Metabolomics by GC-MS

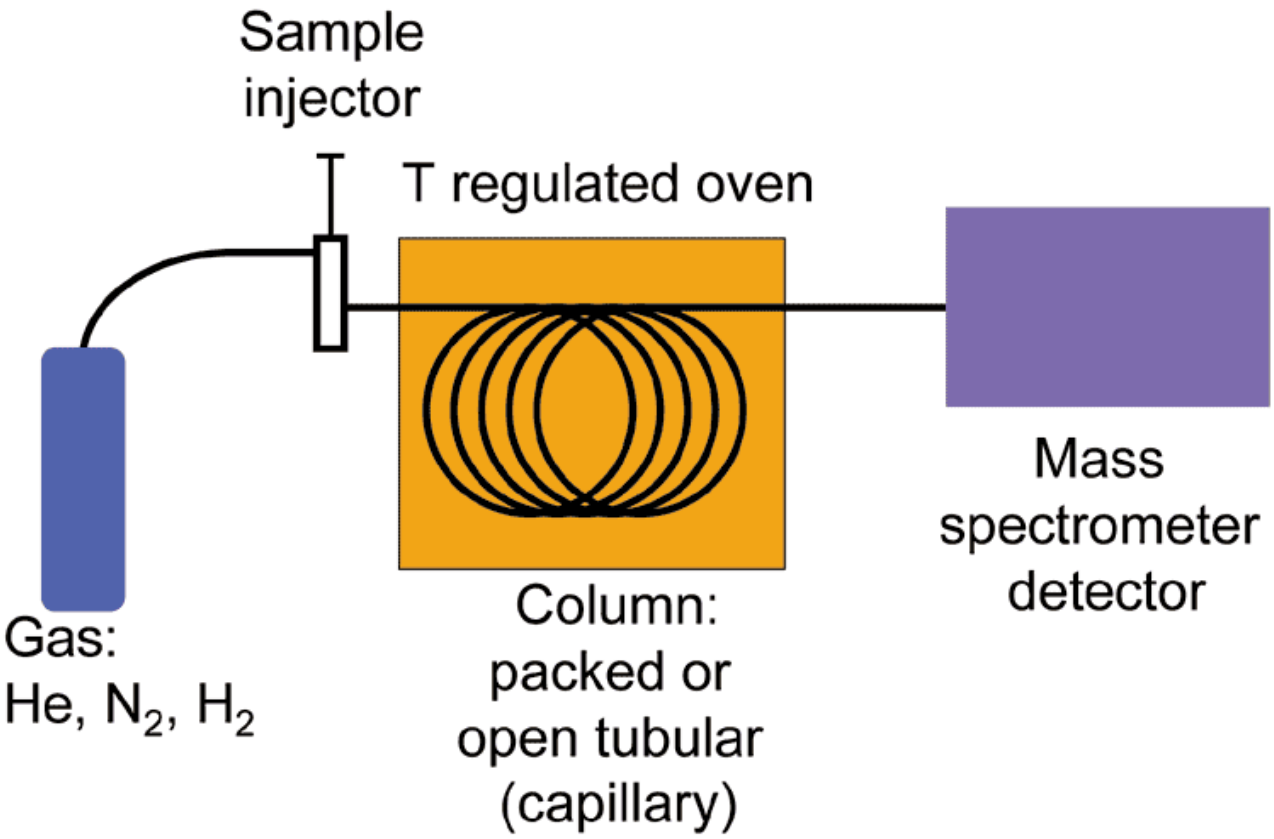
Sara J. Cooper
HudsonAlpha Institute for Biotechnology
Huntsville, AL

January 20, 2016

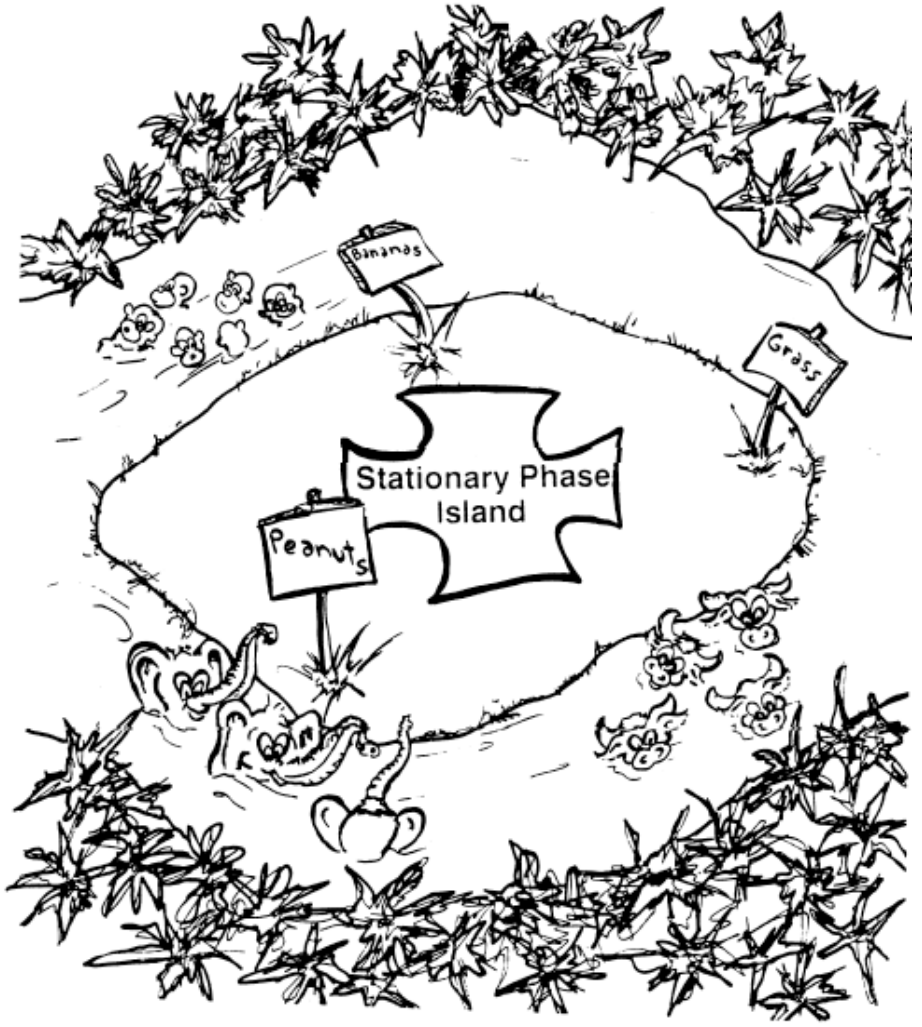
Outline

- 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

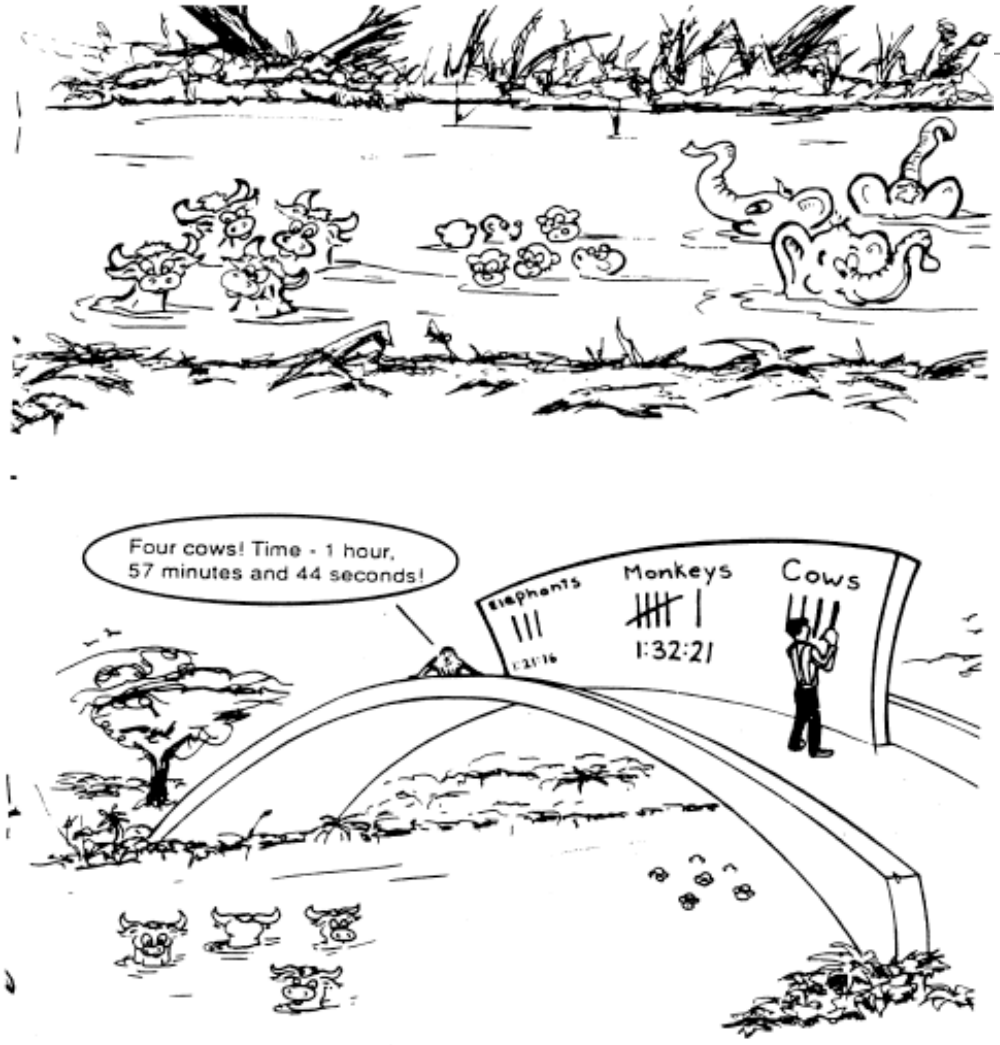
The Nuts and Bolts of GC-MS



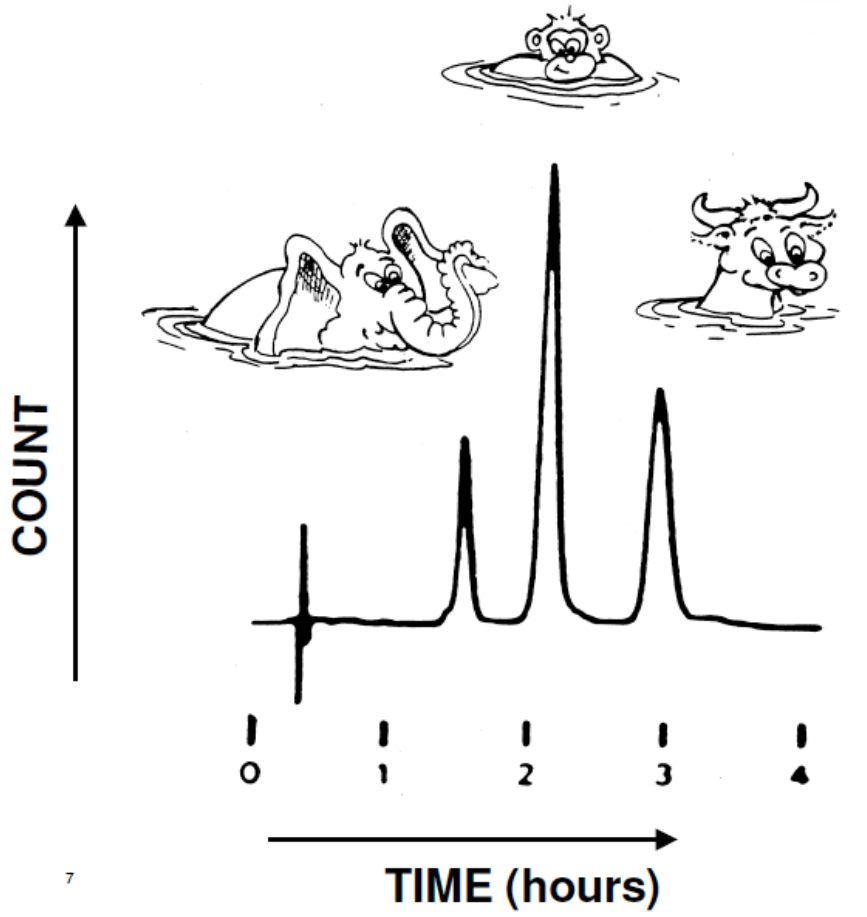
The Principal of GC



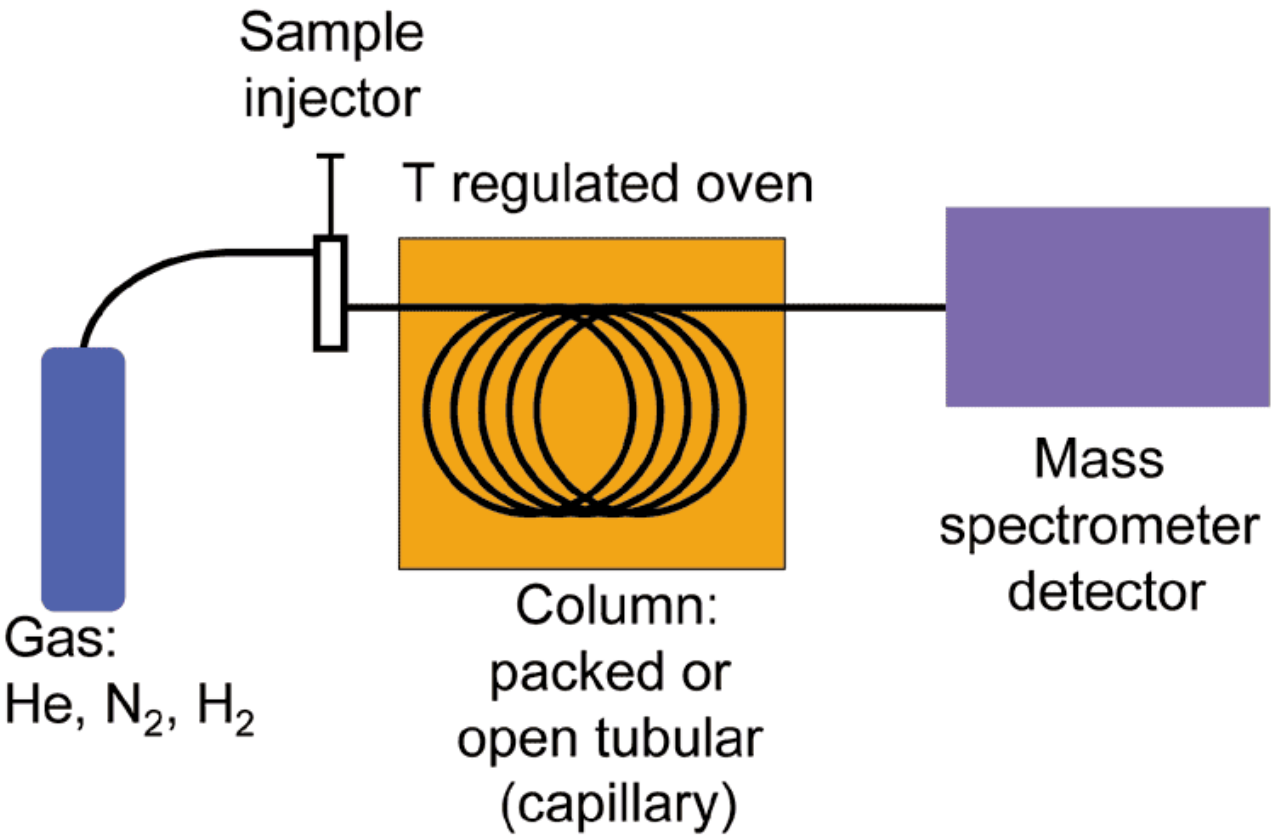
Source-SigmaAldrich ‘thebasicsofgc’



