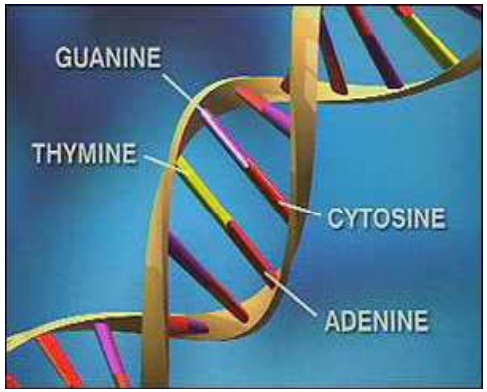
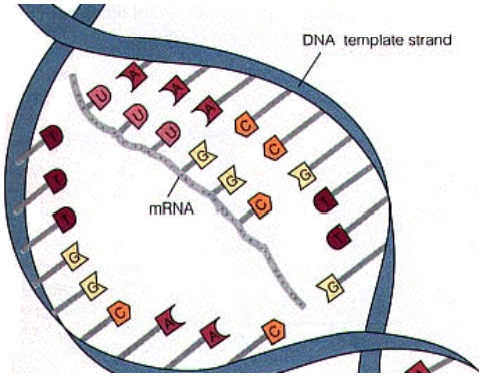
UNIVERSITY OF BIRMINGHAM

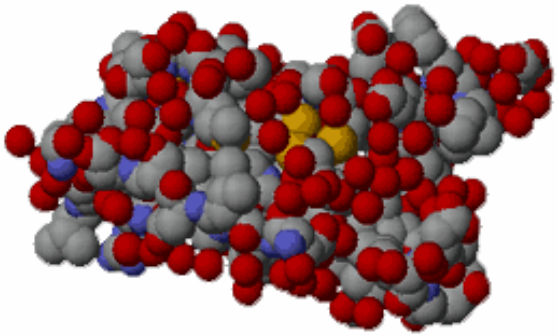
Non-targeted metabolomics in environmental toxicology: workflows, challenges and routes through the maze

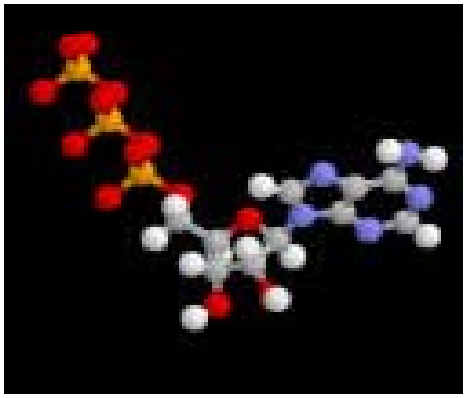
Non-Target-2016, Ascona, Switzerland
30th May 2016

Mark Viant, University of Birmingham, UK



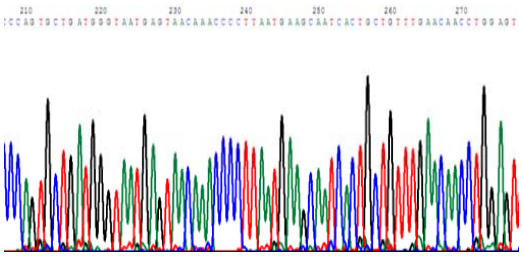


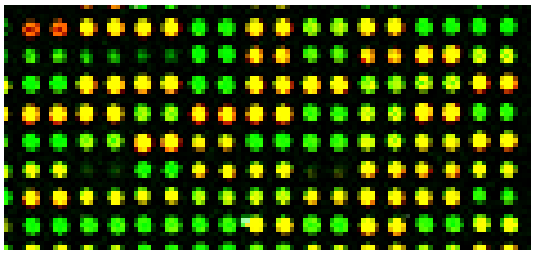


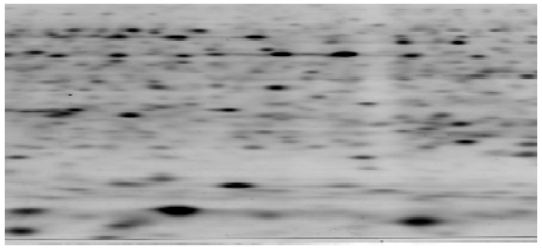


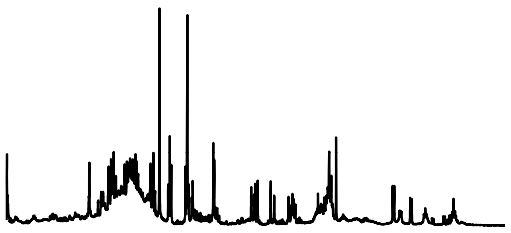
DNA → (Transcription) → mRNA → (Translation) → Proteins ⇄ (Enzyme activity) Metabolites

Genomics	Transcriptomics	Proteomics	Metabolomics

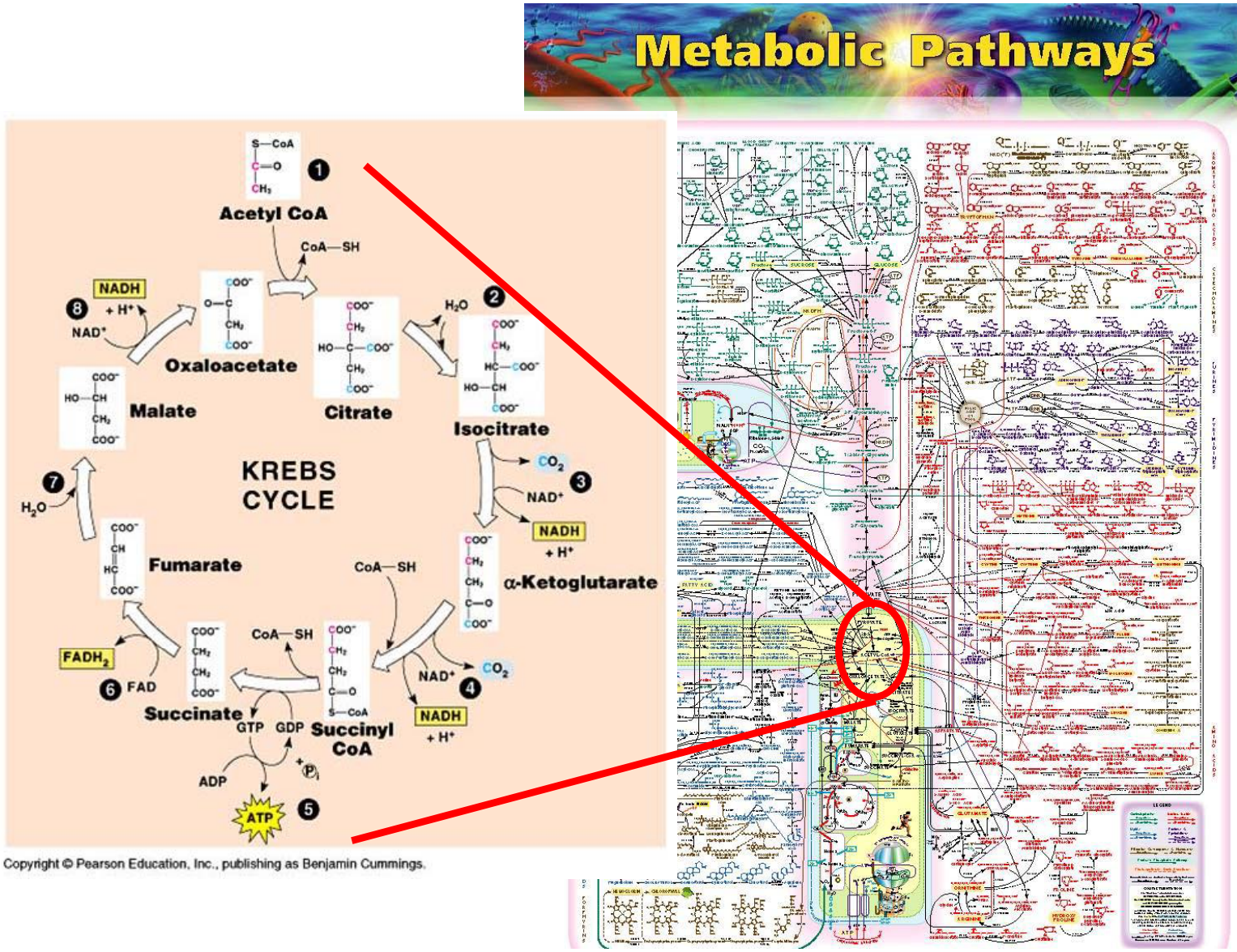








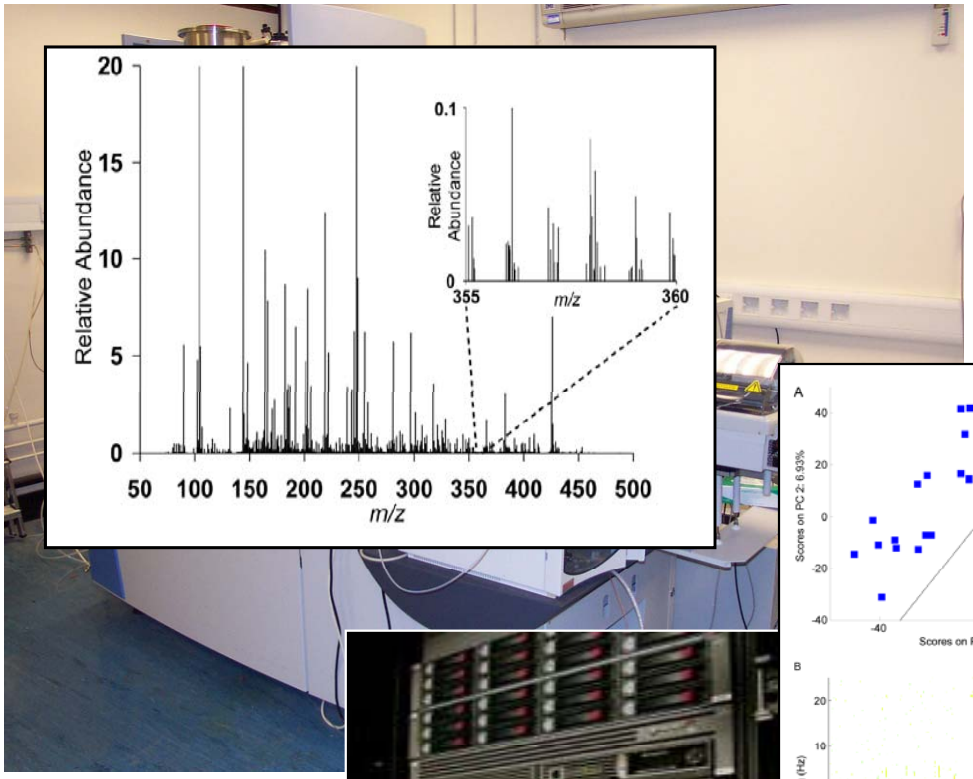
What makes up an organism's metabolome?



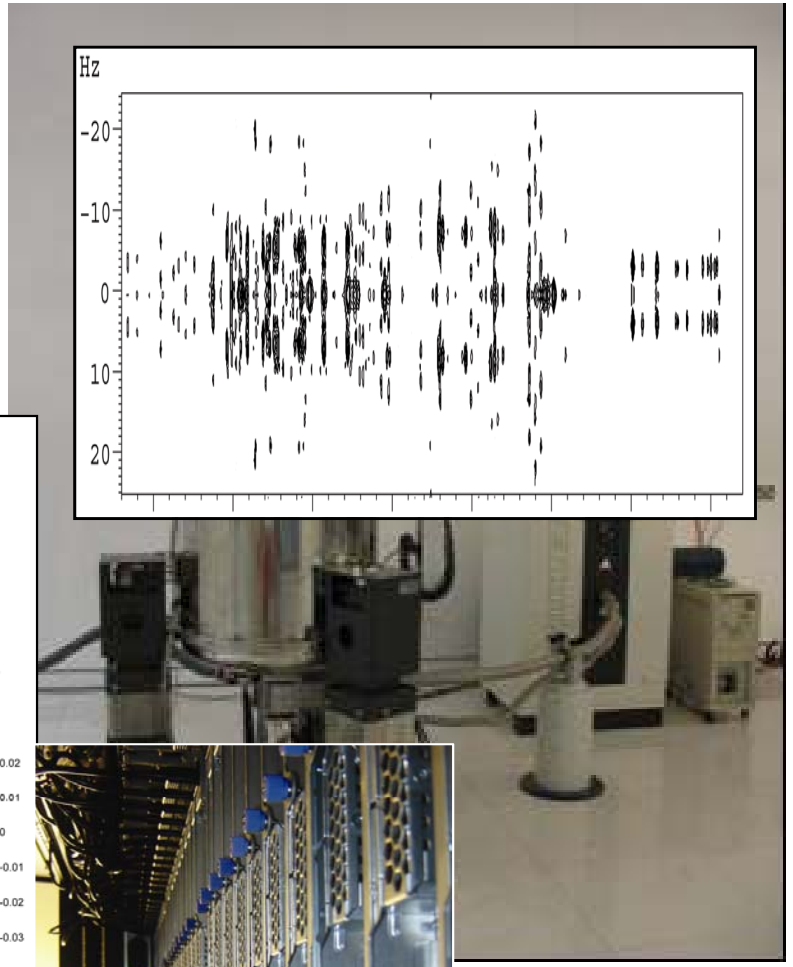
- Amino acids
- Carbohydrates
- Lipids
- Steroids
- Secondary metabolites…
- Estimated to be >10,000 metabolites

