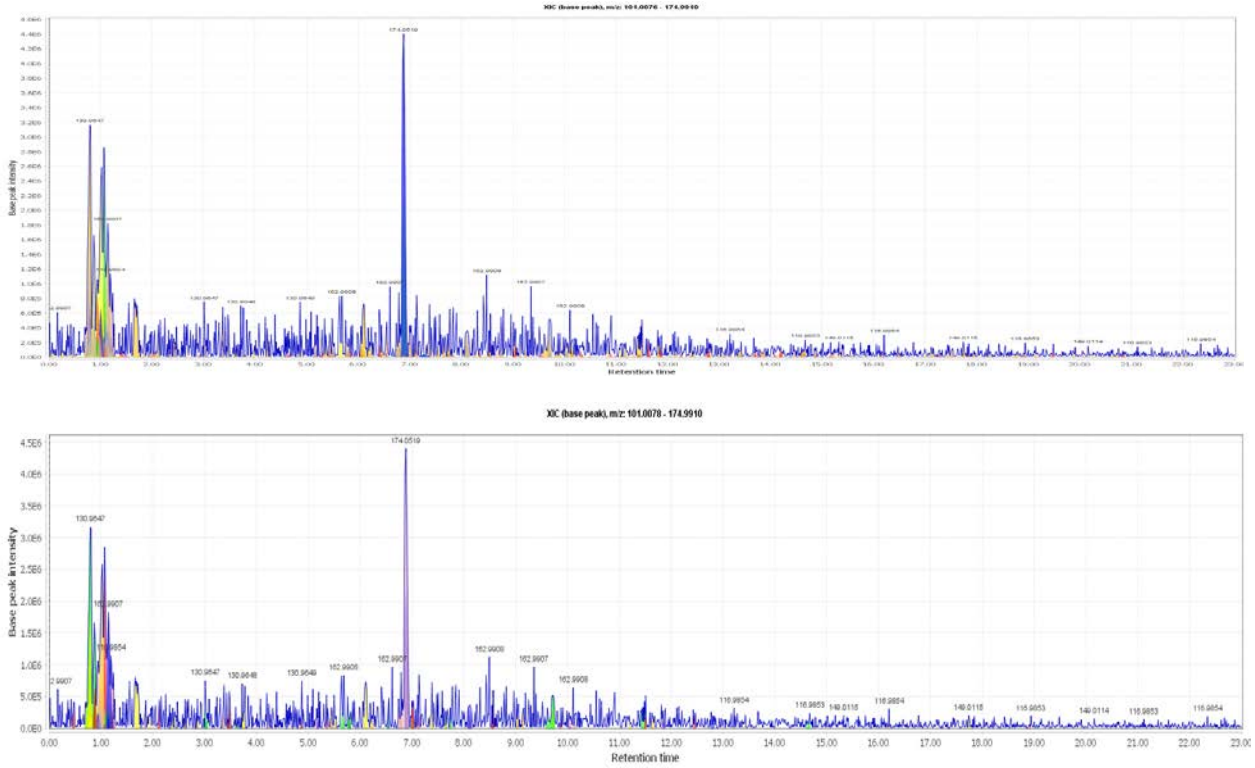


Figure 45. The number of the detected peaks has increased after decovolution for all the problem samples. The first XIC has been created from the peak list before deconvolution, while the second after deconvolution R13_C_1_1 for m/z : 101.0078 -174.9910.

Table 11. Parameters for the chromatogram deconvolution step during data preprocessing for the problem samples.

