Metabolomics: An introduction

Reza Salek [1], Laura Emery [2], Stephan Beisken [3]

- Systems
- Chemical biology
- Beginner
- 1 hour

This course provides a basic introduction into the rapidly emerging field of metabolomics and its importance and applications.

Learning objectives:

- Comprehend the purpose and importance of the field of metabolomics
- Describe some principles of metabolomic study design
- Evaluate advantages and limitations of some analytical techniques used in metabolomics studies
- Discuss some of the modern-day applications of metabolomics
- Access metabolomics resources at the EMBL-EBI

What is metabolomics?

Metabolomics [4] is the large-scale study of small molecules [5], commonly known as metabolites, within cells, biofluids, tissues or organisms. Collectively, these small molecules and their interactions within a biological system are known as the metabolome.

Figure 1 An overview of the four major "omics" fields, from genomics to metabolomics.

Just as genomics is the study of DNA and genetic information within a cell, and transcriptomics [6] is the study of RNA and differences in mRNA expression; metabolomics is the study of substrates and products of metabolism [7], which are influenced by both genetic and environmental factors (Figure 1).

Metabolomics is a powerful approach because metabolites and their concentrations, unlike other "omics" measures, directly reflect the underlying biochemical activity and state of cells / tissues. Thus metabolomics best
Metabolomics: An introduction
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

represents the molecular phenotype [8].

Small molecules

What are small molecules?
A small molecule (or metabolite) is a low molecular weight organic compound, typically involved in a biological process as a substrate or product. Metabolomics [4] usually studies small molecules within a mass range of 50 – 1500 daltons (Da).

Some examples of small molecules include: sugars, lipids, amino acids, fatty acids, phenolic compounds, alkaloids

\[
\text{glucose (a sugar)} \quad \text{stearic acid (a fatty acid)}
\]

\[
\text{lysine (an amino acid)} \quad \text{cholesterol (a lipid)}
\]

Figure 2 Examples of small molecules.

To give you an idea of the variation of metabolites between species, it is estimated there are around 200,000 metabolites across the plant kingdom, and somewhere between 7,000 and 15,000 within an individual plant species (1,2 [9]). By contrast, in humans, there are thought to be around 3,000 endogenous or common metabolites (3 [9]). These estimates are approximations that are likely to be underestimates because it is difficult to detect low-abundance molecules. Nonetheless, it can be concluded that plants are particularly biochemically rich by comparison with many other species. They also typically contain larger numbers of genes.

and many more (Figure 2).

The metabolome and metabolic reactions

The metabolome
The metabolome is the complete set of metabolites within a cell, tissue or biological sample at any given time point. The metabolome is inherently very dynamic: small molecules [5] are continuously absorbed, synthesised, degraded.
and interact with other molecules, both within and between biological systems, and with the environment (Figure 3).

Figure 3 A diagram showing the main different types of metabolic reactions that take place in a cell. These are shown as they are represented in the database Reactome [10].

Metabolomics - a 'snapshot' in time

Many reactions take place continuously within cells, so concentrations of metabolites are considered to be very dynamic, and may change rapidly from one time point to the next. Current analytical techniques used to investigate metabolomics can only take a snapshot in time under a set of defined conditions.

Metabolic reactions

Metabolic pathways are essentially a series of chemical reactions, catalysed by enzymes, whereby the product of one reaction becomes the substrate for the next reaction. These reactions can be divided into anabolic [11] and catabolic [12].

The importance of metabolomics
**Why is metabolomics important?**

The non-invasive nature of metabolomics and its close link to the phenotype make it an ideal tool for the pharmaceutical, preventive healthcare, and agricultural industries, among others. Biomarker discovery and drug safety screens are two examples where metabolomics has already enabled informed decision making. In the future, with the availability of personalised metabolomics, we will potentially be able to track the trends of our own metabolome for personalised drugs and improved treatment strategies. Personalised treatment is likely to be more effective than our current medical population-based approaches.

