Metabolomics by GC-MS

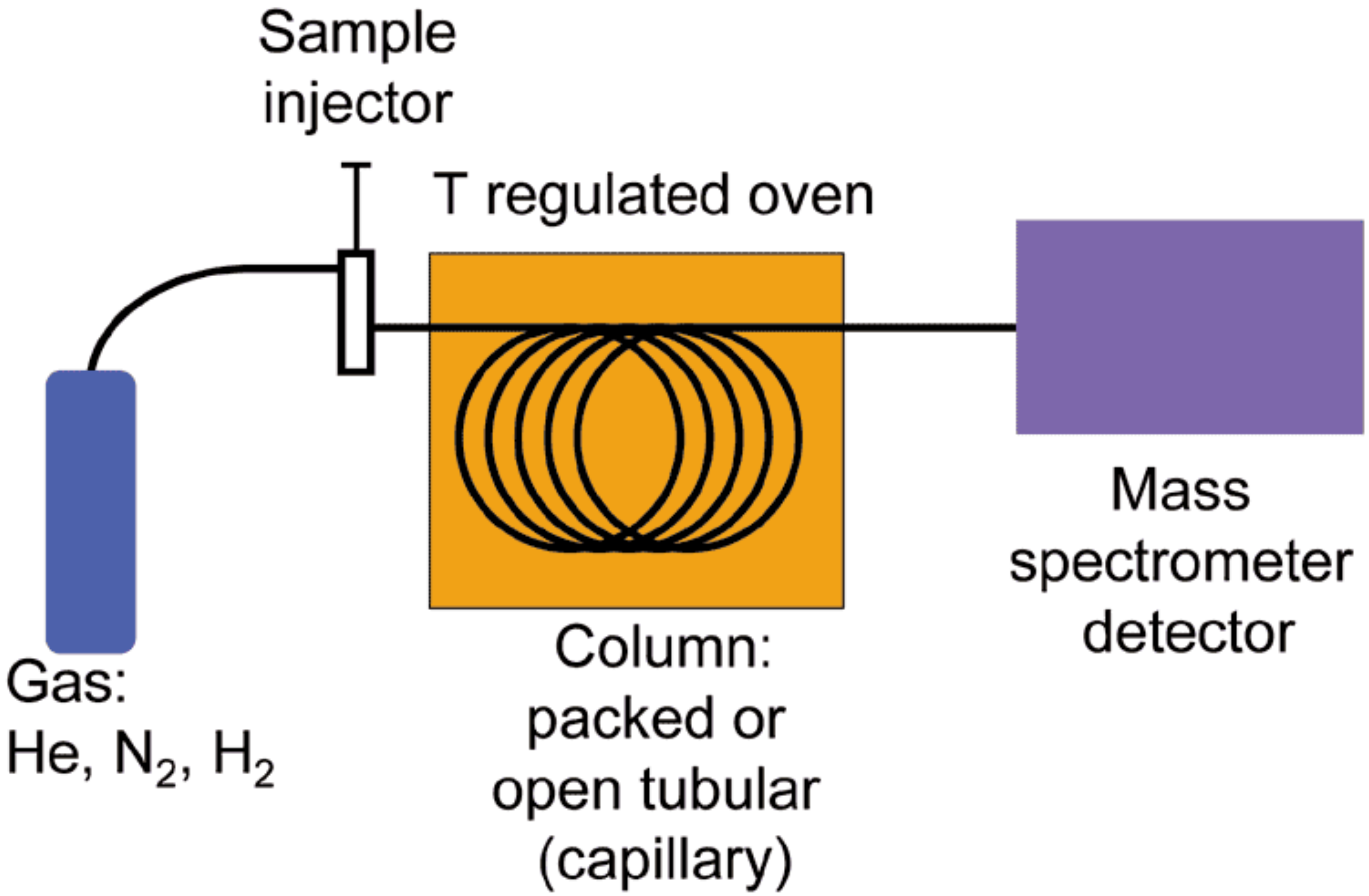
Sara J. Cooper
HudsonAlpha Institute for Biotechnology
Huntsville, AL

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

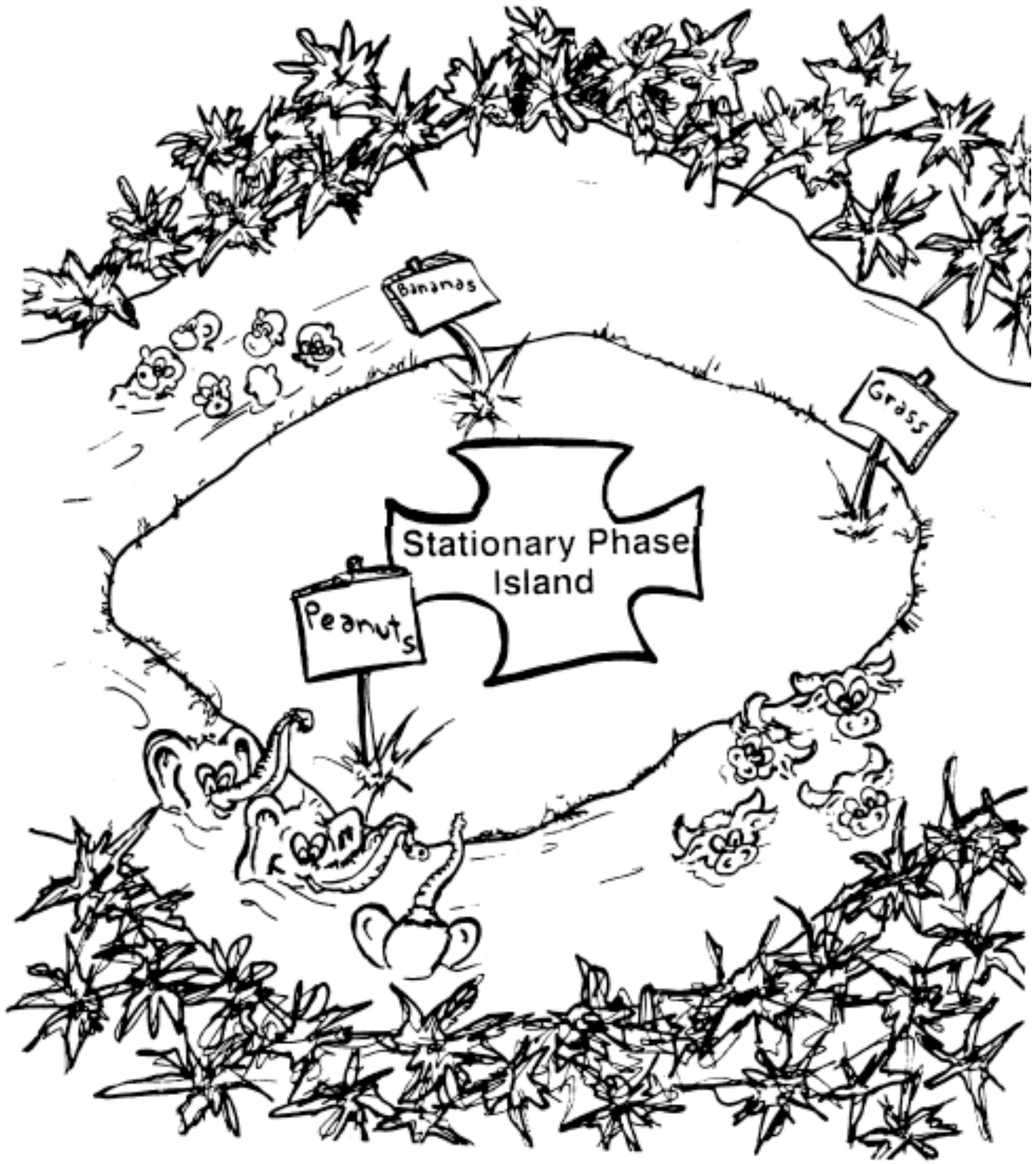
The Nuts and Bolts of GC-MS


Sample injector
T regulated oven
Gas: He, N_2, H_2
Column: packed or open tubular (capillary)
Mass spectrometer detector

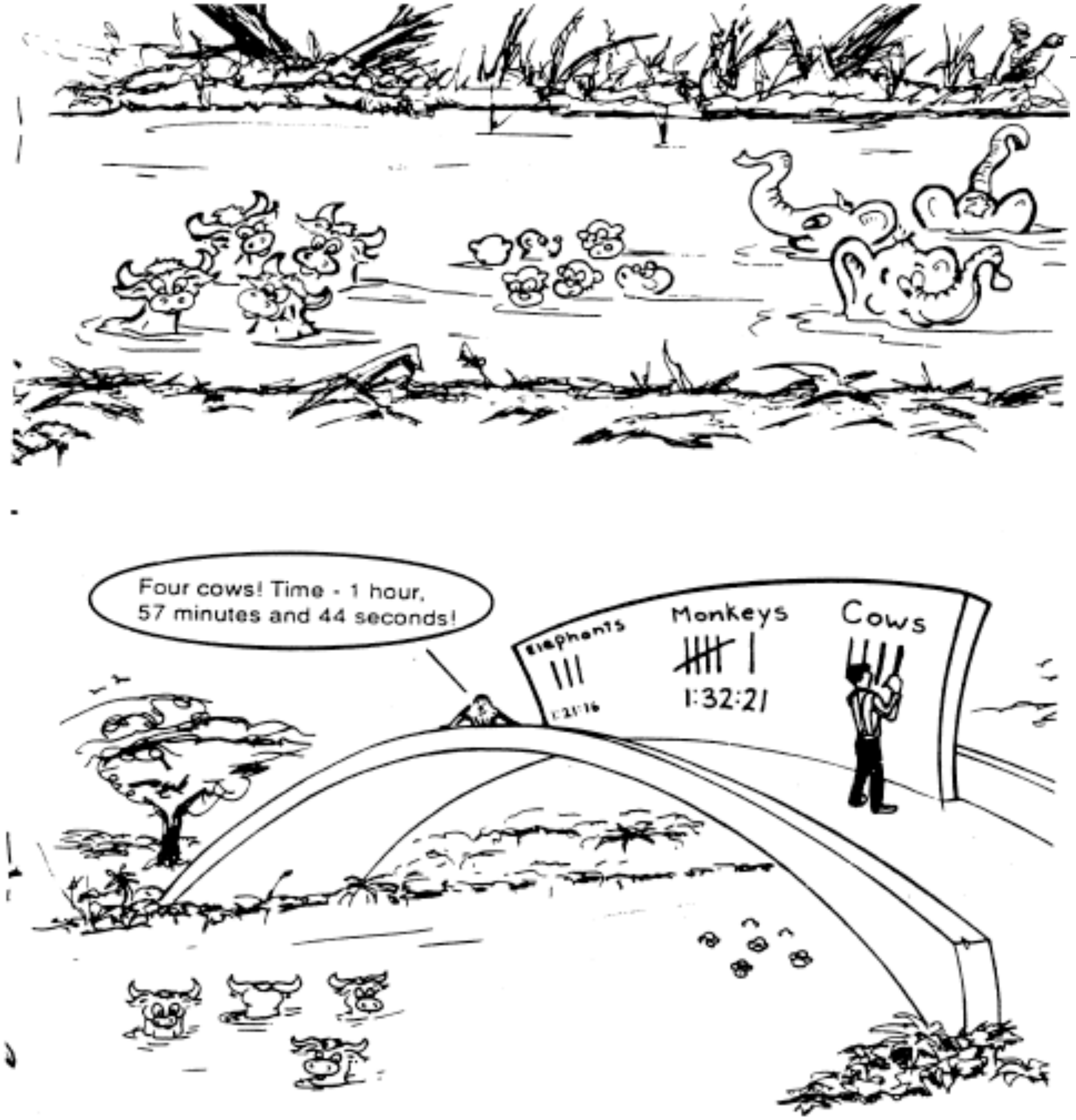
"Gcms schematic" by K. Murray (Kkmurray) - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons