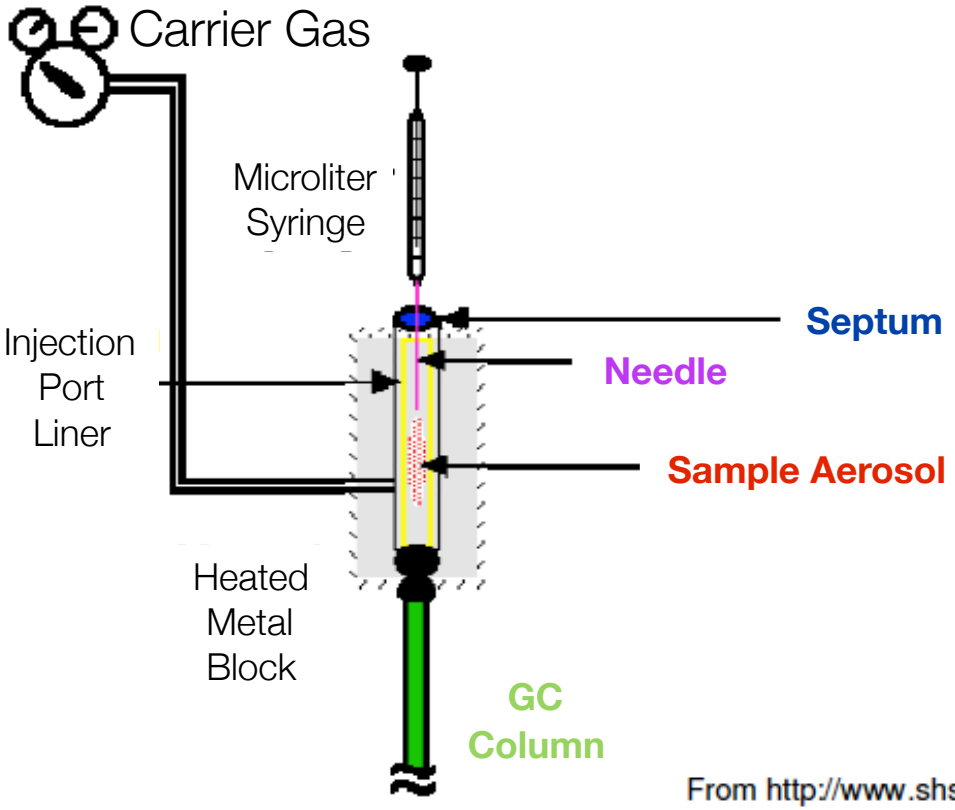
The analysis is now complete.



The Nuts and Bolts of GC-MS

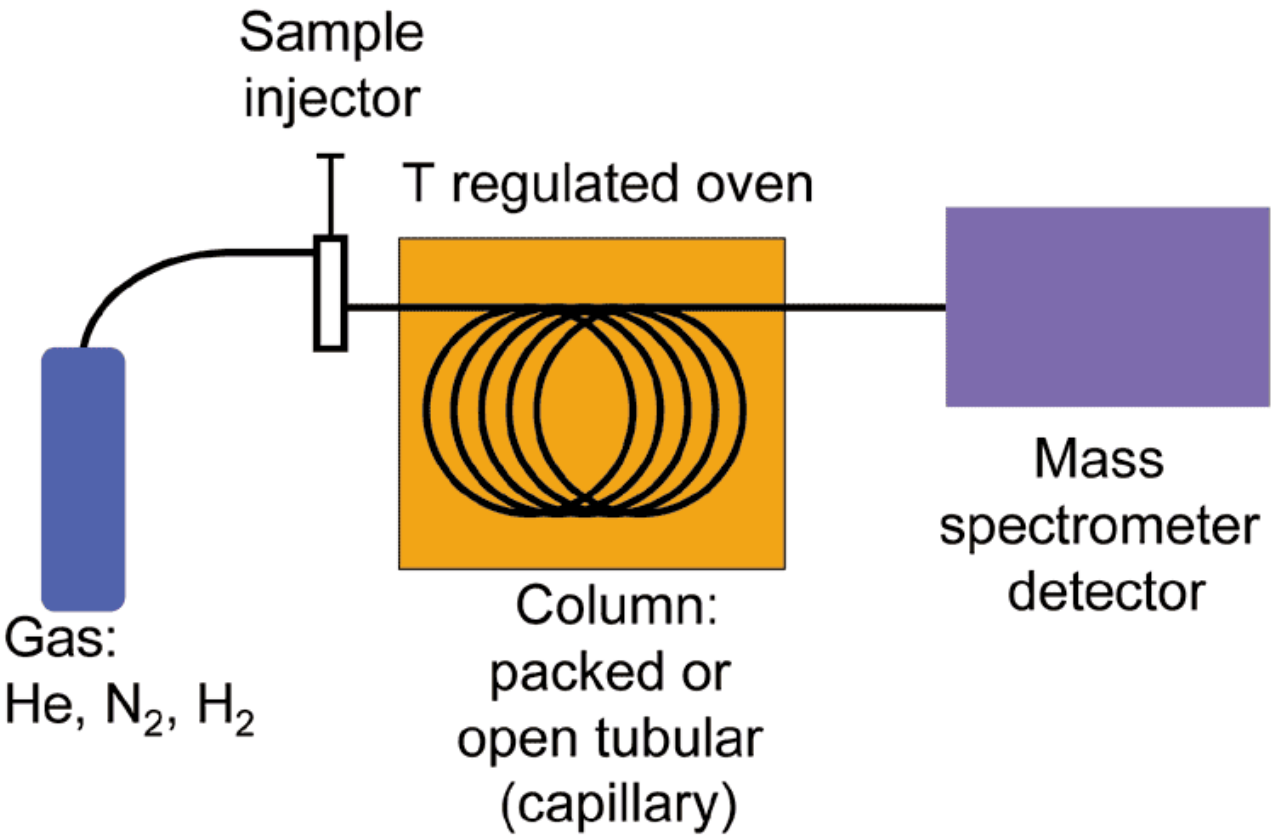


Injection



From http://www.shsu.edu/~chemistry/GC/packed.GIF

The Nuts and Bolts of GC-MS



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

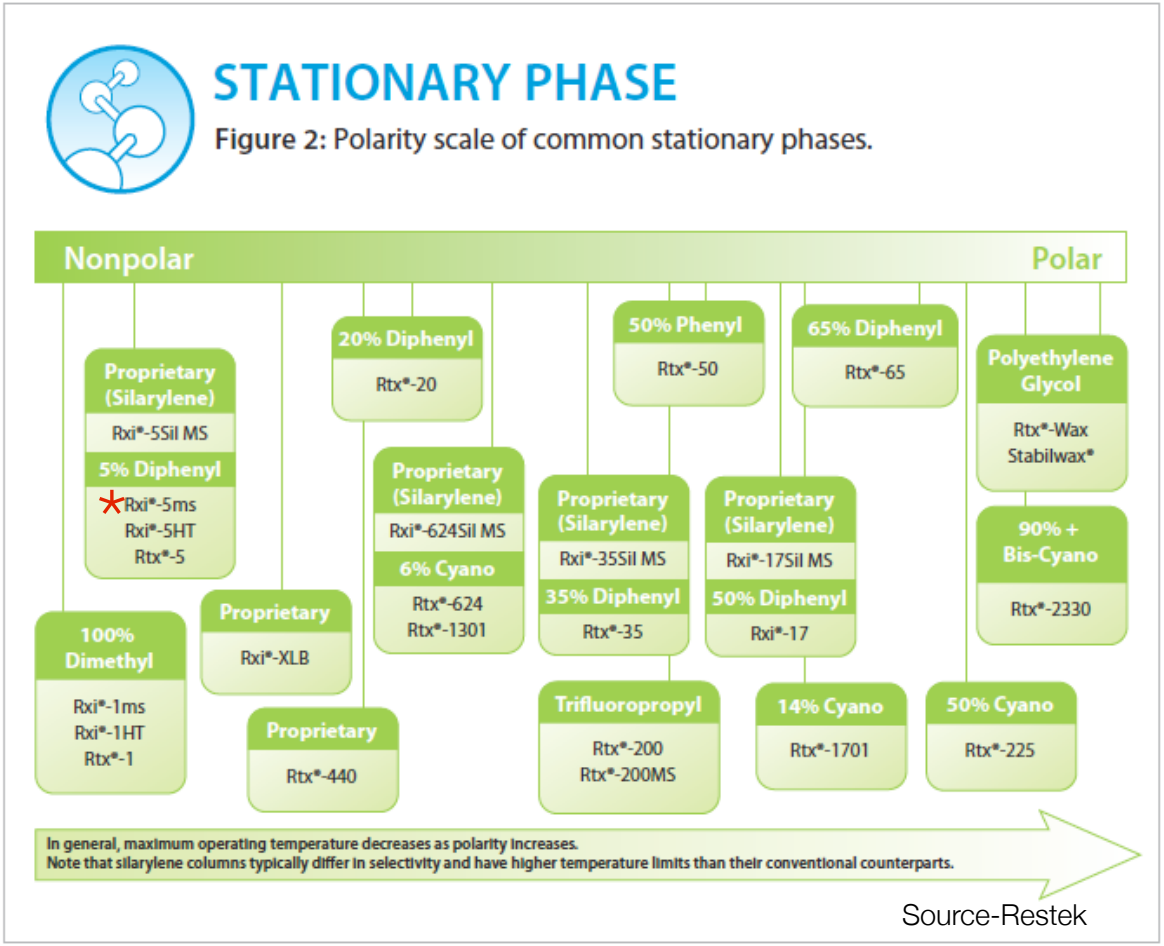
Modern GC column
High efficiency
Usually flexible glass fiber (fused silica) < 1mm ID
Coated phase: organic polymers dissolved in solvent and coated on the inside wall column

Can be 10-30+ meters long
Longer column is better separation, particularly for complex mixtures

Modified from sigmaaldrich.com

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



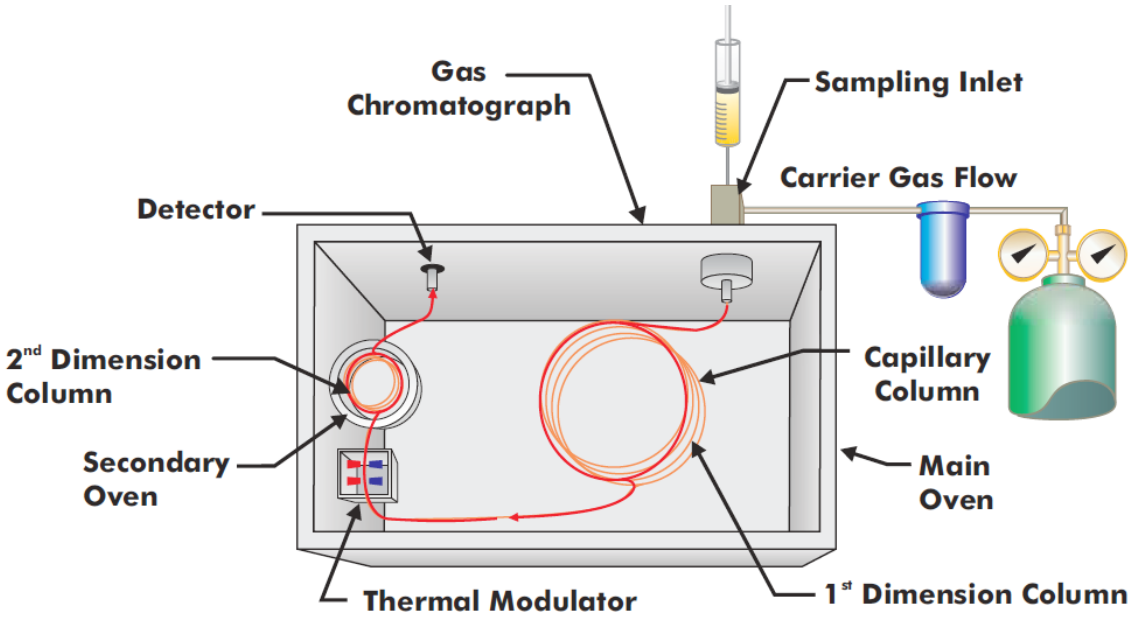
Source-Restek

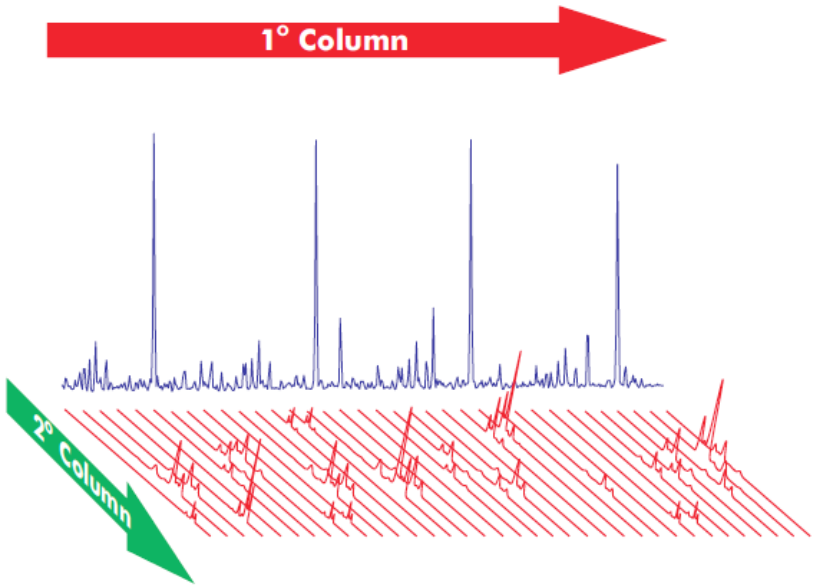
tech tip

Any homologous series of compounds, that is, analytes from the same chemical class (e.g., all alcohols, all ketones, or all aldehydes, etc.) will elute in boiling point order on any stationary phase. However, when different compound classes are mixed together in one sample, intermolecular forces between the analytes and the stationary phase are the dominant separation mechanism, not boiling point.