Sample	Min Absolute height	Max Absolute Height	Minimum Relative height (%)	Search minimum in RT range (min)					
				End	Start	Span	Top	Edge	top/edge
R13A_B_1_3	1.10E5	3.70E6	2.97	14.64	14.56	0.08	3.60E3	2.80E7	0.0001
R13B_B_1_3	1.30E5	2.80E7	0.46	0.96	0.89	0.07	3.60E3	2.80E7	0.0001
R14C_B_1_3	1.50E5	2.80E7	0.54	0.93	0.86	0.07	5.60E3	2.80E7	0.0002
R13A_C_1_3	1.90E5	2.70E7	0.70	1.13	1.07	0.06	2.60E4	2.70E7	0.0010
R13B_C_1_3	2.50E5	2.70E7	0.93	0.94	0.87	0.07	6.70E3	2.70E7	0.0002
R14C_C_1_3	1.50E5	3.00E7	0.50	0.93	0.85	0.08	8.60E3	3.00E7	0.0003
R13A_H_2_3	1.80E5	2.70E7	0.67	0.93	0.87	0.06	5.50E3	2.70E7	0.0002
R13B_H_2_3	1.60E5	2.80E7	0.89	17.94	17.83	0.11	6.20E3	2.80E7	0.0002
R14C_H_1_3	1.30E5	3.10E7	0.42	0.90	0.83	0.07	5.90E3	3.10E7	0.0002
R13A_M1_1_3	1.10E5	2.60E7	0.42	0.91	0.86	0.05	4.30E3	2.60E7	0.0002
R13B_M1_1_3	1.40E5	3.20E7	0.44	0.9	0.84	0.06	4.30E3	3.20E7	0.0001
R14C_M1_2_3	1.60E5	2.80E7	0.57	0.97	0.88	0.09	1.70E4	2.80E7	0.0006
R13A_M2_1_3	1.70E5	2.50E7	0.68	0.91	0.84	0.07	8.40E3	2.50E7	0.0003
R13B_M2_1_3	2.00E5	2.90E7	0.69	0.9	0.83	0.07	2.10E4	2.90E7	0.0007
R14C_M2_1_3	1.50E5	3.10E7	0.48	0.89	0.82	0.07	6.80E3	3.10E7	0.0002

Deisotoping

Isotopic peaks grouper was used as the deisotoping method.

Table 12. Parameters for the chromatogram deisotoping step during data preprocessing for the problem samples.

Sample	Mass error in ppm	Uncertainty in m/z	Min time span		
			End	Start	Span
R13A_B_1_3	4.6204	0.0019	7.78	7.73	0.05
R13B_B_1_3	4.6204	0.0019	10.84	10.81	0.03
R14C_B_1_3	4.8636	0.002	0.25	0.2	0.05
R13A_C_1_3	4.3772	0.0018	18.94	18.9	0.04
R13B_C_1_3	4.8636	0.002	0.06	0.04	0.02
R14C_C_1_3	4.3772	0.0018	5.5	5.44	0.06
R13A_H_2_3	4.3772	0.0018	10.01	9.98	0.03
R13B_H_2_3	4.8636	0.002	13.94	13.91	0.03
R14C_H_1_3	3.6477	0.0015	12.04	12.01	0.03
R13A_M1_1_3	5.1068	0.0021	18.82	18.8	0.02
R13B_M1_1_3	4.6204	0.0019	11.12	11.08	0.04
R14C_M1_2_3	4.3772	0.0018	2.23	2.2	0.03
R13A_M2_1_3	5.1068	0.0021	4.43	4.39	0.04
R13B_M2_1_3	4.6204	0.0019	11.29	11.27	0.02
R14C_M2_1_3	4.3772	0.0018	10.86	10.83	0.03

Alignment

Join aligner algorithm was used to join the rows that are referring to the same metabolite but due to different noise levels or batch effects they are characterized by different m/z and RT values. The m/z tolerance was set at 0.0006 or 1.4591ppm, and the absolute value of the retention time was 0.63 min. Risperidone was used as reference to evaluate the performance of the chosen alignment algorithm.

Gap-Filling

The missing values were minimized after gap-filling using the peak finder, and the values were the same as for alignment.

Instrument Variability and Overall Process Variability

The created list contained 1408 detected metabolites but were 11 metabolites that had been detected only in one out of 15 samples (missing values = 93.33%). To obtain consistent variables these metabolites were excluded from the further bioinformatics analysis. The resulting 2D matrix contained 137 metabolites.

Instrument variability was determined by calculating the median RSD of RT, peak area and mass accuracy for the I.S of the 4 treatment conditions, while overall process variability was determined by calculating the median RSD of RT, peak area and mass accuracy for all the endogenous metabolites of the 4 treatment conditions.

Table 13. Instrument' s and Overall process variability results.