The Principal of GC

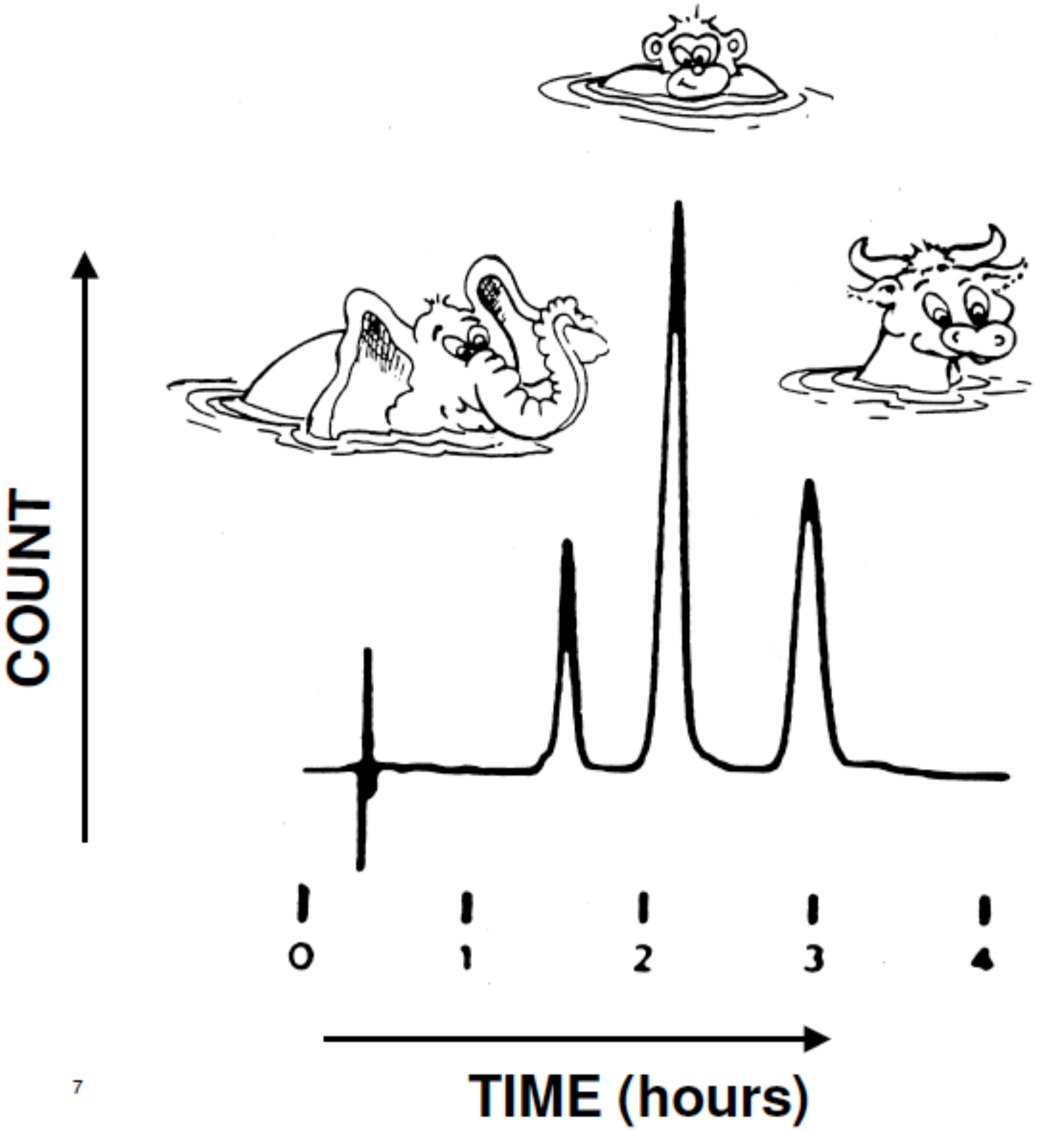
509011-0005



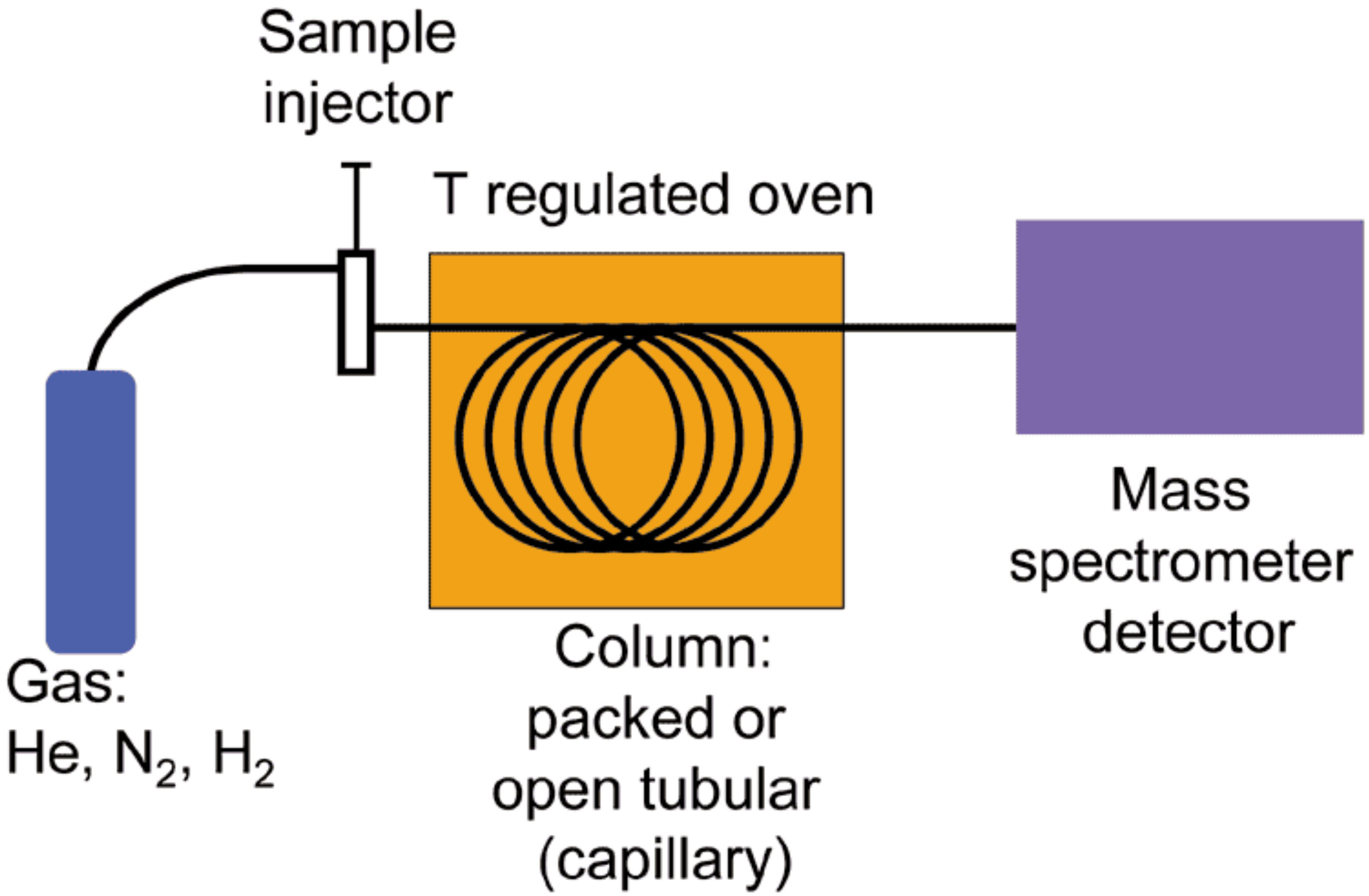
Source-SigmaAldrich ‘thebasicsofgc’



The analysis is now complete.

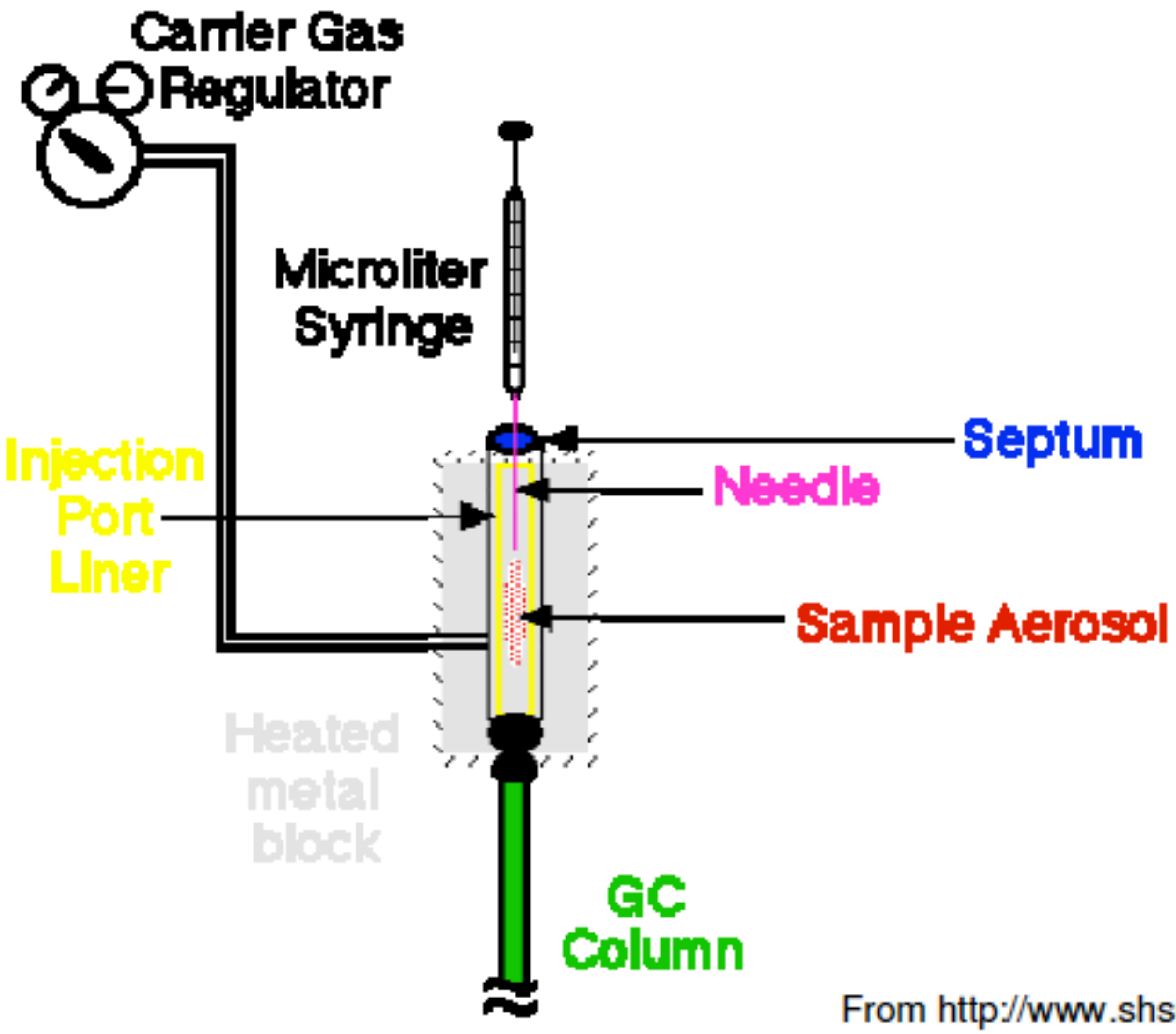


The Nuts and Bolts of GC-MS


Sample injector
T regulated oven
Mass spectrometer detector
Gas: He, N_2, H_2
Column: packed or open tubular (capillary)