Measuring and interrogating the metabolome

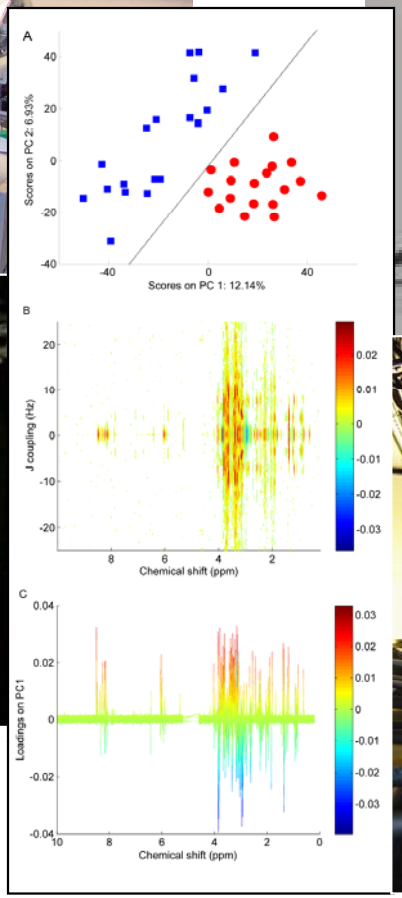
Mass spectrometry



NMR spectroscopy



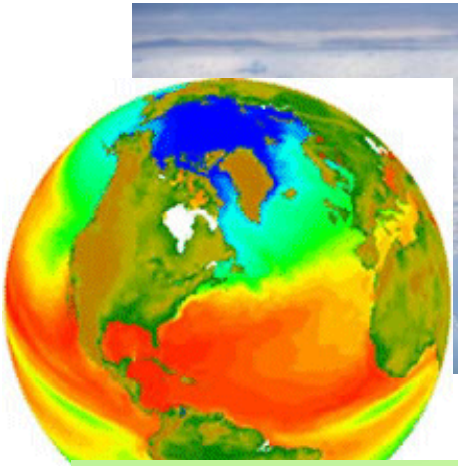
Computational biology

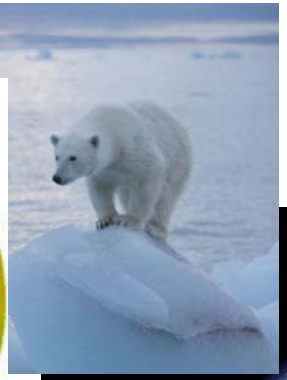


Overview

1. Introduction to environmental metabolomics
2. Workflows
   - Direct infusion mass spectrometry (standardised)
   - Data processing (relatively standardised)
   - Metabolite annotation and identification (not standardised)
3. Examples
   - Endogenous metabolites
   - Xenobiotics within organisms
4. Where next?

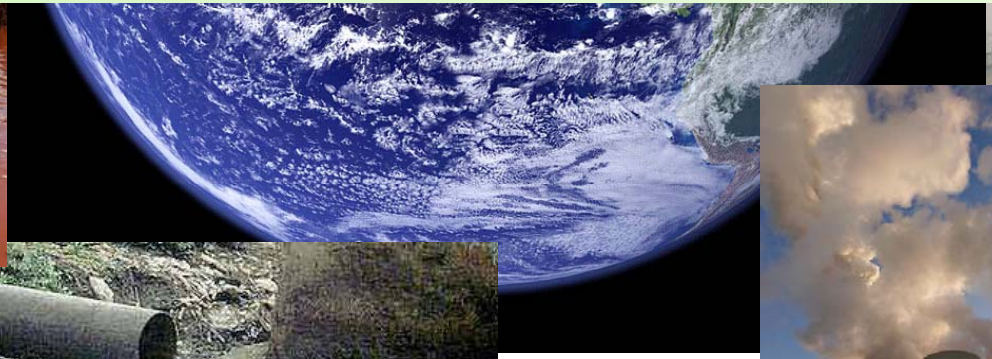
Metabolomics and Environmental stress

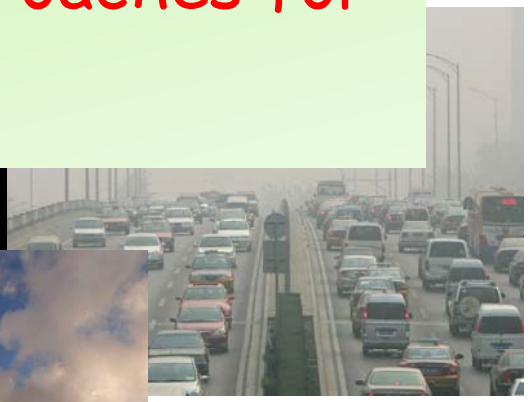




What are the effects of these stressors on living organisms?

Can we develop novel information-rich approaches for environmental regulation?





Water pollution

Air pollution

Environmental monitoring

- Environmental quality assessed (traditionally) by measuring pollutant levels
- European Union has list of ca. 40 priority pollutants

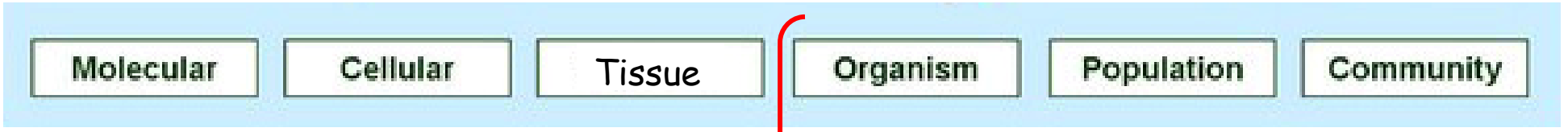
EU WFD (Water Framework Directive, 2005)

- shifted environmental quality assessments towards integrative biological effects monitoring
- currently based on assessment of the composition & abundance of fauna and flora

Need for high throughput, mechanism-based testing strategies to determine environmental health

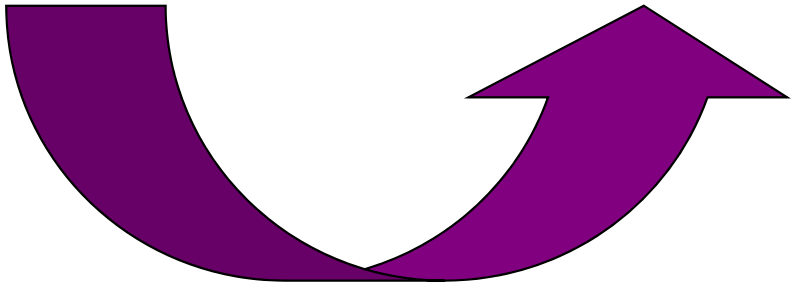
What can molecular biomarkers offer?

Ecological relevance



- Lethality
- Growth
- Reproduction

Basic scientist



Environmental regulator interested in impacts on Darwinian fitness

Molecular markers:

1. Sub-lethal effects
2. Mechanistic understanding
3. Diagnostic fingerprints
4. Early warning indicators

Molecular markers must be strongly linked to higher level responses

Complexity of what we are trying to measure!

Uptake, metabolism & effect of xenobiotics on organism health

Chemical signalling (exometabolome)

Endogenous metabolites (>10,000 forming endometabolome)

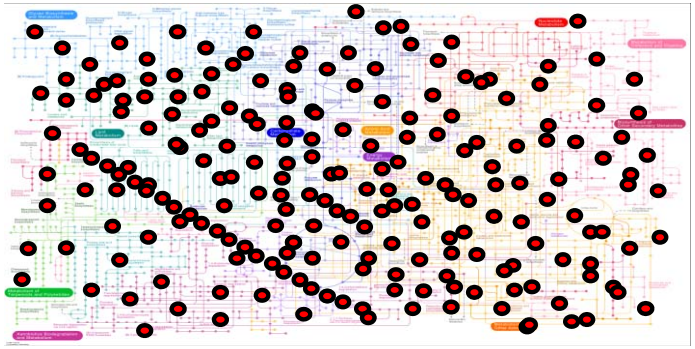
- Biodiversity: 1000's of species, 1000's of metabolomes
- Microbiomes too!

Overview

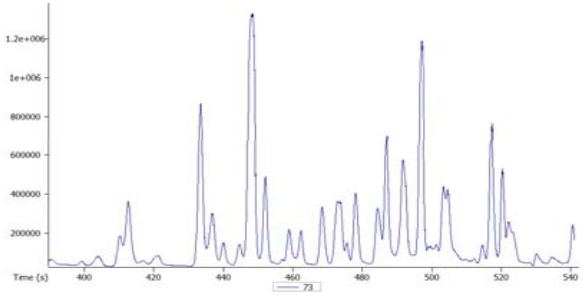
1. Introduction to environmental metabolomics
2. Workflows
   - Direct infusion mass spectrometry (standardised)
   - Data processing (relatively standardised)
   - Metabolite annotation and identification (not standardised)
3. Examples
   - Endogenous metabolites
   - Xenobiotics within organisms
4. Where next?

Non-targeted *vs.* targeted metabolomic studies

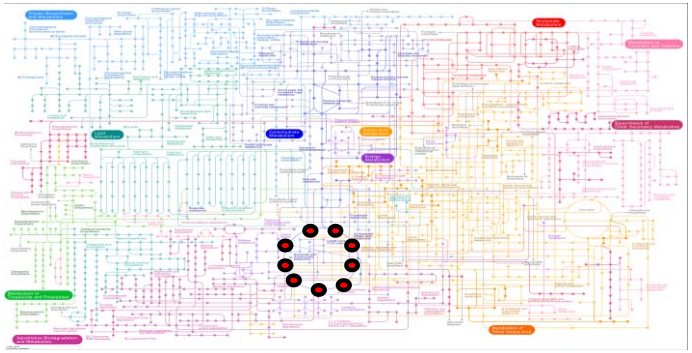
METABOLIC PROFILING or NON-TARGETED ANALYSIS



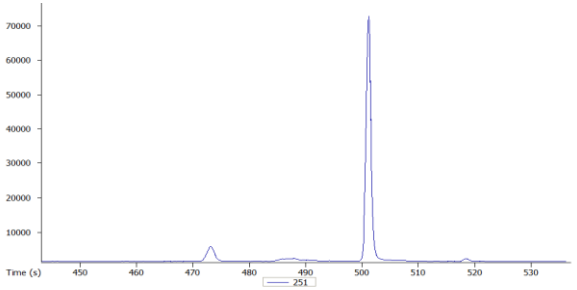
- (Semi)-quantitative detection of a wide range of metabolites
- NMR or GC-MS or LC-MS
- Data acquisition without *a priori* knowledge of biologically interesting metabolites
- Metabolite identification requires post data acquisition
- Discovery/hypothesis generating



TARGETED ANALYSIS



- Quantification of a smaller number of (related) metabolites for
  - generally less than 20
- LC-MS/MS
- Metabolite identity already known
  - no further metabolite identification required
- Hypothesis testing