Instrument Variability based on peaks area						
I.S	RSD_B %	RSD_C %	RSD_H %	RSD_M1 %	RSD_M2 %	RSD_ALL SAMPLES %
Paracetamol	86.44	7.71	8.92	8.06	11.74	0.29
Caffeine	86.60	4.28	1.53	0.76	10.56	0.28
Resperidone	86.60	3.11	3.92	5.68	5.96	0.28
Reserpine	86.86	7.63	1.08	3.36	8.58	0.28
Roxythromycin	86.66	3.78	3.54	4.45	9.19	0.28
Median	86.60	4.28	3.54	4.45	9.19	0.28
Instrument Variability based on peaks mass						
I.S	RSD_B %	RSD_C %	RSD_H %	RSD_M1 %	RSD_M2 %	RSD_ALL SAMPLES %
Paracetamol	5.65E-05	1.26E-05	1.26E-05	8.69E-06	5.02E-06	3.26E-07
Caffeine	3.03E-05	1.19E-05	0	1.35E-05	2.26E-05	2.01E-07
Resperidone	1.39E-05	2.27E-05	7.72E-06	7.42E-06	2.04E-05	1.8E-07
Reserpine	86.60	3.51E-05	5.58E-05	1.9E-05	3.06E-05	0.28
Roxythromycin	86.60	2.76E-05	2.55E-05	4.35E-05	2.48E-05	0.28
Median	5.6539E-05	2.26724E-05	1.26259E-05	1.35473E-05	2.25788E-05	3.25619E-07
Instrument Variability based on peaks RT						
I.S	RSD_B %	RSD_C %	RSD_H %	RSD_M1 %	RSD_M2 %	RSD_ALL SAMPLES %
Paracetamol	0.41	0.20	0.37	0.63	0.57	0.005
Caffeine	2.47	0.06	0.19	0.37	0.31	0.01
Resperidone	2.22	0.06	0.09	0.20	0.39	0.01

Instrument Variability based on peaks area						
I.S	RSD_B %	RSD_C %	RSD_H %	RSD_M1 %	RSD_M2 %	RSD_ALL SAMPLES %
Reserpine	86.60	0.16	0.12	0.15	0.13	0.28
Roxythromycin	86.60	0.05	0.13	0.09	0.21	0.28
Median	2.47	0.06	0.13	0.20	0.31	0.01
Overall Process Variability (considering all metabolites)						
	RSD_B %	Median RSD_C %	MedianRSD_H %	Median RSD_M1 %	Median RSD_M2 %	Median RSD_ALL SAMPLES %
area	28.66	20.93	22.66	22.37	20.73	0.36
mass	2.24751E-05	2.01272E-05	1.7101E-05	1.87235E-05	3.01519E-05	2.99615E-07
RT	2.24	0.92	1.10	1.23	1.23	0.03

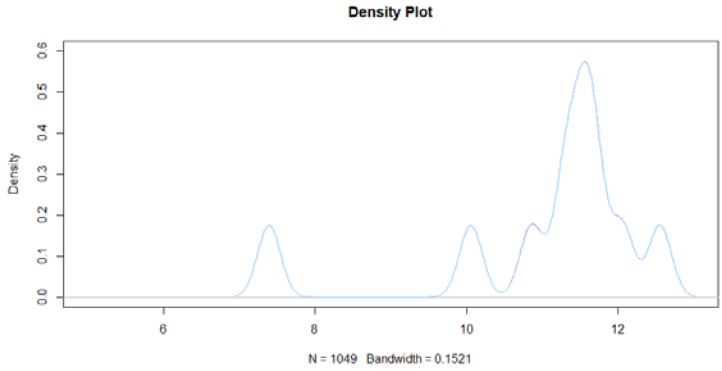
Metabolites Annotation

The internal standards [paracetamol (m/z: 152.0706, rt: 6.27), caffeine (m/z: 195.0871 rt: 7.62), resperidone (m/z: 411.2172, rt: 9.67), reserpine (m/z: 609.2783, rt: 10.87), roxythromycin (m/z: 837.52300, rt: 11.43)]] were excluded from further bioinformatics analysis.

MetaboSearch Tool v.1.2 has been used for metabolites annotation. The used customized databases were the HMDB, Metlin, and Lipid Maps, and the MW tolerance was 5 ppm. The final number of unique identified metabolites was 134/1391.

Normalization

A



B

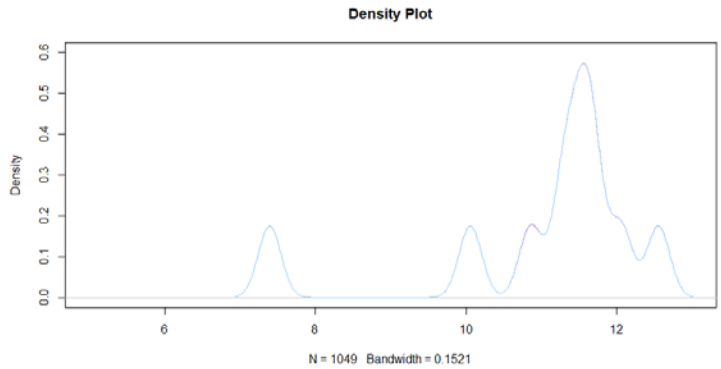


Figure 46. The density plot A illustrates the log transformed raw data from the analysis in the negative ionization mode before normalization, while the density plot B shows the samples distributions after quantile normalization.