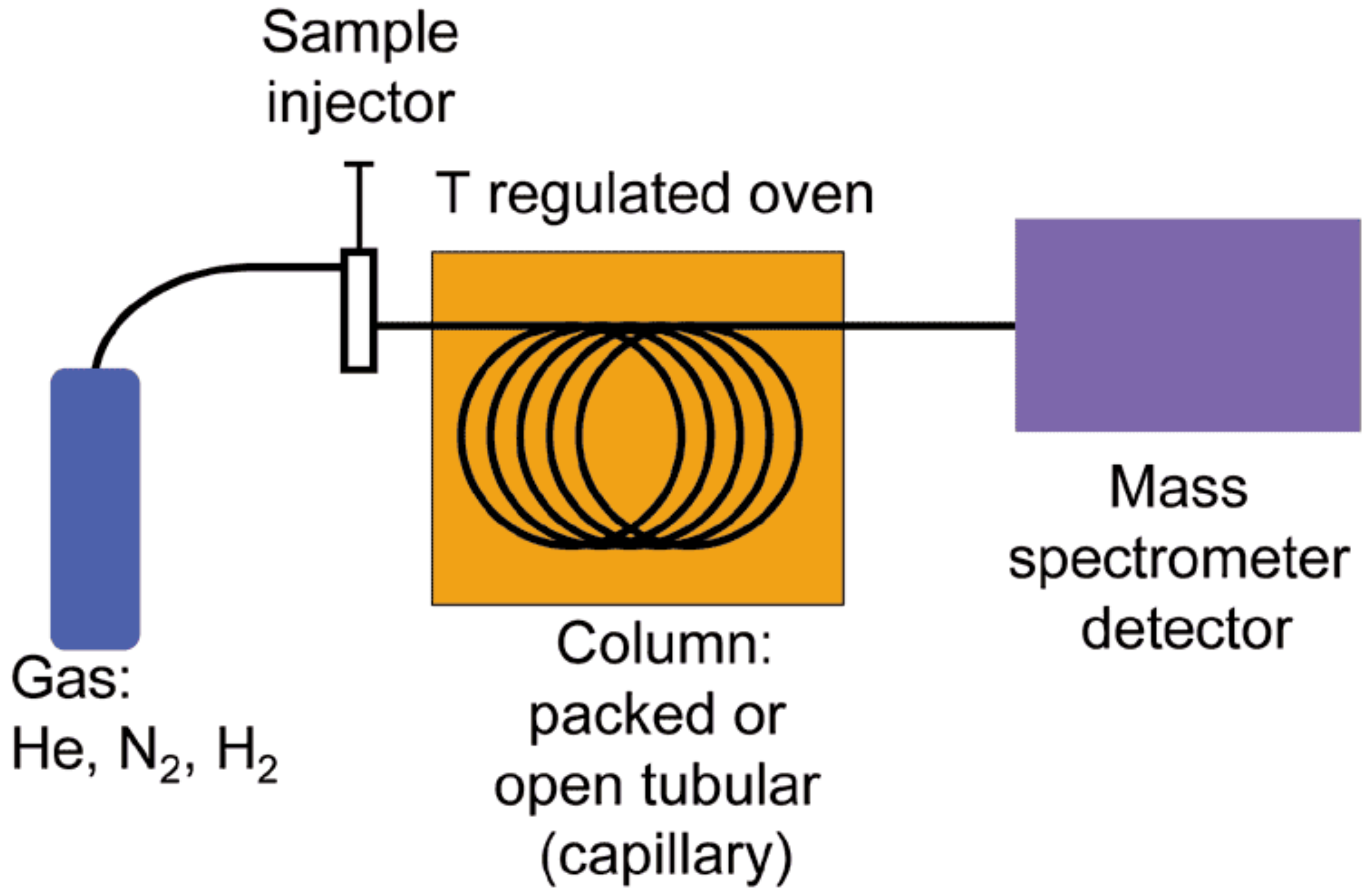
"Gcms schematic" by K. Murray (Kkmurray) - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons

Injection



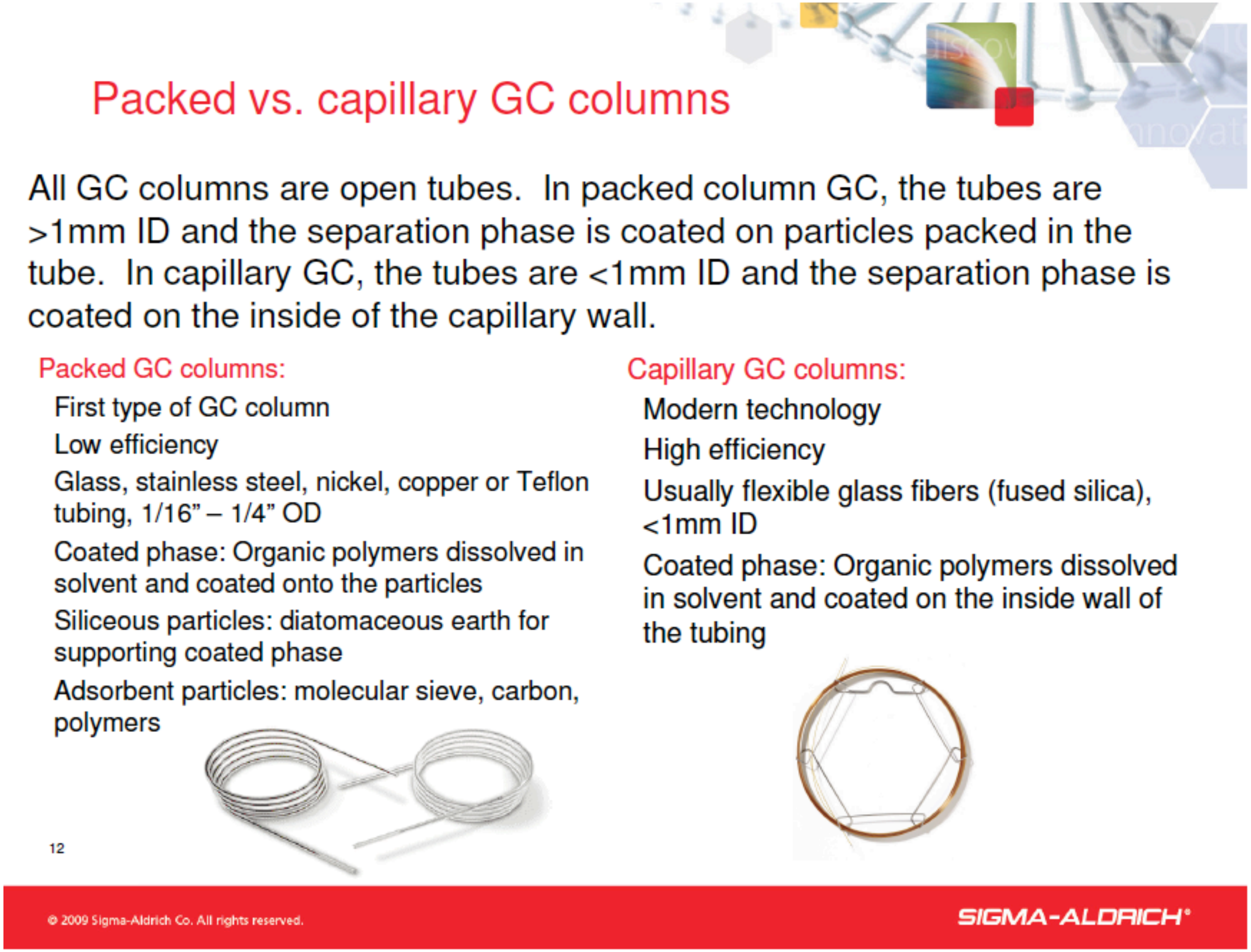
From http://www.shsu.edu/~chemistry/GC/packed.GIF

The Nuts and Bolts of GC-MS


Sample injector
T regulated oven
Mass spectrometer detector
Gas: He, N_2, H_2
Column: packed or open tubular (capillary)

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

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

12

SIGMA-ALDRICH®

Capillary columns can be long (20-100m)

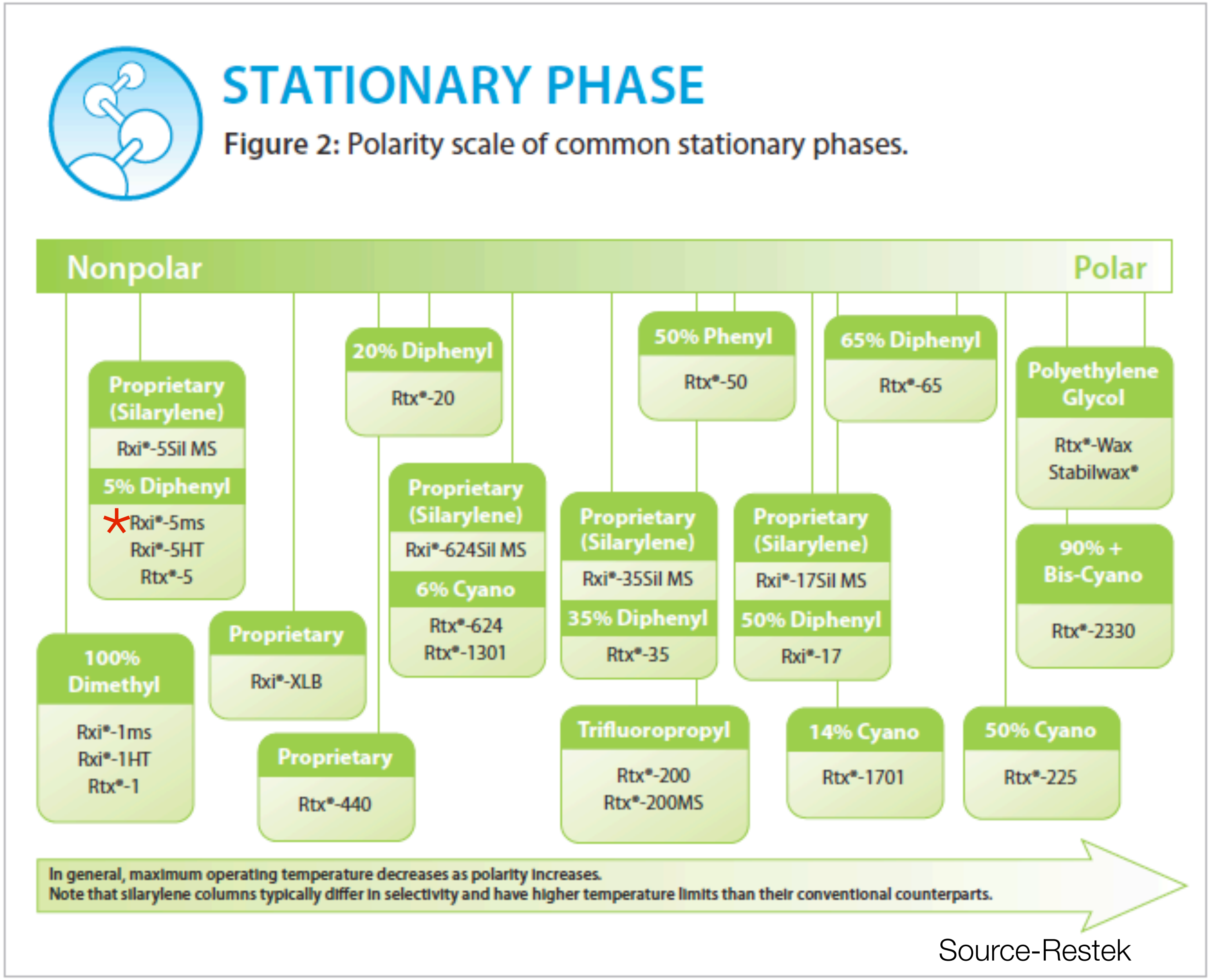
Better separation for complex mixtures

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



STATIONARY PHASE

Figure 2: Polarity scale of common stationary phases.