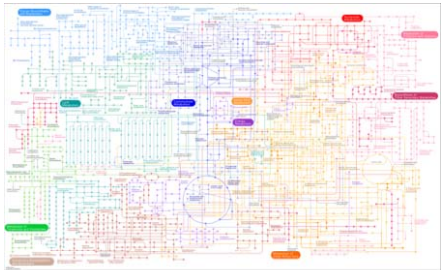
Generic workflow

EXPERIMENTAL DESIGN

BIOLOGICAL EXPERIMENT

ANALYTICAL EXPERIMENT

DATA INTEGRATION, ANALYSIS AND METABOLITE IDENTIFICATION



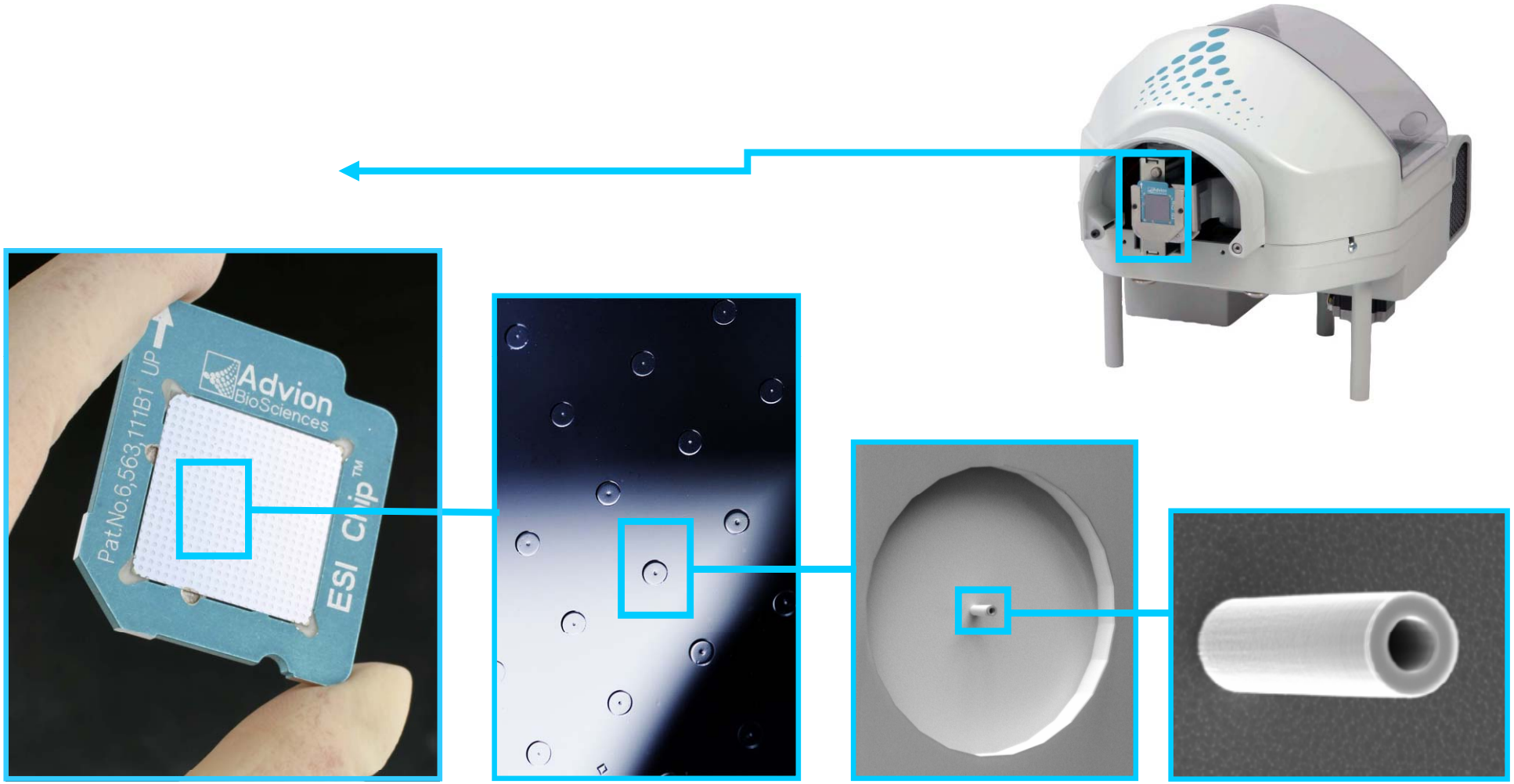
BIOLOGICAL INTERPRETATION

EUREKA!!!!!!!

Why Direct Infusion Mass Spectrometry (DIMS)?

- Non-targeted high-throughput screening approach
- No selection bias due to LC or GC column, yet has high analytical sensitivity
- Potentially low(er) cost than LC-MS
  - Higher sample throughput (few min / sample)
  - Potential savings on consumables
- Extremely high reproducibility of *m/z* data (ppm errors; very small compared to those of LC retention time data)
- But only measures *m/z* (putative annotation of compounds only)
- Potential for ion suppression (but much less of an issue with nano-electrospray ionisation (nESI) than with normal flow rate ESI)

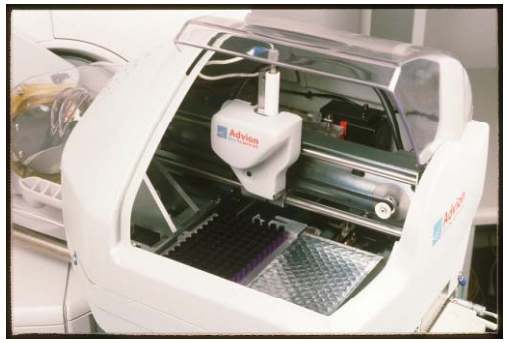
Sample introduction using Triversa chip-based nanoelectrospray system



- Fully automated
- No sample carry-over
- Stable nanoelectrospray (RSDs of few %)

Advion BioSciences

FT-ICR, Orbitrap and Q Exactive mass spectrometers



Triversa nanoelectrospray ion source

Orbitrap spectrometer

Linear ion trap (LTQ) mass spectrometer

Thermo ELECTRON CORPORATION

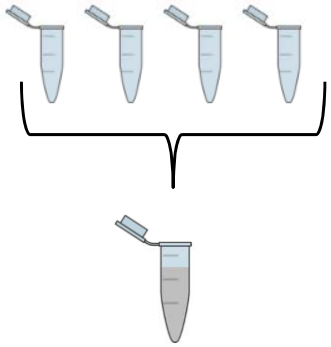
LTQ Orbitrap

DIMS pipeline

DIMS Experiment

Samples

QCs



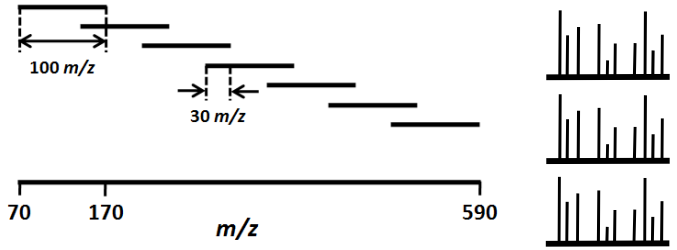
1)

DI nESI high resolution MS

FT-ICR LTQ Orbitrap
LTQ FT Ultra Velos

2)

SIM-Stitching



3)

Statistics

Wide array of methods (see examples)

8)

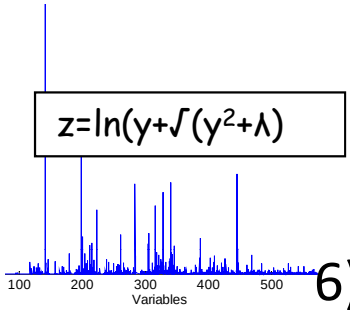
Data processing and quality assessment

Quality assessment

7)

Transformation

$$z=\ln(y+\sqrt{(y^2+\lambda)})$$



6)

Normalisation and missing values

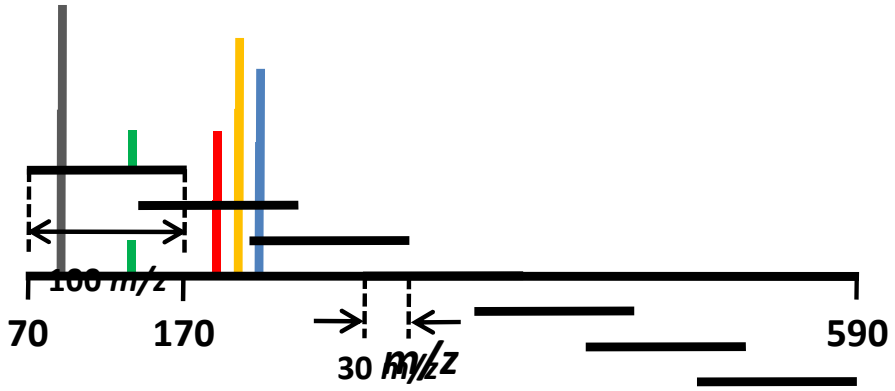
100	200	500	600	
300		800	700	700
500	600	500	600	700
	200	600		800
400	500	800	600	400

5)

Multi-step signal filtering

4)

SIM-Stitching



- Collection of multiple adjacent SIM windows that are stitched together
- An optimized strategy for wide-scan DIMS that increases dynamic range but maintains high mass accuracy (ca. 3000 *m/z* measurements, root mean square mass error of 0.16 ppm and max abs mass error of 0.29 ppm)
- Increases metabolome coverage
- Now applied in other research fields, e.g. petroleomics and organic chemistry *

Southam et al. Anal Chem. 79:4595-4602 (2007)
Weber et al. Anal Chem. 83:3737-43 (2011)

* Chainet et al. *Anal. Chem.* 2012, 84 (9):3998–4005
* Kujawinski et al. *Annual Reviews of Marine Science.* 2011, 3: 567-599

Optimized SIM-stitching parameters

Table 1. Parameters for DI SIM-stitching implemented on a LTQ FT Ultra (FT-ICR)

parameter	LTQ FT Ultra
AGC target	1×10^6
SIM scan range	m/z 100[a]
overlap of SIM scans	m/z 30 (m/z 15 removed from each end)
time for SIM scan (transients)	15 s (10)
total range	m/z 70–590
total number of overlapping SIM scans	7
total acquisition time per sample[b]	2 min 15 s

[a] Scan mode: wide SIM. [b] Including a 30 s start delay of dummy scans.

Southam et al. Anal Chem. 79:4595-4602 (2007)
Weber et al. Anal Chem. 83:3737-43 (2011)
Southam et al. Nature Protocols (under review)

Overview

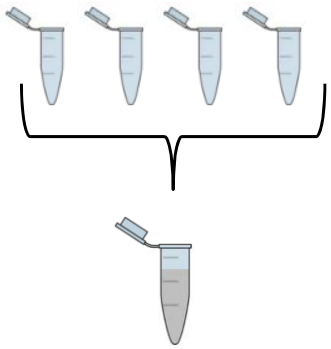
1. Introduction to environmental metabolomics
2. Workflows
   - Direct infusion mass spectrometry (standardised)
   - Data processing (relatively standardised)
   - Metabolite annotation and identification (not standardised)
3. Examples
   - Endogenous metabolites
   - Xenobiotics within organisms
4. Where next?

Data processing and quality assessment

DIMS Experiment

Samples

QCs



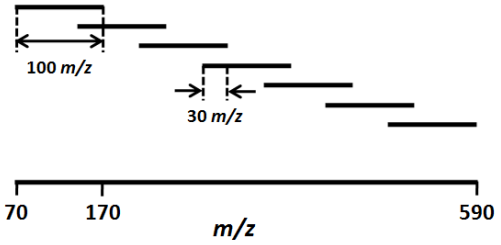
1)

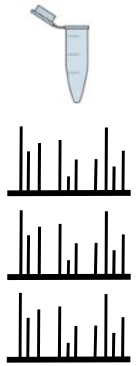
DI nESI high resolution MS

FT-ICR LTQ Orbitrap
LTQ FT Ultra Velos

2)

SIM-Stitching





3)

Statistics

Wide array of methods (see examples)

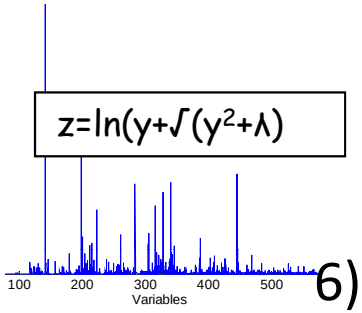
8)

Data processing and quality assessment

Quality assessment

7)

Transformation



6)

Normalisation and missing values

100	200	500	600	
300		800	700	700
500	600	500	600	700
	200	600		800
400	500	800	600	400

5)

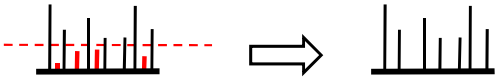
Multi-step signal filtering

4)

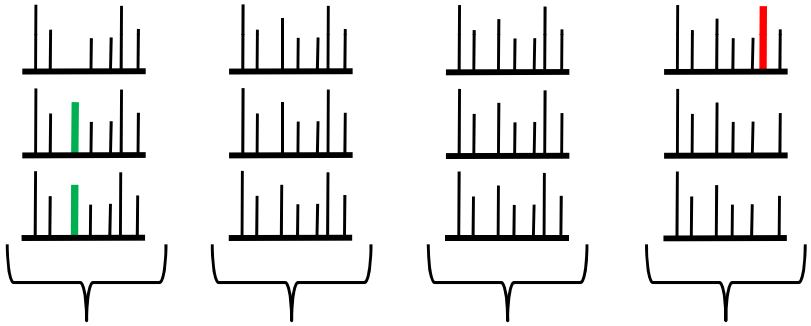
Multi-step signal filtering

SIM-Stitching

(1) SNR threshold
(>3.5)



(2) Replicate filter
peaks in 2-out-of-3



(3) Sample filter
peaks in >=50% of biological samples

	m/z measurments									
Samples										

Payne et al., J Am Soc Mass Spectrom. 20:1087-95 (2009)

Handling missing values – k-nearest neighbour

- Compared missing value imputation methods; we found that k-nearest neighbour (KNN) to be superior
- Uses samples with similar characteristics to impute the missing values
- Intensity matrix:

