Metabolomics by GC-MS

Sara J. Cooper
HudsonAlpha Institute for Biotechnology
Huntsville, AL

January 23, 2015
Outline

• Basics of GC-MS
  • How it works
  • How it is different from other platforms

• Applications of GC-MS for human health research
  • Designing an experiment
  • Analyzing the data (tools and tricks)
  • Signatures of Disease
  • Integrative analysis
The Nuts and Bolts of GC-MS

Sample injector

T regulated oven

Gas: He, N₂, H₂

Column: packed or open tubular (capillary)

Mass spectrometer detector

"Gcms schematic" by K. Murray (Kkmurray) - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons
The analysis is now complete.
The Nuts and Bolts of GC-MS

"Gcms schematic" by K. Murray (Kkmurray) - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons
Injection

From http://www.shsu.edu/~chemistry/GC/packed.GIF
The Nuts and Bolts of GC-MS

"Gcms schematic" by K. Murray (Kkmurray) - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons

Thursday, January 22, 15
Columns

**Packed vs. capillary GC columns**

All GC columns are open tubes. In packed column GC, the tubes are >1mm ID and the separation phase is coated on particles packed in the tube. In capillary GC, the tubes are <1mm ID and the separation phase is coated on the inside of the capillary wall.

**Packed GC columns:**
- First type of GC column
- Low efficiency
- Glass, stainless steel, nickel, copper or Teflon tubing, 1/16” – 1/4” OD
- Coated phase: Organic polymers dissolved in solvent and coated onto the particles
- Siliceous particles: diatomaceous earth for supporting coated phase
- Adsorbent particles: molecular sieve, carbon, polymers

**Capillary GC columns:**
- Modern technology
- High efficiency
- Usually flexible glass fibers (fused silica), <1mm ID
- Coated phase: Organic polymers dissolved in solvent and coated on the inside wall of the tubing

Capillary columns can be long (20-100m)

Better separation for complex mixtures
Selecting a column

A nonpolar stationary phase is used for separation of polar analytes. Thickness of the stationary phase affects retention time and column capacity. Inner diameter affects separation and retention times.

**Figure 2: Polarity scale of common stationary phases.**

In general, maximum operating temperature decreases as polarity increases. Note that silylene columns typically differ in selectivity and have higher temperature limits than their conventional counterparts.

Source: Restek
Two-dimensional chromatography

- GC Columns function in series to improve resolution of chemically similar analytes
Mass Spectrometer - Ionization and mass measurement

- Ionization
  - Electron Ionization (Standard -70keV)
    - Fragmentation
  - Chemical Ionization (less common)

- Detection
  - Time-of-flight mass spectrometry
    - mass calculated based on time from ionization to reaching detector
  - High-Resolution TOF
    - offers higher mass resolution for metabolite identification
Example data output-Chromatogram
Signal Deconvolution

Source: Leco
Principles of Deconvolution

• Generally implemented in AMDIS

• Goal: computationally separate chromatographically overlapping peaks

Source: Du and Zeisel 2013
Principles of Deconvolution
Principles of Deconvolution
Data projected into two dimensions

Glutamate

asparagine
Metabolite Identification

- reproducible fragmentation has generated libraries of known compounds

- Calculating similarity

  - Retention indices are routinely used to confirm metabolite identification based on relative retention times. (Kovats index)

  - Using a dot-product based metric, analytes can be assigned an ID based on similarity to known compounds

source: Schauer et al 2005
Metabolite ID advances

• Generation of publicly or commercially available databases
  • NIST
  • Golm
  • Fiehn ($)
• Metabolite structure prediction algorithms
  • Using clustering, modeling
• Improved algorithms for database searches
<table>
<thead>
<tr>
<th>Metabolites</th>
<th>GC</th>
<th>LC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolite ID</td>
<td>Libraries</td>
<td>Inferred composition by accurate mass</td>
</tr>
<tr>
<td>Chromatography</td>
<td>Highly reproducible-Retention indices</td>
<td>Less critical</td>
</tr>
<tr>
<td>Polarity</td>
<td>Requires derivitization to reduce polarity</td>
<td>Better for polar nucleotides, lipids (including long)</td>
</tr>
<tr>
<td>Size</td>
<td>Small</td>
<td>Medium to Large</td>
</tr>
<tr>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