Two-dimensional chromatography

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

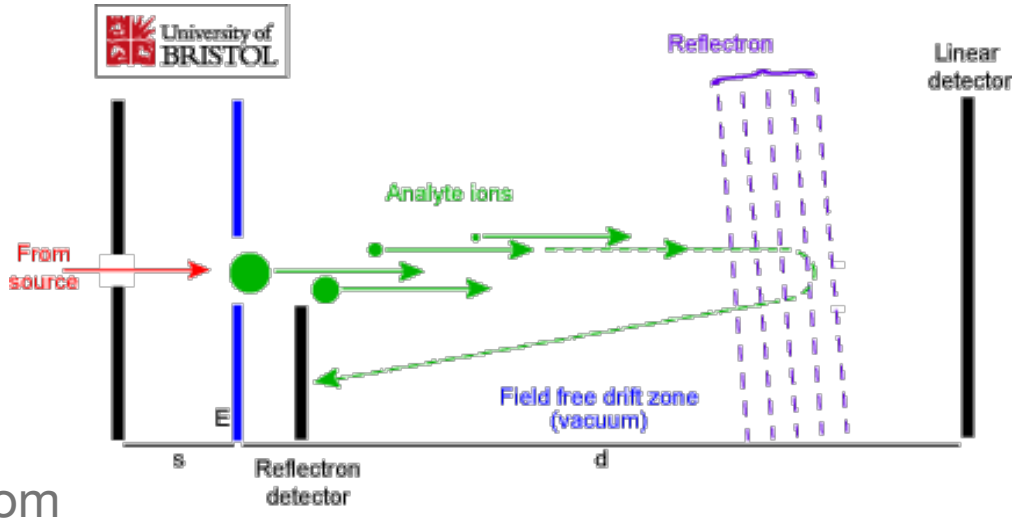




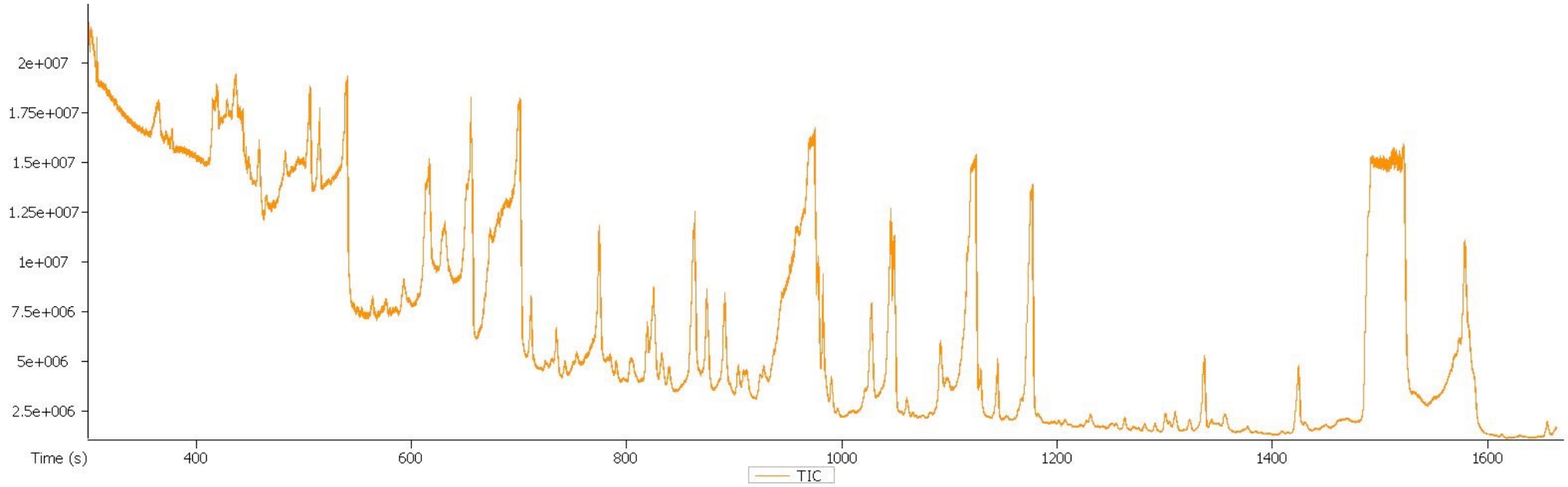
Source: Leco Corp

Mass Spectrometer - Ionization and mass measurement

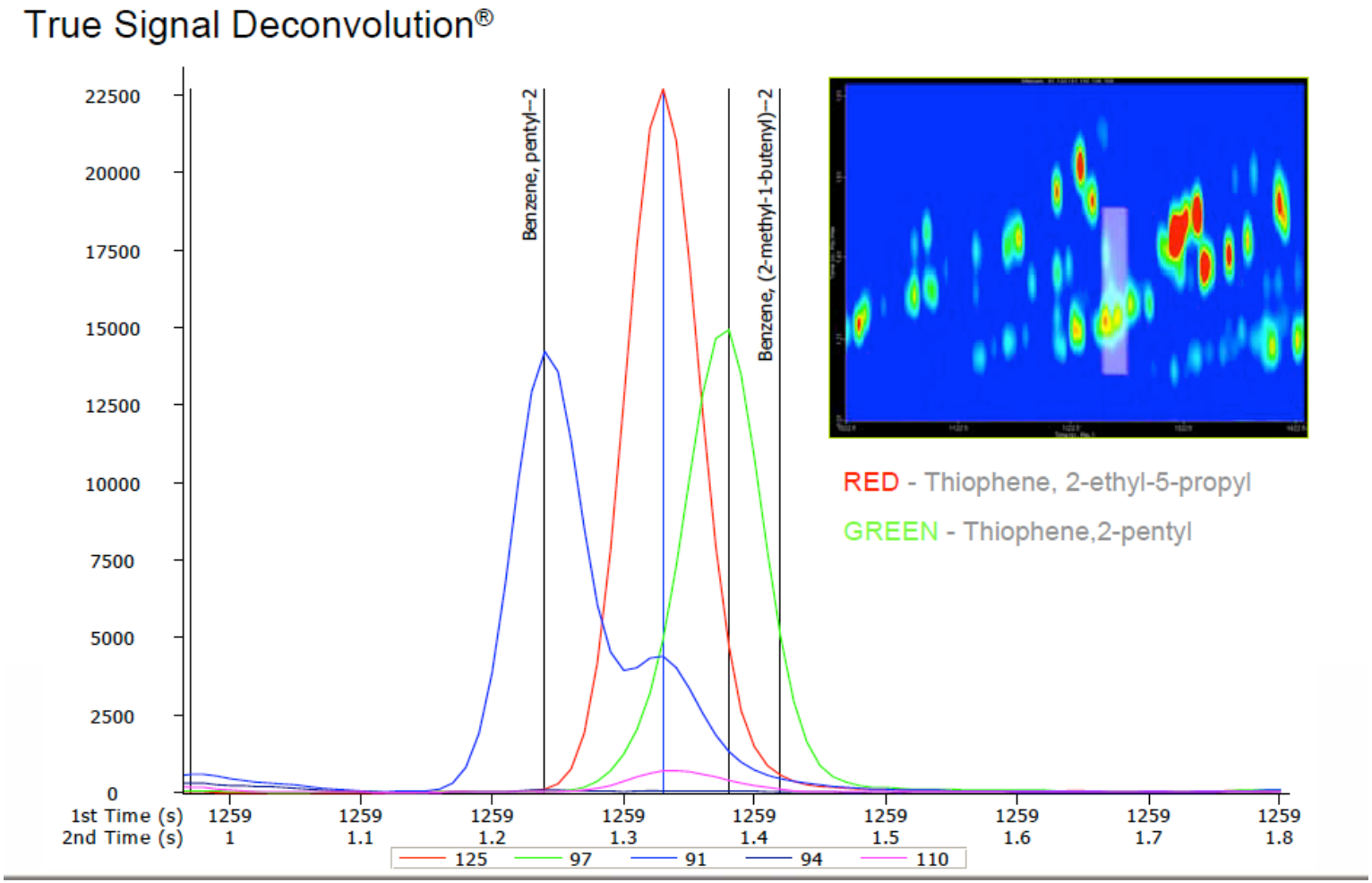
- Ionization
  - Electron Ionization (Standard -70keV)
    - Fragmentation
  - Chemical Ionization (less common)
- Detection
  - Time-of-flight mass spectrometry
    - mass calculated based on time from ionization to reaching detector
  - High-Resolution TOF
    - offers higher mass resolution for metabolite identification



Example data output-Chromatogram



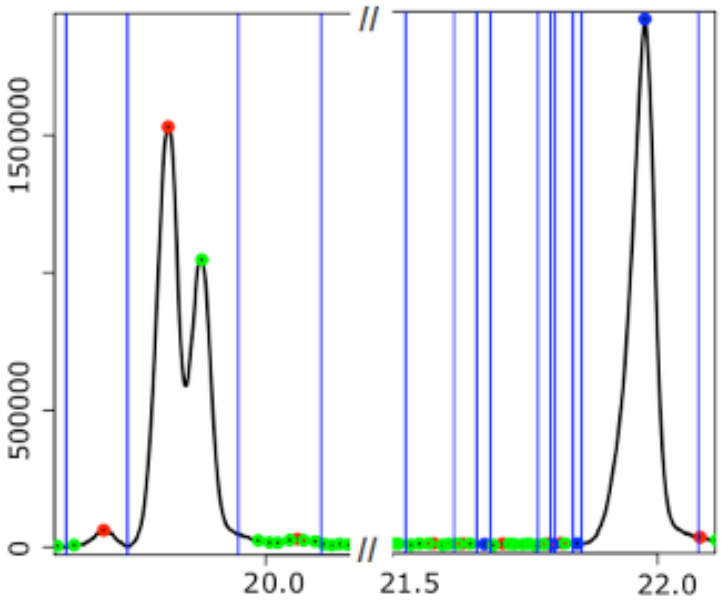
Signal Deconvolution



Source: Leco

Principles of Deconvolution

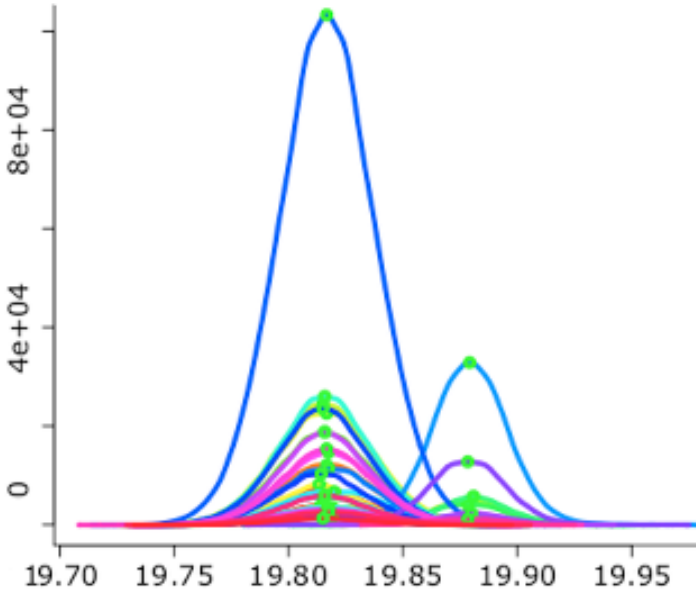
- Generally implemented in AMDIS
- Goal: computationally separate chromatographically overlapping peaks