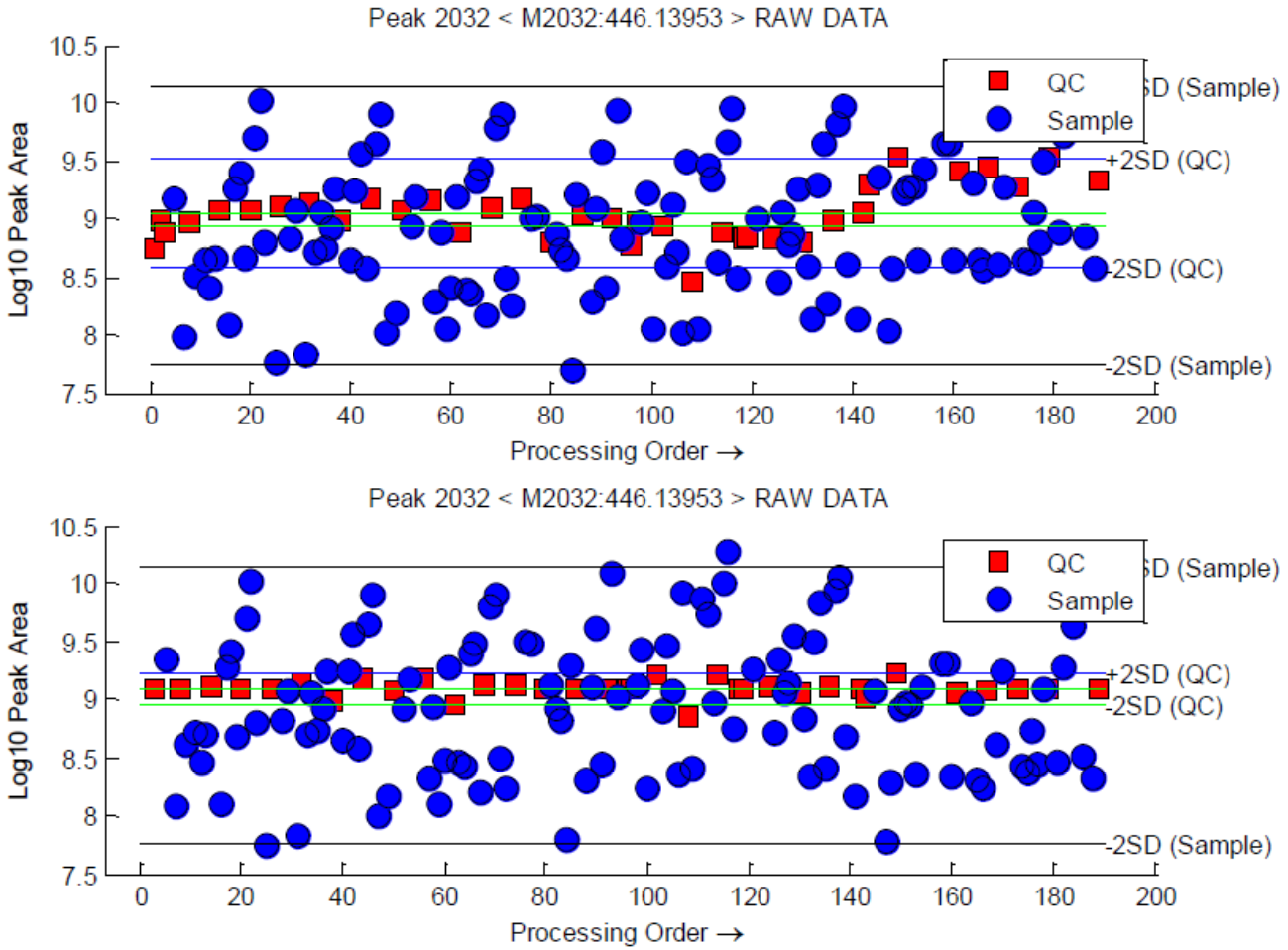
m/z →

Samples ↓

Sample 1	100	200	300	400	500
Sample 2	100	200		400	500
Sample 3	500	200	600	800	100
Sample 4	500	500	600		100
Sample 5	100	200	300	400	500

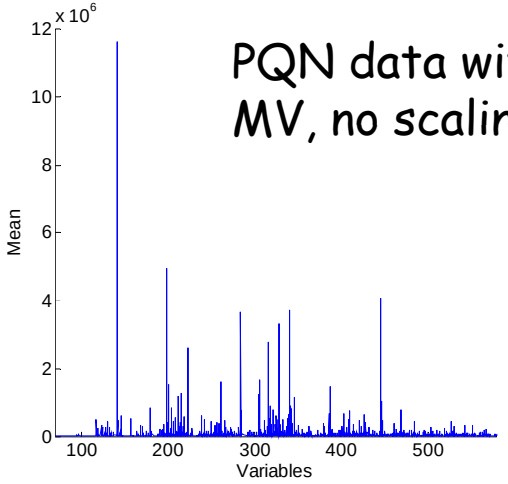
Hrydziuszko et al. Metabolomics 8:161-174 (2012)

Batch (or drift) correction



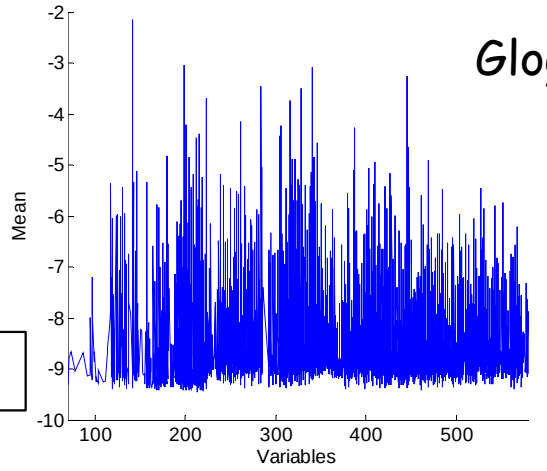
Kirwan et al. Anal Bioanal Chem 405:5147-5157 (2013)

Generalized logarithm (glog) transform



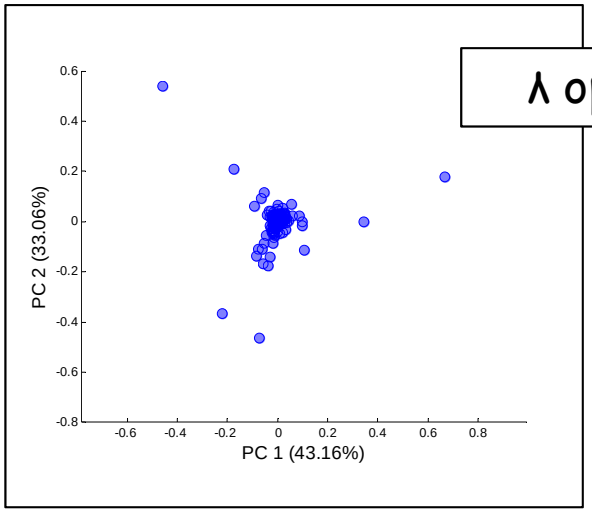
PQN data with KNN
MV, no scaling

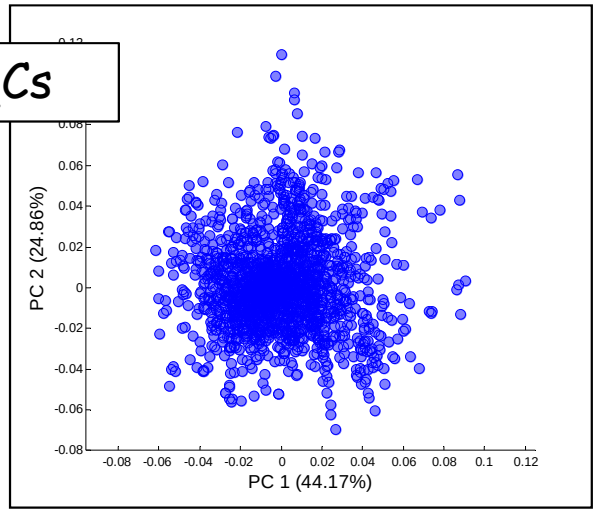
$z=\ln(y+\sqrt{(y2+\lambda)})$



Glog'ed data

λ optimised using QCs



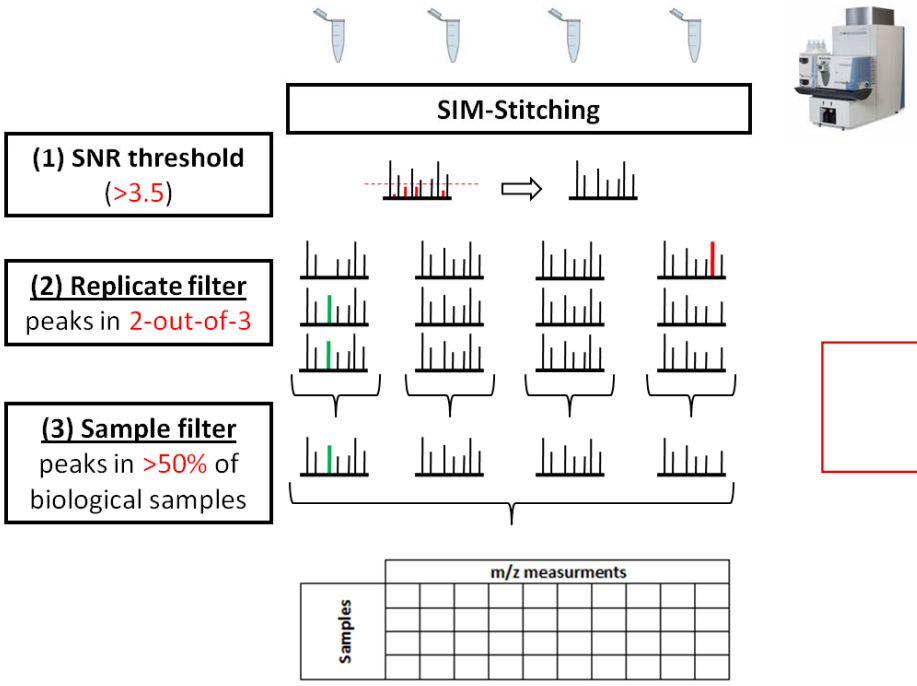


Loadings plot dominated by very few high intensity peaks

Most peaks now make some contribution to the loadings plot

Glog transformation stabilises the technical variance of the peaks

Quality assessment



Median RSD (across 3 technical replicates) calculated for each sample for QA/QC

Normalisation (PQN)

Median RSD for each peak for QA/QC

Missing values (KNN)

Variance scaling (glog transform)

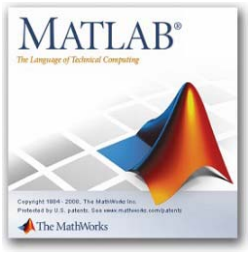
Statistical analysis

Assess QC drift, median RSD for QA/QC, reported

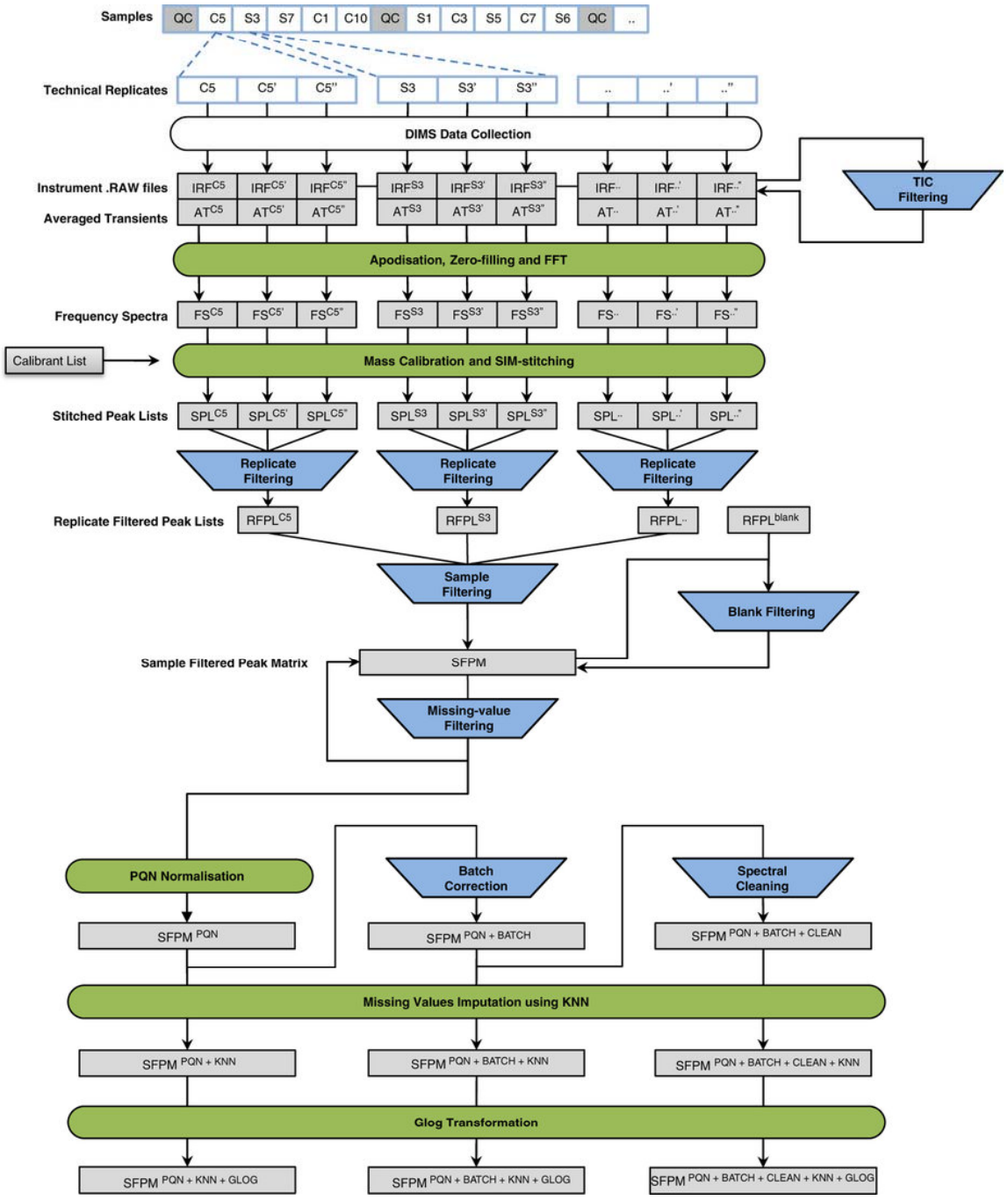
Parsons et al. Analyst 134:478-485 (2009)

DIMS pipeline

> dozen papers published
10 yrs of development



python™



Kirwan et al., Scientific Data. 1, no. 140012 (2014)
Southam et al., Nature Protocols (under review)

Making workflow available...