Nonpolar → Polar

100% Dimethyl: Rxi®-1ms, Rxi®-1HT, Rtx®-1

Proprietary (Silarylene): Rxi®-5Sil MS

5% Diphenyl: *Rxi®-5ms, Rxi®-5HT, Rtx®-5

Proprietary: Rxi®-XLB

Proprietary: Rtx®-440

20% Diphenyl: Rtx®-20

Proprietary (Silarylene): Rxi®-624Sil MS

6% Cyano: Rtx®-624, Rtx®-1301

Proprietary (Silarylene): Rxi®-35Sil MS

35% Diphenyl: Rtx®-35

Trifluoropropyl: Rtx®-200, Rtx®-200MS

50% Phenyl: Rtx®-50

Proprietary (Silarylene): Rxi®-17Sil MS

50% Diphenyl: Rxi®-17

14% Cyano: Rtx®-1701

65% Diphenyl: Rtx®-65

50% Cyano: Rtx®-225

Polyethylene Glycol: Rtx®-Wax, Stabilwax®

90% + Bis-Cyano: Rtx®-2330

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

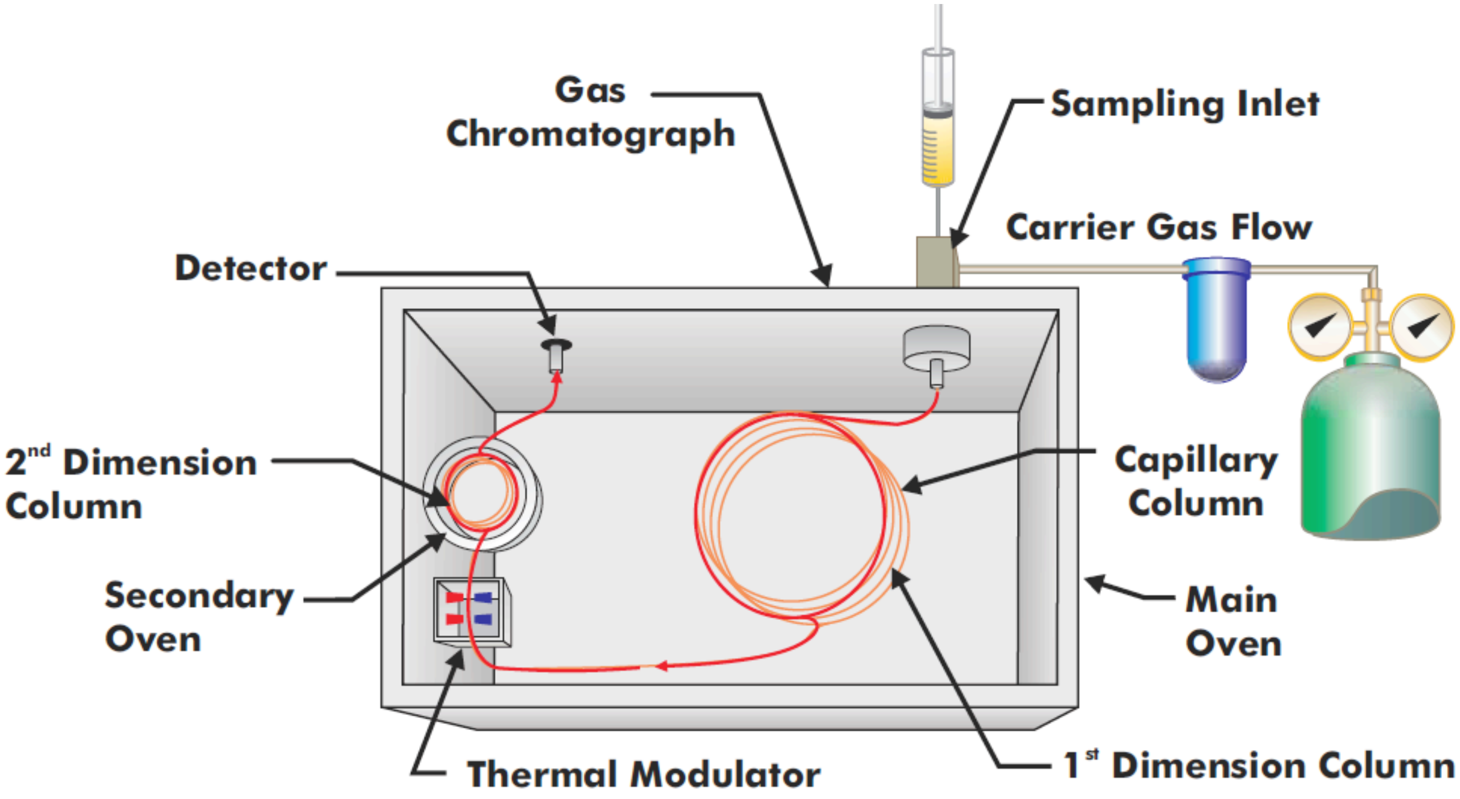
Source-Restek

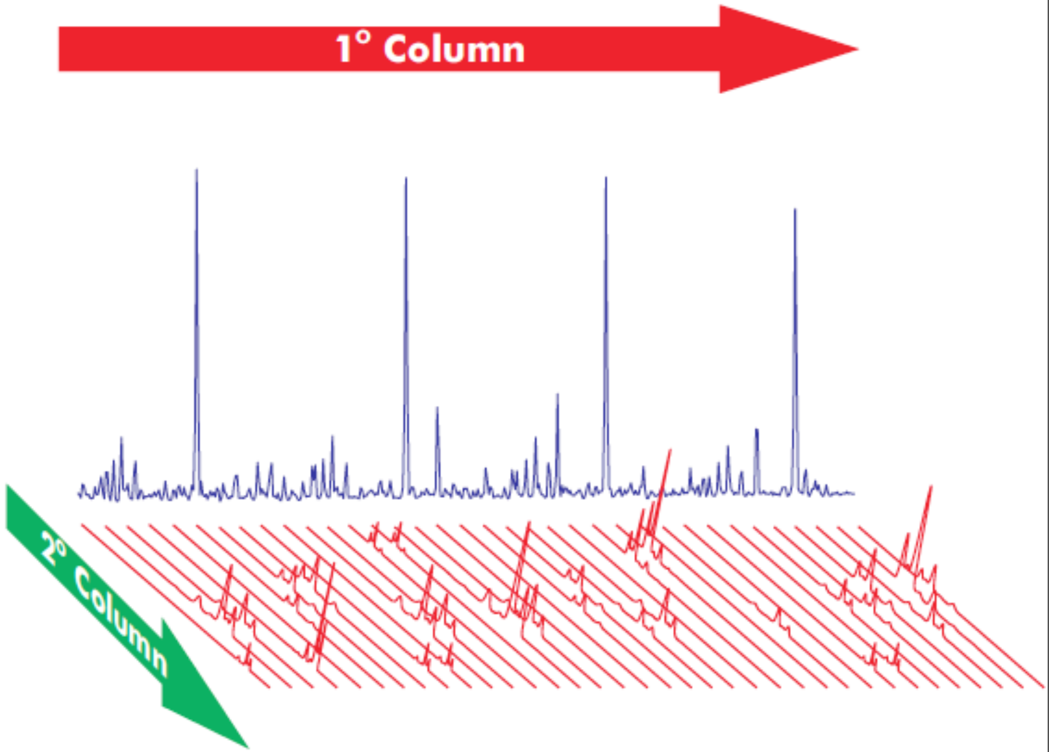
tech tip

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

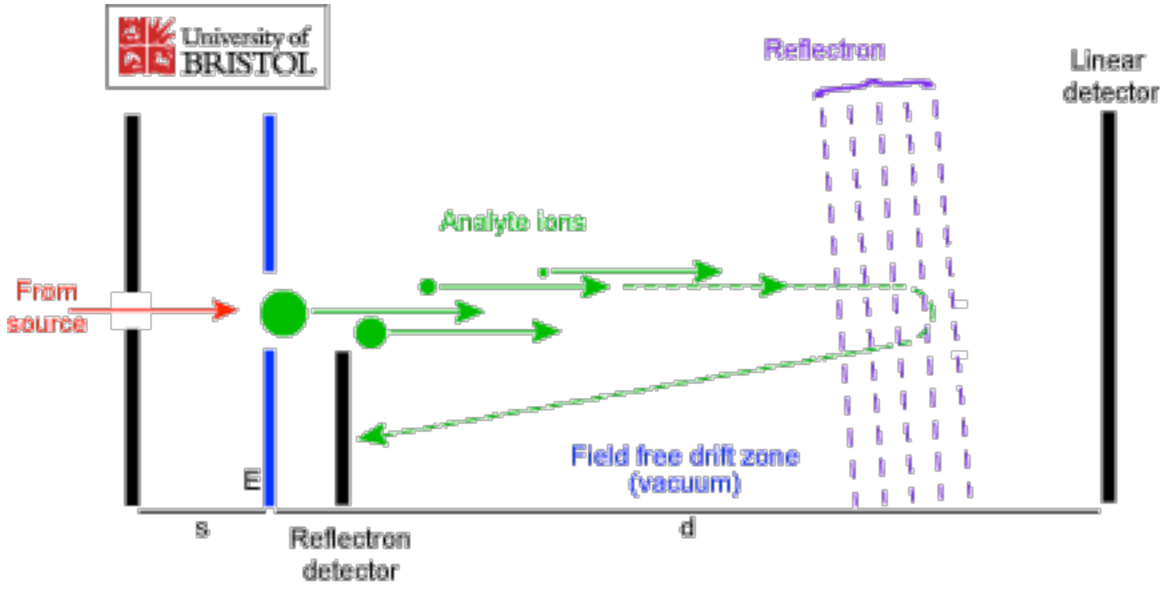




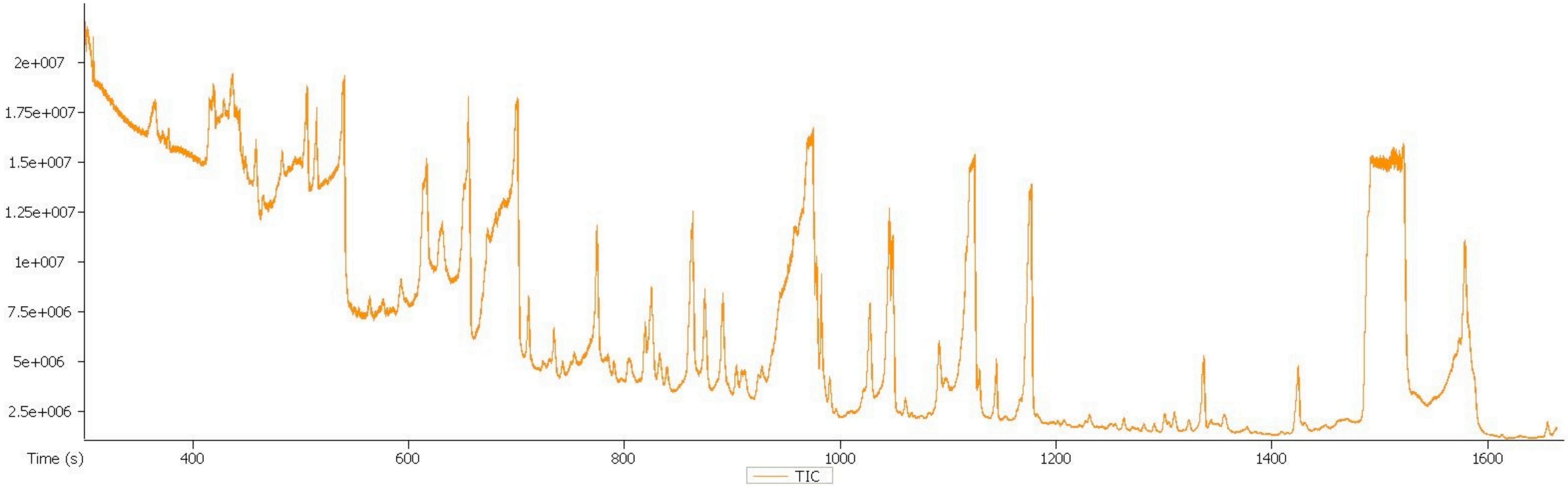