**Why do GC-MS?**

- **GC**
  - Size: Small
  - Polarity: Requires derivitization to reduce polarity
  - Metabolites: a.a., organic acids, fatty acids (short-medium)
  - Chromatography: Highly reproducible-Retention indices
  - Metabolite ID: Libraries

- **LC**
  - Size: Medium to Large
  - Polarity: Better for polar nucleotides, lipids (including long)
  - Metabolites: ...
  - Chromatography: 
  - Metabolite ID:

---

Thursday, January 22, 15
Applications for GC-MS

- Petroleum and Biodiesel
- Biofluids and tissues
- Breath
- Pesticides
- Pollutants in air, soil and water
- Yeast for brewing and wine-making
So you’ve decided to do GC...what to expect

• Experimental Design!! What question(s) do you want to answer

• Sample preparation

• Data collection

• Preliminary Data analysis

  • tools

• Metabolite identification
Sample procurement/preparation

- Samples should be snap frozen as quickly as possible after extraction and stored frozen until extraction.

- Cultured cells should be grown in a minimal media if possible.
  - Avoid conditions where there are media/solvent components present at high concentration.
    - e.g. Urine samples may be treated with urease.
  - Aspiration is the best way to remove media efficiently before freezing.

- Extraction should be done under cold conditions when possible.
Gas Chromatography for Metabolomics

- Gas chromatography requires all analytes to be volatile
- Common procedure for biological samples is derivatization
- Most common method is methoximation + silylation
- Basic Protocol:
  - Dry all analytes by centrivap
  - Add methoxamine (stabilize ketones)
  - TMS reagent (generate volatile compounds)
Data collection

• You can expect anywhere from 500-5000 unfiltered peaks depending on extraction method, sample complexity and concentration

• Typical number of quantified metabolites found in the majority of samples:
  • Yeast: 150-200
  • Serum: 200-250
  • Urine: 350-500
  • Tissue: 200-300
Analyzing the Data

• Most instruments utilize proprietary software to do peak deconvolution

• Raw data can be analyzed as well and there are tools out there to analyze raw data (e.g. Metlin)

• ChromaTOF (Leco’s peak calling and deconvolution software) Output:
  • List of peaks
  • Determination of Quant Mass for each peak (unique mass, typically)
  • Quantification of metabolite (either relative to reference or absolute)
  • Library Matches for Metabolite ID
Steps to analyzing Metabolomics Data

1. Filtering Peaks
2. Alignment
3. Missing Values (Typical Data set is up to 2%)
4. Normalization
5. Statistical Analysis
Data Analysis: Filtering

Filter peaks originating from derivitization reagents or from solvent
Data Analysis: Alignment

• For each sample, determine whether every measured metabolite (from every other sample) is present

• Complex, Computationally intense problem

• Use all available information: Retention Index, (RT1 and RT2 for 2D-GC), and Spectral Match

  • MetPP, Guineu (2D GC) or MetAlign (e.g.) for GC

• Typical Result: 200-400 peaks are present in ~80% of samples-Missing values 2-5% of data
Data Analysis: Missing Values

- Conservative Filter: only consider metabolites present in the VAST majority of the samples (~95%)

- Assuming missing values are below detectable levels (0.5x lowest value for that metabolite)

- Assume missing values are present at an average or median level

- K nearest neighbor estimation-characterizes what values are present in other samples with the most highly correlated values for other metabolites to estimate a likely concentration

Limited to small number of metabolites (High Confidence)

Can skew results if there are a large number of missing values

Conservative, but can skew data

Moderately conservative, but not possible if missing data is abundant
Data Analysis: Normalization

