Metabolomics by GC-MS

Sara J. Cooper
HudsonAlpha Institute for Biotechnology
Huntsville, AL

January 20, 2016
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

Column: packed or open tubular (capillary)

Gas: He, N₂, H₂

Mass spectrometer detector

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

Carrier Gas

Microliter Syringe

Injection Port Liner

Heated Metal Block

Septum

Needle

Sample Aerosol

GC Column

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

Gas: He, N₂, H₂

Sample injector

T regulated oven

Column: packed or open tubular (capillary)

Mass spectrometer detector
Columns: Packed v. Capillary

Packed GC Columns
“Original” GC column
Low efficiency
Coated phase: organic polymers dissolved in solvent and coated onto particles in the tube

Capillary GC Columns
Modern GC column
High efficiency
Usually flexible glass fiber (fused silica) < 1mm ID
Coated phase: organic polymers dissolved in solvent and coated on the inside wall column

Can be 10-30+ meters long
Longer column is better separation, particularly 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.
Two-dimensional chromatography

- GC Columns function in series to improve resolution of chemically similar analytes

Source: Leco Corp
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

True 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 validate or improve 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
Library matching

Unknown spectrum

Asparagine
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
## Why do GC-MS?

<table>
<thead>
<tr>
<th></th>
<th>GC</th>
<th>LC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>Small</td>
<td>Medium to Large</td>
</tr>
<tr>
<td>Polarity</td>
<td>Requires derivitization to reduce polarity</td>
<td>Better for polar</td>
</tr>
<tr>
<td>Metabolites</td>
<td>a.a., organic acids, fatty acids (short-medium)</td>
<td>nucleotides, lipids (including large)</td>
</tr>
<tr>
<td>Chromatography</td>
<td>Highly reproducible-Retention indices</td>
<td>Less critical</td>
</tr>
<tr>
<td>Metabolite ID</td>
<td>Libraries-good for knowns (Some HRT now)</td>
<td>Inferred composition by accurate mass - good for unknowns</td>
</tr>
</tbody>
</table>
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 or filtering 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 (based on our typical 2D-GC protocol but it varies depending on column configuration and data collection speeds):
  • 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 open source tools to analyze raw data (e.g. Metlin, XCMS)

• 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
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 from high quality raw data: 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 T-tests, 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.

- **Joint Pathway Analysis**
  - To perform joint metabolic pathway analysis on results.

- **Other Utilities**
  - This module contains some utility functions commonly used by MetaboAnalyst.
## 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>3.842E-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)
How to design my own experiment - words of wisdom

• Replicates are critical because:
  • Alignment algorithms are not perfect, so you may have missing data
  • Deconvolution is not perfect, so quantification can be noisy in a complex sample
  • Statistics require at least 3 of each sample to do ANYTHING
  • Biological replicates are better than technical replicates (decide based on how difficult it is to get biological replicates and importance of interpretation

• Sample preparation is critical
  • If possible, prepare your samples as a single batch. If not possible, make sure each batch contains more than one type so you can use methods that allow for statistical correction for batch effects

• Sample number - more is better!

• Decide before you begin about whether there are specific metabolites you want to make sure to quantitate. Determine whether they are measurable with this technology and run standards if possible.
Resources for GC-MS

- Restek Column Selection guide www.restek.com/
- Leco
- Agilent
- Sigma 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)
  - “Mass Spectrometry based metabolomics” Dettmer 2007 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1904337/
- Analysis
  - Metaboanalyst.ca
  - impala.molgen.mpg.de
  - hmdb.ca
  - golm database: gmd.mpimp-golmmpg.de
  - metlin.scripps.edu
  - xcmsonline.scripps.edu
Break for Questions???

Thank you
Integrated genomic and metabolomic analysis reveals key metabolic pathways in pancreatic cancer

Sara J. Cooper
HudsonAlpha Institute
Pancreatic Cancer Statistics

1) Early detection is unusual
2) Limited treatment options for advanced stage cancer (no cures)
3) Resistant to chemotherapy

Statistics from cancer.gov

<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>
Metabolic alteration in pancreatic cancer

- Glutamine addiction (PaCa, small cell lung, AML)
  - mTor signaling is affected by glutamine
  - Myc regulates glutamine metabolism

- K-ras is a driver mutation: >90% of PaCa patients have an activating mutation
  - K-ras activates metabolic changes via mtor pathway/Akt
Metabolic reprogramming in pancreatic cancer

1. Detect a metabolic shift in serum and urine from pancreatic cancer patients

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

3. Explore whether alterations in metabolic pathways correlate with outcome
Pancreatic Cancer - Integrating Metabolomics and Genomics

Serum from pancreatic cancer patients

Metabolic changes in pancreatic cancer

Common Pathways

Identification of pathways important to tumor growth and patient survival

Samples

Genes

Gene expression changes in pancreatic cancer

Tumor tissue

Normal tissue
Initial metabolomic analysis reveals altered amino acid metabolism.

Glycine has previously been shown by Mootha et al. to correlate with proliferation in NCI-60 panel & survival in breast cancer patients.