Davidson *et al. GigaScience* (2016) 5:10
DOI 10.1186/s13742-016-0115-8

GigaScience

TECHNICAL NOTE Open Access

Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data

Robert L. Davidson[1,2†], Ralf J. M. Weber[2†], Haoyu Liu[2], Archana Sharma-Oates[2] and Mark R. Viant[2*]

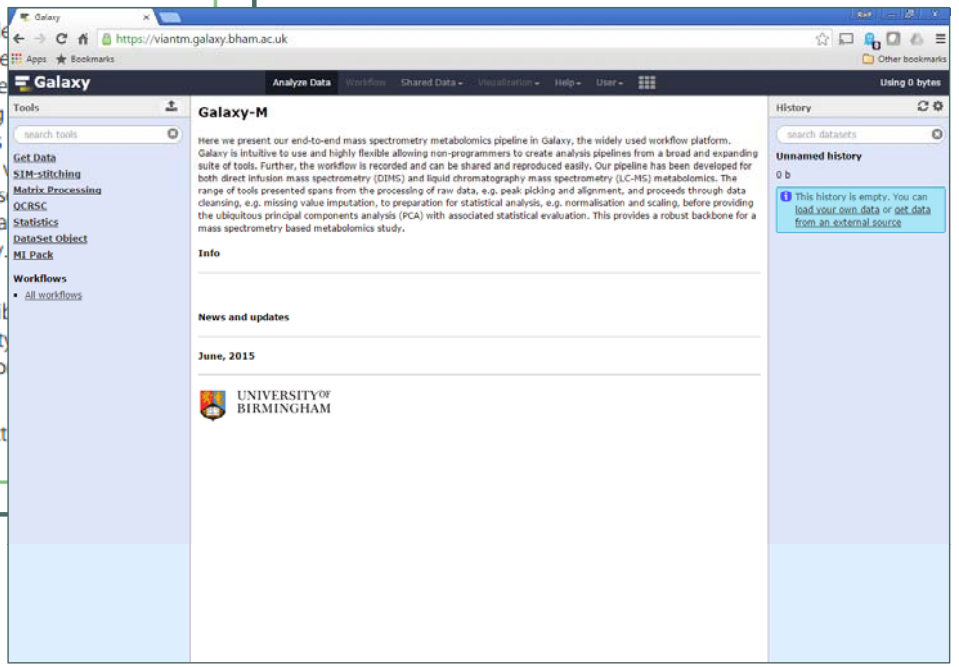
Abstract

Background: Metabolomics is increasingly recognized as an invaluable tool in the biological, medical and environmental sciences yet lags behind the methodological maturity of other omics fields. To achieve its full potential, including the integration of multiple omics modalities, the accessibility, standardization and reproducibility of computational metabolomics tools must be improved significantly.

Results: Here we present our end-to-end mass spectrometry metabolomics workflow in the wide
Galaxy. Named Galaxy-M, our workflow has been developed for both direct infusion mass spe
and liquid chromatography mass spectrometry (LC-MS) metabolomics. The range of tools pre
processing of raw data, e.g. peak picking and alignment, through data cleansing, e.g. missing
preparation for statistical analysis, e.g. normalization and scaling, and principal components
associated statistical evaluation. We demonstrate the ease of using these Galaxy workflows
DIMS and LC-MS datasets, and provide PCA scores and associated statistics to help other us
they can accurately repeat the processing and analysis of these two datasets. Galaxy and da
pre-installed in a virtual machine (VM) that can be downloaded from the GigaDB repository.
code, executables and installation instructions are available from GitHub.

Conclusions: The Galaxy platform has enabled us to produce an easily accessible and reproducit
metabolomics workflow. More tools could be added by the community to expand its functionality
that Galaxy-M workflow files are included within the supplementary information of publicatio
metabolomics studies to achieve greater reproducibility.

Keywords: Metabolomics, Lipidomics, Workflow, Pipeline, Liquid chromatography mass spect
Fourier transform ion cyclotron resonance, FT-ICR, Galaxy project, Reproducibility



Galaxy is intuitive to use and highly flexible allowing non-programmers to create workflows

Accessibility, Standardisation & Reproducibility

Davidson et al., Gigascience 5:10 (2016)

Overview

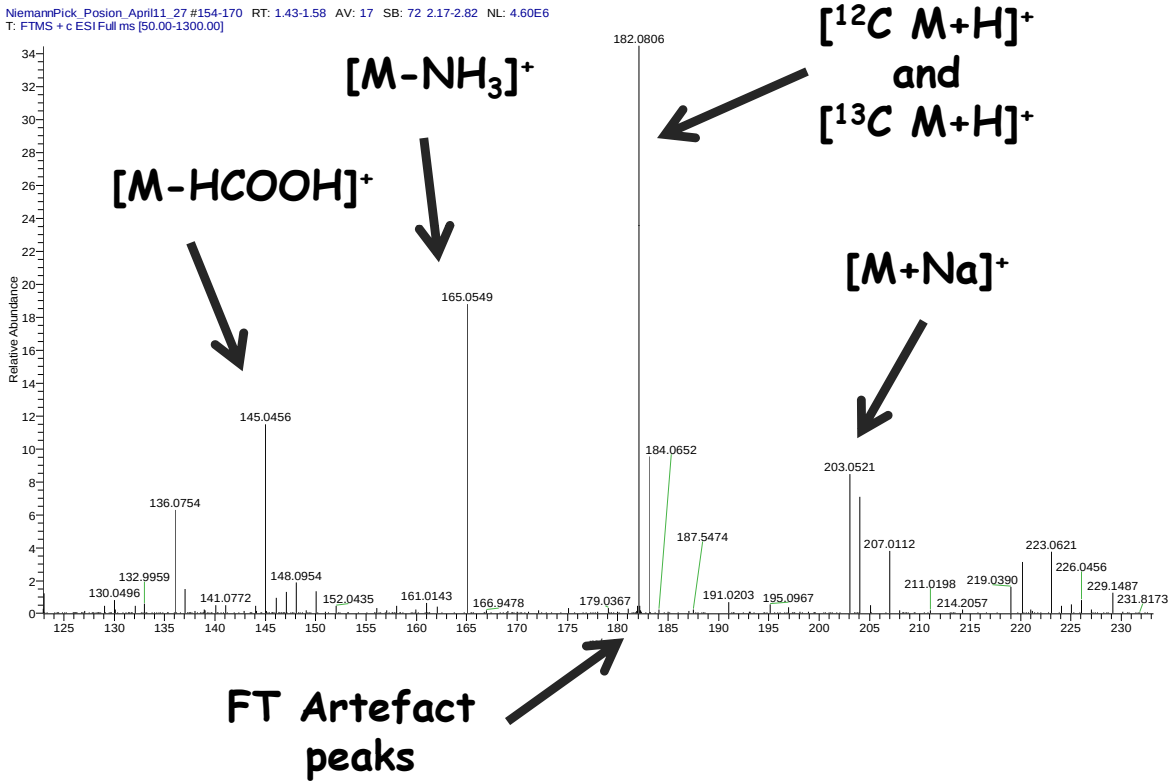
1. Introduction to environmental metabolomics
2. Workflows
   - Direct infusion mass spectrometry (standardised)
   - Data processing (relatively standardised)
   - Metabolite annotation and identification (not standardised)
3. Examples
   - Endogenous metabolites
   - Xenobiotics within organisms
4. Where next?

Metabolite identification – A BOTTLENECK IN METABOLOMICS

For metabolomics to be successful it is essential to derive biological knowledge from analytical data - a view emphasised by a Metabolomics ASMS Workshop Survey 2009 which found that the biggest bottlenecks in metabolomics were thought to be identification of metabolites (35%) and assignment of biological interest (22%)

http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics-Survey-2009

Mass spectral data includes another level of complexity



In one sample set there were 20 different "ion types"

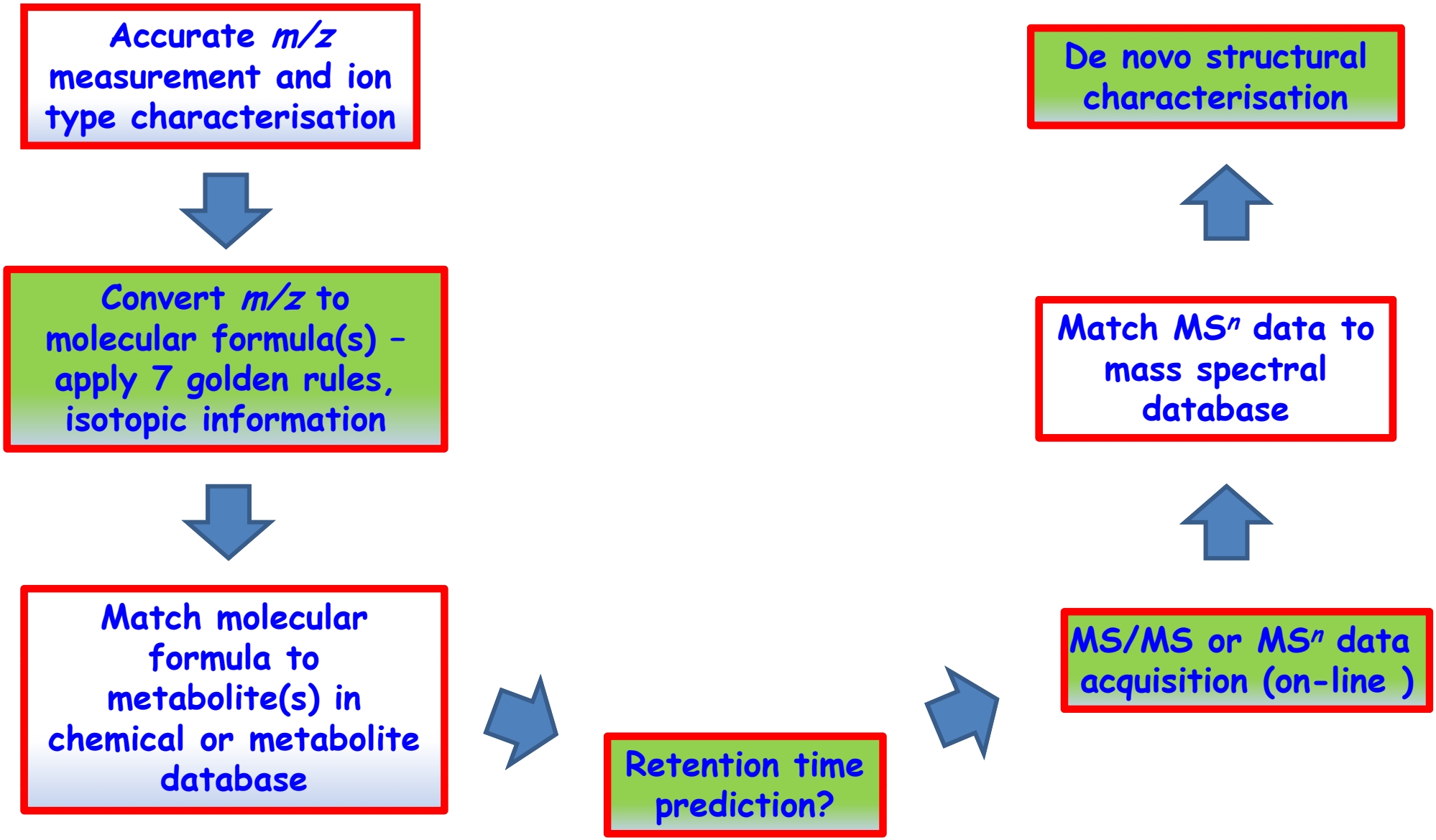
<u>Protonated and deprotonated ions</u>
$[M+H]^+$ and $[M-H]^-$

<u>Fragment ions</u>
$[M-HCOOH]^+$ and $[M-NH_3]^+$

<u>Adduct ions</u>
$[M+HCOOK]^+$, $[M+HCOONa]^+$
$[M+3HCOONa]^+$
$[M+NaCl+HCOONa]^+$
$[M+3NaCl]^+$
$[M+Fe]^{2+}$
$[M+Cu]^{2+}$

There is structure to the data - Apply RT, response correlation, *m/z* difference to group metabolite features of same metabolite

Typical workflows for metabolite annotation & identification (DIMS, LC-MS…)



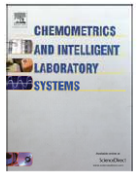
MI-Pack – Metabolite Identification Package