**How is metabolomics used?**

We benefit from metabolomics on various levels: from product and stress testing in food industries, e.g. control of pesticides and identification of potentially harmful bacterial strains, to research in agriculture (crop protection and engineering), medical diagnostics in healthcare, and future applications in personalised medicine resulting in personliased treatment strategies.

**Some applications of metabolomics**

**Agricultural**

The development of new pesticides is critical to meet the growing demands on farming. Metabolomics [4] enables us to improve genetically modified plants, and helps us to estimate associated risks by allowing us to get a glimpse of their complex biochemistry via informative snapshots acquired at different time points during plant development.

Plant metabolomics is particularly interesting because of the range and functions of primary and secondary metabolites in plants. About 300 distinct metabolites could be routinely identified per sample a decade ago, and the number is gradually increasing over time.

**Biomarker discovery**

Biomarker discovery is another area where metabolomics informs decision making. Biomarkers are "objective indications of medical state observed from outside the patient - which can be measured accurately and reproducibly" (Kyle et al. What are Biomarkers?). In metabolomics, biomarkers are small molecules [5] (metabolites) that can be used to distinguish two groups of samples, typically a disease and control group. For example, a metabolite reliably present in disease samples, but not in healthy individuals would be classed as a biomarker. Samples of urine, saliva, bile, or seminal fluid contain highly informative metabolites, and can be readily analysed through metabolomics fingerprinting or profiling, for the purpose of biomarker discovery.

**Personalised medicine**

Personalised medicine, the ultimate customisation of healthcare, requires metabolomics for quick medical diagnosis to identify disease. In healthcare, we currently use classical biochemical tests to measure individual metabolite concentrations to identify disease states (e.g. the blood-glucose level in the case of diabetes). Metabolomics offers the potential for the rapid indentification of hundreds of metabolites, enabling us to identify these disease states much earlier.

**Designing a metabolomics study**

The two main approaches that can be used in metabolomics [4] are untargeted and targeted approaches.
The approach chosen will determine how you design your experiment, prepare your samples, and what analytical techniques you use.

**Untargeted (global) approach**: This method measures as many metabolites as possible from a range of biological samples without any (intended) bias.

**Targeted approach**: This method is used when you want to measure sets of metabolites and have a specific biochemical question that you want to answer.

This approach is often used in pharmacokinetic studies of drug metabolism [7] and when looking at the effect of therapeutics or genetic modifications on a specific enzyme.

### Key stages of a metabolomics study

#### Overview of complete analysis workflow

Both targeted and untargeted metabolomics [4] studies follow a similar pipeline. An example of this pipeline for mass spectrometry [13]-based metabolomics studies is shown in Figure 4 below.
Next, we will go through each of the main steps in more details.

**Design experiment**

The study design, also known as ‘experimental design’, is of paramount importance for every study. It is essential to make sure that the samples collected reflect and represent the biology in question. In order to determine and examine the most influential factors that are relevant for the hypothesis under investigation, external factors that can affect the experiment have to be eliminated or identified so that they can be accounted for during data analysis.

In the study design, factors like sample size, randomisation, and storage must all be taken into account to guarantee reproducible and successful experiments that minimise erroneous variability, and yet highlight the metabolites of interest and their potential interactions (Figure 5).
Noise (or error) is an important consideration to factor in because it distorts the signals in your data. There are two types of noise:

- Random noise - this results from contaminants and general technological limitations. It produces signal spikes and discontinuous data that could be mistaken for meaningful data.
- Systematic noise - this results from external factors that are not relevant for the study. Baseline drift is one example of systematic noise and is a common problem in liquid chromatography-mass spectrometry (LC-MS [14]) where the gradient of the mobile phase [15] causes the chromatographic baseline to be irregular.

Sample preparation

Sample preparation usually involves the following steps (Figure 6):

- collection;
- storage;
- extraction;
- preparation;
custom preparation for individual measurement systems, e.g. derivatisation [16] for gas chromatography.