Pancreatic cancer cells are characterized by their “glutamine addiction.”

Glycine has previously been shown by Mootha et al. to correlate with proliferation in NCI-60 panel & survival in breast cancer patients.
Leveraging gene expression information to focus on vital metabolic pathways

- Is there evidence of metabolic reprogramming 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?
Integrated analysis of tumor v. normal genomic and metabolomic data

<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>Sig. Gene Overlap (Total)</th>
<th>Sig. Metabolite Overlap (Total)</th>
<th>Joint Q Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triacylglycerol Degradation</td>
<td>9 (15)</td>
<td>4 (14)</td>
<td>1.53E-02</td>
</tr>
<tr>
<td>Gly, Ser, Thr Metabolism</td>
<td>35 (78)</td>
<td>5 (22)</td>
<td>6.14E-03</td>
</tr>
<tr>
<td>Sphingomyelin Met./Ceramide Salv.</td>
<td>4 (8)</td>
<td>4 (13)</td>
<td>5.50E-02</td>
</tr>
<tr>
<td>Val, Leu, Ile, Metabolism</td>
<td>31 (44)</td>
<td>0 (41)</td>
<td>2.14E-04</td>
</tr>
</tbody>
</table>

Tumor v. Normal clustering using Ser/Gly/Thr genes
Glycine pathway gene expression associated with poor prognosis
Integrated analysis of tumor v. normal genomic and metabolomic data reveals role for fatty acids
Fatty acid gene expression favors lipogenic processes

Lipid Related Transcript Expression

- Red: Lipolytic/Free Fatty Acid Generating
- Blue: Lipogenic/Free Fatty Acid Degrading

Log(Expression Fold Change Normal/Tumor)
Testable hypotheses

- How does alteration of lipid metabolism or amino acid metabolism affect cell proliferation and migration in pancreatic cancer cell models?
- Does metabolic programming in pancreatic cancer rely on K-ras activation?
Analysis of patient survival and gene expression

323 genes associated with patient survival

- Short-survivors
- Long-survivors
Pathway enrichment of transcripts over-represented in survival analysis

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Number Overlapping (Total)</th>
<th>Q - Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digestion of Dietary Lipid</td>
<td>5(5)</td>
<td>0.00381</td>
</tr>
<tr>
<td>Pancreatic Secretion</td>
<td>21(96)</td>
<td>0.00381</td>
</tr>
<tr>
<td>Retinoid Metabolism</td>
<td>9(42)</td>
<td>0.267</td>
</tr>
<tr>
<td>Triacylglycerol Degradation</td>
<td>5(15)</td>
<td>0.269</td>
</tr>
</tbody>
</table>

- Many of these genes are related to exocrine function
- Previous report suggests an exocrine subtype
A correlation between survival and lipase expression
Pancreatic cancer patient survival based on lipase gene expression

Kaplan-Meier plot

- Low Lipase
- Medium Lipase
- High Lipase

Months
Replication in independent samples

TCGA

ICGC Australian Cohort Survival

RSEM Counts

CEL CLPS PNLIP PNLIP1 PNLIP2 LPL

Survived > 3 yrs
Survived < 1 yr
HOW do these genes confer altered survival?

Genes with altered expression in long-survivors and drug sensitive-cell lines:

<table>
<thead>
<tr>
<th>Gene</th>
<th>Short-survivors/resistant cancer cell lines</th>
<th>Long-survivors/sensitive cancer cell lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTSH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B3GNT5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANGPTL4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMD9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KLF6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCUBE1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WFDC1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What role do these genes have in survival?

PPAR → ANGPTL4 → LPL

**Bar graph:**
- X-axis: Time (Hours) [24, 48, 72]
- Y-axis: Luminescence
- Colors:
  - Red: Non-Targeted
  - Blue: ANGPTL4 #1
  - Purple: ANGPTL4 #2

Significance levels:
- *: p < 0.05
- **: p < 0.01
- ***: p < 0.001
Future Directions

- What explains differential survival?
- Is there a role for K-ras?
- What role do regulators of lipid metabolism play in prognosis?
Potential regulators

Kras?

\[ \downarrow \]

GATA6 \[ \rightarrow \] RBPJL

with PTF1-J complex

PPARG

GATA6

\[ \star \]

RBPJL

PPARG cofactors

PPARG

RXRalpha

PPARGC1a
Acknowledgements

S. Cooper Lab
Karl Ackerman (GGS-UAB)
Bobbi Johnston
Karin Bosma (Alum)
Ryne Ramaker (MSTP-UAB)
Rebecca Hauser (Biotrain and future UAB GGS student)

Pancreatic Cancer Team
Marie Kirby (HudsonAlpha)
Rick Myers (HudsonAlpha)

Greg Bowersock (UAB)
Don Buchsbaum (UAB)
Bill Grizzle (UAB)
Jim Mobley (UAB)
Patsy Oliver (UAB)
James Posey (UAB)
Kathy Sexton (UAB)
Selwyn Vickers (UAB)
Mel Wilcox (UAB)