Source: Leco Corp

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

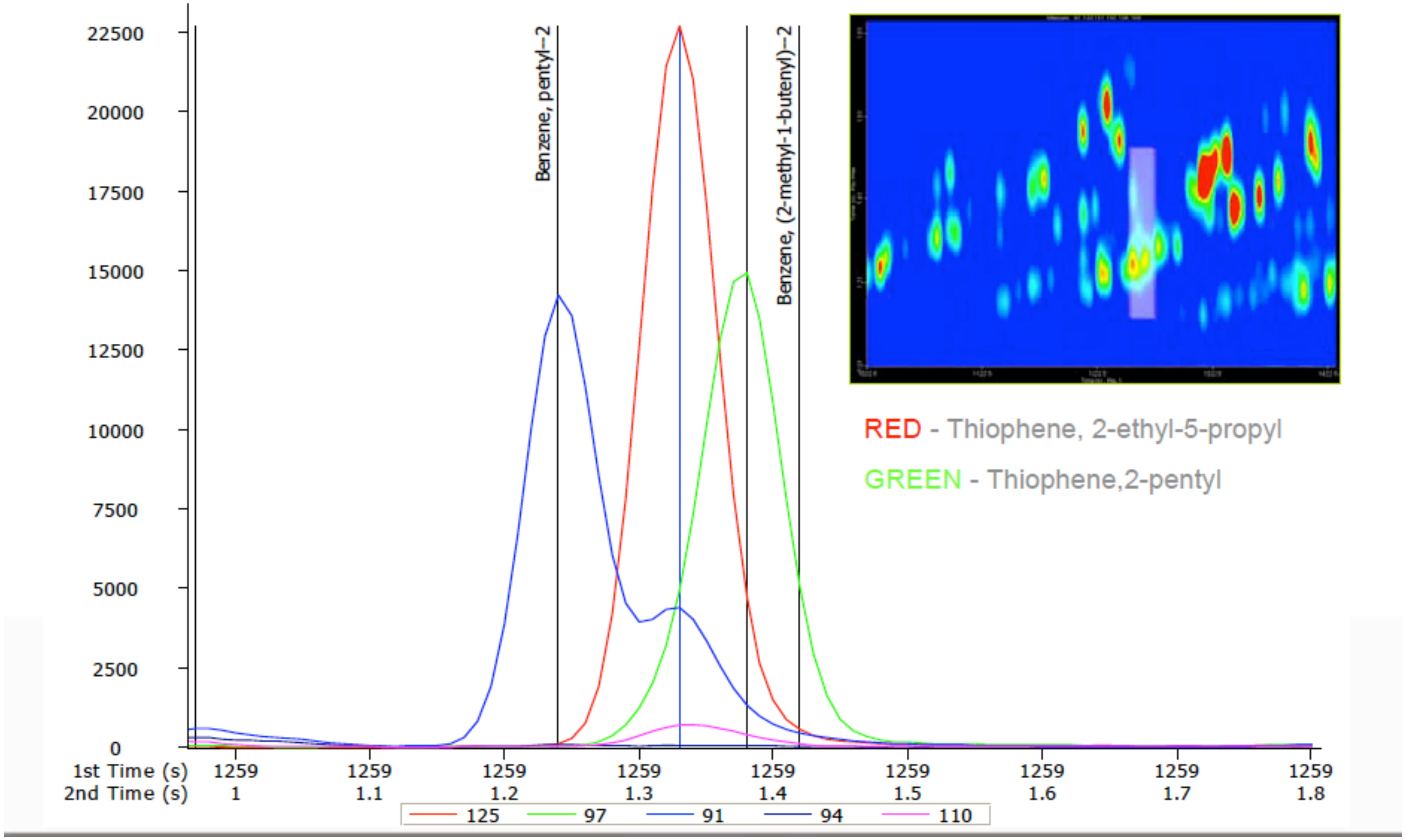


Example data output-Chromatogram



Signal Deconvolution

True Signal Deconvolution®



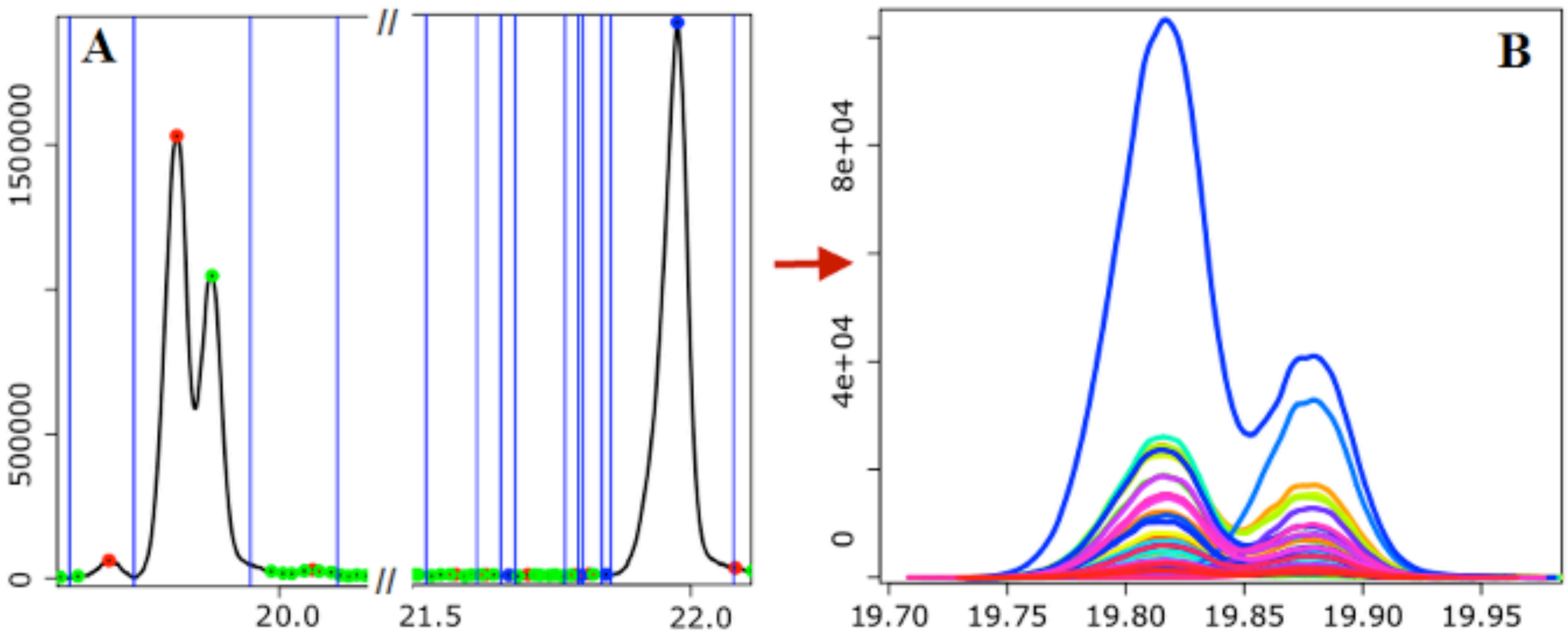
RED - Thiophene, 2-ethyl-5-propyl

GREEN - Thiophene,2-pentyl

Source: Leco

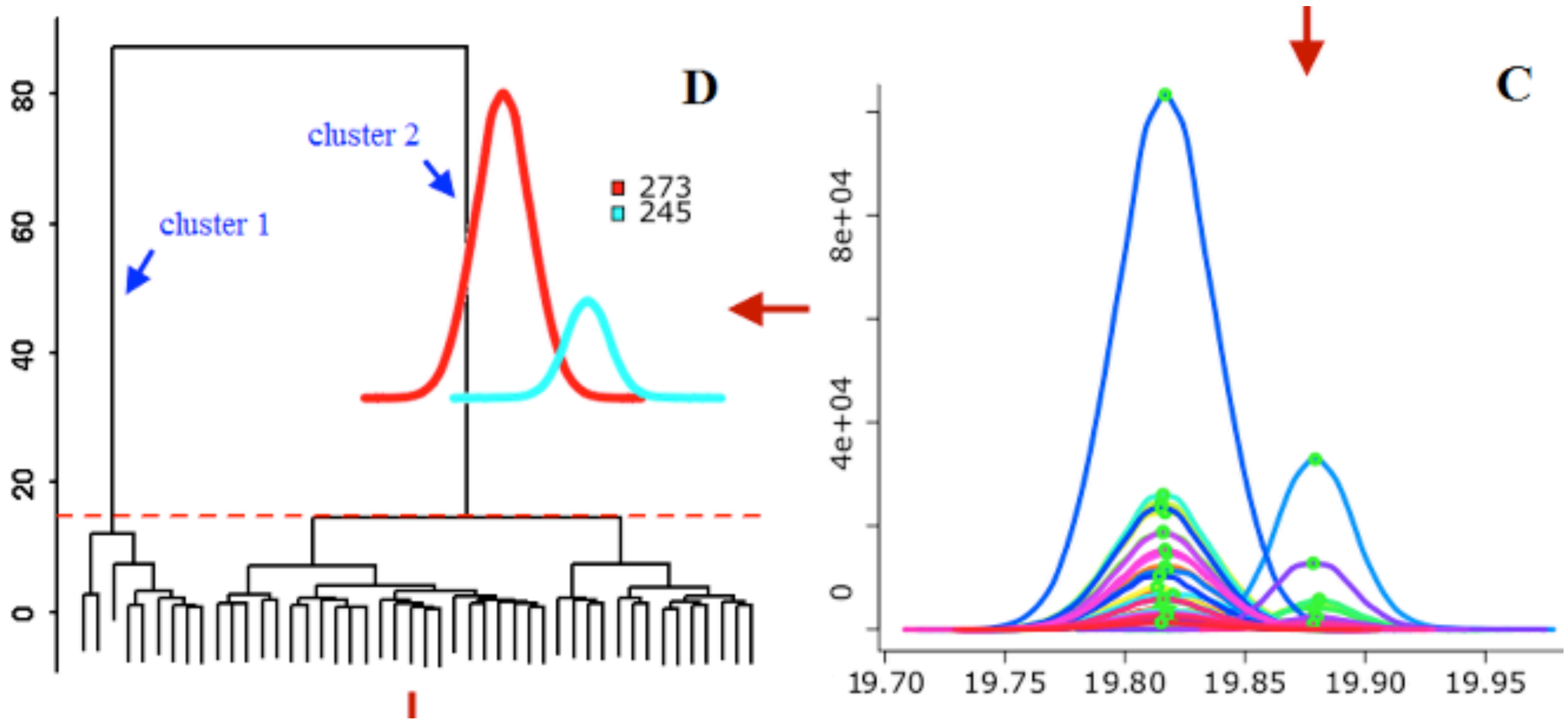
Principles of Deconvolution

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

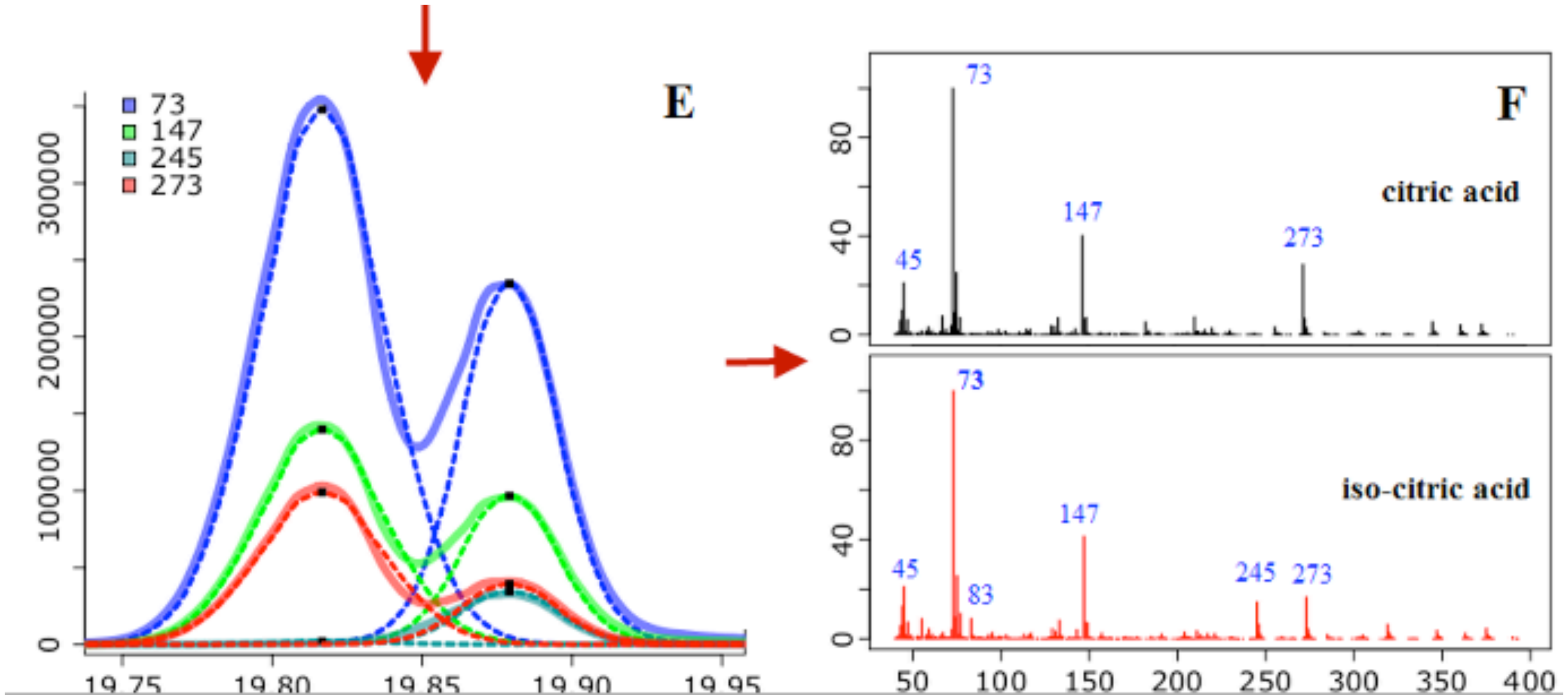


Source: Du and Zeisel 2013

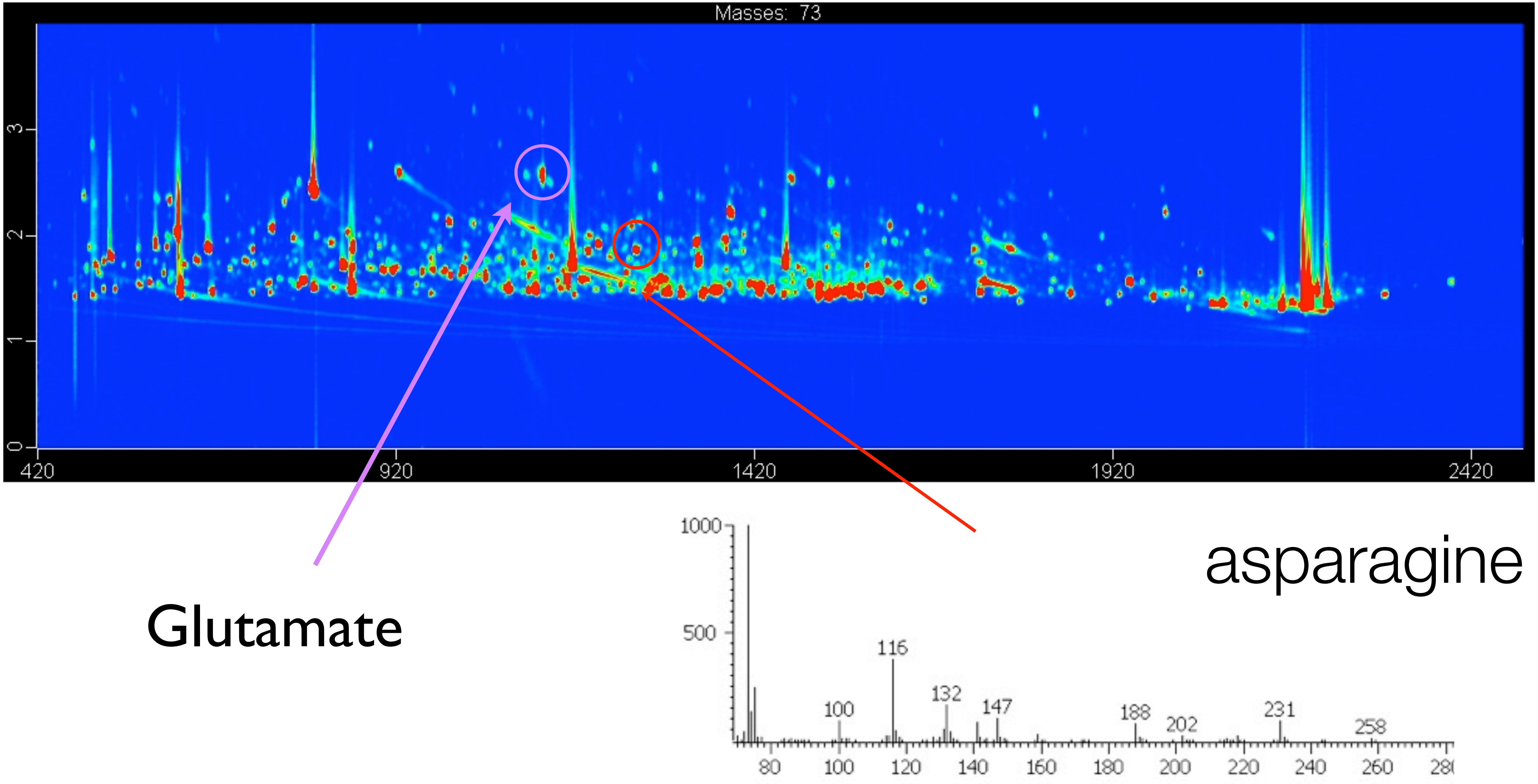
Principles of Deconvolution



Principles of Deconvolution



Data projected into two dimensions

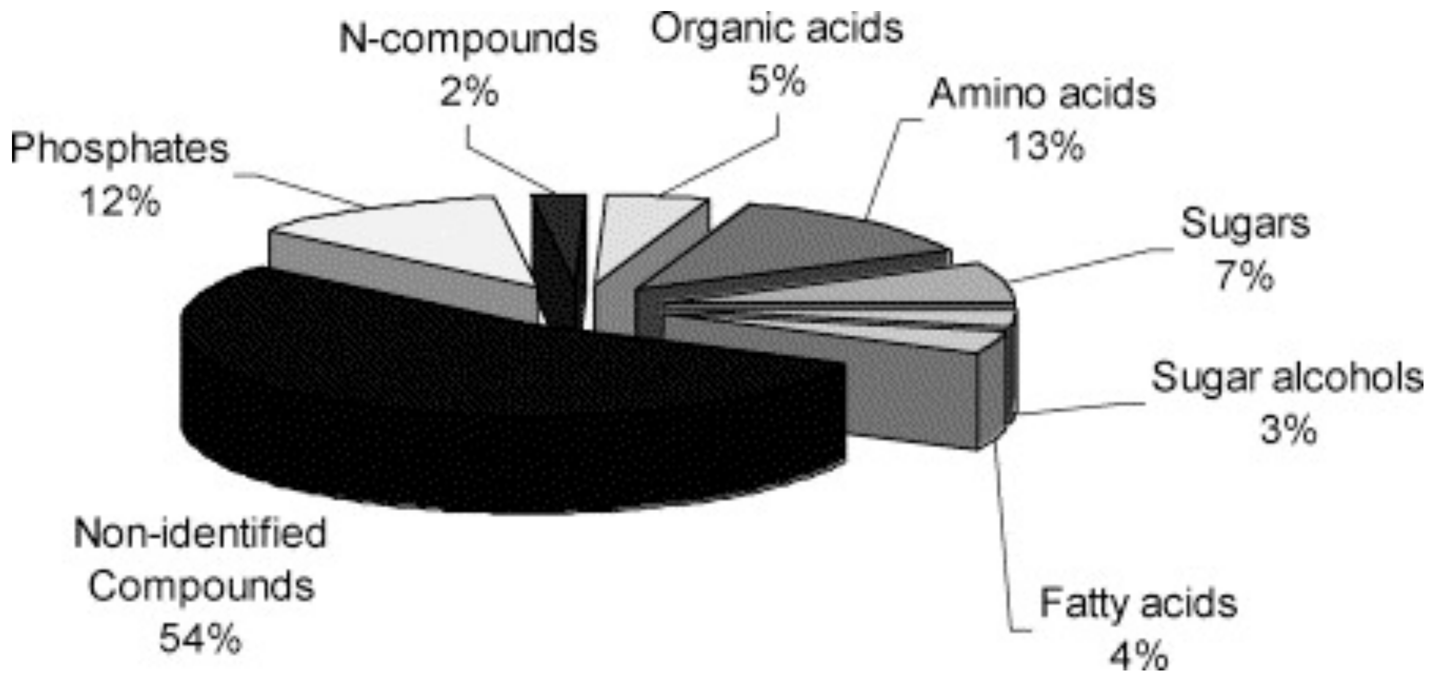


asparagine

Glutamate

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

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 long)
Chromatography	Highly reproducible- Retention indices	Less critical