**Extraction techniques**

**Solid-phase extraction (SPE)**

SPE is a process whereby compounds which are dissolved or suspended in a liquid mixture are separated from other compounds according to their chemical and physical properties. SPE is often used in metabolomics [4] laboratories to concentrate and purify a sample.

**Chromatography**

Chromatography is an important step in metabolomics experiments to separate individual metabolites from a mixture. The most common technologies used in mass spectrometry [13] are gas and liquid chromatography. Through interactions of analytes [17] with a mobile and stationary phase [18], compounds are separated and elute [19] off the chromatographic column at different time points, based on their physiochemical properties.

**Mass spectrometry**

Mass spectrometry [13] (MS) is an analytical technique used to measure small molecules [5]. The small molecules may be either directly injected into the mass spectrometer (direct infusion) or through a coupled chromatographic system. The analytes [17] are ionised at an ion source before they can be detected in a coupled mass detector (Figure 7). The resulting data typically consists of mass-to-charge [20] (m/z), time, and intensity triplets that describe - for every detected ion mass - the strength of the ion beam and the time it is detected by the spectrometer.

Find out more about mass spectrometry [21], including an interactive animation of how a mass spectrometer works [22] (4 [9]).
1. Sample inlet: The port through which samples enter the mass spectrometer. A mass spectrometer can be combined with a chromatographic technique or used via direct infusion without prior separation of analytes.

2. Ion source: Ionisation techniques are grouped into hard and soft. Hard ionisation, such as electron impact ionisation (EI), heavily fragments a compound by creating high energy electrons that interact with an analyte. In contrast, soft ionisation, such as electron spray ionisation (ESI), ionises a compound but creates only a few fragments.

3. Mass analyser: Generated ions are separated by their m/z ratio in the mass analyser where – for simplicity – charge is often assumed to be equal to one. Consequently a m/z ratio approximately equals the molecular mass of an ion. All mass analysers exploit the mass and electrical charge properties of ions but use different separation methods.

4 & 5. Detector and recorder: Separated ions are detected by a mass detector that scans a pre-defined mass range at close intervals. The chromatographic profile of an ion, i.e. the generated continuous ion beam, is recorded across multiple scans at discrete time intervals.

Nuclear magnetic resonance (NMR)

NMR is an analytical technique that is used to measure organic and some inorganic compounds inside biological samples (as solid tissue or extracted metabolite). When a sample is exposed to a magnetic field and radio frequency (rf) pulse, the nuclei absorb and re-emit this electromagnetic radiation. The energy that is emitted has a specific resonating frequency, which depends on several factors including the magnetic properties of the atoms’ isotopes and the strength of the magnetic field (usually referred to as chemical shifts). In the case of metabolomics [4], proton atoms from small molecules [5] are usually investigated (1H-NMR). For further information and animations on NMR concepts, see this website [23] (4 [9]).

NMR-based metabolomics is a non-invasive and non-destructive technique with high reproducibility, making it a powerful tool for searching new and novel biomarkers. For NMR, we measure the resulting signal from small molecules’ protons resonating within a magnetic field. One of the first usages of NMR was to detect metabolites in unmodified biological samples (5 [9]).

Comparison of NMR and MS

The two most common techniques used in data acquisition are nuclear magnetic resonance [24] and mass spectrometry [13]. Table 1 shows some of the key differences between the two techniques.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Contrasting some of the advantages and disadvantages of NMR and MS.</th>
<th>Nuclear magnetic resonance (NMR)</th>
<th>Mass spectrometry</th>
</tr>
</thead>
</table>