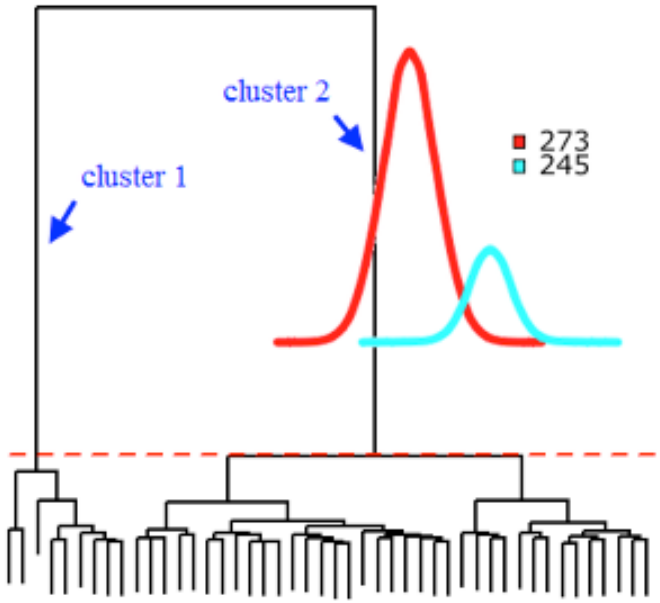


TIC

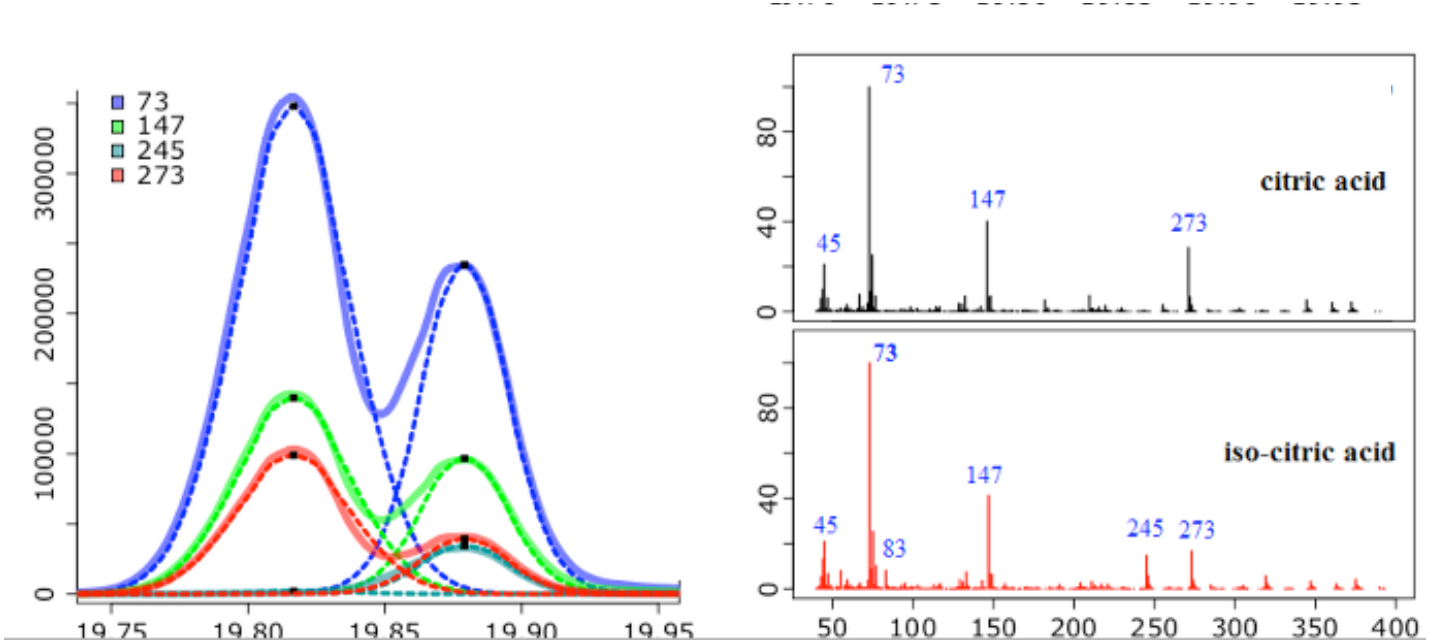
Source: Du and Zeisel 2013

Principles of Deconvolution

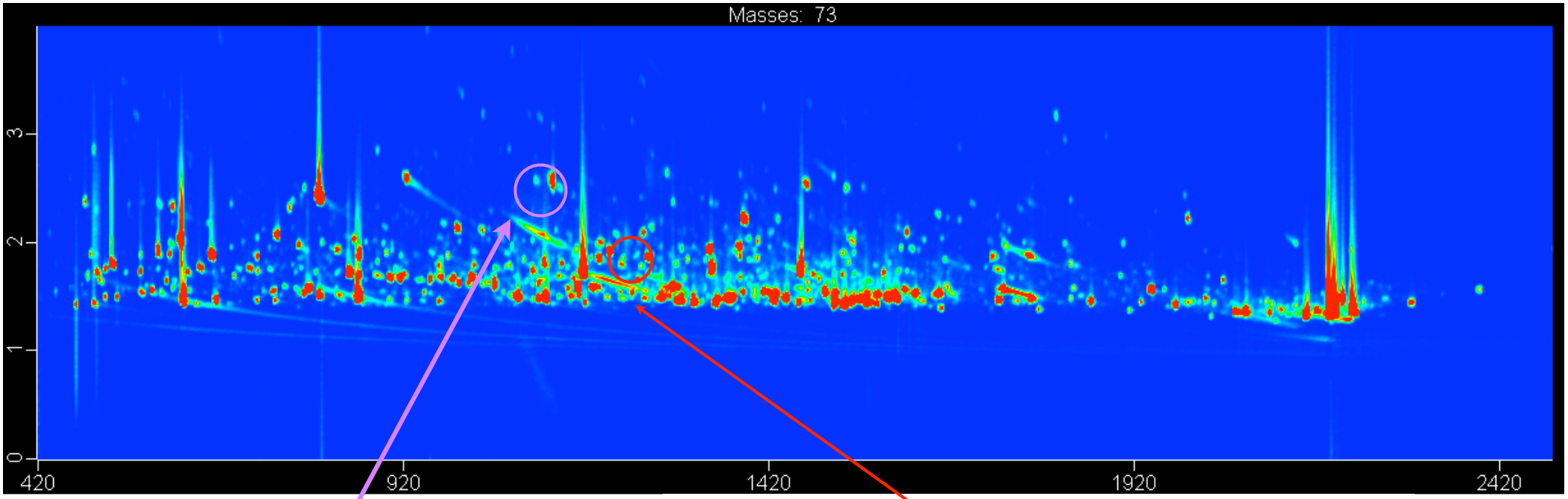




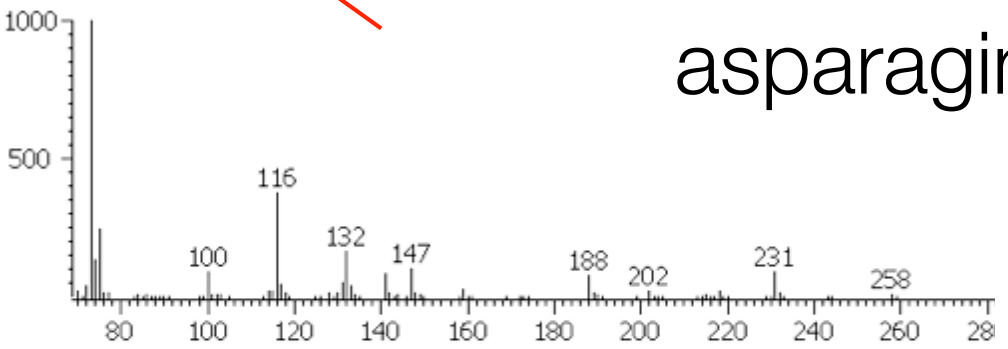
Principles of Deconvolution



Data projected into two dimensions



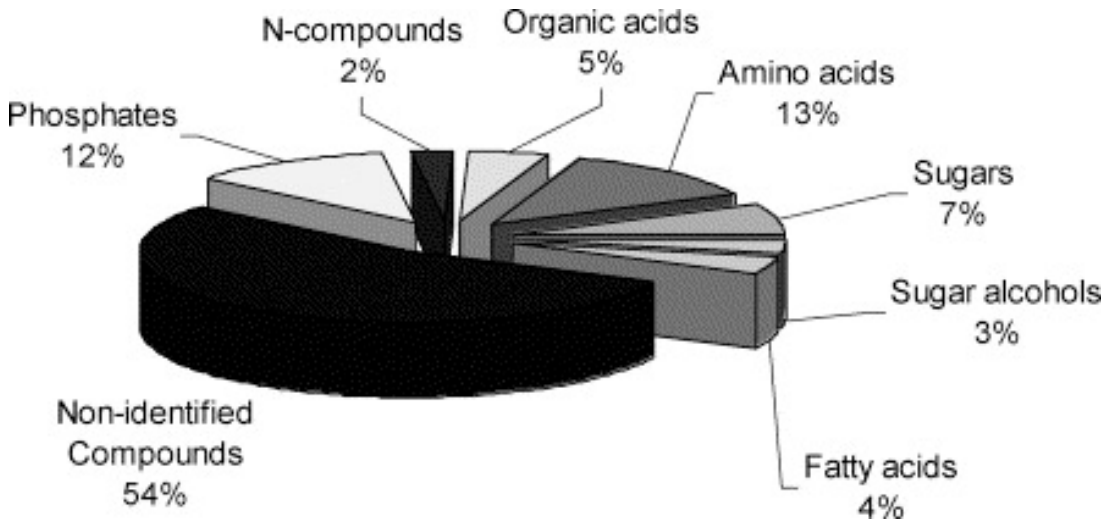
Glutamate



asparagine

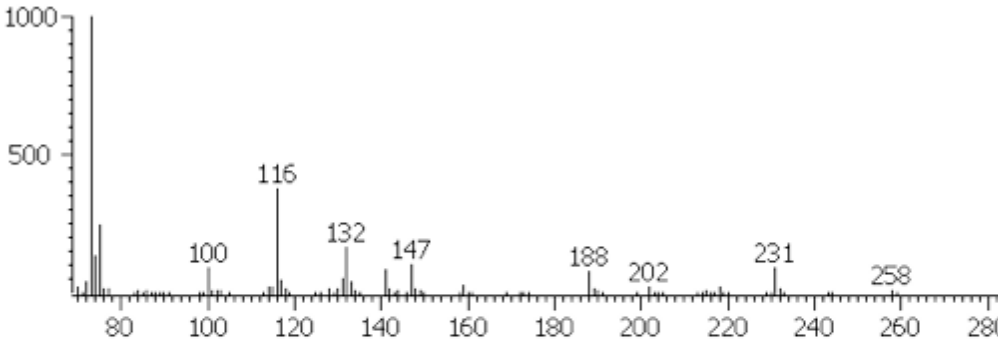
Metabolite Identification

- 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

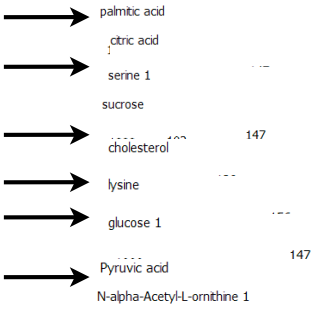


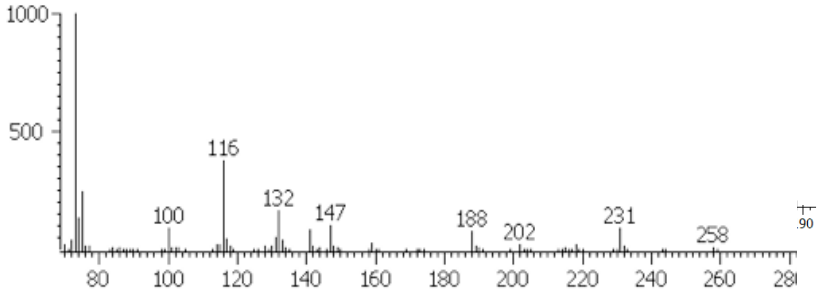
source: Schauer et al 2005

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?

	GC	LC
Size	Small	Medium to Large
Polarity	Requires derivitization to reduce polarity	Better for polar
Metabolites	a.a., organic acids fatty acids (short-medium)	nucleotides, lipids (including large)
Chromatography	Highly reproducible- Retention indices	Less critical
Metabolite ID	Libraries- good for knowns (Some HRT now)	Inferred composition by accurate mass - good for unknowns

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

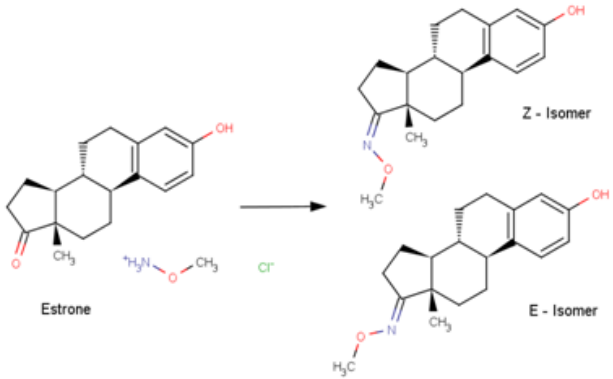
- Experimental Design!! What question(s) do you want to answer?
- Sample preparation
- Data collection
- Preliminary Data analysis
  - tools
- Metabolite identification

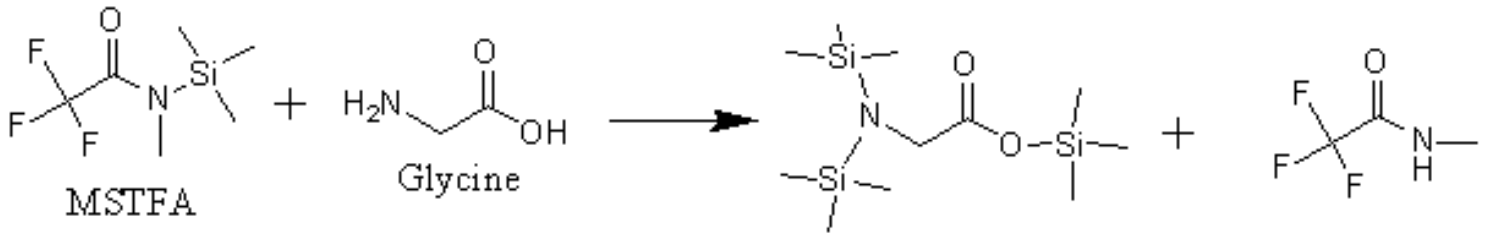
Sample procurement/preparation

- 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 are 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

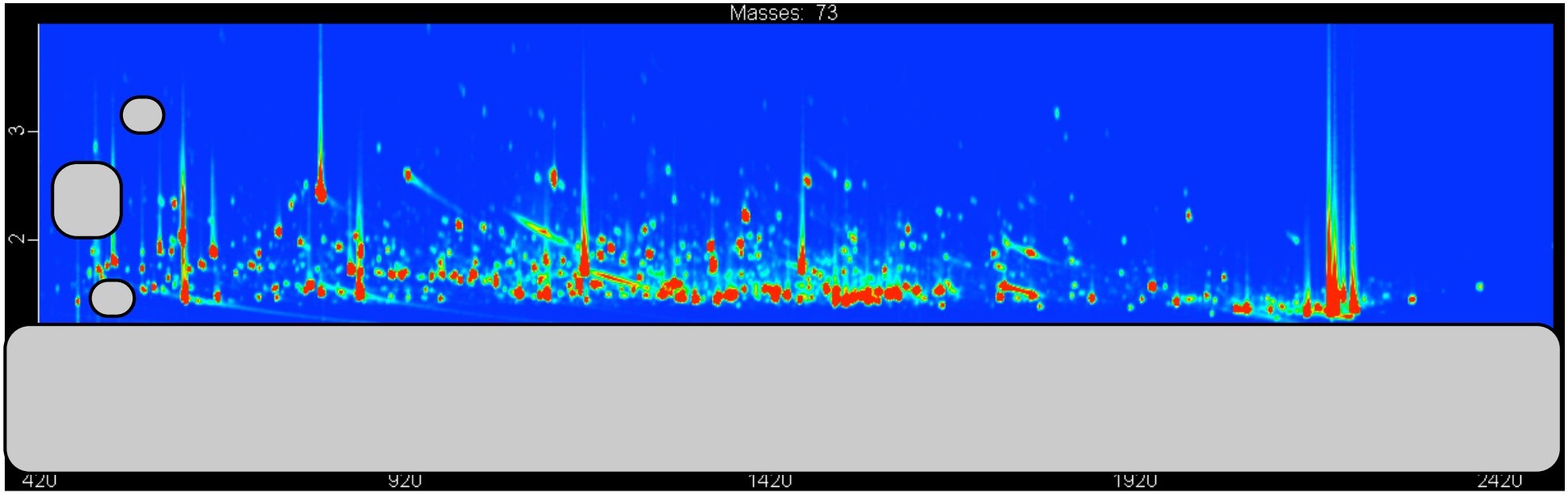
- Most instruments utilize proprietary software to do peak deconvolution
- Raw data can be analyzed as well and there are open source tools to analyze raw data (e.g. Metlin, XCMS)
- ChromaTOF (Leco's peak calling and deconvolution software) Output:
  - List of peaks
  - Determination of Quant Mass for 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%)	Limited to small number of metabolites (High Confidence)
• Assuming missing values are below detectable levels (0.5x lowest value for that metabolite)	Can skew results if there are a large number of missing values
• Assume missing values are present at an average or median level	Conservative, but can skew data
• 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	Moderately conservative , but not possible if missing data is abundant