Metabolite ID	Libraries	Inferred composition by accurate mass

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

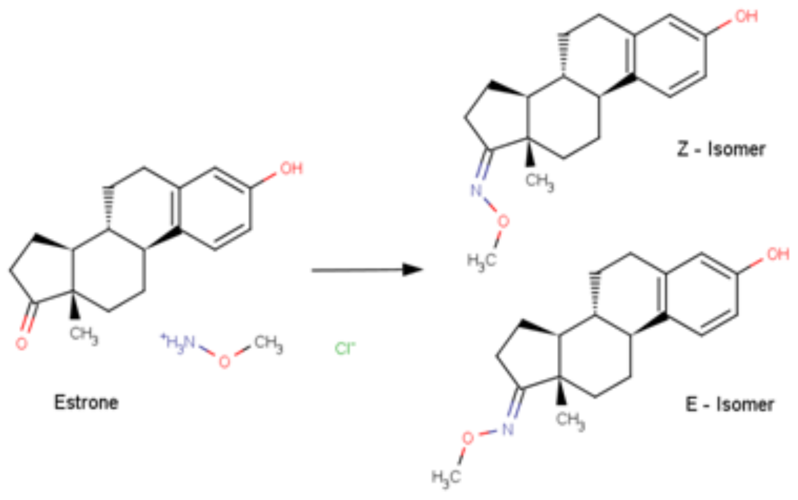
- Experimental Design!! What question(s) do you want to answer
- Sample preparation
- Data collection
- Preliminary Data analysis
  - tools
- Metabolite identification

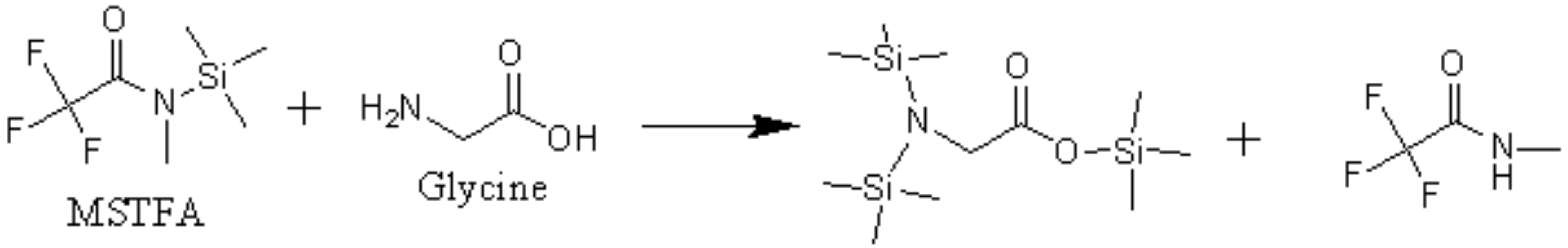
Sample procurement/preparation

- 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 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

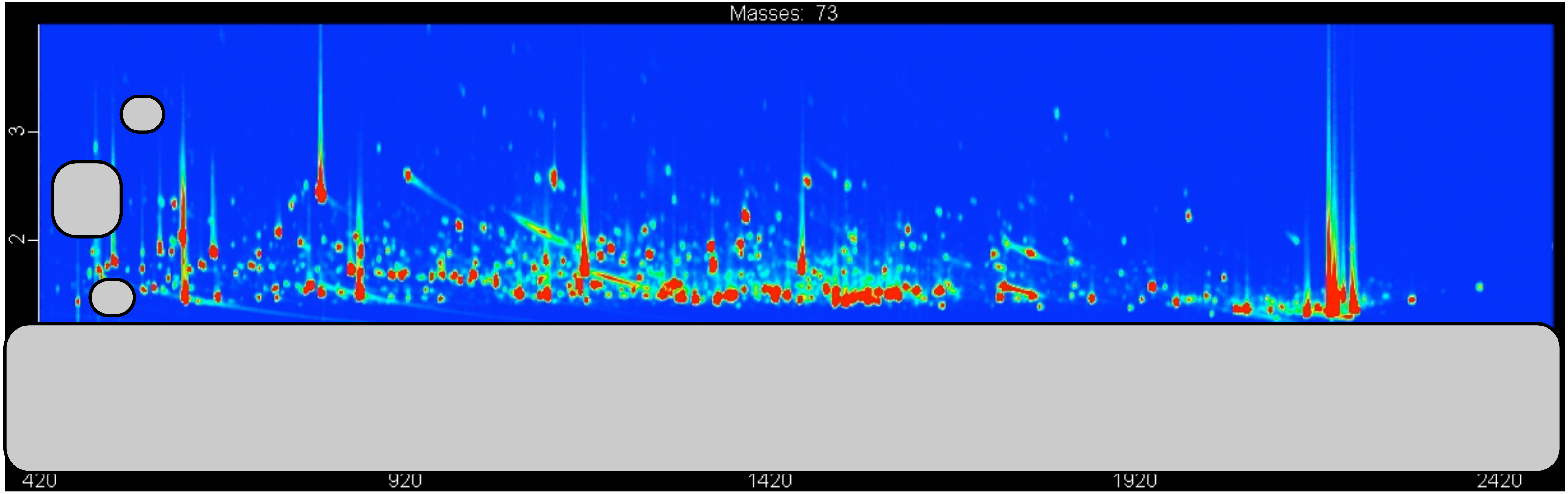
- Most instruments utilize proprietary software to do peak deconvolution
- Raw data can be analyzed as well and there are tools out there to analyze raw data (e.g. Metlin)
- ChromaTOF (Leco's peak calling and deconvolution software) Output:
  - List of peaks
  - Determination of Quant Mass for each peak (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%)