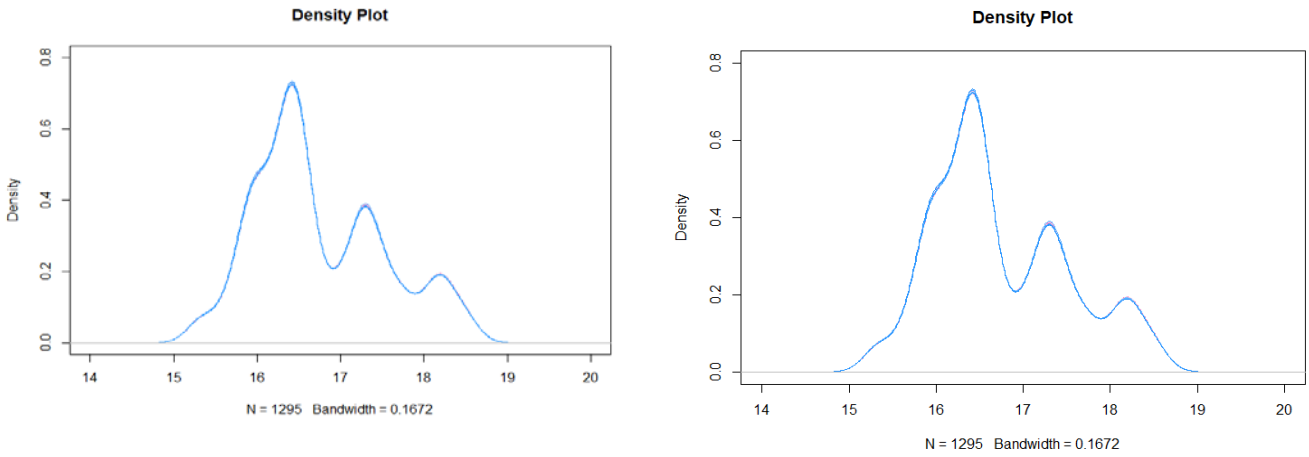


Figure 47. The density plot on the left illustrates the log transformed raw data from the analysis in the positive ionization mode before normalization, while the density plot on the right shows the samples distributions after quantile normalization.

The value of the bandwidth is extremely low due to the fact that the dataset is characterized by low standard deviation.

The formula for the calculation of standard bandwidth is given below:

$$bw = 0.9n^{-1/5} \cdot \min(sd(x), \frac{IQR(x)}{1.34}$$

where n is the sample size, sd() denotes the sample standard deviation and IQR() is the interquartile range function, i.e. it calculates the difference between the upper and the lower quartile: $q_{0.75}-q_{0.25.}$

According to the density plot the samples have common distributions, that means the distributions are not being driven by some technical variables.