Chemometrics and Intelligent Laboratory Systems 104 (2010) 75–82

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

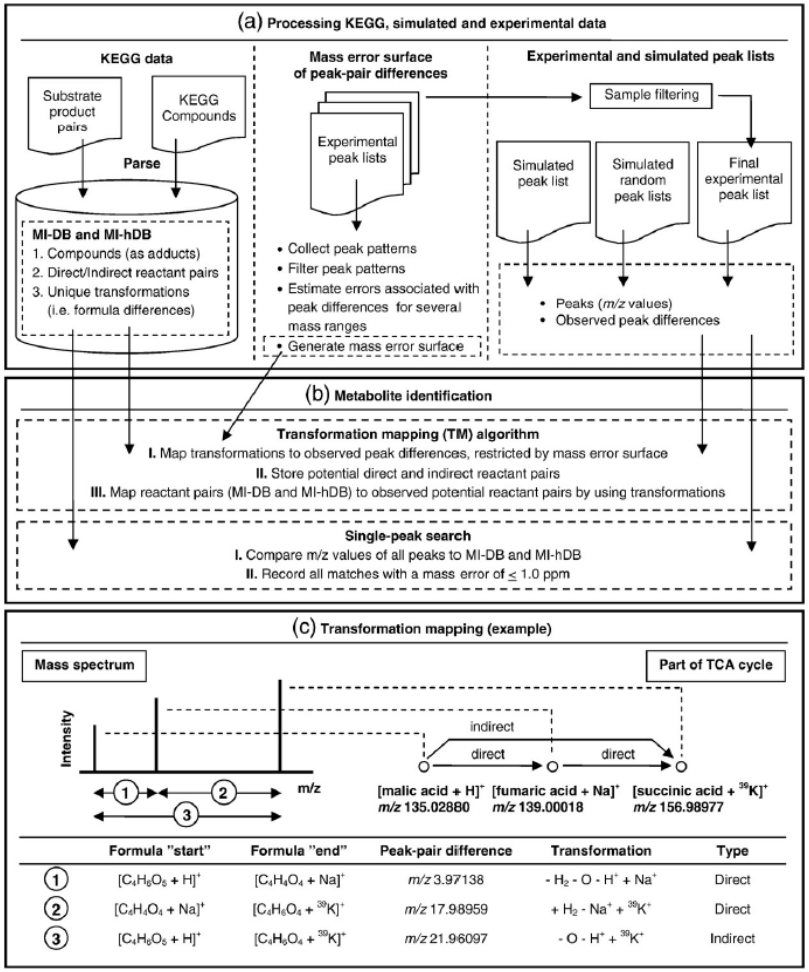


MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways

Ralf J.M. Weber [a], Mark R. Viant [a,b,*]

[a] Centre for Systems Biology, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom
[b] School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

Additionally can use knowledge of metabolic pathways



PutMetID - Conversion to molecular formula and then metabolite

Apply RT, response correlation, *m/z* difference to group features of same metabolite

BIOINFORMATICS **ORIGINAL PAPER** Vol. 27 no. 8 2011, pages 1108–1112
doi:10.1093/bioinformatics/btr079

Systems biology Advance Access publication February 16, 2011

Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets

Marie Brown[1], David C. Wedge[2], Royston Goodacre[2,3], Douglas B. Kell[2], Philip N. Baker[4], Louise C. Kenny[5], Mamas A. Mamas[1,6], Ludwig Neyses[1,6] and Warwick B. Dunn[1,2,3,7,*]

[1]School of Biomedicine, The University of Manchester, Manchester M13 9PT, [2]School of Chemistry, [3]Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, University of Manchester, Manchester M1 7DN, UK, [4]Department of Obstetrics and Gynecology, Faculty of Medicine and Dentistry, University of Alberta, 2J2.01 WMC, Edmonton AB T6G 2R7, Canada, [5]The Anu Research Centre, Department of Obstetrics and Gynaecology, University College Cork, Cork University Maternity Hospital, Cork, Ireland, [6]Manchester Heart Centre, Central Manchester University Hospitals NHS Foundation Trust, Manchester Royal Infirmary and [7]Centre for Advanced Discovery and Experimental Therapeutics, York Place (off Oxford Road), Central Manchester University Hospitals NHS Foundation Trust, Manchester M13 9WL, UK

Associate Editor: John Quackenbush

ANNOTATION OF ALL FEATURES BASED ON ACCURATE MASS DIFFERENCES, RETENTION TIME AND CORRELATION ANALYSIS

↓

MATCHING OF ACCURATE MASS TO MOLECULAR FORMULA(E) IN REFERENCE FILE

↓

MATCHING OF MOLECULAR FORMULA(E) TO METABOLITE(S) IN A REFERENCE FILE (E.G. MMD)

Annotation vs. Identification

- **Identification** = two orthogonal properties (RT, MS/MS) compares to authentic chemical standard under identical analytical conditions

- **Annotation** = one (or more) orthogonal property match to databases (not necessarily acquired under identical analytical conditions)

Salek *et al. GigaScience* 2013, **2**:13
http://www.gigasciencejournal.com/content/2/1/13

COMMENTARY Open Access

The role of reporting standards for metabolite annotation and identification in metabolomic studies

Reza M Salek[1,2], Christoph Steinbeck[1], Mark R Viant[3], Royston Goodacre[4] and Warwick B Dunn[3*]

Abstract

The application of reporting standards in metabolomics allow data from different laboratories to be shared, integrated and interpreted. Although minimum reporting standards related to metabolite identification were

Salek et al., Gigascience 2:13 (2013)

Four levels of confidence

- Sumner et al. Proposed minimum reporting standards for chemical analysis, Metabolomics, 2007, 3:211-221
- Currently, four levels of metabolite identifications can be reported
- Not defining how to perform metabolite identification but defining how to report it

Level	Confidence of Identity	Level of Evidence
1	Confidently identified compounds.	Comparison of two or more orthogonal properties with an authentic chemical standard analysed under identical analytical conditions.
2	Putatively annotated compounds	Based upon physicochemical properties and/or spectral similarity with public/commercial spectral libraries, without reference to authentic chemical standards.
3	Putatively annotated compound classes	Based upon characteristic physicochemical properties of a chemical class of compounds, or by spectral similarity to known compounds of a chemical class.
4	Unknown compounds	Although unidentified and unclassified, these metabolites can still be differentiated and quantified based upon spectral data.

Overview

1. Introduction to environmental metabolomics
2. Workflows
   - Direct infusion mass spectrometry (standardised)
   - Data processing (relatively standardised)
   - Metabolite annotation and identification (not standardised)
3. Examples
   - Endogenous metabolites
   - Xenobiotics within organisms
4. Where next?

Experimental design

Individual
Daphnia

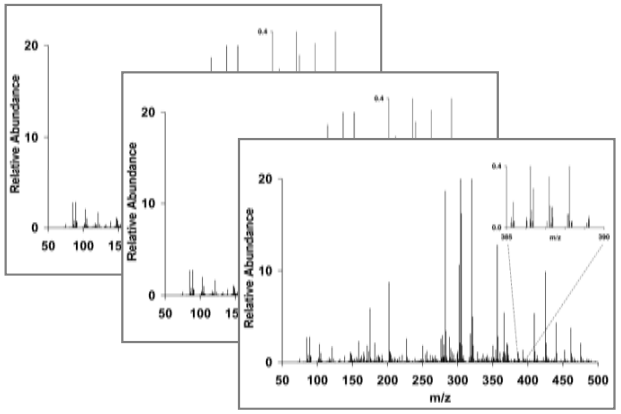
Chemical exposures

Cadmium
Propranolol
Dinitrophenol (DNP)

Measure reproductive fitness

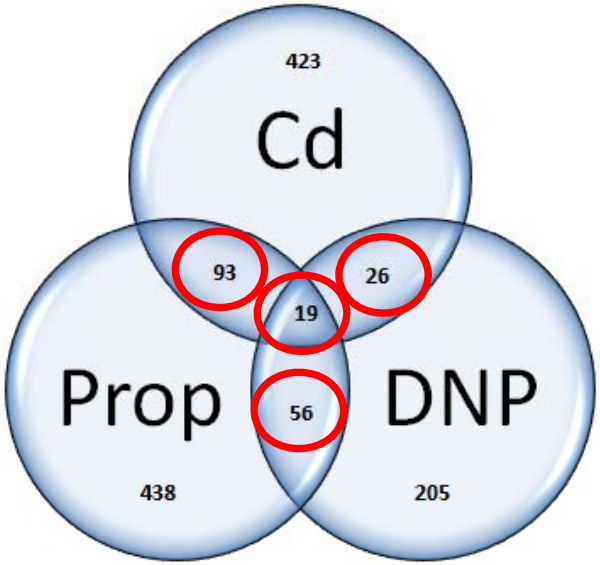
Measure metabolism

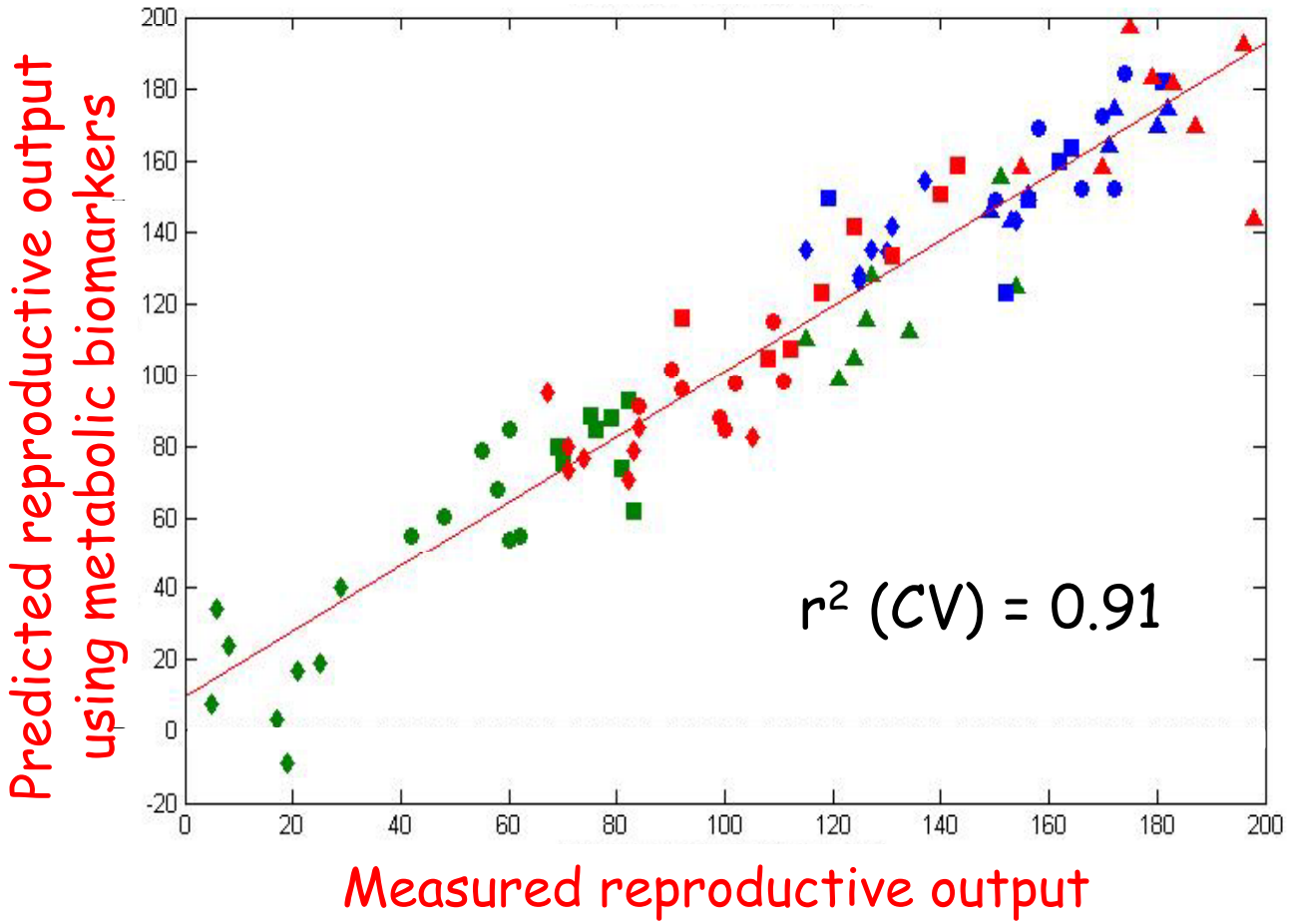
of individual
Daphnia



Multivariate PLS regression to determine whether metabolites can predict reproductive fitness

Metabolic biomarkers can also predict reproductive fitness in response to all 3 toxicants





Optimal PLS regression model: 49 peaks derived using forward selection

Which metabolites predict reproductive fitness?

Putative annotation of 49-biomarker signature using MI-Pack

Metabolite identification confirmed by MS/MS of metabolite sample compared to pure standard

Measured m/z	Empirical formula(e)	Ion form	Mass error (ppm)	Putative metabolite name(s)
175.02480	C6H8O6	[M-H]-	-0.08	Ascorbic acid (confirmed by MS/MS)
243.00911	C7H10O7	[M+37Cl]-	0.01	Methylcitrate or homocitrate
258.05642	C8H15NO6	[M+37Cl]-	0.11	N-Acetyl-D-hexosamine
etc…				

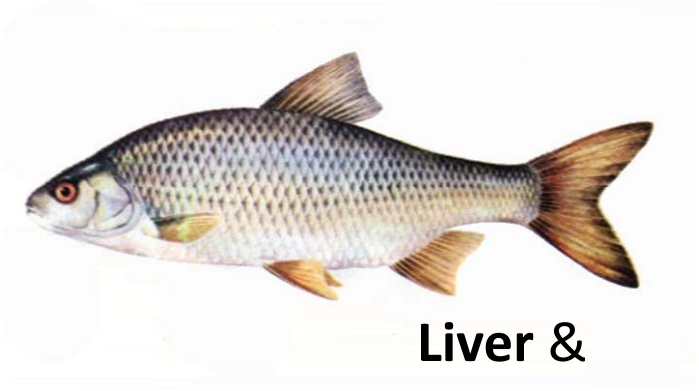
"Ascorbic acid has long been associated with fertility"
Luck et al., Biol. Reprod. 52, 262-266 (1995)

"We conclude that ascorbic acid is a leading nutrient in reproductive tissue functions [in teleost fish]" Dabrowski & Ciereszko, Aquacult. Res. 32, 623-638 (2001)

Overview

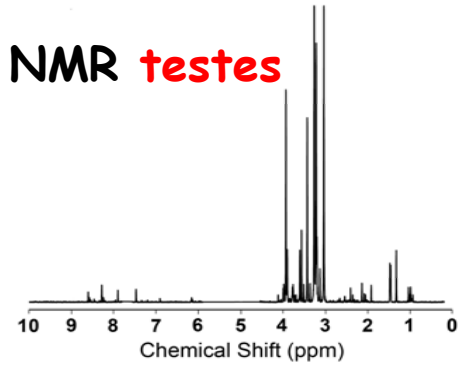
1. Introduction to environmental metabolomics
2. Workflows
   - Direct infusion mass spectrometry (standardised)
   - Data processing (relatively standardised)
   - Metabolite annotation and identification (not standardised)
3. Examples
   - Endogenous metabolites
   - Xenobiotics within organisms
4. Where next?

