Basics of Metabolomics

Vladimir Tolstikov, Ph.D.

TOXI-LATIN 2014
1st Latin America Congress of Clinical and Laboratorial Toxicology
(April 27- May 2, 2014)
Universidade Federal do Pampa, Porto Alegre, Brazil
Metabolomics

PubMed shows an exponential growth in the number of publications
What is a Metabolite?

- Any organic molecule detectable in the organism with a MW < 2000 Da
- Includes amino acids, sugars, nucelosides, nucleotides, organic acids, fatty acids, ketones, aldehydes, amines, peptides, oligonucleotides, lipids, steroids, alkaloids, foods, food additives, toxins, pollutants, drugs and drug metabolites
- Includes human & microbial products
- Endogenous metabolites - produced by the host organism
- Exogenous metabolites – not produced by the host organism
What is a Metabolome?

• The complete collection of small molecule metabolites in a cell, organ, tissue or organism.

• Includes endogenous and exogenous molecules as well as transient molecules.

• In many cases defined by the detection technology.
Metabolic pathways
Omics by the numbers

- **25,000 Genes**
- **1,000,000 Proteins**
- **9,000 Chemicals**

Environmental Influence

- **Genomics**
- **Proteomics**
- **Metabolomics**
Systems Biology

- DNA
  - Genomics
- RNA
  - Epigenomics
  - Transcriptomics
- Proteins/Enzymes
  - Proteomics
- Metabolites
  - Metabolomics
  - Lipidomics

Environment
### Table 1. High-Throughput “Omics” Approaches in Integrative Biology

<table>
<thead>
<tr>
<th>Technology</th>
<th>Molecule</th>
<th>Knowledge</th>
<th>Limits</th>
<th>Clinical applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomics</td>
<td>DNA</td>
<td>Genetic polymorphisms, haplotypes, full genome sequence through next-generation sequencing</td>
<td>Restricted to genetic determinants, ignores the environment</td>
<td>Genome-wide association studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Personal genomics</td>
</tr>
<tr>
<td>Transcriptomics</td>
<td>RNA</td>
<td>Expression patterns</td>
<td>Expression does not necessarily match protein abundance</td>
<td>Blood and biopsy transcript signatures</td>
</tr>
<tr>
<td>Proteomics</td>
<td>Proteins</td>
<td>Targeted and untargeted protein abundance profiles</td>
<td>Protein abundance does not mean function</td>
<td>Numerous clinical applications for measurement of circulating endogenous protein markers</td>
</tr>
<tr>
<td>Metabonomics</td>
<td>Metabolites</td>
<td>Intermediary phenotypes related to metabolism in the absence of genetic information</td>
<td>Structural assignment can be challenging</td>
<td>Metabolome-wide association studies and metabolomic GWAS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Observation of effects from host genetics, lifestyle, and environment (including microbiome)</td>
<td>Trade-off between targeted and untargeted approaches (ie, precise knowledge of detected metabolites vs extensive coverage of the metabolome)</td>
<td>Patient stratification through pharmacometabonomic approaches</td>
</tr>
</tbody>
</table>

---

**REVIEWS IN BASIC AND CLINICAL GASTROENTEROLOGY AND HEPATOLOGY**

**Metabolic Phenotyping and Systems Biology Approaches to Understanding Metabolic Syndrome and Fatty Liver Disease**

Marc-Emmanuel Dumas,¹ James Kinross,¹,² and Jeremy K. Nicholson¹
Human Metabolomes

- 3100 (T3DB) - Toxins/Env. Chemicals
- 1000 (DrugBank) - Drug metabolites
- 30000 (FooDB) - Food additives/Phytochemicals
- 1450 (DrugBank) - Drugs
- 8500 (HMDB) - Endogenous metabolites

Concentration units: M, mM, μM, nM, pM, fM
Why Metabolomics is Difficult

- 4 Bases
- 20 Amino acids
- $9 \times 10^3$ Chemicals

Chemical Diversity

Metabolomics → Proteomics → Genomics
Limitations of metabolomics

• Sensitivity to environmental impact. High biological variance in metabolite levels (i.e., the variation between genetically identical organisms grown in the same conditions). Biological replicas are needed for statistics.