Annex 3

K-means clustering analysis

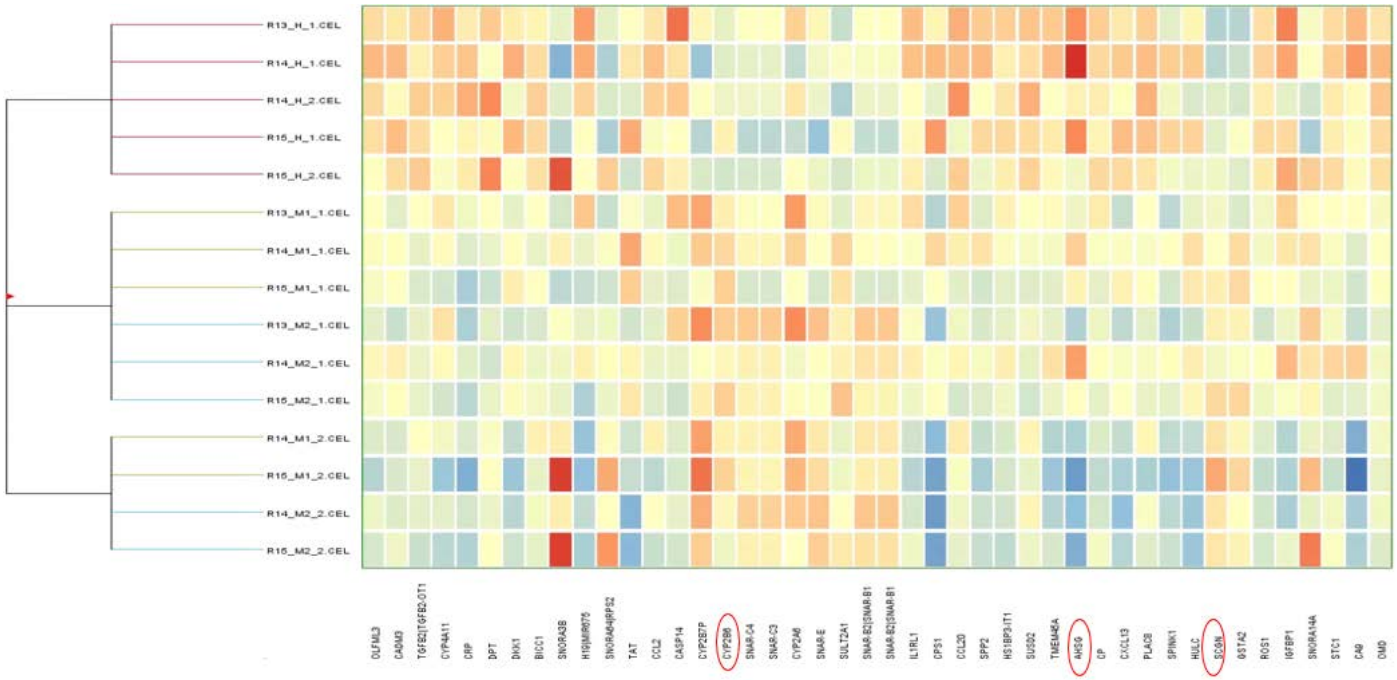


Figure 48. K-means Clustering analysis of significant differential expressed genes.