- Common Practice:
  - Injection Control (A known amount of substance is injected with each sample. Those peaks should have the same area each time)
  - Normalization by SUM (total area under the curve). Normalizes for overall sample concentration
  - Clinical samples: normalization by creatinine or other specific analytes (not ideal for research, but sometimes necessary depending on application)
Data Analysis: Statistical Analysis

• A wide variety of tools and packages available

• Metaboanalyst is a great place to start (R-package in web-based app)
  • Upload your aligned data in .csv or .txt format. It goes through the normalization, missing data and filtering steps and then allows a variety of analysis
    • Heatmaps, Clustering
    • PCA
    • PLS-DA
    • T-tests (paired and unpaired)
    • Some pathway analysis
    • etc.

www.metaboanalyst.ca
Please choose a functional module to proceed:

- **Statistical Analysis**
  This module offers various commonly used statistical and machine learning methods from Mests, ANOVA to PCA and PLS-DA. It also provides clustering and visualization such as dendrogram, heatmap, K-means, as well as classification based on random forests and SVM.

- **Enrichment Analysis**
  This module performs metabolite set enrichment analysis (MSEA) for human and mammalian species based on several libraries containing ~6300 groups of biologically meaningful metabolite sets. Users can upload a list of compounds, a list of compounds with concentrations, or a concentration table.

- **Pathway Analysis**
  This module supports pathway analysis (integrating enrichment analysis and pathway topology analysis) and visualization for 21 model organisms, including Human, Mouse, Rat, Cow, Chicken, Zebrafish, Arabidopsis thaliana, Rice, Drosophila, Malaria, Budding yeast, E.coli, etc., with a total of ~1600 metabolic pathways.

- **Time Series Analysis**
  This module supports data overview (PCA and heatmaps), two-way ANOVA, multivariate empirical Bayes time-series analysis for detecting distinctive temporal profiles across different experimental conditions, and ANOVA-simultaneous component analysis (ASCA) for identification of major patterns associated with each experimental factor.

- **Power Analysis**
  This module allows you to upload a pilot data set to calculate the minimum number of samples required to detect the existence of a difference between two populations with a given degree of confidence.

- **Biomarker Analysis**
  To perform various ROC curve based biomarker analysis. It supports classical single biomarker analysis, multivariate biomarker analysis, and manual biomarker selection and evaluation.

- **Other Utilities**
  This module contains some utility functions commonly
Input test dataset (Cancer patients Cachexic v. control)

<table>
<thead>
<tr>
<th>Compounds</th>
<th>p-value</th>
<th>FC</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uracil</td>
<td>3.84E-04</td>
<td>3.4154</td>
<td>0.024204</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>0.0011396</td>
<td>2.9432</td>
<td>0.035898</td>
</tr>
<tr>
<td>Acetone</td>
<td>0.0051404</td>
<td>2.289</td>
<td>0.10795</td>
</tr>
<tr>
<td>Succinate</td>
<td>0.013088</td>
<td>1.8831</td>
<td>0.1502</td>
</tr>
<tr>
<td>4-Hydroxyphenylacetate</td>
<td>0.013611</td>
<td>1.8661</td>
<td>0.1502</td>
</tr>
<tr>
<td>Hypoxanthine</td>
<td>0.015669</td>
<td>1.805</td>
<td>0.1502</td>
</tr>
<tr>
<td>Methylguanidine</td>
<td>0.016881</td>
<td>1.7726</td>
<td>0.1502</td>
</tr>
<tr>
<td>Pantothenate</td>
<td>0.019073</td>
<td>1.7196</td>
<td>0.1502</td>
</tr>
<tr>
<td>Glucose</td>
<td>0.038618</td>
<td>1.4132</td>
<td>0.25269</td>
</tr>
<tr>
<td>Creatine</td>
<td>0.04011</td>
<td>1.3967</td>
<td>0.25269</td>
</tr>
</tbody>
</table>
Sample Data-top25 features by Ttest
Pathway Analysis