<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>Very high</td>
<td>Average</td>
</tr>
<tr>
<td>Number of detectable metabolites</td>
<td>30-100</td>
<td>300-1000+ (depending on whether GC-MS or LC-MS is used)</td>
</tr>
<tr>
<td>Targeted analysis</td>
<td>Not optimal for targeted analysis</td>
<td>Better for targeted analysis</td>
</tr>
<tr>
<td>Sample preparation</td>
<td>Minimal sample preparation required</td>
<td>More complex sample preparation</td>
</tr>
<tr>
<td>Tissue extraction</td>
<td>Not required. Tissues can be analysed directly</td>
<td>Requires tissue extraction</td>
</tr>
<tr>
<td>Sample analysis time</td>
<td>Fast. The whole sample can be analysed in one measurement</td>
<td>Takes longer than NMR. Requires different chromatography techniques depending on type of metabolites</td>
</tr>
<tr>
<td>Instrument Cost</td>
<td>More expensive and occupies more space than MS</td>
<td>Cheaper and occupies less space</td>
</tr>
<tr>
<td>Sample Cost</td>
<td>Low cost per sample</td>
<td>High cost per sample</td>
</tr>
</tbody>
</table>

### Data processing

Data processing aims to extract [25] biologically relevant information from the acquired data. It includes many steps that are similar for MS and NMR. A good understanding of the steps involved is important in order to minimise the risk of skewed or false results. Typically, the endpoint of MS and NMR metabolomics [4] studies is an (annotated) feature matrix (Figure 8). A feature is typically a peak or signal that represents a chemical compound. Thus, a feature matrix contains the intensities or (relative) abundances of relevant signals for every sample, describing the metabolomics fingerprint. Ultimately, this feature list would become a list of identified metabolites with semi-quantified or quantified values.
Figure 8 Example of an MS feature matrix. Transpositions of the matrix are also common.

To compile a feature matrix, noise reduction and background correction are essential before features can be extracted via peak picking (Figure 8). This process greatly tidies up the data. Extracted features of individual samples are then aligned across samples to compensate for drifts in the chemical shift (NMR) or retention time (MS) (see Figure 9). Aligned features can then be aggregated in a feature matrix: a feature has a characteristic chemical shift (NMR) or mass (MS) that can be used as column header. The rows represent individual samples.
Figure 9 A summary of components contributing to signal distortions. (a) Random noise adds variation to a signal around the mean (zero). (b) Systematic noise, e.g. baseline drifts, introduces a systematic drift or bias in the data that needs to be removed before data analysis. Systematic noise can impact heavily on signal intensities and derived signal areas. (c) The actual signal follows – in theory – a Gaussian distribution. Deviations from this distribution reflect external factors. (d) Overlay of components (a), (b), and (c), and the resulting 'measured' signal (black line).

Noise (or error) is an important consideration to factor in because it distorts the signals in your data. There are two types of noise:

- Random noise - this results from contaminants and general technological limitations. It produces signal spikes and discontinuous data that could be mistaken for meaningful data.
- Systematic noise - this results from external factors that are not relevant for the study. Baseline drift is one example of systematic noise and is a common problem in liquid chromatography-mass spectrometry (LC-MS [14]) where the gradient of the mobile phase [15] causes the chromatographic baseline to be irregular (Figure 9d).

Analysis and interpretation

In the context of metabolomics [4], the most common statistical analysis approaches are grouped into univariate
and multivariate methods. Each method offers unique insights into the data structure. Multivariate analysis works on a matrix of variables and highlights characteristics based on the relationships between all variables. Univariate analysis takes only one variable into account, resulting in differently weighted results.

The goal of statistical analysis is the categorisation and prediction of sample properties through generation of models that capture the information contained in data matrices. In mass spectrometry [13], the m/z ratio and signal intensity are the two most important variables. In NMR we select integrated signals of interest for data analysis.
Metabolomics: An introduction
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

Figure 10 (a) Example PCA plot of three batches (red, green, blue). The red batch exhibits a strong batch effect. (b) Pearson’s correlation heatmap.