• Unlike nucleic acids and proteins, metabolites have a vast range of chemical structures and properties. Their molecular weights span two orders of magnitude (20–2000 Da). Therefore no single extraction protocol and/or analytical method works for all metabolites. (Unlike DNA sequencing, microarrays, MS analysis of proteins – all are general methods.)

• The concentrations of various metabolites can vary dramatically from mM to pM concentrations.

• Some metabolites are labile and won’t survive extraction and analysis.

• Some transient metabolites are extremely low abundant and hardly can be detected, until appropriate downstream enzyme activity is inhibited.
Metabolomics Technologies

ASMS Metabolomics Workshop Survey

- LC/MS: 173
- GC/MS: 87
- CE/MS: 25
- NMR: 21
- Other: 9
- LC/NMR: 4
Technology & Sensitivity

Knowns

Unknowns

# Metabolites or Features detected (\(\log_{10}\))

Sensitivity or LDL

M, mM, \(\mu\)M, nM, pM, fM
Modern NMR for Metabolomics

The high reproducibility of NMR-based techniques gives this method a number of advantages over other analytical techniques in large-scale and long-term metabolomic studies, such as epidemiological studies.

High-Resolution NMR Spectrometers:
300 MHz (7Tesla)-900 MHz (21Tesla).
Cryo-probe, flow through sampling.

In vivo MRI/MRS Scanners:
1.5 Tesla (80 MHz) – 7 Tesla (300 MHz)
4.7 Tesla (200 MHz) – 14 Tesla (600MHz)

What are NMR limitations? SENSITIVITY!!!!!
-NMR detects only high-abundant metabolites (micromole to millimole range);
-NMR suffers from signal overlap between individual metabolites;
Modern Mass Spectrometry For Metabolomics

- GC-MS \( \rightarrow \) GC-EI/CI-HRMS \(< 1\text{ppm}\)
- CE-MS \( \rightarrow \) CE-HRMS \(< 5\text{ppm}\)
  Inductive ionization (no contact)
- LC-MS \( \rightarrow \) LC-HRMS \(< 5\text{ppm}\)
- LC-MS/MS systems \( \rightarrow \) 10-50 folds more sensitive.
- HPLC \( \rightarrow \) UPLC and HILIC.
- Lipidomics \( \rightarrow \) infusion-shotgun
- LC-MS-NMR \( \rightarrow \) LC-HRMS-HRNMR

What are MS limitations? - A great deal of information.
Metabolomics Workflow

1. Biological or Tissue Samples
2. Extraction
3. Biofluids or Extracts
4. Data Analysis
5. Chemical Analysis
Gas Chromatography coupled to Mass Spectrometry

Chemical derivatization is needed for molecules to be volatile.