Data Analysis: Normalization

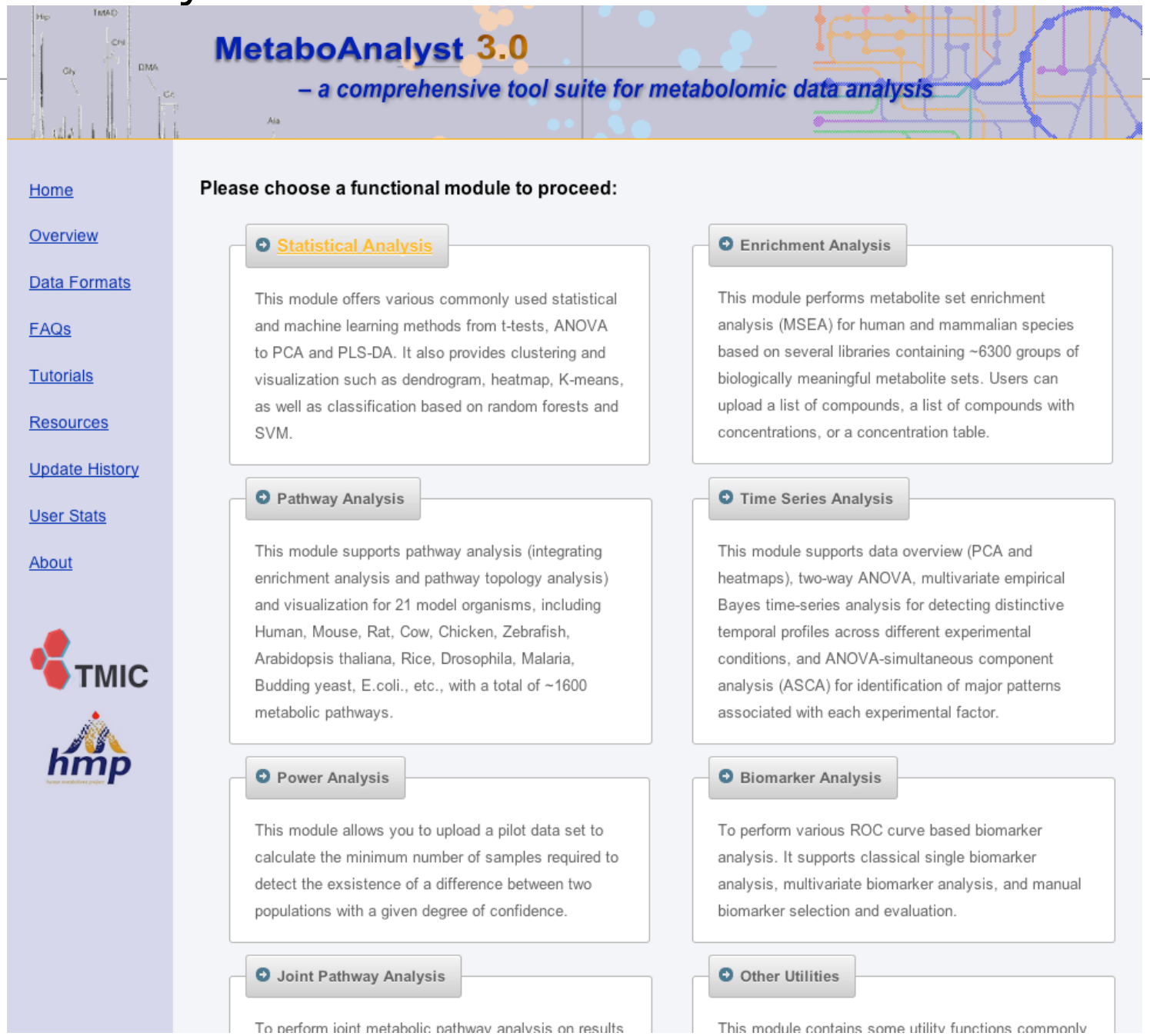
- Common Practice:
  - Injection Control (A known amount of substance is injected with each sample. Those peaks should have the same area each time)
  - Normalization by SUM (total area under the curve). Normalizes for overall sample concentration
  - Clinical samples: normalization by creatinine or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

Data Analysis: Statistical Analysis

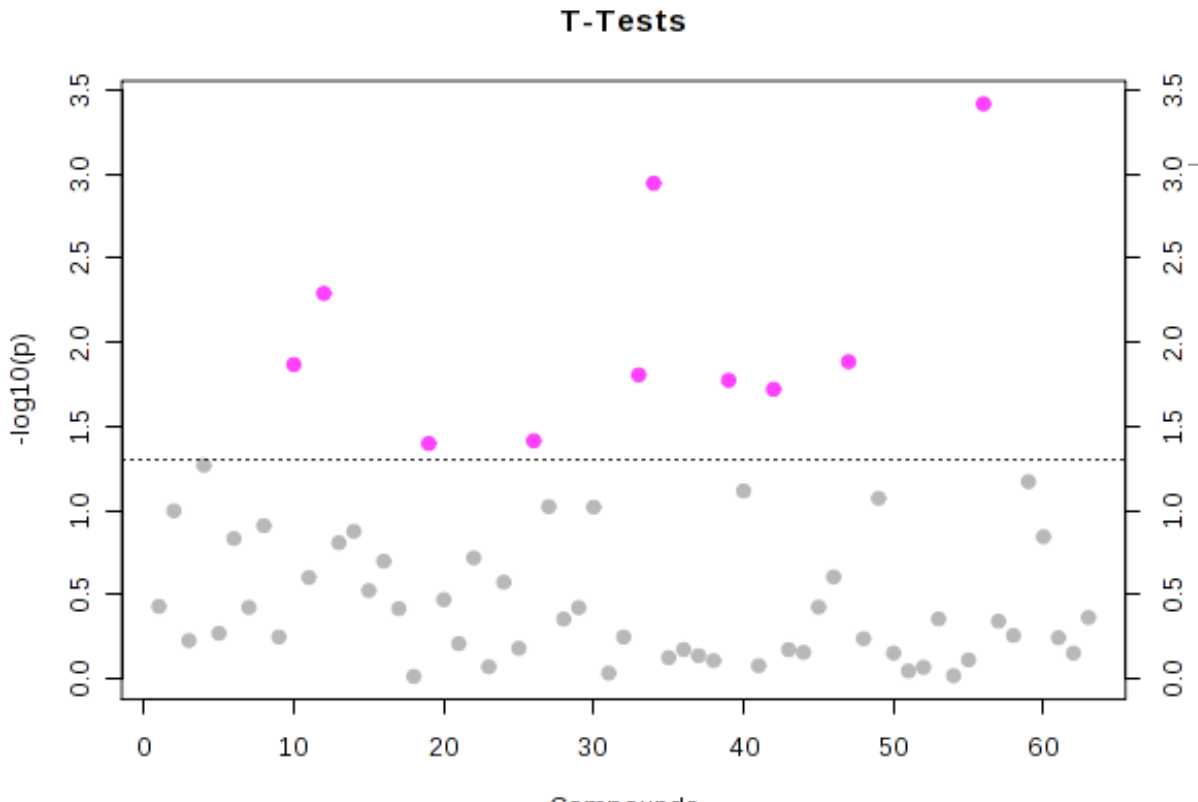
- A wide variety of tools and packages available
- Metaboanalyst is a great place to start (R-package in web-based app)
  - Upload your aligned data in .csv or .txt format. It goes through the normalization, missing data and filtering steps and then allows a variety of analysis
  - Heatmaps, Clustering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

www.metaboanalyst.ca

Metaboanalyst

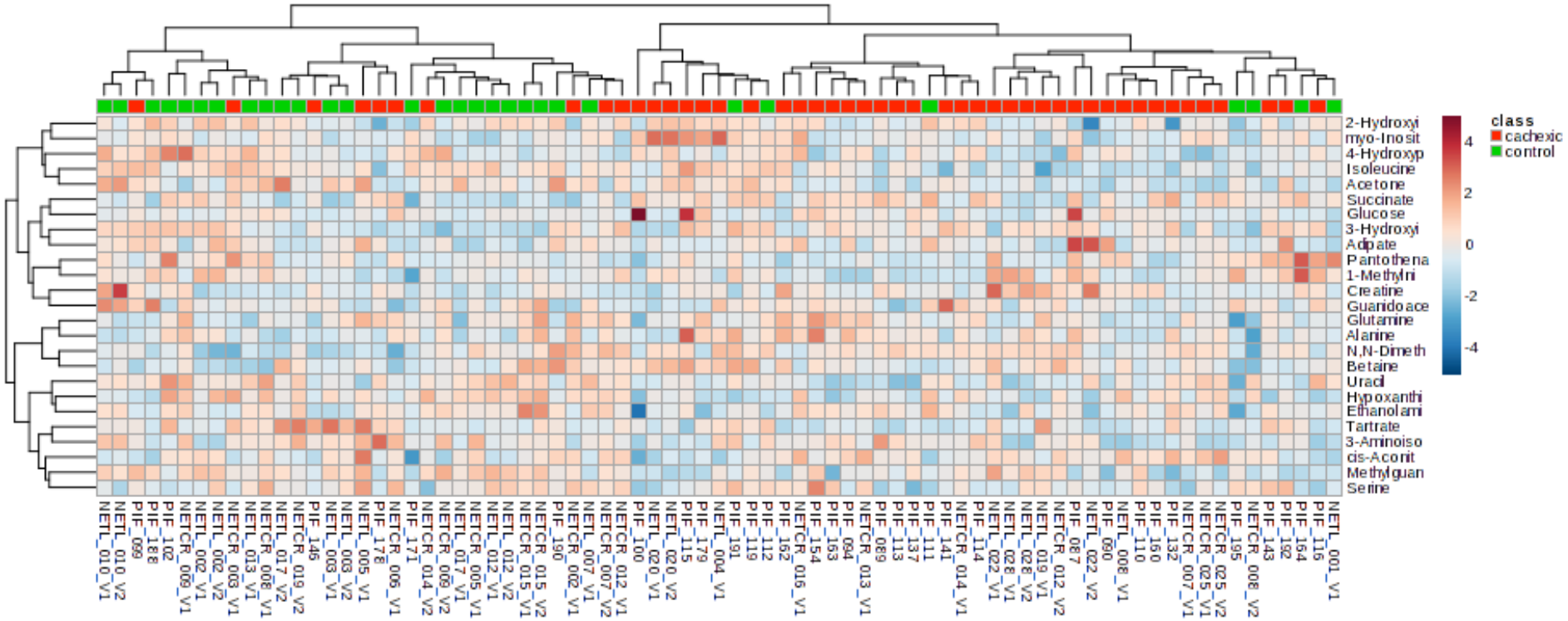


Input test dataset (Cancer patients Cachexic v. control)

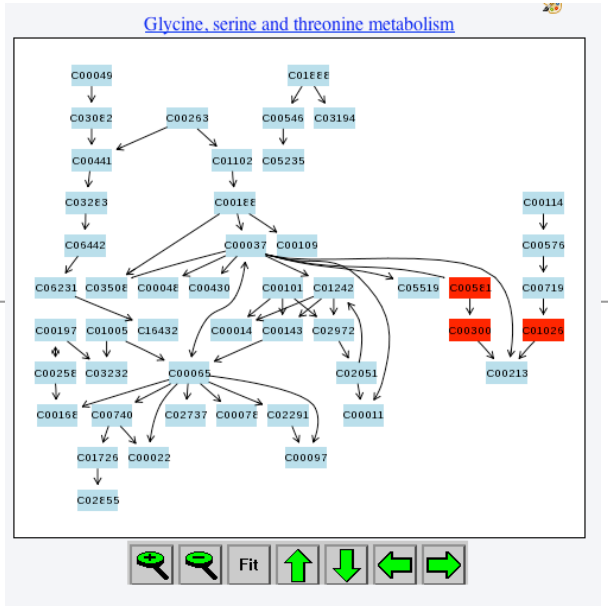


	p-value	FC	FDR
Uracil	3.842E-04	3.4154	0.024204
Isoleucine	0.0011396	2.9432	0.035898
Acetone	0.0051404	2.289	0.10795
Succinate	0.013088	1.8831	0.1502
4-Hydroxyphenylacetate	0.013611	1.8661	0.1502
Hypoxanthine	0.015669	1.805	0.1502
Methylguanidine	0.016881	1.7726	0.1502
Pantothenate	0.019073	1.7196	0.1502
Glucose	0.038618	1.4132	0.25269
Creatine	0.04011	1.3967	0.25269

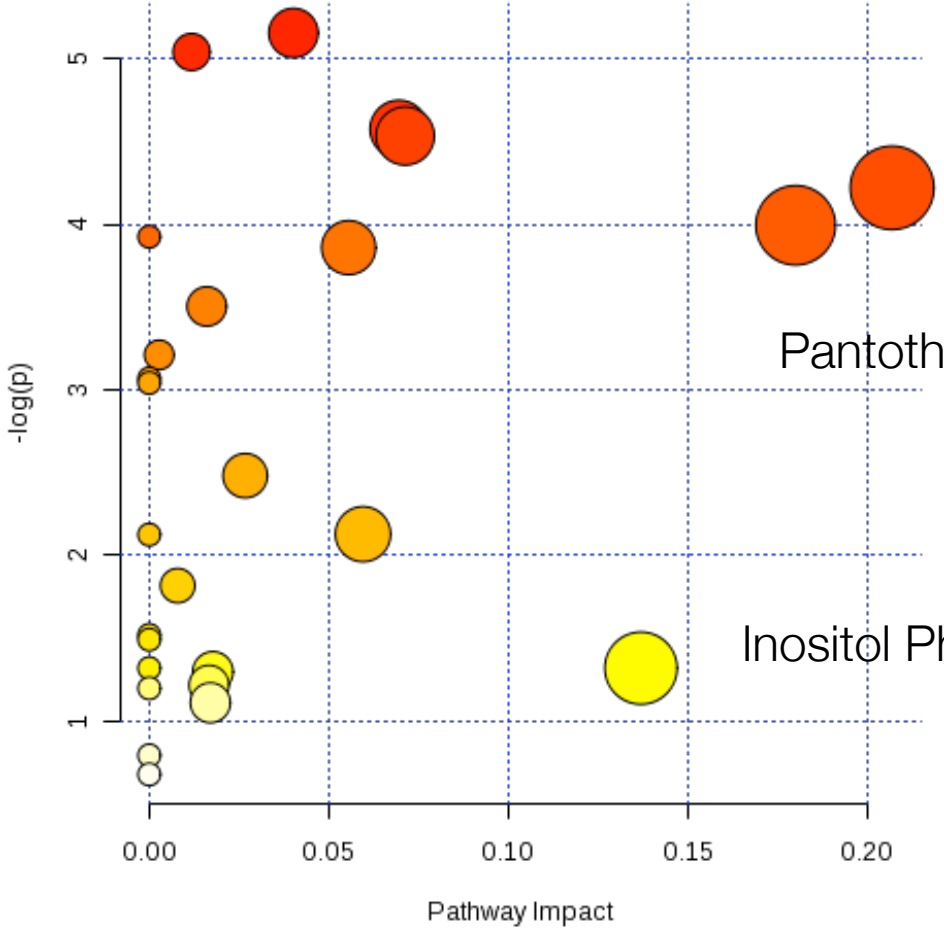
Sample Data-top25 features by Ttest



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/
  - http://www.restek.com/pdfs/GNBR1724-UNV.pdf
- 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