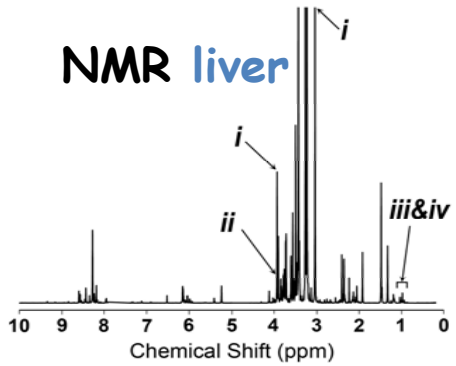
Complexity of one exogenous (xenobiotic) compound

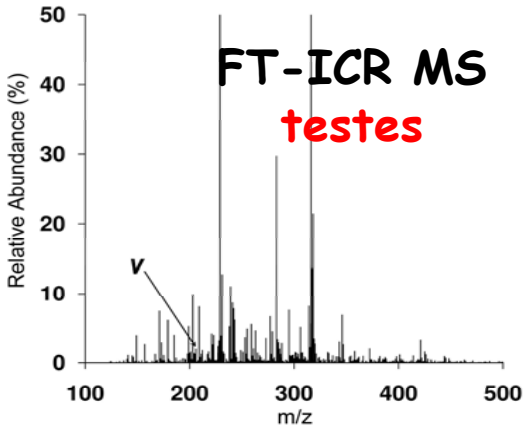


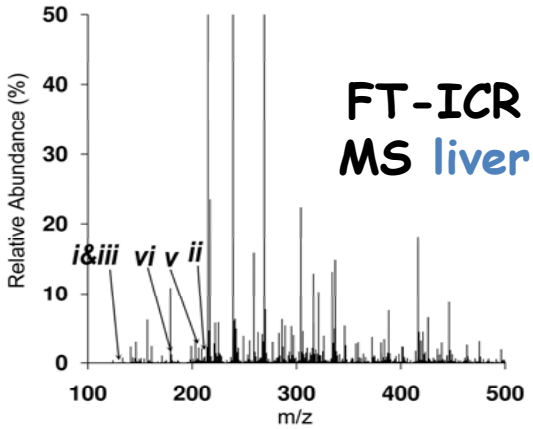
Liver & testes

Non-targeted metabolomics



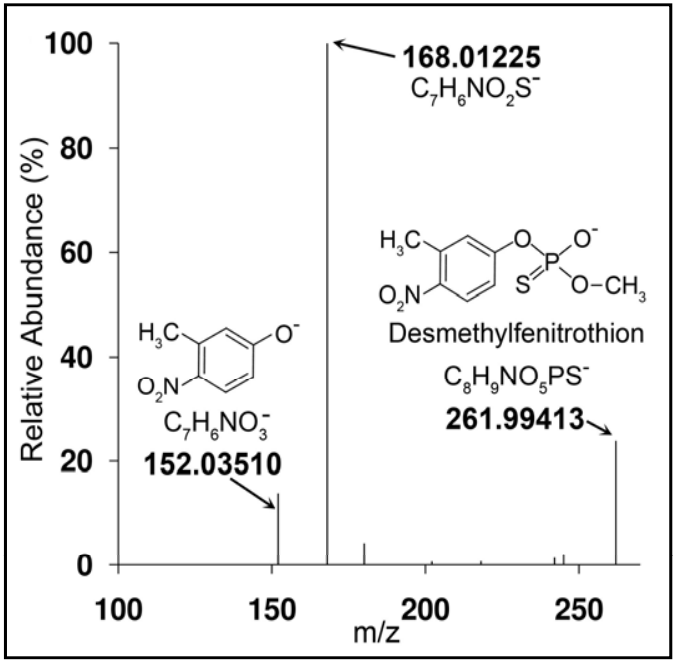


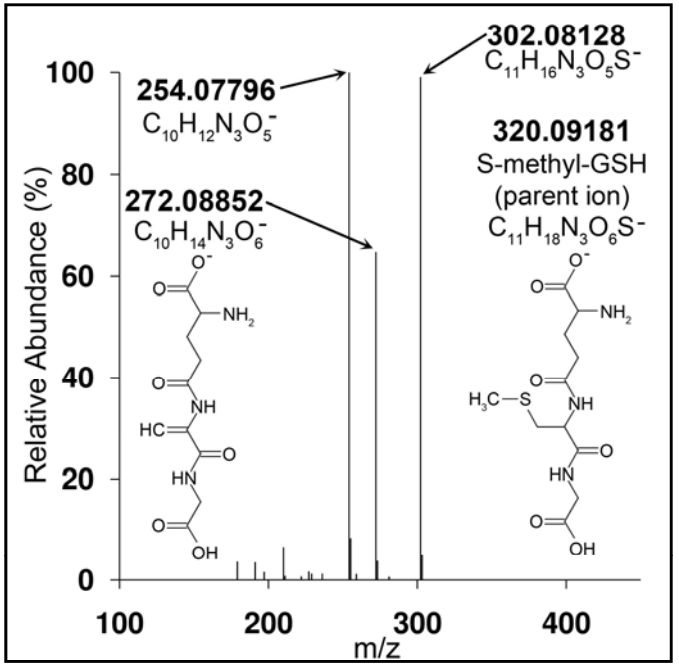




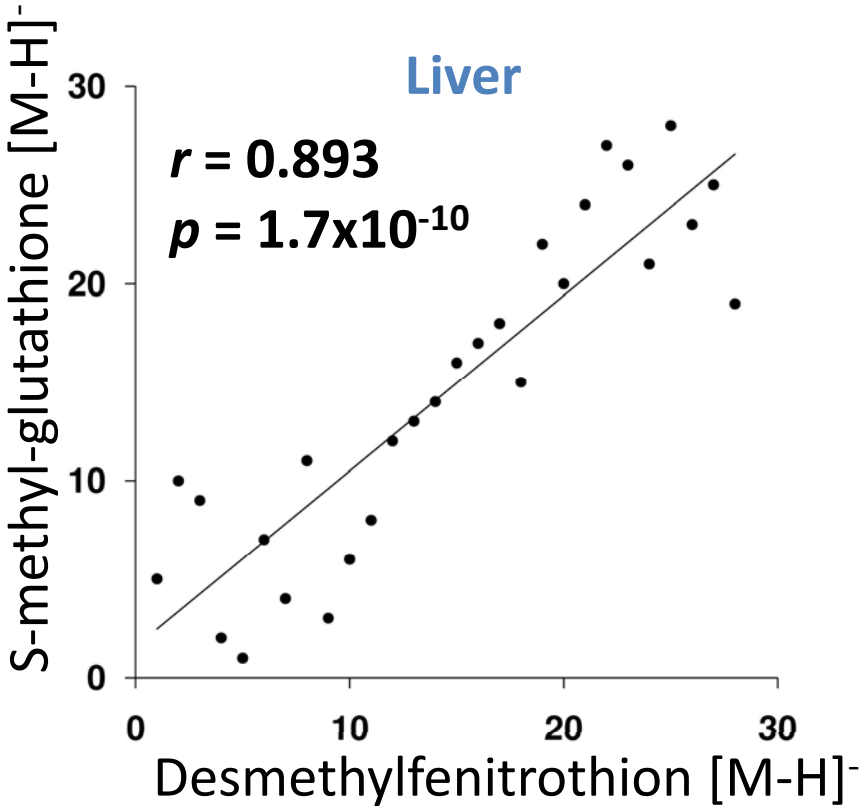
Fenitrothion Metabolism (1)

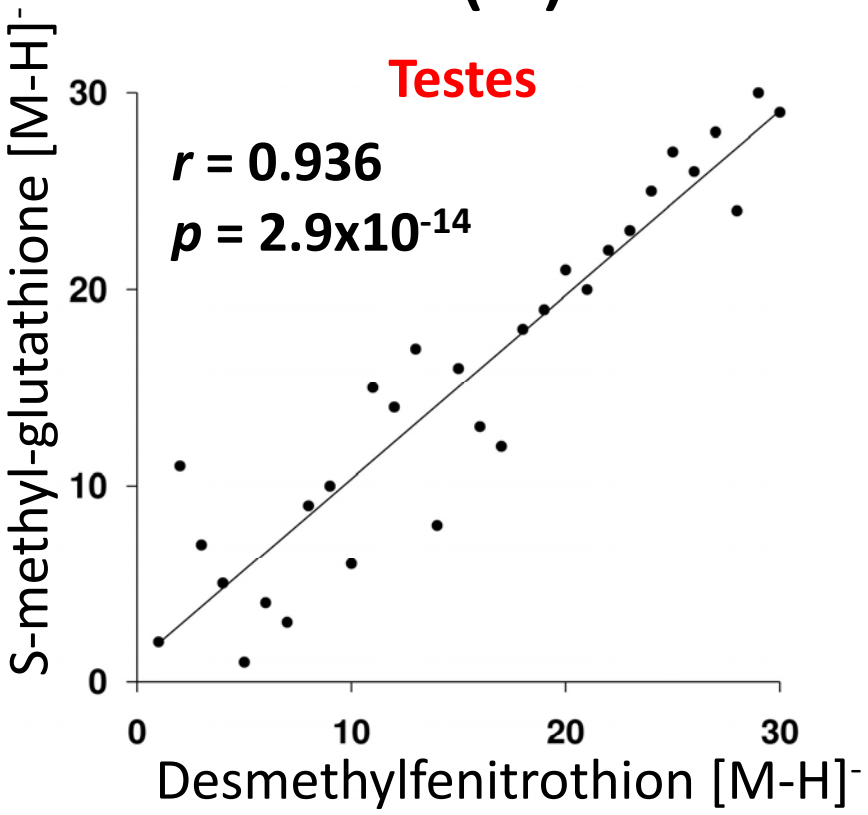
Observed *m/z*	Tissue	*p*-value	Fold change		Empirical formulae	Metabolite identification
			HD/SC	HD/LD		
261.99450	**Liver**	1.10×10^{-15}	∞	24.83	$C_8H_{10}NO_5PS$ **[M-H]$^-$**	**Desmethyl-fenitrothion**
261.99444	**Testes**	2.61×10^{-12}	∞	33.82		
320.09224	**Liver**	2.09×10^{-14}	73.44	15.32	$C_{11}H_{19}N_3O_6S$ **[M-H]$^-$**	**S-methyl-glutathione**
320.09240	**Testes**	3.14×10^{-9}	9.82	9.41		



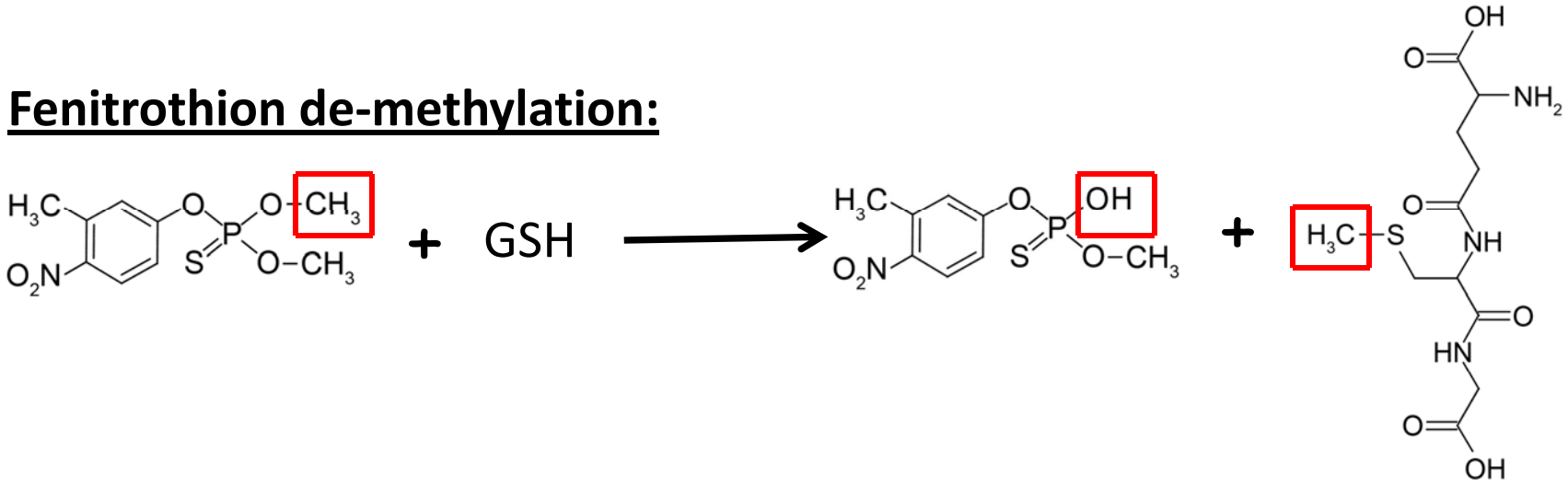


Fenitrothion Metabolism (2)





Fenitrothion de-methylation:



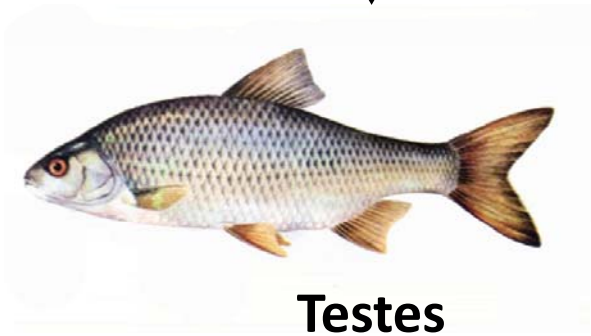
Findings from fenitrothion study

1) Endogenous metabolism: fenitrothion significantly disrupts acetylcholine, disrupts key steroids, affects energy metabolism and disrupts phenylalanine metabolism

2) Xenobiotic metabolism: O-demethylation observed as the major route of fenitrothion detoxification in roach.