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

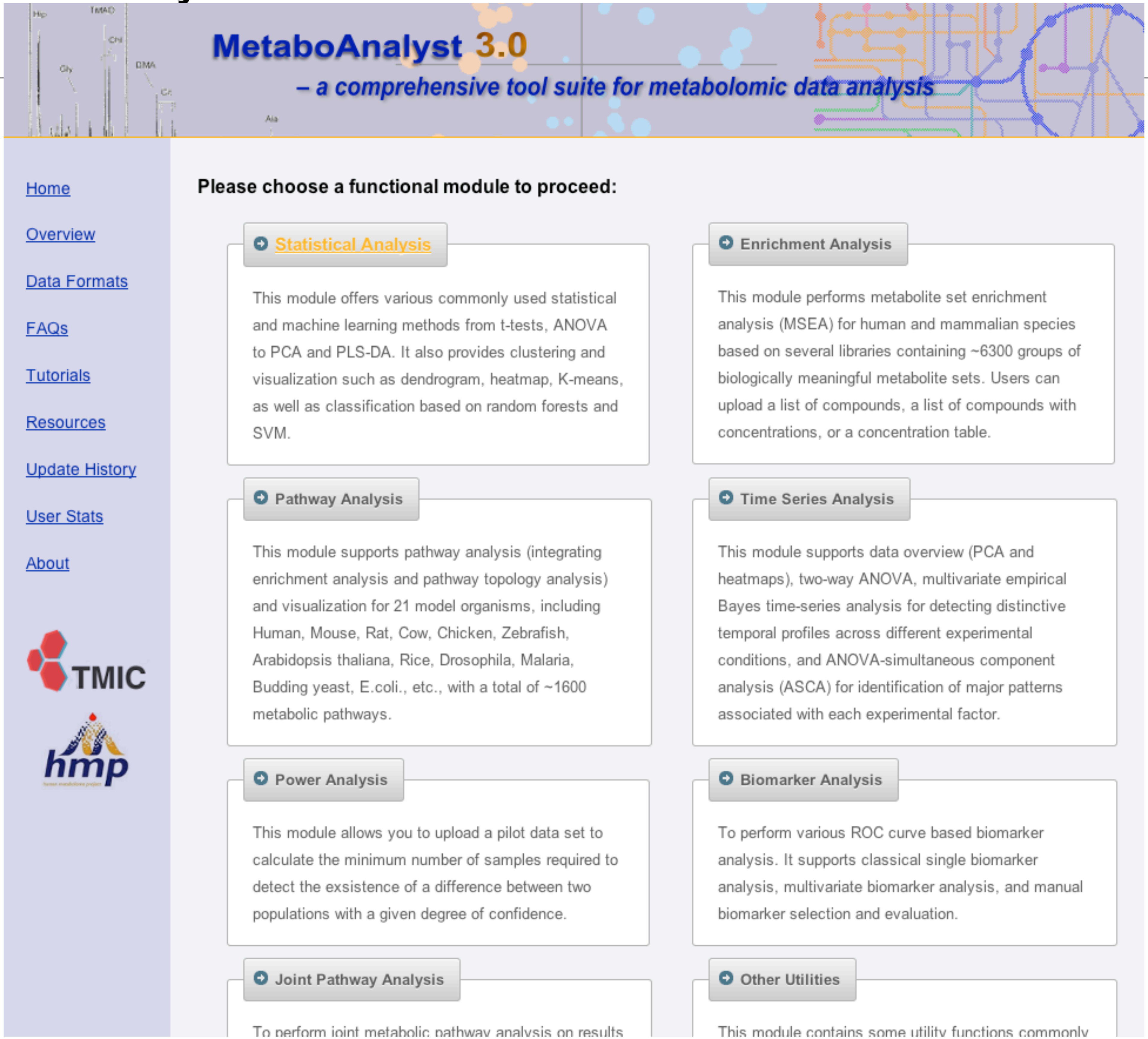
- Common Practice:
  - Injection Control (A known amount of substance is injected with each sample. Those peaks should have the same area each time)
  - Normalization by SUM (total area under the curve). Normalizes for overall sample concentration
  - Clinical samples: normalization by creatinine or 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



MetaboAnalyst 3.0

– a comprehensive tool suite for metabolomic data analysis

- Home
- Overview
- Data Formats
- FAQs
- Tutorials
- Resources
- Update History
- User Stats
- About

TMIC

hmp

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 exsistence 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.

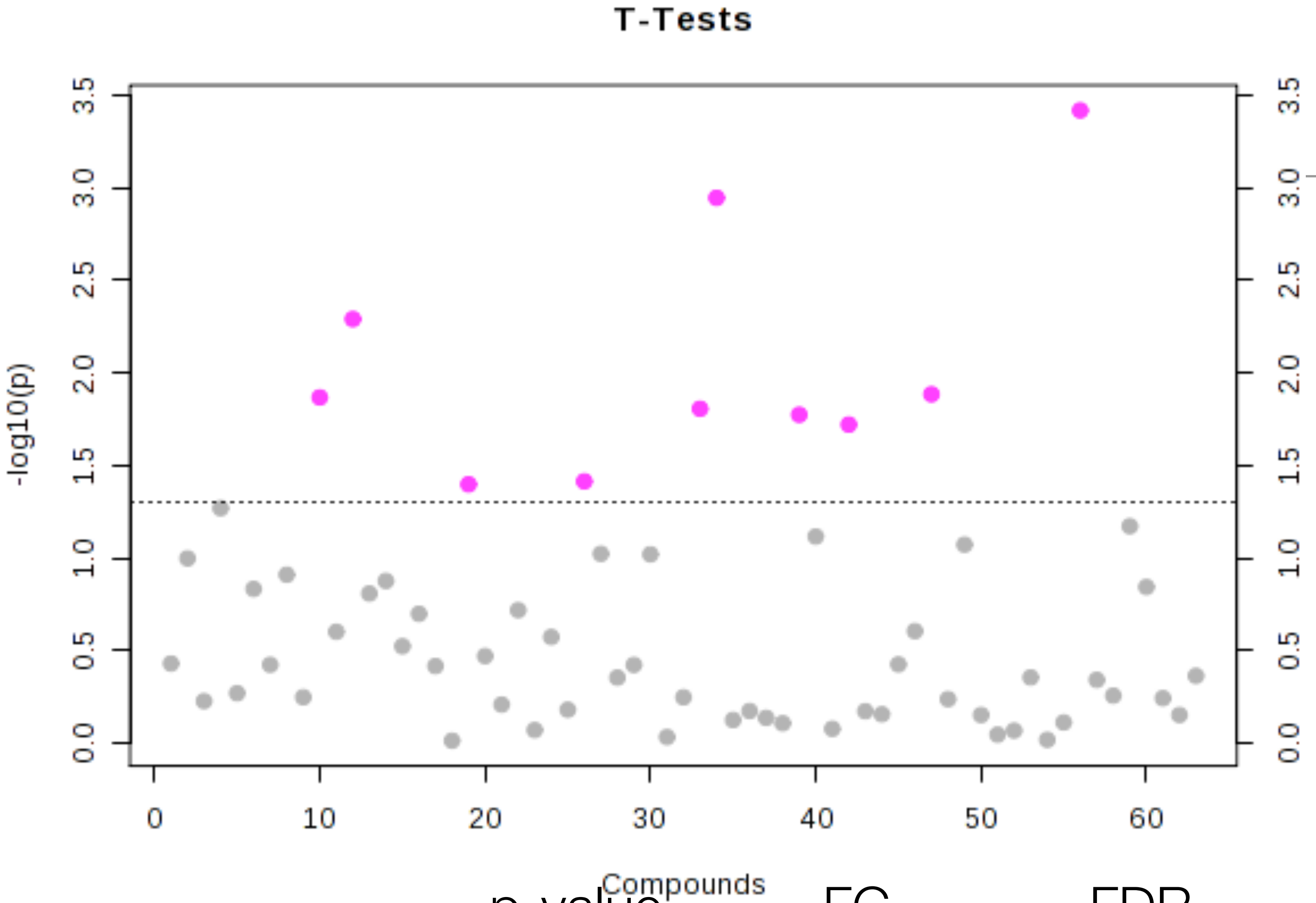
Joint Pathway Analysis

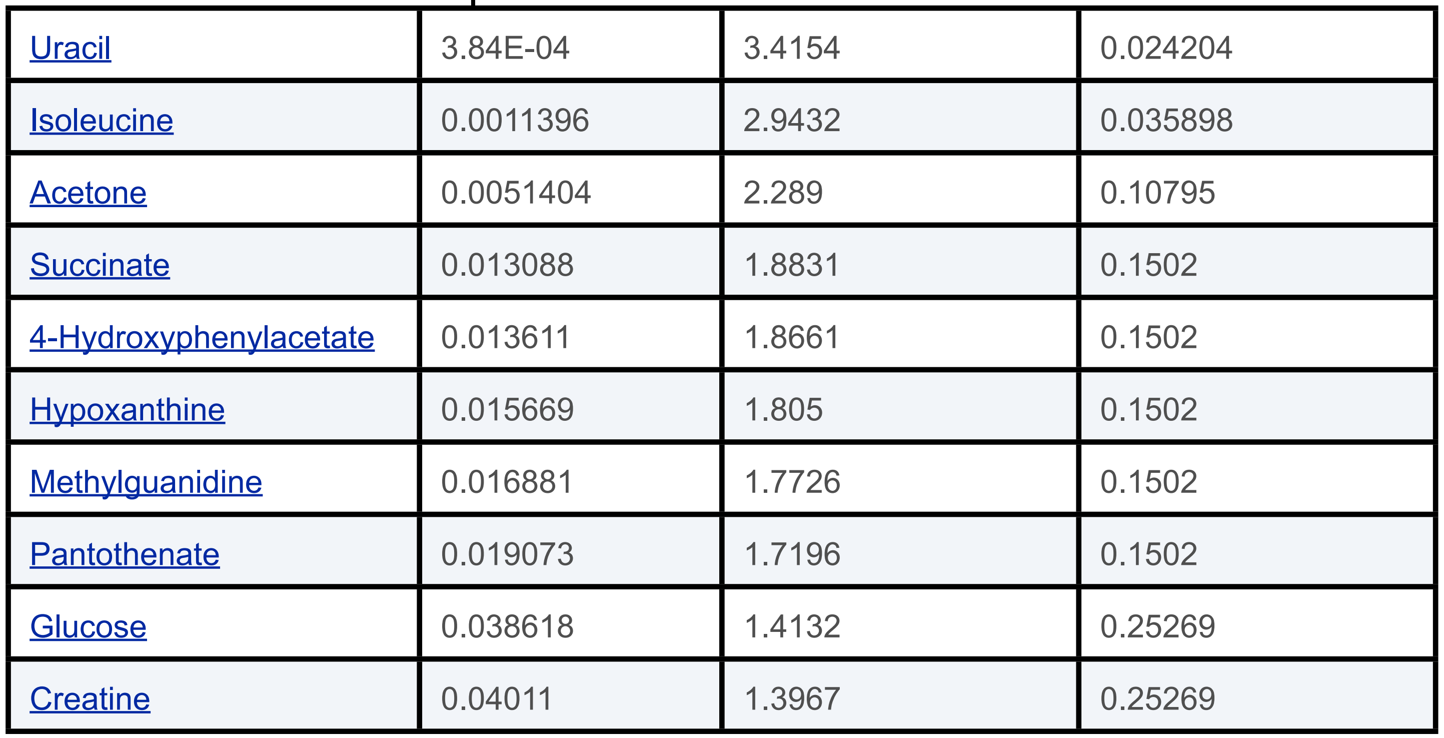
To perform joint metabolic pathway analysis on results

Other Utilities

This module contains some utility functions commonly

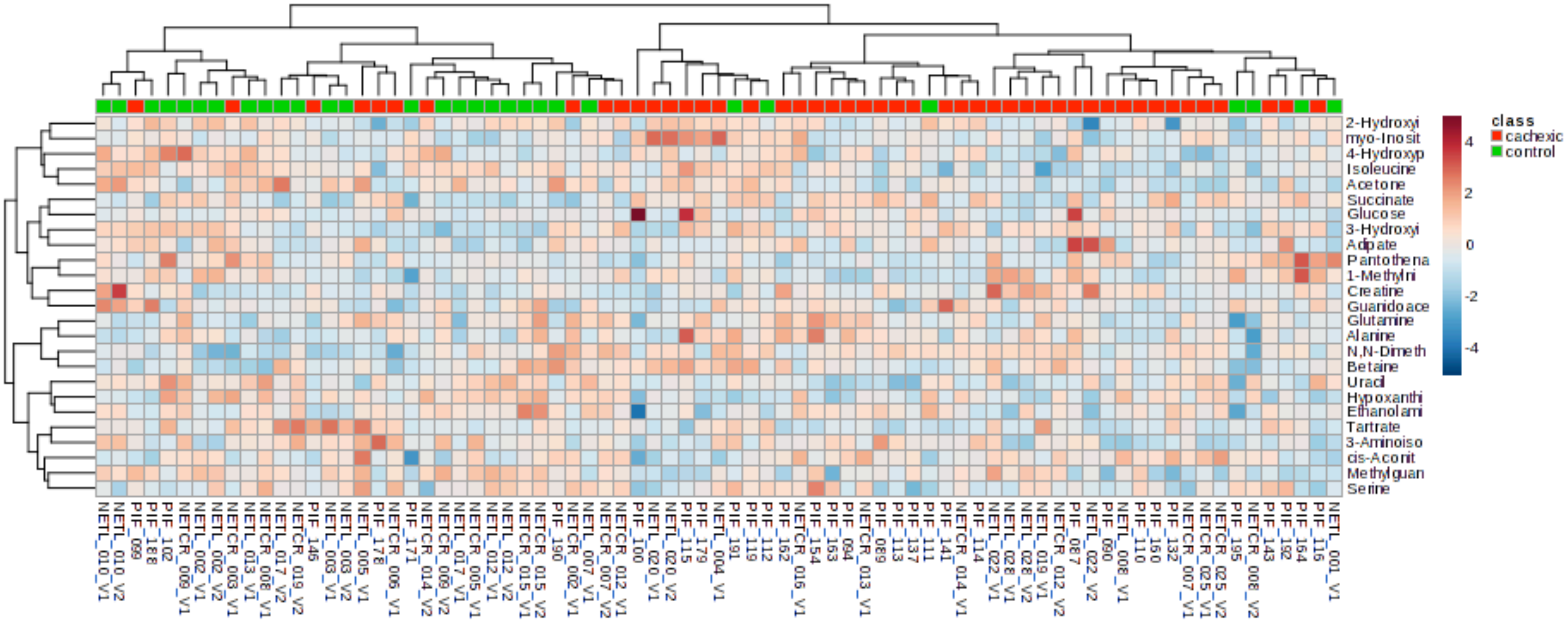
Input test dataset (Cancer patients Cachexic v. control)