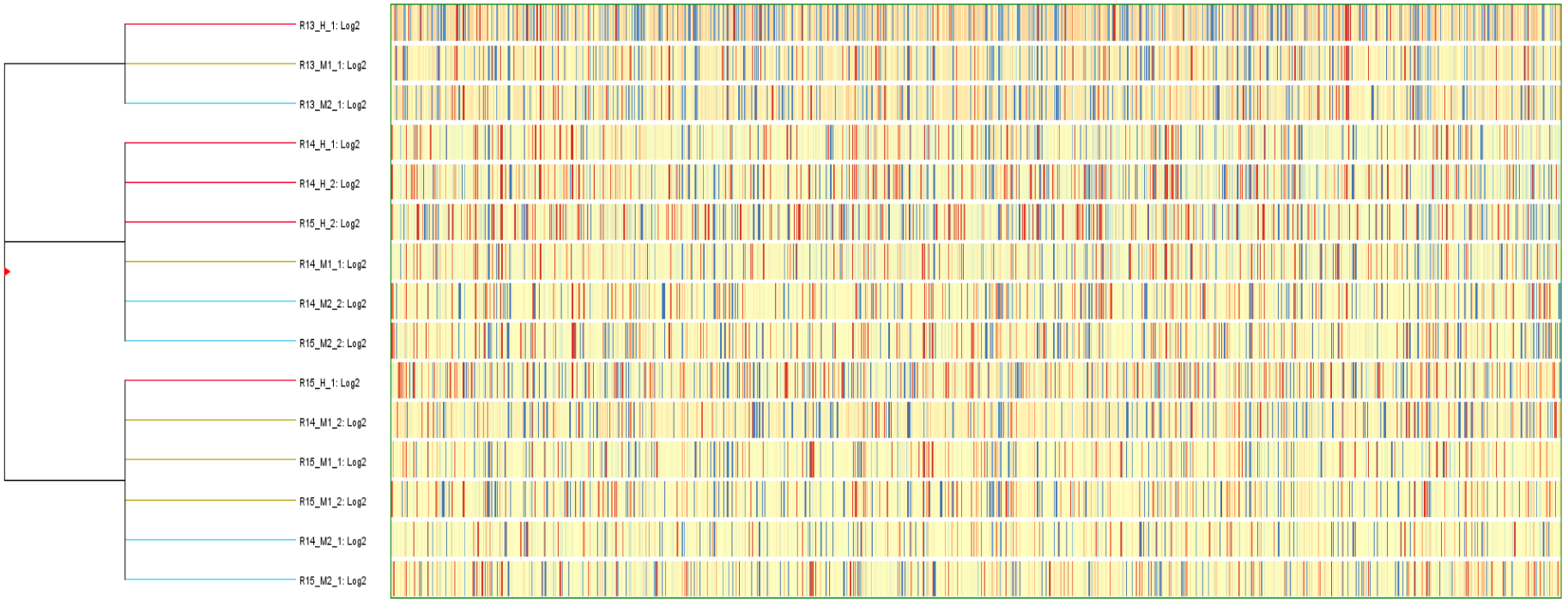


Figure 49. K-means Clustering analysis of significant differential expressed proteins.

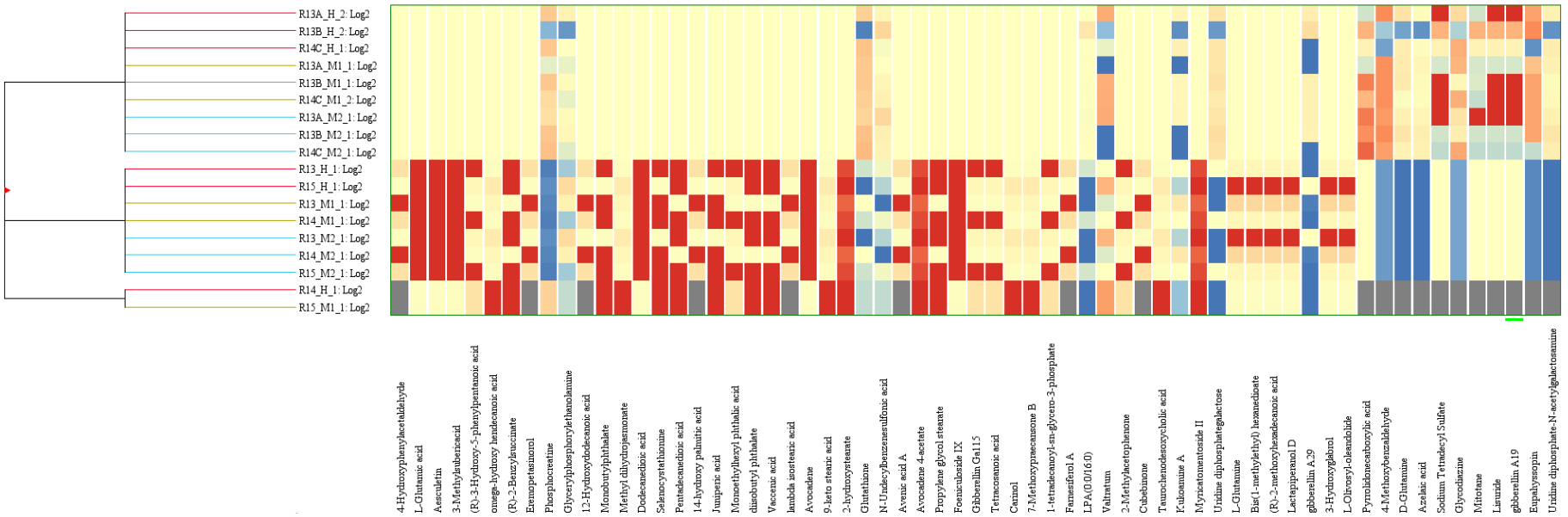


Figure 50. K-means Clustering analysis of significant differential expressed metabolites.

Self-Organizing Map (SOM)