Without venturing too much into the area of statistics, principal component analysis (PCA, figure 10a) and partial least squares (PLS) are established methods for multivariate analysis of metabolomics data. PCA is a method that enables us to reduce the dimensionality of our data into inferred variables, thus helping us to identify major trends and features.

The dimensionality-reduction methods can be used in classification, regression, and prediction exercises. The quality of the statistical models that we infer depends significantly on the data pre-processing, scaling and normalisation methods used. Therefore successful data analysis requires careful investigation of multiple models for consensus building (i.e. don’t rely on a single model!).

Figure 10b shows a correlation heatmap of a feature matrix, typically used in metabolomics analysis.

More information about data analysis and interpretation of metabolomics studies can be found in the materials from the first three days of our 2014 EMBO Practical Course on Metabolomics Bioinformatics for Life Scientists.

Metabolomics resources at the EBI

EMBL-EBI has three main resources that cover the field of metabolomics (Figure 11):

1. **MetaboLights** [30] is a general purpose, open access repository for metabolomics studies, their raw experimental data and associated metadata [32].

2. **ChEBI** [33] is a freely available database of Chemical Entities of Biological Interest, containing manually annotated small molecular entities (molecules not encoded by the genome)

3. **Reactome** [34] is a freely available curated database of pathways and reactions. You can browse Reactome and gain a detailed view of the molecular processes of human biology. Reactome represents biological processes as interconnected molecular events or ‘reactions’, covering all types of biological molecules, primarily proteins and small molecules [5].
Summary

- **Metabolomics** [4] is the large-scale study of small molecules [5] (metabolites) within cells, tissues or organisms.
- Applications of metabolomics are found within the pharmaceutical, healthcare, and agricultural industries, among others.
- There are two main approaches used in metabolomic studies: untargeted (global) and targeted (specific).
- Careful planning and design of experiments is of paramount importance in metabolomic studies.
- **Nuclear magnetic resonance** [24] (NMR) and **mass spectrometry** [13] (MS) are two of the most commonly used analytical methods in metabolomic studies.
- EMBL-EBI has three main resources that cover the field of metabolomics: **Metabolights** [30], **ChEBI** [35] and **Reactome** [34].

Quiz: Metabolomics quiz

Questions: 8
Attempts allowed: Unlimited
Available: Always
Pass rate: 75%
Backwards navigation: Allowed

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Learn more

Find out more

Review articles:

Recommended courses

Train online courses available:

- Metabolights: quick tour [38]
- ChEBI [39]: quick tour [40]
- Reactome [10]: quick tour [41]
- Reactome: exploring and analysing biological pathways [42]

Face to face workshops:

EMBL-EBI hosts an annual EMBO Practical Course on Metabolomics Bioinformatics for Life Scientists. Check the Training page for more information about this course [43].

References


3. The Human Metabolome Project [46].


Contributors
Reza Salek [1]

EMBL-EBI
Project Manager (COSMOS), MetaboLights Senior Scientific Database Curator: Cheminformatics and metabolism

Reza Salek was awarded his PhD in Molecular Biophysics and Biochemistry from University College London, UK. Currently, he is managing and coordinating a large EU infrastructure project on metabolomics standards COSMOS (Coordination of Standards in Metabolomics) as well MetaboLights, the first open-access, general-purpose metabolomics repository curation efforts. COSMOS aims to set and promote community standards that will make it easier to disseminate metabolomics data through life science e-infrastructures. In the past Reza Salek has worked as scientific investigator at the Medical Research Council HNR unit in the Lipid Signalling group. Reza Salek is also a research visitor and former research associate at the University of Cambridge, Biochemistry Department. In addition, he is member of the Cambridge Systems Biology Centre, Cambridge Neuroscience and Cambridge Cancer Centre. In the past he has been involved in large national and international consortia and metabolomics projects on breast cancer biomarker discovery based on integrated NMR and mass spectrometry metabolomics (MetaCancer- [49]), animal models and clinical trials on Type 2 diabetes (GSK) and animal models of neurological disorders (Babraham Institute) during his work at Cambridge University, gaining insight working both within academic and industrial environments. Reza is also the main organizer of the "EMBO Practical Course on Metabolomics Bioinformatics for Life Scientists" for 2013 and 2014.