Stage at diagnosis	Stage distribution %	5-year survival (%)
Localized	8	23.3
Regional (spread to lymph nodes)	27	8.9
Distant (metastatic)	53	1.8
Unknown	12	3.9

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

Metabolic alteration in pancreatic cancer

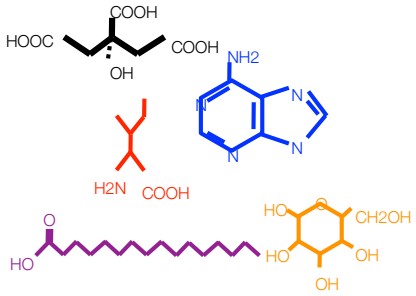
- Glutamine addiction (PaCa, small cell lung, AML)
  - mTor signaling is affected by glutamine
  - Myc regulates glutamine metabolism
- K-ras is a driver mutation: >90% of PaCa patients have an activating mutation
  - K-ras activates metabolic changes via mtor pathway/Akt

Metabolic reprogramming in pancreatic cancer

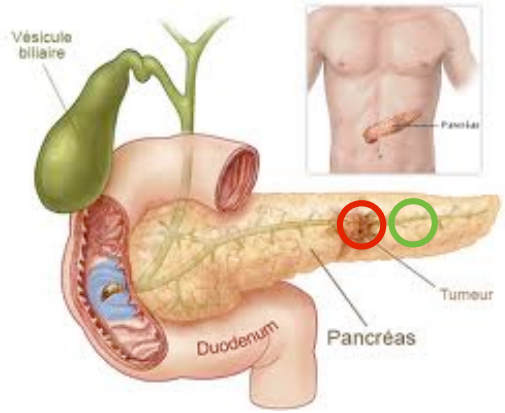
1. Detect a metabolic shift in serum and urine from pancreatic cancer patients
2. Determine whether those alterations represent metabolic changes in the pancreatic tumor
3. Explore whether alterations in metabolic pathways correlate with outcome

Pancreatic Cancer-
Integrating Metabolomics and Genomics

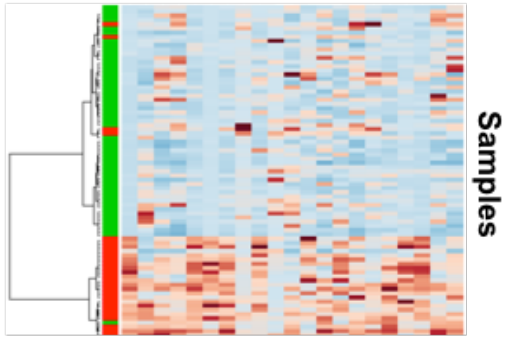
Serum from pancreatic cancer patients



Metabolic changes in pancreatic cancer



Tumor tissue
Normal tissue



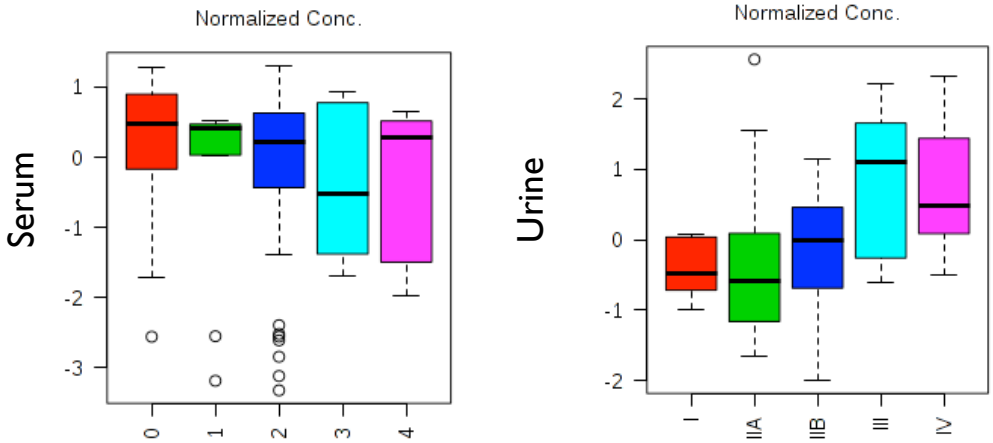
Gene expression changes in pancreatic cancer

Common Pathways

Identification of pathways important to tumor growth and patient survival

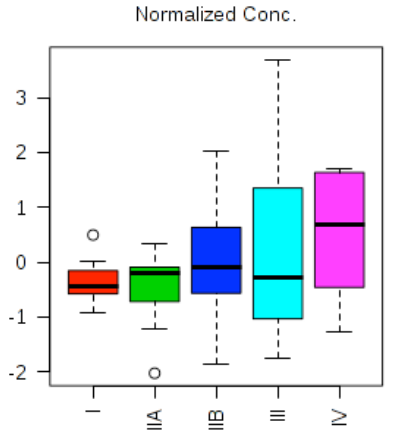
Initial metabolomic analysis reveals altered amino acid metabolism

Glutamine



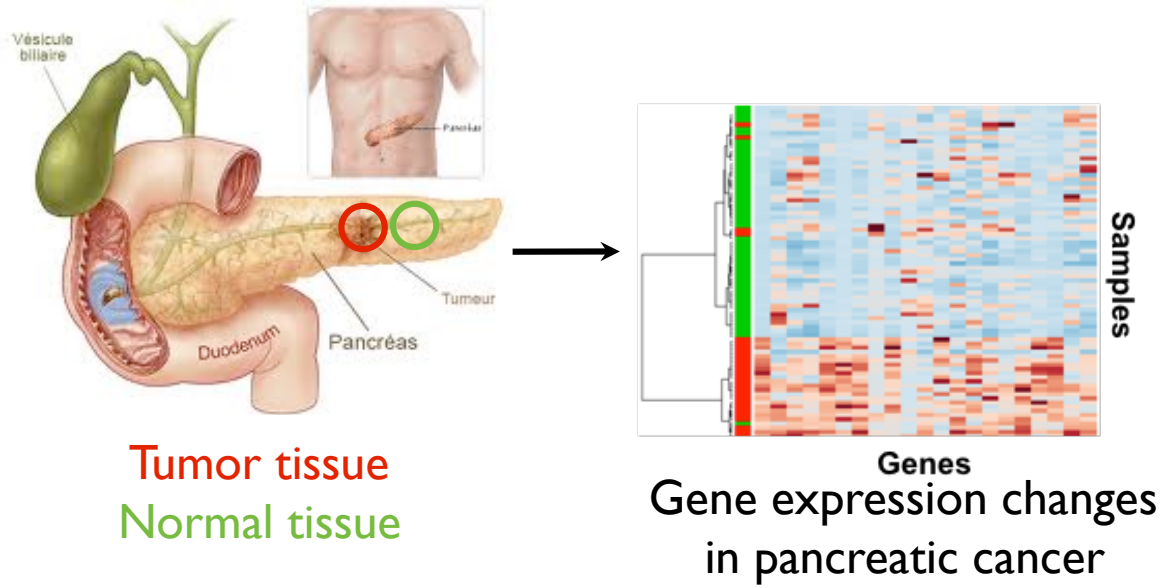
Pancreatic cancer cells are characterized by their "glutamine addiction"

Glycine



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



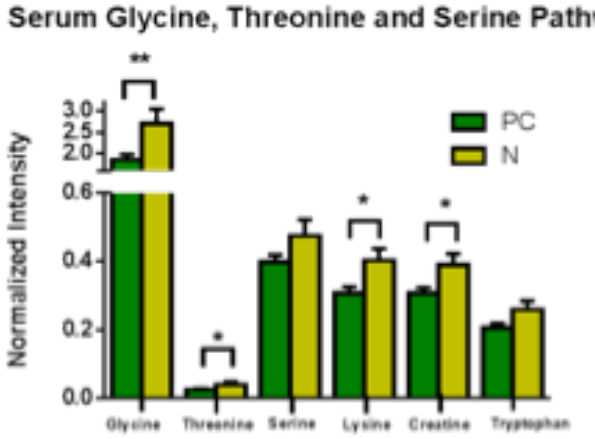
Tumor tissue
Normal tissue

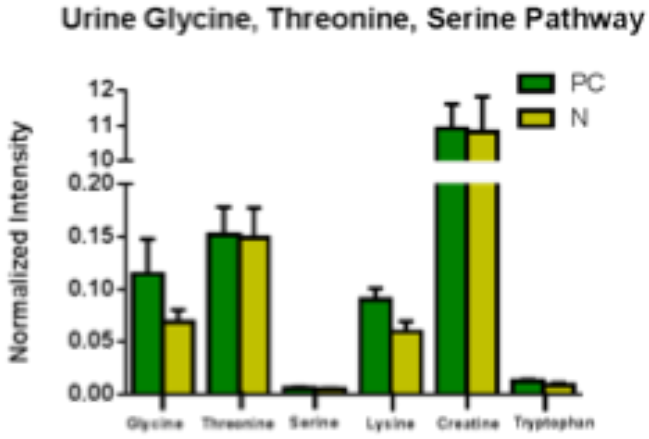
Gene expression changes in pancreatic cancer

- 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?

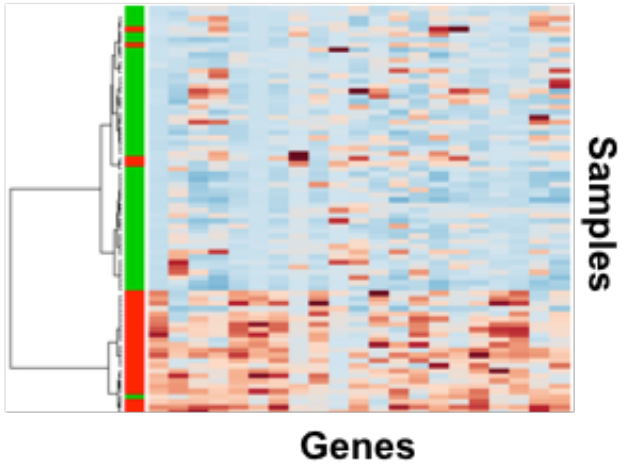
Integrated analysis of tumor v. normal genomic and metabolomic data