Glycine, Serine, Threonine

Alanine/Aspartate

Pantothenate and CoA

Inositol Phosphate
Data Analysis: Biological Understanding

• Web-based tools for pathway analysis
  • KEGG (KEGGMapper) (all organisms)
  • HMDB (Human Metabolome Database)
    • Serum, urine, metabolome databases
  • Yeast- Biochemical Pathways at yeastgenome.org
    • ymdb (yeast metabolome database)
• Integrated analysis with genomic, proteomic data
  • IMPaLA (similar to GO enrichment but specific to metabolic pathways)
  • Ingenuity ($$$)
  • Metaboanalyst (new)
Resources for GC-MS

- Leco
- Agilent
- Sigma [https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Aldrich/Bulletin/1/the-basics-of-gc.pdf](https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Aldrich/Bulletin/1/the-basics-of-gc.pdf)
- Books, Chapters, Reviews:
  - *Metabolomics* by Wofram Weckwerth (Methods and Protocols)
- Analysis
  - Metaboanalyst.ca
  - impala.molgen.mpg.de
  - hmdb.ca
  - golm database: gmd.mpimp-golmmpg.de
  - metlin.scripps.edu
  - xcmsonline.scripps.edu
BREAK for questions
Biology’s central dogma

DNA → RNA → Protein

DNA sequencing → RNA sequencing → Proteins: proteomics

Metabolomics: Mass Spectrometry, NMR
Small molecules as sensors

DNA → RNA → Protein

Amino Acids

Sugars

Lipids

Organic Acids

Nucleotides

Translation

Signaling

Energy source

Energy storage

Cellular structure

ATP Synthesis

Phenotype

Replication

Transcription

Translation

Signaling

Energy source

Energy storage

Cellular structure

ATP Synthesis

Replication

Transcription

Phenotype
Part II: Using Metabolomics in biological research

• Yeast Phenomics

• Pancreatic Cancer
Yeast phenomics
Integrating data

7078 phenotypes correlated to at least one other phenotype
Genetic associations

Manhattan plot of significantly associated SNPs with peptides and transcripts
A metabolite with heritable variation: Ribose
Summary

• Integrating metabolomics with genomics and proteomics data-a model for integrated human studies

• Applying metabolomics to improve understanding of pancreatic cancer
The Role of Metabolism in Pancreatic Cancer
Using genomics and metabolomics to improve human health

Healthy

Prevention:
What causes disease

Diagnosis:
What happens early in disease

Sick

Treatment:
How to treat people (individuals)
Pancreatic Cancer

Rare cancer, but accounts for 4th most cancer deaths in US

- 43,920 new cases in 2012
- 37,390 deaths
- Only cancer whose incidence and death rate is increasing

Thursday, January 22, 15
## Pancreatic Cancer Statistics

<table>
<thead>
<tr>
<th>Stage at diagnosis</th>
<th>Stage distribution %</th>
<th>5-year survival (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Localized</td>
<td>8</td>
<td>23.3</td>
</tr>
<tr>
<td>Regional (spread to lymph nodes)</td>
<td>27</td>
<td>8.9</td>
</tr>
<tr>
<td>Distant (metastatic)</td>
<td>53</td>
<td>1.8</td>
</tr>
<tr>
<td>Unknown</td>
<td>12</td>
<td>3.9</td>
</tr>
</tbody>
</table>

Statistics from cancer.gov

- Extremely aggressive
  1) Early detection is unusual
  2) Limited treatment options for advanced stage cancer (no cures)
  3) Resistant to chemotherapy
Use genomic technologies to improve Pancreatic Cancer patient outcomes

Solutions:

1) Better diagnostic markers
2) Improved and/or personalized treatment options
A role for metabolism in pancreatic cancer