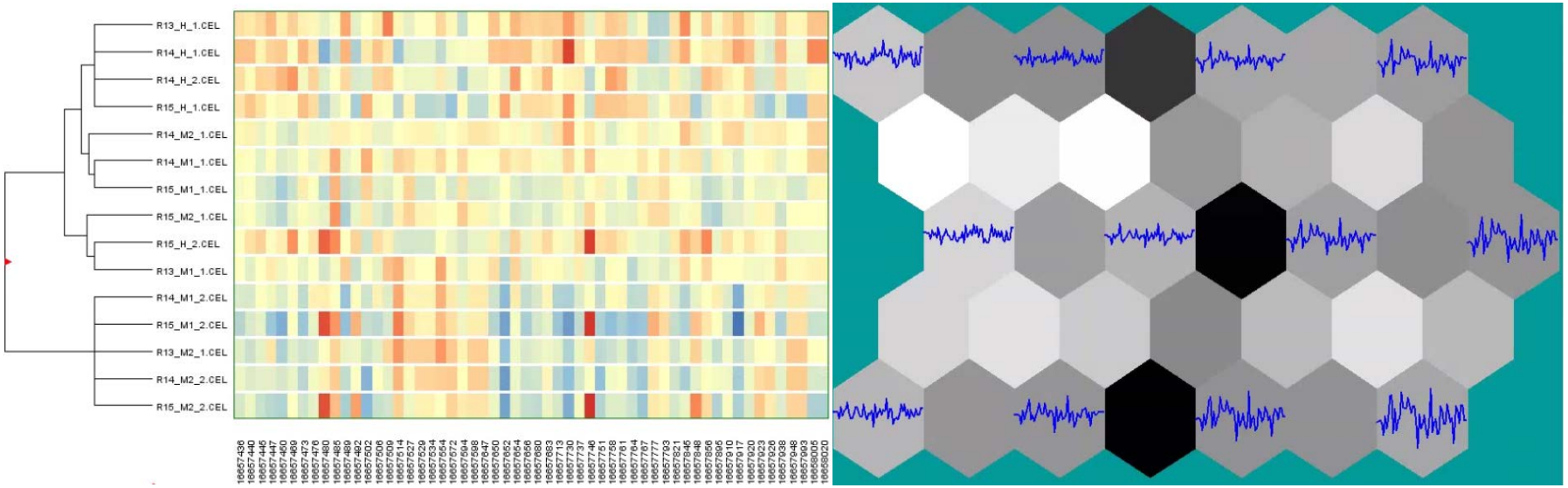


Figure 51. Clustering analysis of significant differential expressed genes using Self - Organizing Map. The results are presented in a cluster tree on the left and in a UMatrix diagram on the right.

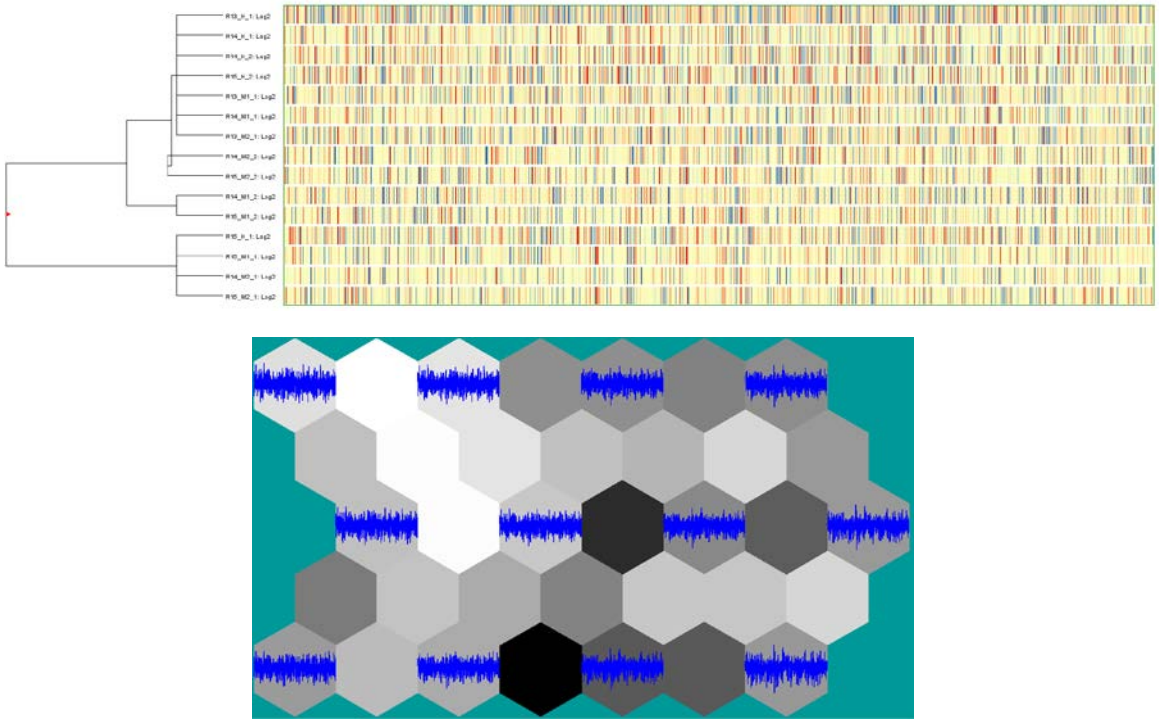


Figure 52. Clustering analysis of significant differential expressed proteins using Self - Organizing Map. The results are presented in a cluster tree on the left and in a UMatrix diagram on the right.