Quantitation: peak area
GS-MS mass chromatogram of human blood plasma

More than decade data collection in NIST MS database >100000 spectra

Table III: Representative compounds in NIST human plasma

<table>
<thead>
<tr>
<th>Peak</th>
<th>Name</th>
<th>Formula</th>
<th>R.T. (s)</th>
<th>Area</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Phosphate (3 TMS)</td>
<td>C₆H₂₂O₄PSi₃</td>
<td>533</td>
<td>10740930</td>
<td>857</td>
</tr>
<tr>
<td>2</td>
<td>Threonine (3 TMS)</td>
<td>C₁₃H₂₃NO₂Si₂</td>
<td>583</td>
<td>1694554</td>
<td>894</td>
</tr>
<tr>
<td>3</td>
<td>5-Oxo-proline (2 TMS)</td>
<td>C₁₃H₁₉NO₂Si₂</td>
<td>639</td>
<td>8477106</td>
<td>912</td>
</tr>
<tr>
<td>4</td>
<td>Creatinine (3 TMS)</td>
<td>C₁₃H₁₃N₃O₃Si₃</td>
<td>652</td>
<td>2302177</td>
<td>866</td>
</tr>
<tr>
<td>5</td>
<td>Citric acid (4 TMS)</td>
<td>C₁₈H₄₀O₇Si₄</td>
<td>739</td>
<td>824486</td>
<td>851</td>
</tr>
<tr>
<td>6</td>
<td>Glucose (NMOSO, 5 TMS)</td>
<td>C₂₅H₅₉NO₅Si₈</td>
<td>805</td>
<td>20744340</td>
<td>829</td>
</tr>
<tr>
<td>7</td>
<td>Inositol (6 TMS)</td>
<td>C₂₄H₆₀O₆Si₆</td>
<td>823</td>
<td>1093541</td>
<td>895</td>
</tr>
<tr>
<td>8</td>
<td>Arachidonic acid (TMS)</td>
<td>C₂₃H₄₀O₂Si</td>
<td>898</td>
<td>597056</td>
<td>848</td>
</tr>
<tr>
<td>9</td>
<td>α-Tocopherol (TMS)</td>
<td>C₃₂H₅₆O₂Si</td>
<td>1148</td>
<td>266286</td>
<td>849</td>
</tr>
<tr>
<td>10</td>
<td>Cholesterol (TMS)</td>
<td>C₃₆H₅₄O₄Si</td>
<td>1164</td>
<td>11656743</td>
<td>915</td>
</tr>
</tbody>
</table>

Liquid Chromatography coupled to Mass Spectrometry

Quantitation: peak area
HILIC versus RP separations

LC-MS General Profiling
High Dimension Data

Data extraction: peaks alignment, noise reduction, redundant data removal
Data preprocessing: normalization, transformation, missing values and filtering
**Unknown metabolite** identification is a very complex process which requires at least MS and NMR data of high resolution and high quality. Lucky match with the described in literature/library chemical has probability score.

1. Physicochemical properties: retention time, formation of positive and negative ions, adducts formation.
4. Parent ion fragmentation for positive and negative ions. Adducts fragmentation.
5. $^1$H-NMR.
6. NMR correlation experiments.
MS/MS fragmentation for identification

PC (34:2) [M+H]+ m/z 758.57
parent ion

MI = 758.57 Da

MI = 502.33 Da

MI = 520.34 Da

daughter ions

MI = 496.34 Da

MI = 478.33 Da
LC-HRMS - Identification

Carnitines

daughter ions

parent ion
Targeted Metabolomics

Identification: retention time and parent ion/daughter ion
Quantitation: peak area

- Parent ion
  - Q1 selects and passes only precursor of interest
  - Q2 selected precursor is fragmented
  - Q3: monitor specific fragment ions

- Daughter ion

Extracted Ion Chromatograms
Fig. 1. Endogenous metabolite concentrations in serum/plasma determined by targeted and non-targeted approaches with LC-MS/MS.
Metabolomics Data

- Preprocessing: normalization, transformation, missing values and filtering
- Statistical analysis: unsupervised techniques, supervised techniques
Statistical analysis in Metabolomics

• Metabolomics experiments usually result in a large quantity of data. Univariate and multivariate analysis techniques are routinely used to extract relevant information from the data with the aim of providing biological knowledge on the problem studied. Statistical tools like the t-test, analysis of variance, principal component analysis, and partial least squares discriminant analysis are major components of metabolomics data analysis.

• The challenge is to identify those metabolites that are related to the phenomenon, and isolate them from the ones whose variation is not related to the phenomenon.

• When only one variable is analyzed at a time (in omics disciplines usually one out of a panel of many measured), a so-called univariate analysis is performed. Univariate methods include tests to compare different sets of samples such as t-test or ANOVA.