Stephan Beisken [3]

Laura Emery [2]

EMBL-EBI
Scientific Training Officer - Training

Laura completed her undergraduate degree in Genetics (BSc first class hons) at the University of Nottingham. After dabbling in experimental work on a summer placement at the Wellcome Trust Sanger Institute, she found joy in using computer-based analyses to explore biological questions. Laura completed her PhD in Evolutionary Genetics at the University of Edinburgh’s Institute of Evolutionary Biology. There she used a bioinformatics approach to investigate patterns of codon usage in Archaea while under the supervision of Prof Paul Sharp FRS.

She also worked as a postdoc at the University of Manchester, where she explored molecular co-evolution among interacting proteins. Having obtained qualified teacher status (QTS), Laura joined EMBL-EBI as a Scientific Training Officer in 2012 and is responsible for the development and delivery of training courses.
Metabolomics: An introduction
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

Richard Grandison [50]
EMBL-EBI
Scientific Training Officer - eLearning

Richard is the Scientific Training Officer (eLearning) for the Training Team at EMBL-EBI. He is responsible for monitoring and maintaining EMBL-EBI’s Train online eLearning portal and developing new online training courses in collaboration with subject-matter experts.

Richard has a BSc in Zoology and a PhD in the biology of ageing from University College London. Before joining the team, he worked at the Royal Society of Chemistry to develop eLearning resources and websites for chemistry students and teachers. Richard has also worked as an online editor for Faculty of 1000 and as a medical writer.

Source URL: https://www.ebi.ac.uk/training/online/course/introduction-metabolomics

Links
[1] https://www.ebi.ac.uk/training/online/trainers/reza.salek
[2] https://www.ebi.ac.uk/training/online/trainers/laura.emery
[3] https://www.ebi.ac.uk/training/online/trainers/SBeisken
[4] https://www.ebi.ac.uk/training/online/glossary/metabolomics
[5] https://www.ebi.ac.uk/training/online/glossary/small-molecules
[6] https://www.ebi.ac.uk/training/online/glossary/transcriptomics
[7] https://www.ebi.ac.uk/training/online/glossary/metabolism
[8] https://www.ebi.ac.uk/training/online/glossary/phenotype
[9] https://www.ebi.ac.uk/training/online/course/introduction-metabolomics/references
[10] https://www.ebi.ac.uk/training/online/glossary/reactome
[11] https://www.ebi.ac.uk/training/online/glossary/anabolic
[12] https://www.ebi.ac.uk/training/online/glossary/catabolic
[13] https://www.ebi.ac.uk/training/online/glossary/mass-spectrometry
[14] https://www.ebi.ac.uk/training/online/glossary/lc-ms
[15] https://www.ebi.ac.uk/training/online/glossary/mobile-phase
[16] https://www.ebi.ac.uk/training/online/glossary/derivatisation
[17] https://www.ebi.ac.uk/training/online/glossary/analytes
[18] https://www.ebi.ac.uk/training/online/glossary/stationary-phase
[19] https://www.ebi.ac.uk/training/online/glossary/elute
[20] https://www.ebi.ac.uk/training/online/glossary/mass-charge
[23] http://www.rsc.org/learn-chemistry/collections/spectroscopy/introduction#NMRSpectroscopy
[24] https://www.ebi.ac.uk/training/online/glossary/nuclear-magnetic-resonance
[25] https://www.ebi.ac.uk/training/online/glossary/extract
[26] https://www.ebi.ac.uk/training/online/glossary/pre-processing
[27] https://www.ebi.ac.uk/training/online/glossary/normalisation-methods
[28] https://www.ebi.ac.uk/training/online/glossary/heatmap