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

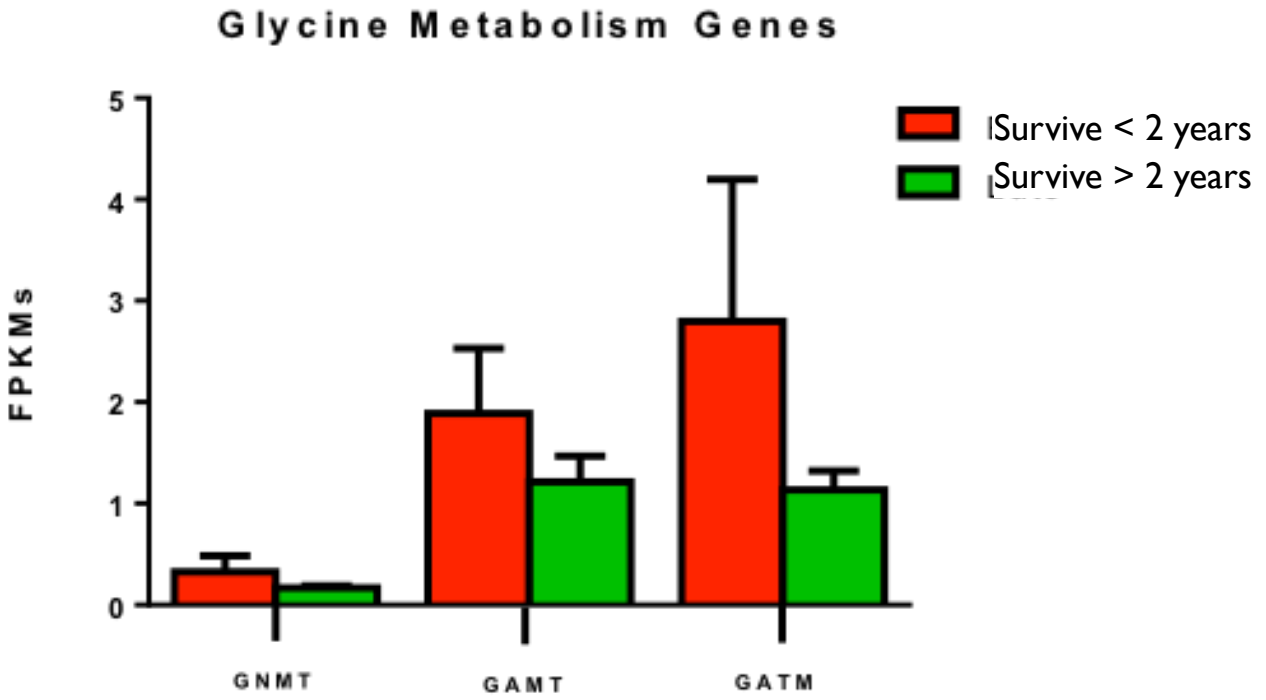


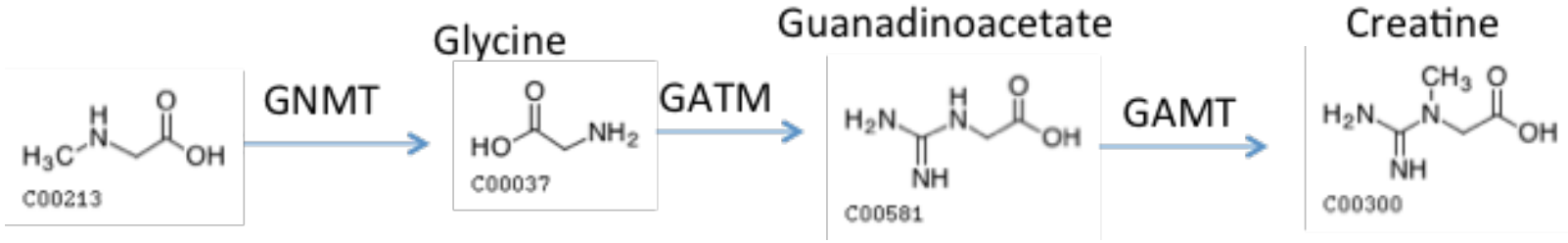


Tumor v. Normal clustering using Ser/Gly/Thr genes

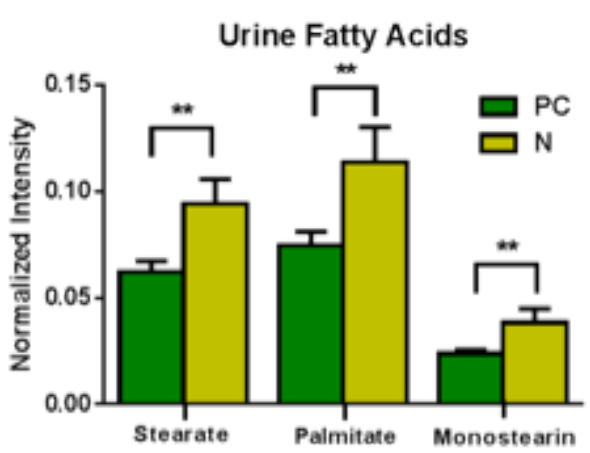


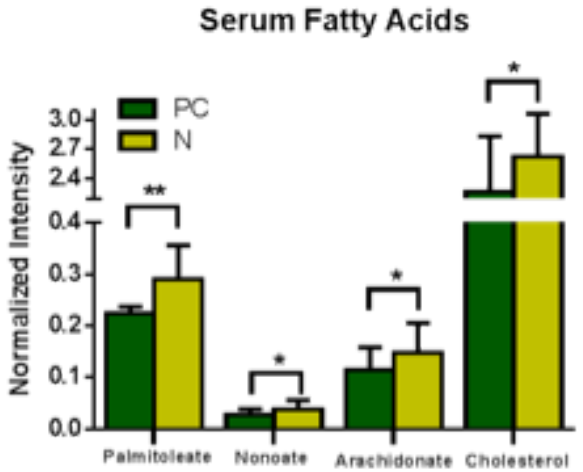
Glycine pathway gene expression associated with poor prognosis

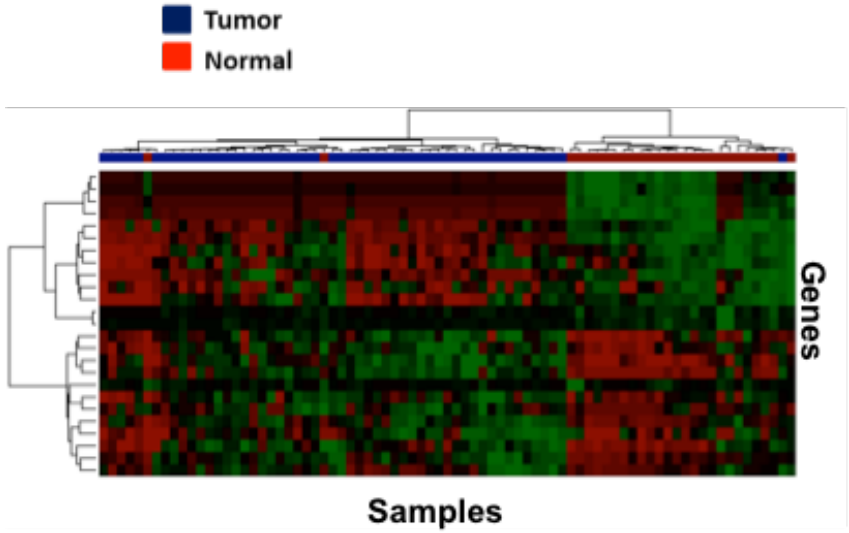




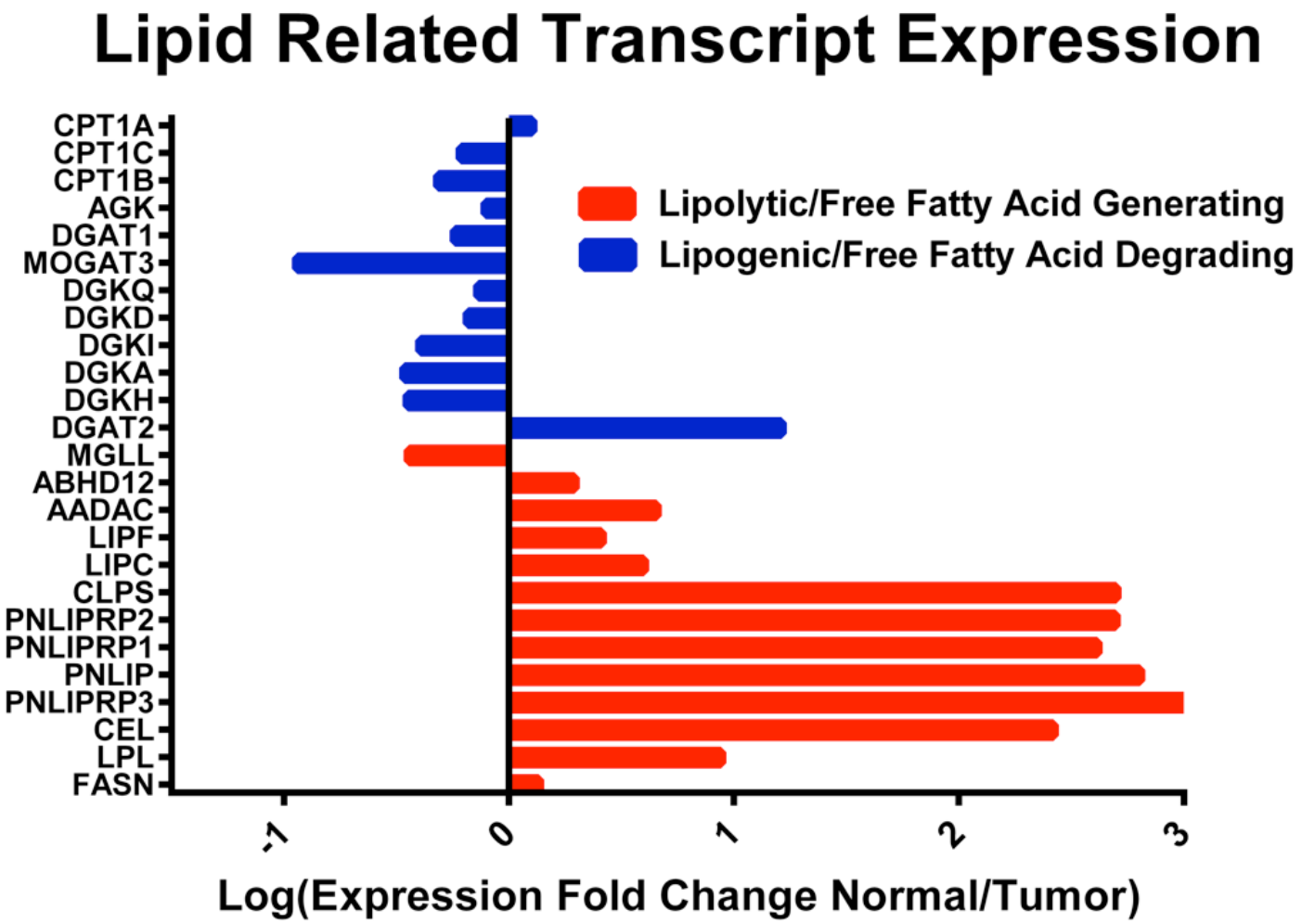
Integrated analysis of tumor v. normal genomic and metabolomic data reveals role for fatty acids







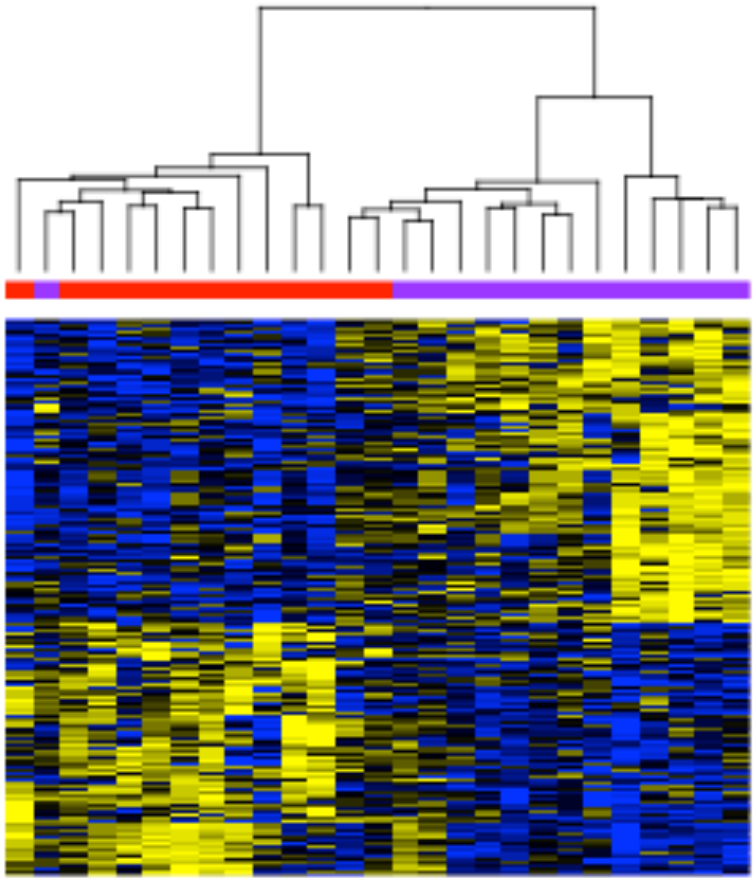
Fatty acid gene expression favors lipogenic processes