• When in the course of an experiment two or more variables are measured the resulting data are multivariate data. Principal component analysis (PCA), and partial least squares discriminant analysis (PLS-DA) are routinely applied in metabolomics.

We have to distinguish between quality assurance as best scientific practice for within-laboratory use and the regulatory use, which will typically require formal validation and acceptance of the method and a formal quality regime such as Good Laboratory Practice. Quality systems (QS) were developed originally for industrial production but they were increasingly introduced in many other sectors and organizations to establish a formal structure for establishing quality criteria.

Quality control procedures also include the estimation of the stability of the analytical procedure, the use of standards, error estimation of data reproducibility and criteria for data inclusion (and exclusion).

Metabolic biomarkers

- Metabolic biomarkers – Metabolites reporting on a biological state, i.e. "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention."

- The term metabolomics has been recently introduced to address the global analysis of all metabolites in a biological sample. In general disease/drug/chemical impacts some of the metabolic pathway, this parameter can be used as a marker.

- Example: Serotonin production pathway activated in alcoholic drinking person it can be metabolic marker of recent alcohol consumption.
Metabolic biomarkers

- Clinical Study in Over 4,000 Subjects Finds Blood TMAO Levels Linked to Increased Risk of Heart Disease, Even In the Absence of Known Cardiovascular Risks.
- The current study is an extension of Dr. Hazen’s previous work, in which he found that a chemical byproduct called trimethylamine N-oxide (TMAO) is produced when intestinal bacteria digest the nutrient phosphatidylcholine, commonly known as lecithin. The prior research showed that TMAO levels in the blood were associated with heart disease. Dr. Hazen and colleagues have now confirmed that gut flora are essential in forming TMAO in humans and demonstrated a relationship between TMAO levels and future cardiac events like heart attack, stroke, and death—even in those with no prior evidence of cardiac disease risk.
Metabolic biomarkers


• Biomarkers of preclinical disease will be critical to the development of disease-modifying or even preventative therapies. Unfortunately, current biomarkers for early disease, including cerebrospinal fluid tau and amyloid-β levels, structural and functional magnetic resonance imaging and the recent use of brain amyloid imaging or inflammaging, are limited because they are either invasive, time-consuming or expensive. Blood-based biomarkers may be a more attractive option, but none can currently detect preclinical Alzheimer's disease with the required sensitivity and specificity. Herein, we describe our lipidomic approach to detecting preclinical Alzheimer's disease in a group of cognitively normal older adults. We discovered and validated a set of ten lipids from peripheral blood that predicted phenoconversion to either amnestic mild cognitive impairment or Alzheimer's disease within a 2–3 year timeframe with over 90% accuracy. This biomarker panel, reflecting cell membrane integrity, may be sensitive to early neurodegeneration of preclinical Alzheimer's disease.
Case Study: PDAC
UC Davis Genome Center, USA

Discovery of *metabolic biomarkers* for pancreatic ductal adenocarcinoma (PDAC) diagnostic test development. Blood plasma profiling.

- Perform Metabolomics study on selected group cancer patients and control volunteers.