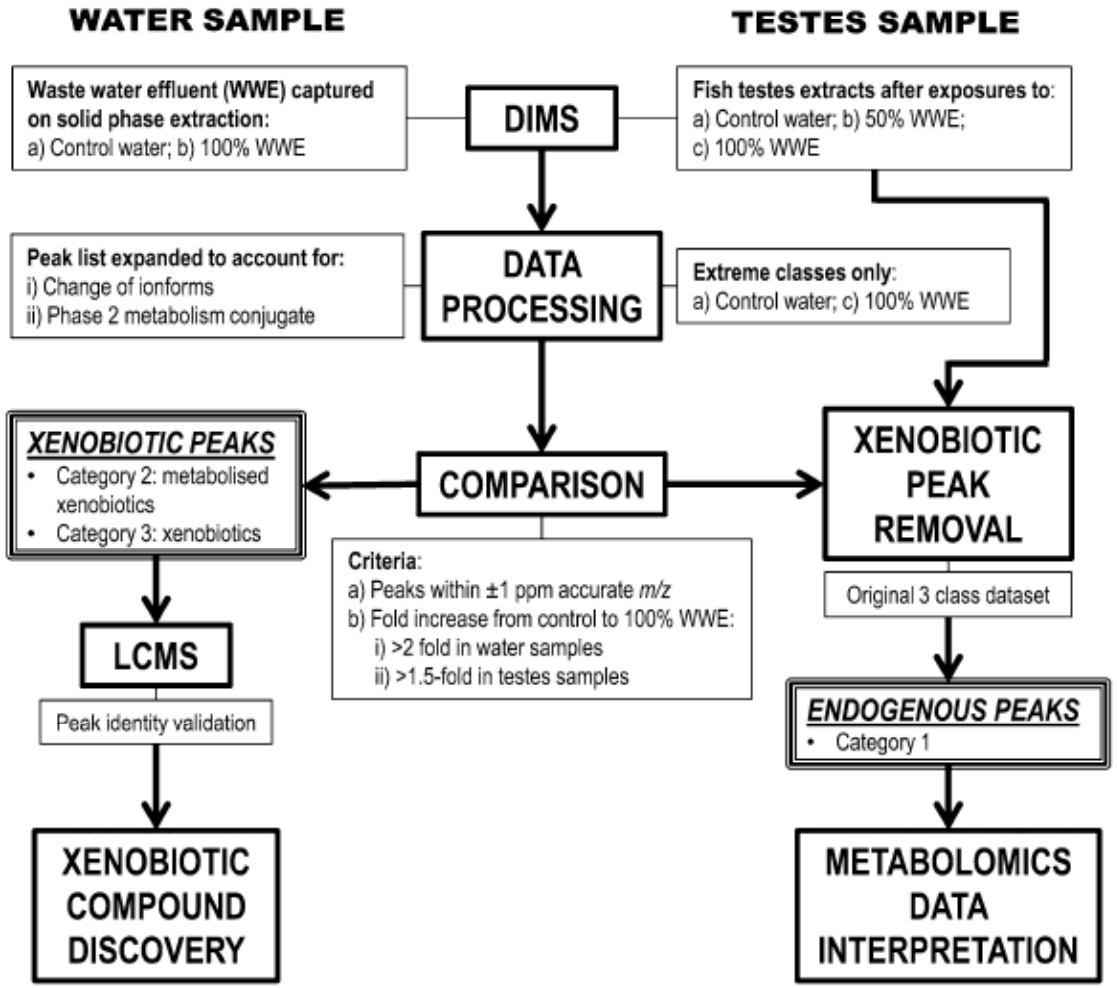
Southam et al., Environ Sci Technol 45:3759-3767 (2011)

Complexity of complex, uncharacterised mixture of exogenous (xenobiotic) compounds



Testes

Non-targeted metabolomics



Xenobiotics and metabolised xenobiotics discovered in fish testes

Table 2 UHPLC-QTOF MS based identification of a selection of peaks computationally predicted as being of xenobiotic or metabolised xenobiotic origin in the direct infusion MS datasets (Tables S5, S6)

UHPLC-QTOF MS validation		Waste water effluent			Testes extract						
Name	Confirmation type	*m/z*	Peak intensity		*m/z*	Extract phase	Peak intensity		*q*	Peak modification in testes	ppm error
			Dilution water	Effluent			Dilution water	Effluent exposed			
Chloroxylenol	RT	155.02697	628	8,967	155.02696	Lipid	0	2731	1.6×10^{-5}	None	−0.033
Chlorophene	RT & MS/MS	217.04252	0	47,570	217.04279	Lipid	0	23202	5.2×10^{-6}	None	0.897
Chlorophene (^{13}C)	RT & MS/MS	218.04591	1157	12,637	218.04614	Lipid	0	3754	2.0×10^{-6}	None	0.704
Triclosan	RT & MS/MS	286.94392	0	49,887	366.90074	Polar	0	3917	9.1×10^{-10}	$+SO_4$	−0.011
Triclosan sulfate	MS/MS	366.90064	0	104,973						None	0.286
Triclosan (^{37}Cl)	RT & MS/MS	288.94099	0	58,851	368.89780	Polar	0	4271	8.7×10^{-9}	$+SO_4$	−0.019
Triclosan sulfate (^{37}Cl)	MS/MS	368.89768	0	97,743						None	0.333
Triclosan ($2 \times {}^{37}Cl$)	RT & MS/MS	290.93798	0	22,895	370.89480	Polar	0	1551	2.7×10^{-6}	$+SO_4$	0.003
Triclosan sulfate ($2 \times {}^{37}Cl$)	MS/MS	370.89481	0	31,439						None	−0.024
Linear alkylbenzene sulfonate (LAS) metabolite	MS/MS	357.14504	0	20,627	387.15539	Polar	68	925	2.5×10^{-3}	$+OCH_3–H$	−0.553
Linear alkylbenzene sulfonate (LAS) metabolite	MS/MS	327.13410	0	29,889						$[M–H]^-$ to $[M + OAc]^-$	0.413

Peaks were confirmed with standard compounds utilising UHPLC-QTOF MS retention times (RT), tandem mass spectrometry (MS/MS), or both (Table S7). The *q* values correspond to *p* values that has been FDR corrected

Southam et al., Metabolomics 10:1050-1058 (2014)

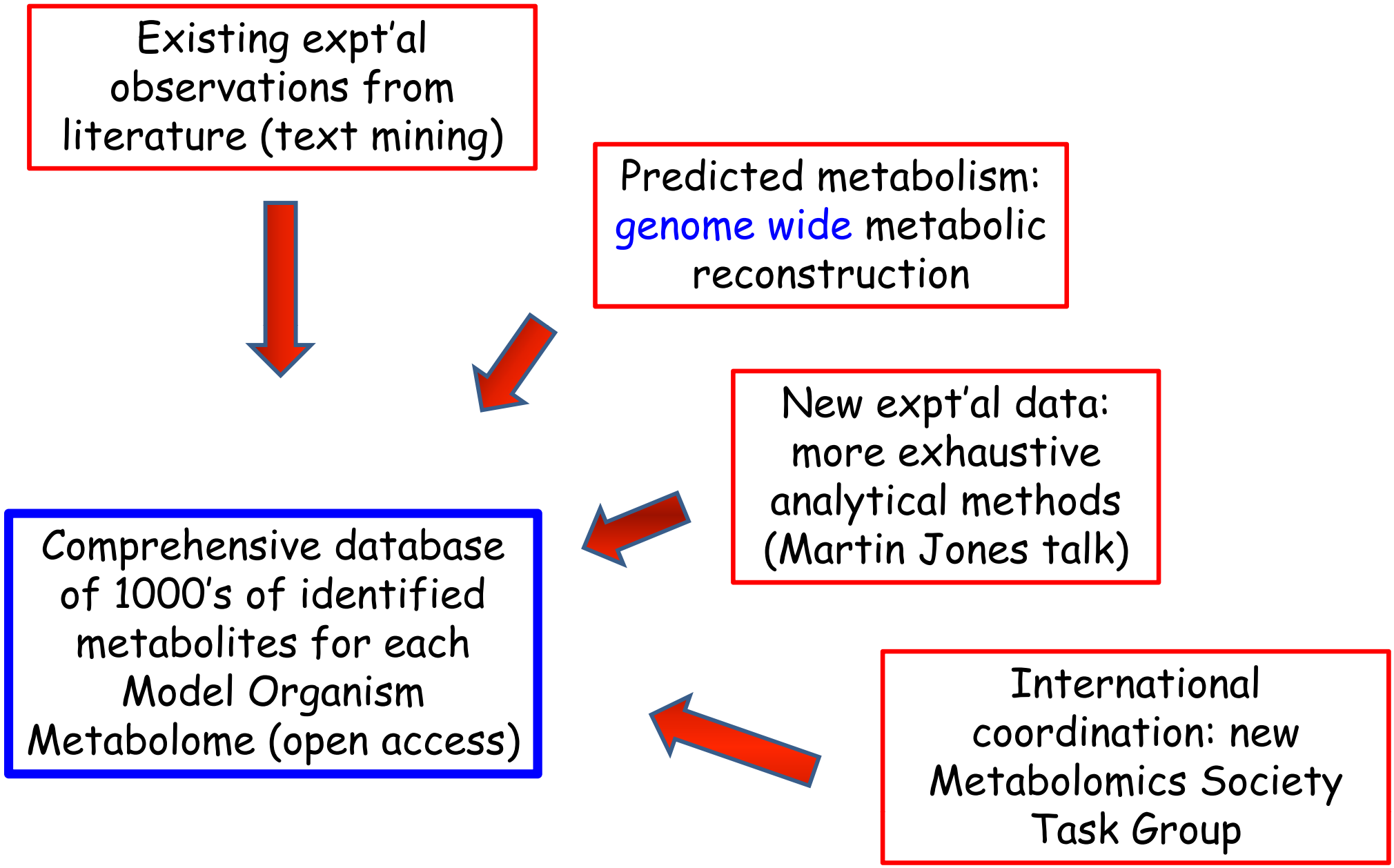
Overview

1. Introduction to environmental metabolomics
2. Workflows
   - Direct infusion mass spectrometry (standardised)
   - Data processing (relatively standardised)
   - Metabolite annotation and identification (not standardised)
3. Examples
   - Endogenous metabolites
   - Xenobiotics within organisms
4. Where next?

Reflect on current status…

1. Metabolomics workflows, both analytical and computational, are improving and there is increasing trend towards harmonisation
2. Metabolomics community is generally in favour of open access / data sharing etc.
3. Yet metabolite identification remains a huge challenge
4. How do we accelerate research into metabolite identification?

Focus on Model Organism Metabolomes



Daphnia Deep Metabolome Annotation Project (Martin Jones' talk this Thursday)

- Multi-platform characterisation: extensive extraction & fractionation chemistries, chromatography (LC, GC,...), detectors (mass spectrometry, NMR spectroscopy...)
- Databases: new local database, mzCloud, and MetaboLights
- Part of University of Birmingham's Technology Alliance Partnership with Thermo Fisher Scientific

Thermo SCIENTIFIC

metabolites

MDPI

Communication

The Time Is Right to Focus on Model Organism Metabolomes

Arthur S. Edison [1], Robert D. Hall [2], Christophe Junot [3], Peter D. Karp [4], Irwin J. Kurland [5], Robert Mistrik [6], Laura K. Reed [7], Kazuki Saito [8], Reza M. Salek [9], Christoph Steinbeck [9], Lloyd W. Sumner [10] and Mark R. Viant [11,*]

Metabolites **2016**, *6*, 8; doi:10.3390/metabo6010008

Environmental Metabolomics @ University of Birmingham

Martin Jones
Dr Ulf Sommer
Dr Jaspreet Sihra
Dr Adam Hines
Dr Andrew Southam
Dr Nadine Taylor
Dr Ralf Weber
Tom Lawson
Dr Cate Winder
Dr Warwick Dunn
Prof Charles Tyler (Exeter)
Prof Elizabeth Hill (Sussex)
Prof Kevin Chipman
Dr Michael Cunliffe (Plymouth)

http://www.birmingham.ac.uk/research/activity/metabolomics

NATURAL ENVIRONMENT RESEARCH COUNCIL

BBSRC bioscience for the future

Thermo SCIENTIFIC