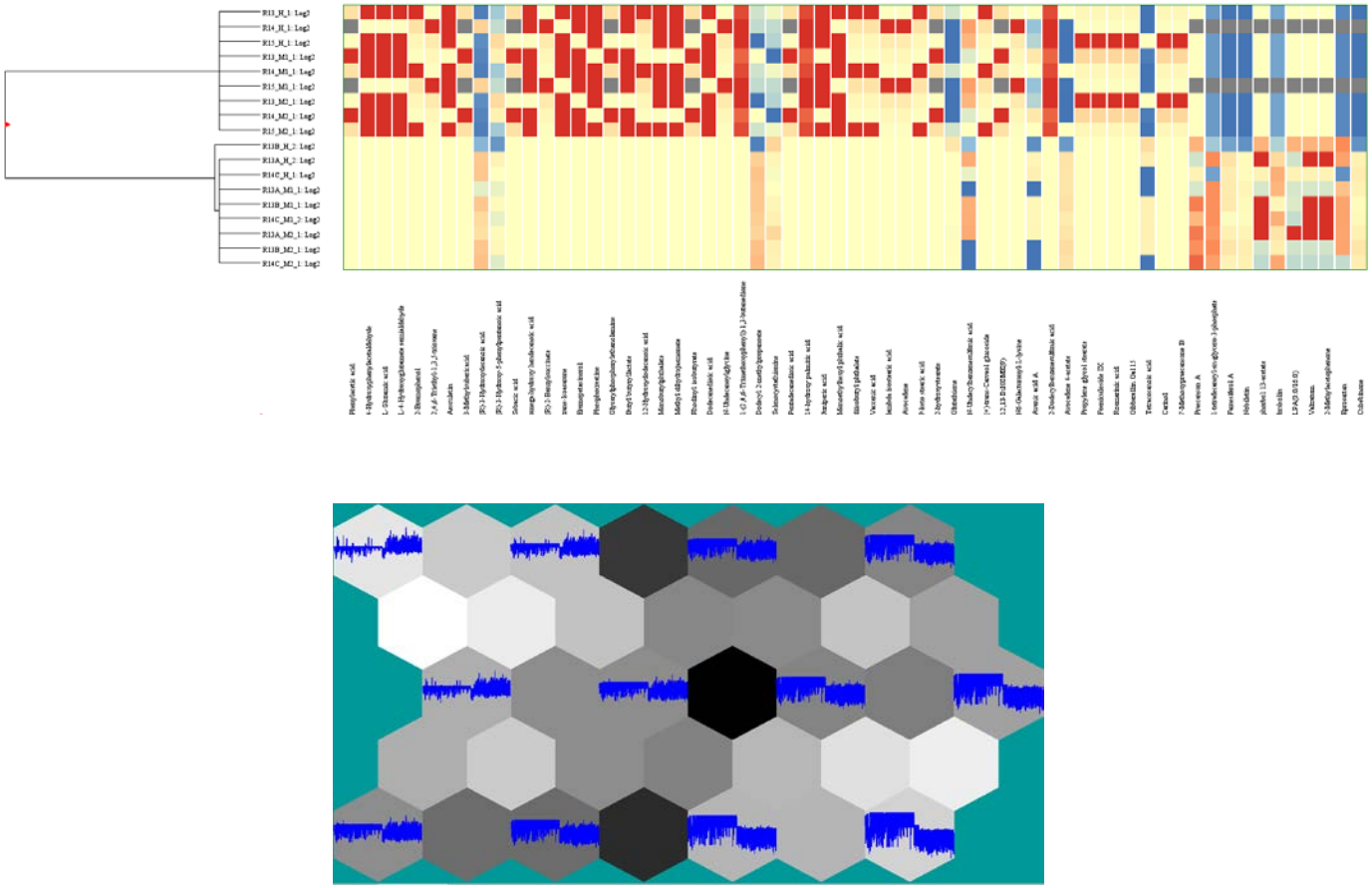


Figure 53. Clustering analysis of significant differential expressed metabolites using Self - Organizing Map.