### Table 1. Clinical characteristics of pancreatic ductal adenocarcinoma (PDAC) patients and controls

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Sex</th>
<th>Race</th>
<th>Plasma sampling</th>
<th>Diagnosis</th>
<th>Jaundice</th>
<th>Cancer stage</th>
<th>CA19-9 (U/mL)</th>
<th>Diabetes &lt;3yrs</th>
<th>Smoking</th>
<th>Alcohol</th>
<th>BMI (kg/m²)</th>
<th>Family history of cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>209</td>
<td>32</td>
<td>M</td>
<td>C</td>
<td>6/5/2006</td>
<td>Normal</td>
<td>No</td>
<td>n/a</td>
<td>n/a</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>n/a</td>
<td>No</td>
</tr>
<tr>
<td>218</td>
<td>55</td>
<td>M</td>
<td>C</td>
<td>6/22/2006</td>
<td>PDAC</td>
<td>No</td>
<td>IIB</td>
<td>17182</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>28.1</td>
<td>No</td>
</tr>
<tr>
<td>Case 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>211</td>
<td>52</td>
<td>F</td>
<td>C</td>
<td>6/8/2006</td>
<td>CP</td>
<td>No</td>
<td>n/a</td>
<td>n/a</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>23.0</td>
<td>Personal breast cancer</td>
</tr>
<tr>
<td>214</td>
<td>57</td>
<td>F</td>
<td>C</td>
<td>6/14/2006</td>
<td>PDAC</td>
<td>No</td>
<td>IIB</td>
<td>1023</td>
<td>No&gt;3yrs</td>
<td>No</td>
<td>No</td>
<td>28.8</td>
<td>No</td>
</tr>
<tr>
<td>Case 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>219</td>
<td>37</td>
<td>F</td>
<td>C</td>
<td>7/17/2006</td>
<td>CP</td>
<td>No</td>
<td>n/a</td>
<td>n/a</td>
<td>No</td>
<td>No</td>
<td>Rare</td>
<td>28.6</td>
<td>No</td>
</tr>
<tr>
<td>222</td>
<td>58</td>
<td>M</td>
<td>C</td>
<td>7/27/2006</td>
<td>PDAC</td>
<td>No</td>
<td>IIB</td>
<td>n/a</td>
<td>No</td>
<td>10pk/yr</td>
<td>Rare</td>
<td>21.7</td>
<td>No</td>
</tr>
<tr>
<td>Case 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>236</td>
<td>57</td>
<td>F</td>
<td>C</td>
<td>10/2/2006</td>
<td>Normal</td>
<td>No</td>
<td>n/a</td>
<td>n/a</td>
<td>No</td>
<td>30pk/yr</td>
<td>No</td>
<td>20.5</td>
<td>No</td>
</tr>
<tr>
<td>227</td>
<td>74</td>
<td>F</td>
<td>C</td>
<td>8/4/2006</td>
<td>PDAC</td>
<td>No</td>
<td>IIB</td>
<td>35</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>n/a</td>
<td>No</td>
</tr>
<tr>
<td>Case 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>243</td>
<td>71</td>
<td>F</td>
<td>C</td>
<td>11/17/2006</td>
<td>CP</td>
<td>No</td>
<td>n/a</td>
<td>n/a</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>24.7</td>
<td>No</td>
</tr>
<tr>
<td>230</td>
<td>76</td>
<td>M</td>
<td>C</td>
<td>8/25/2006</td>
<td>PDAC</td>
<td>No</td>
<td>IIB</td>
<td>16940</td>
<td>No</td>
<td>55pk/yr</td>
<td>No</td>
<td>28.4</td>
<td>No</td>
</tr>
</tbody>
</table>
Blood Plasma Profiling PDAC

- 20 PDAC patients and 20 volunteers

HILIC-LC-MS

- Year 2006-2008 unmatched samples
- Year 2006 matched samples
Prospective Cohort Study: PDAC

• Perform Metabolomics study on larger group cancer patients (>100) and control volunteers (>100).

The metabolite specific multivariate and ROC analyses revealed specific features such as elaidic acid, uric acid, 2,3-propanediol, arachidonic acid, docosahexanoic acid, 5-oxo-EET, lysine, LysoPC(18:2), 9(10)-EpOME, LysoPC(16:0), sphingosine-1-phosphate and others as strong discriminators for recent-onset diabetic patients with pancreatic cancer. Such set yielded AUC of 0.964 with 15 features.

Conclusions

• Metabolomics evolved from routine phenotyping assay into advanced tool in biomedical research.
• Metabolomics is gaining confidence.
• Metabolomics needs advanced bioinformatics support.
• New developments and initiatives become available due to significant progress in analytical techniques and data handling methodologies.