1. Identify metabolic changes in serum and urine from pancreatic cancer patients

2. Determine whether those metabolic changes represent metabolic changes in the pancreatic tumor

3. Determine whether alterations in metabolic pathway correlate with outcome
Measuring metabolites

Analysis

Sample 1  Sample 2

3 1
1 1
1 3
Analyses

• Directed-Known pathways PC v. Normal
• Unbiased-most significant differences between classes
• Metabolites/pathways changing with
  • stage
  • metastasis
TCA cycle

- Warburg effect
- Known mutations occurring in cancer
  - isocitrate dehydrogenase
  - fumarate hydratase
  - pyruvate kinase
  - succinate dehydrogenase

Review: Wu and Zhao 2012
Urine-TCA cycle

p=0.24

p=0.001

p=6.6x10^-4

p=0.38

p=0.01
Most significant effects

Glutamine

Glycine

Pancreatic Cancer cells are characterized by their “glutamine addiction”

Glycine has previously been shown by Mootha et al to correlate to proliferation in NCI-60 panel & survival in breast cancer patients.
Multi- “Omics” approach

• RNA-Seq was performed on tumor tissues and neighboring normal/benign tissue

• Revealed over 6000 significantly changing genes between tumor and normal tissue

• Which of these is important???
Leveraging gene expression information to focus on vital metabolic pathways

- Is there evidence of altered metabolic pathways in gene expression data?
- Are the same pathways we identified in blood and urine changing in tumor samples?
- What do we learn by intersecting these data?
Pancreatic Cancer—Integrating Metabolomics and Genomics

Metabolic changes in pancreatic cancer

Common Pathways
Identification of pathways important to tumor growth and patient survival

Gene expression changes in pancreatic cancer

Samples

Tumor tissue
Normal tissue

Gene expression changes in pancreatic cancer

Serum from pancreatic cancer patients
Glycine, Threonine, Serine Synthesis

Serum Glycine, Threonine and Serine Pathway

Urine Glycine, Threonine, Serine Pathway

Samples

Genes
Glycine pathway gene expression associated with poor prognosis

**Glycine Metabolism Genes**

- **GNMT**: Survive < 2 years
- **GAMT**: Survive > 2 years

**Chemical Structures**
- Glycine
- Guanadinoacetate
- Creatine
Fatty Acid Biosynthesis

Urine Fatty Acids

Serum Fatty Acids

Samples

Genes

Thursday, January 22, 15
A correlation between survival and lipase expression

Early vs. Late Survivors

Lipase Transcripts

Survive < 2 years
Survive > 2 years

FPKM values for LPL, CEL, CLPS, CLPS, PNLIP, PNLIPRP1
Replication in independent samples

ICGC Australian Cohort Lipase Survival

TCGA Lipase Survival

Survived < 1 year
Survived > 2 years
A pathway to link these pathways: Sphingolipid Biosynthesis
Is the link important to disease?

• If ceramide/sphingolipid biosynthesis is essential for apoptosis in cancer cells, they should reduce ceramide production, perhaps through downregulation of lipase genes.

• We can test in vitro whether apoptosis in cancer cell lines is sensitive to fatty acid concentration, and whether apoptosis requires ceramide production.
Future Directions

• We are at the beginning:
  • Thousands of differentially expressed genes
  • Dozens of differentially abundant metabolites
  • How is it all connected: regulation???
  • Lipase genes (and other fatty acid biosynthesis genes) are regulated by
Potential key regulators

PPAR Gamma Expression

GATA 2 Expression
Acknowledgements

S. Cooper Lab
Karl Ackerman (UAB)
Bobbi Johnston
Karin Bosma (Alum)
Ryne Ramaker (UAB)
Rebecca Hauser

Pancreatic Cancer
Jim Mobley
Greg Bowersock
Marie Kirby
Rick Myers
Don Buchsbaum
Bill Grizzle
Mel Wilcox

YRC Collaborators
Maitreya Dunham
Emily Mitchell
*Dan Skelly
Josh Akey
Stan Fields
Trisha Davis
Eric Mueller