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

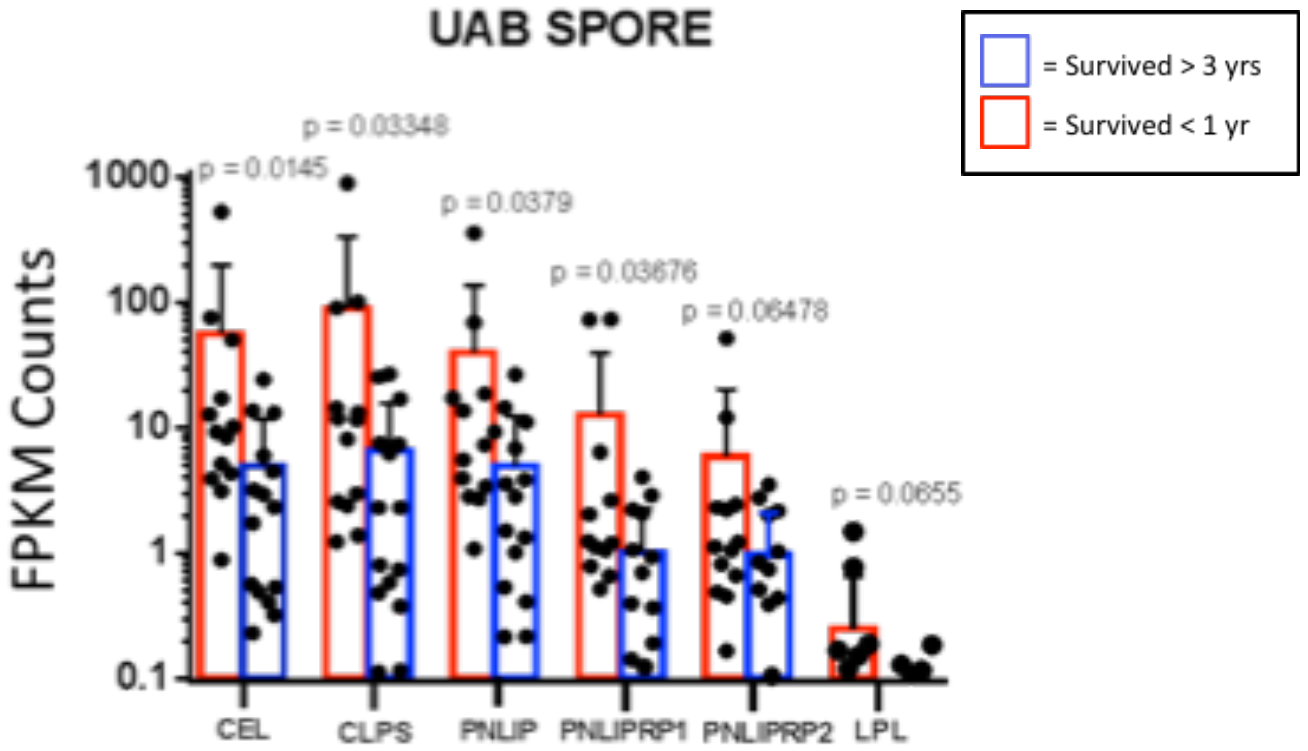
Long-survivors

Pathway enrichment of transcripts over-represented in survival analysis

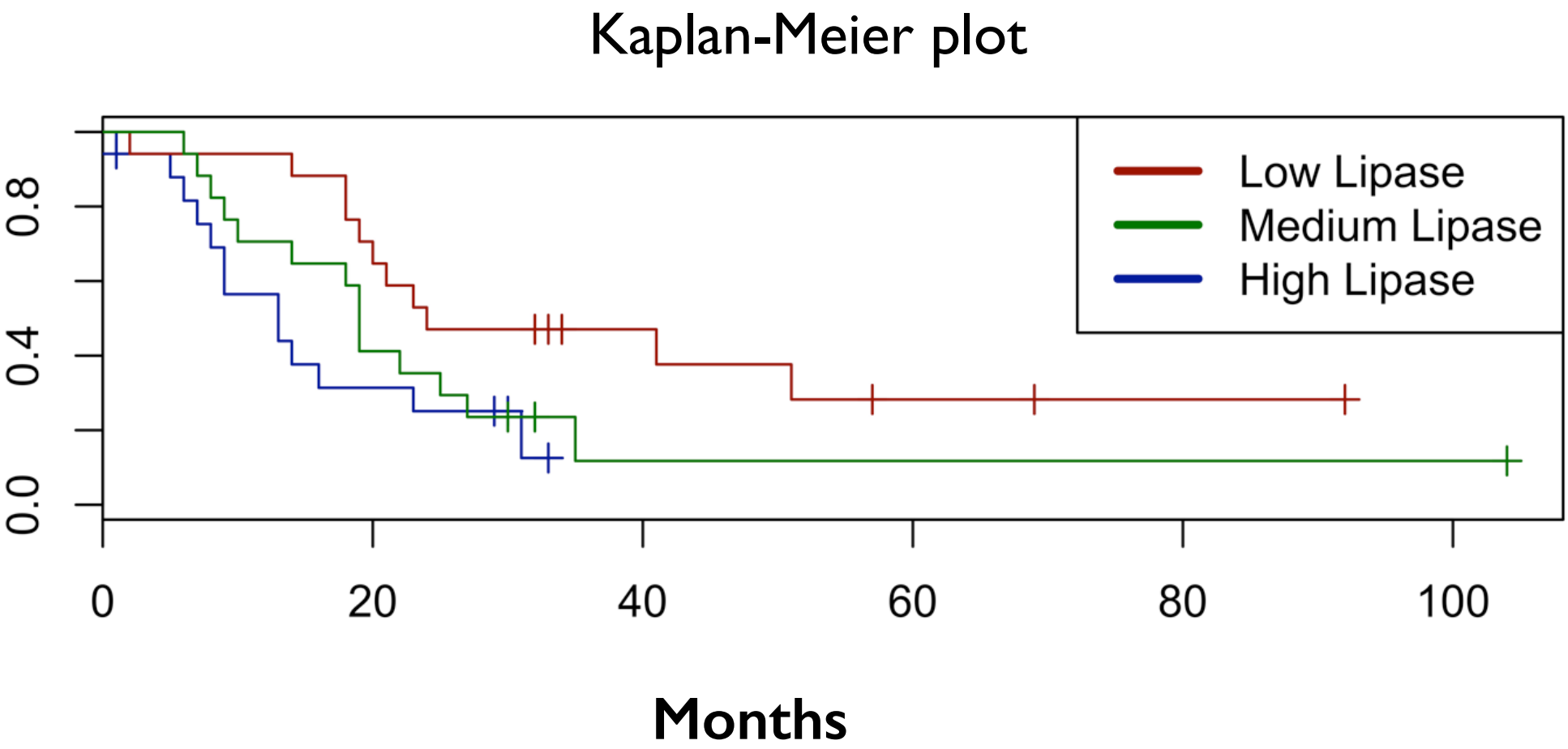
Pathway	Number Overlapping (Total)	Q - Value
Digestion of Dietary Lipid	5(5)	0.00381
Pancreatic Secretion	21(96)	0.00381
Retinoid Metabolism	9(42)	0.267
Triacylglycerol Degradation	5(15)	0.269

- Many of these genes are related to exocrine function
- Previous report suggests an exocrine subtype

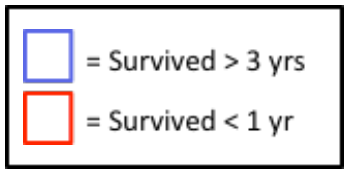
A correlation between survival and lipase expression

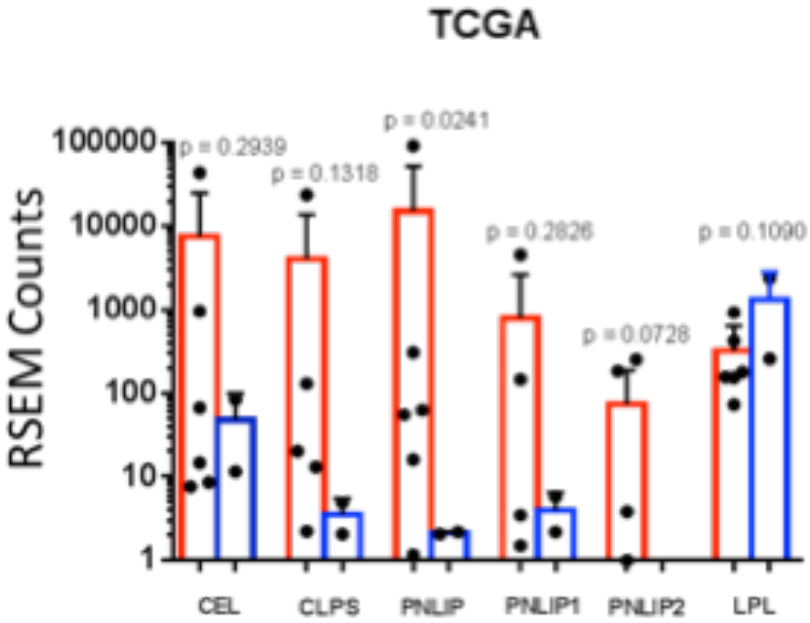


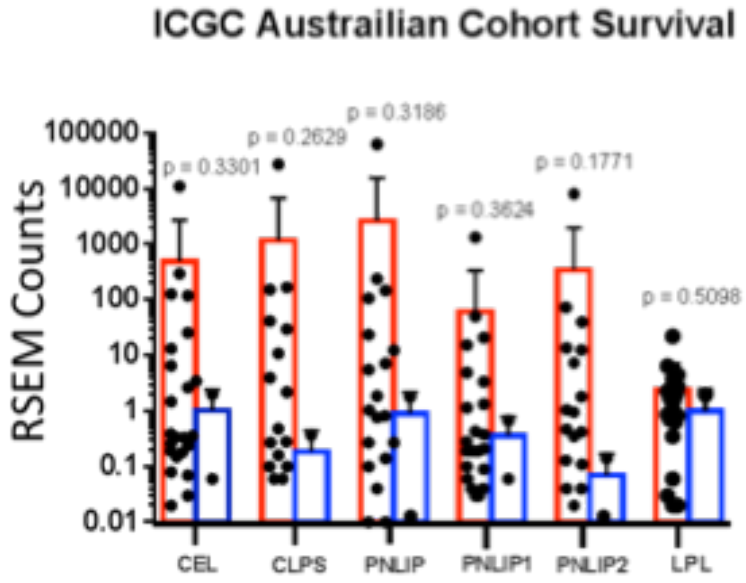
Pancreatic cancer patient survival based on lipase gene expression



Replication in independent samples





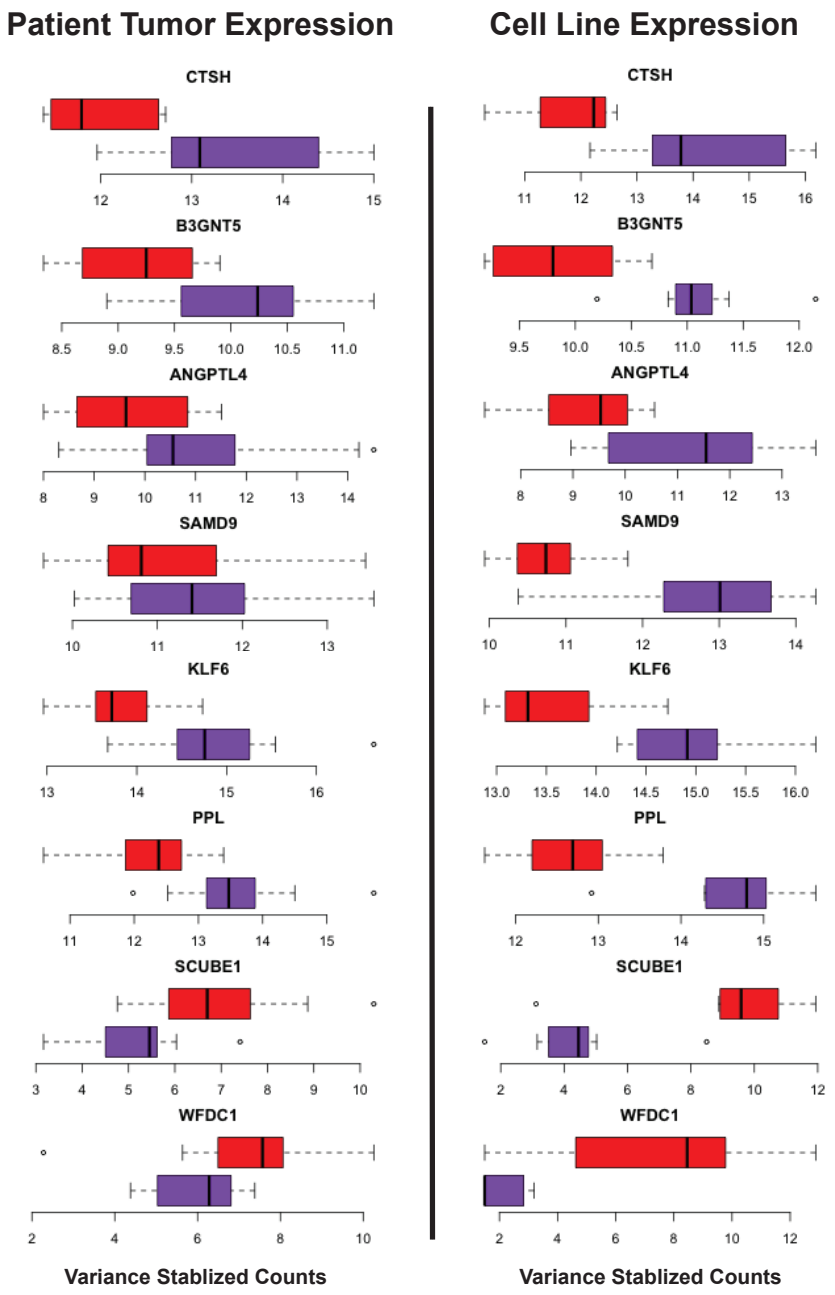


HOW do these genes confer altered survival?

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

Short-survivors/resistant cancer cell lines

Long-survivors/sensitive cancer cell lines



What role do these genes have in survival?

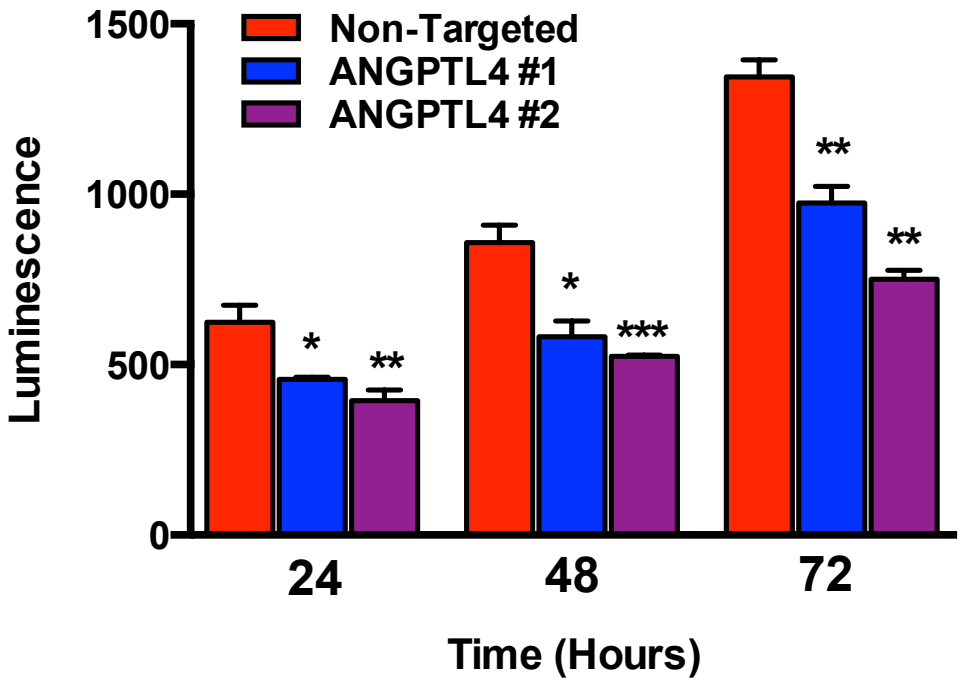
PPAR

↓

ANGPTL4

⊣

LPL



Future Directions

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

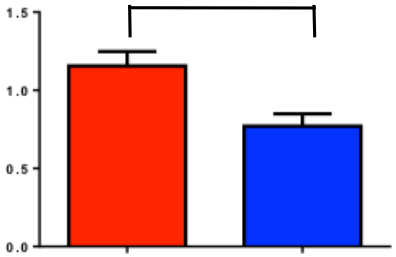
Potential regulators

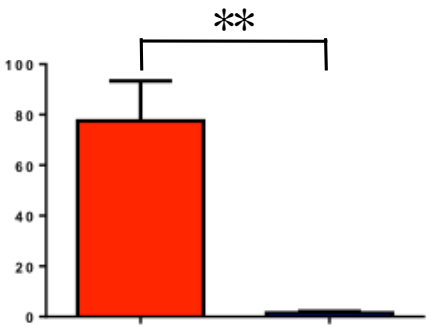
Kras?

↓

GATA6 → RBPJL

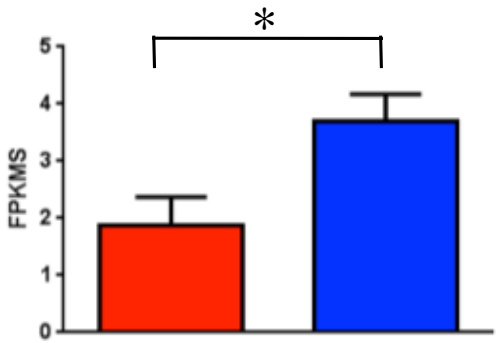
with PTF1-J complex

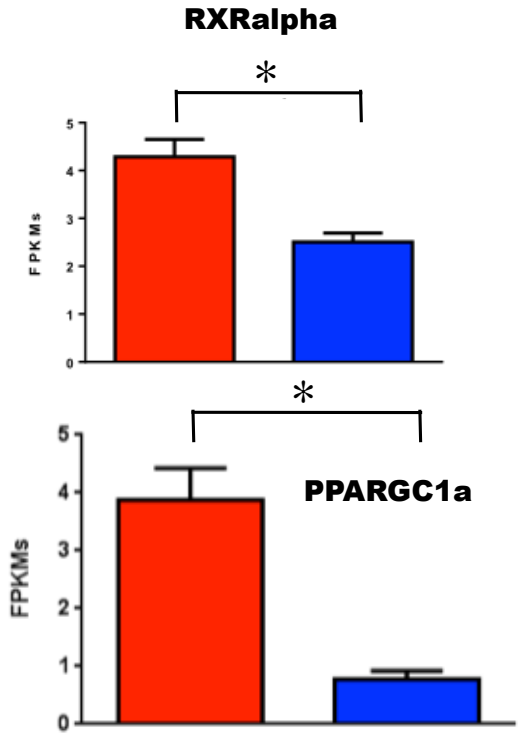




PPARG cofactors

PPARG





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)