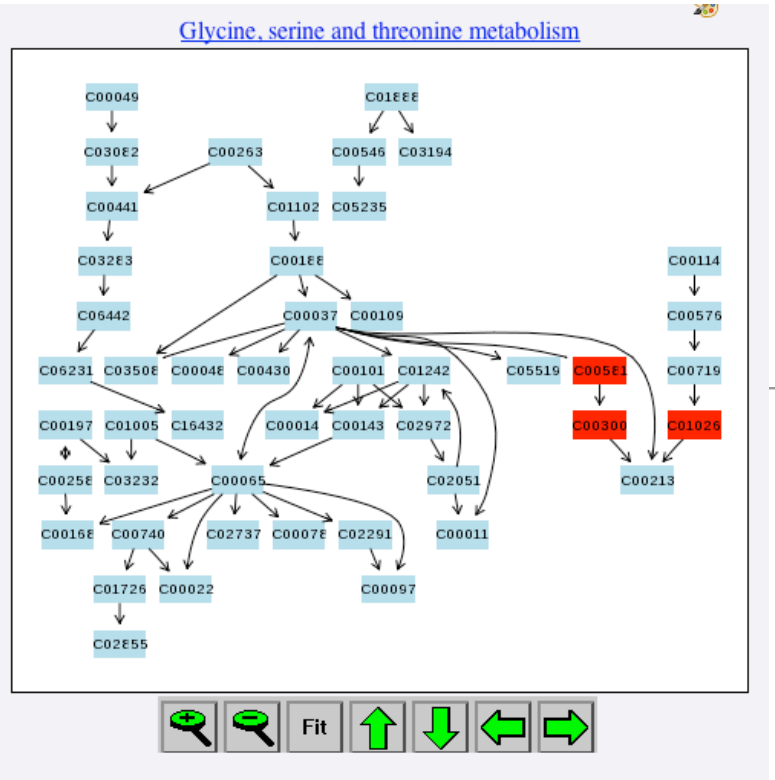


	p-value	FC	FDR
Uracil	3.84E-04	3.4154	0.024204
Isoleucine	0.0011396	2.9432	0.035898
Acetone	0.0051404	2.289	0.10795
Succinate	0.013088	1.8831	0.1502
4-Hydroxyphenylacetate	0.013611	1.8661	0.1502
Hypoxanthine	0.015669	1.805	0.1502
Methylguanidine	0.016881	1.7726	0.1502
Pantothenate	0.019073	1.7196	0.1502
Glucose	0.038618	1.4132	0.25269
Creatine	0.04011	1.3967	0.25269

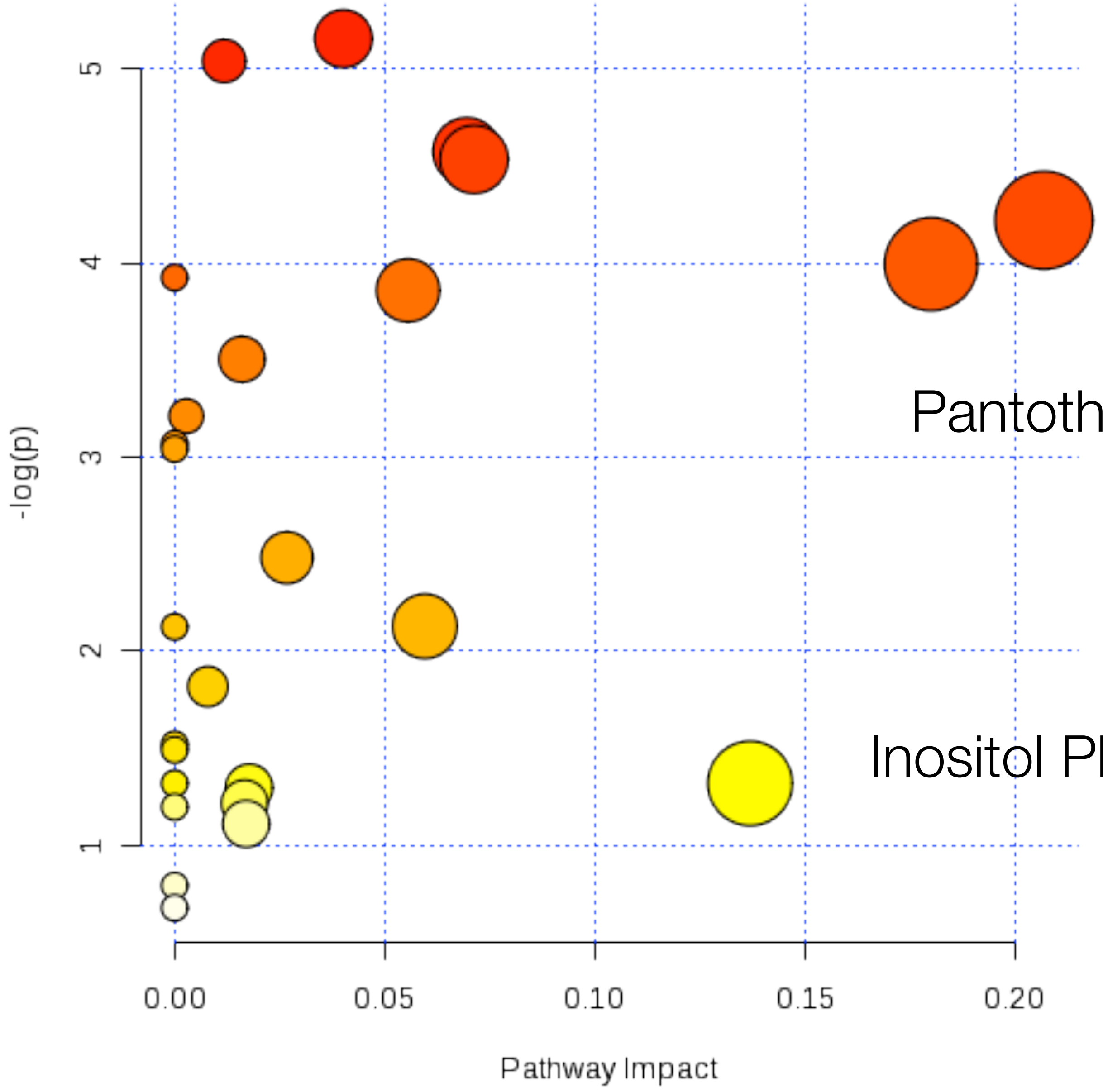
Sample Data-top25 features by Ttest



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)

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

Biology’s central dogma

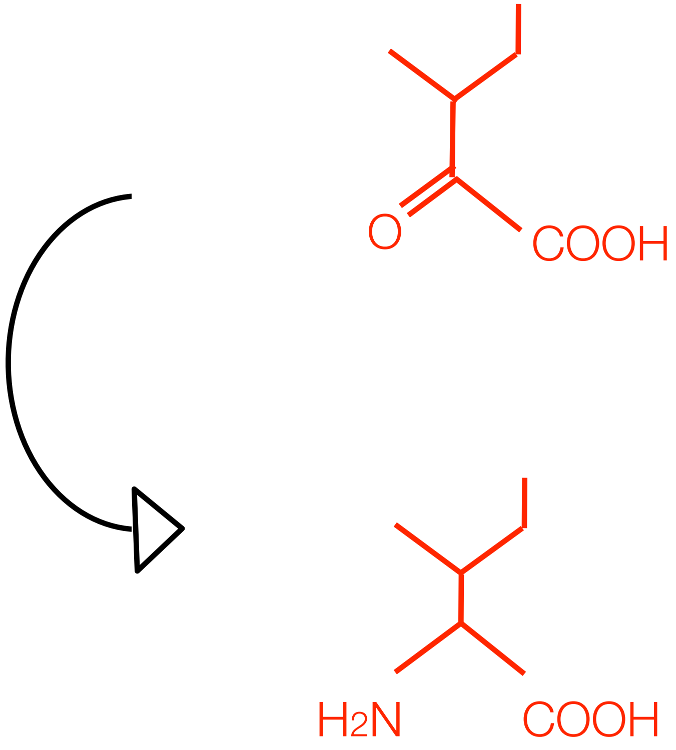
DNA
sequencing

→

RNA
sequencing

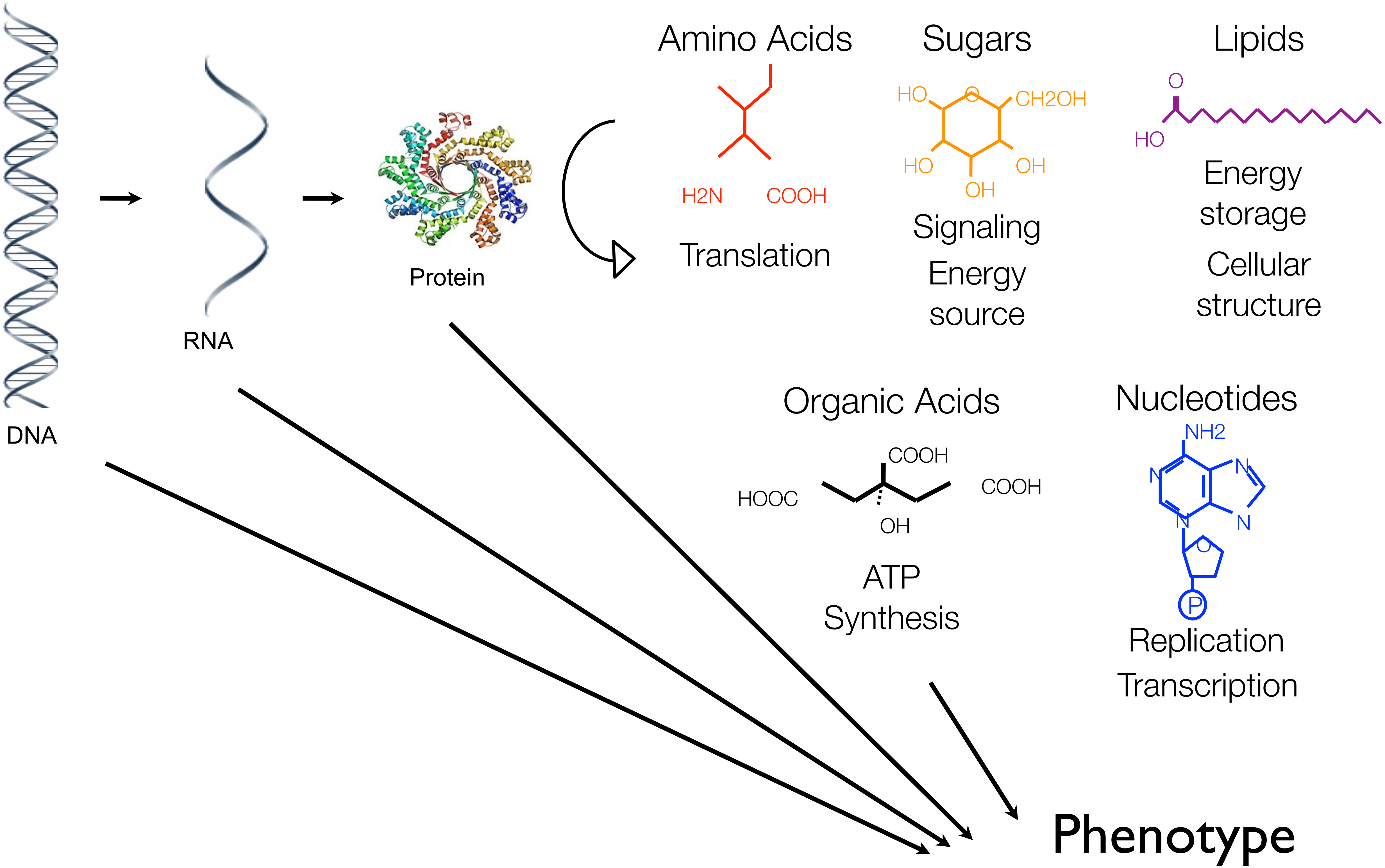
→

Proteins:
proteomics



Metabolomics
Mass Spectrometry
NMR

Small molecules as sensors



DNA

RNA

Protein

Amino Acids

Translation

Sugars

Signaling

Energy source

Lipids

Energy storage

Cellular structure

Organic Acids

ATP Synthesis

Nucleotides

Replication

Transcription

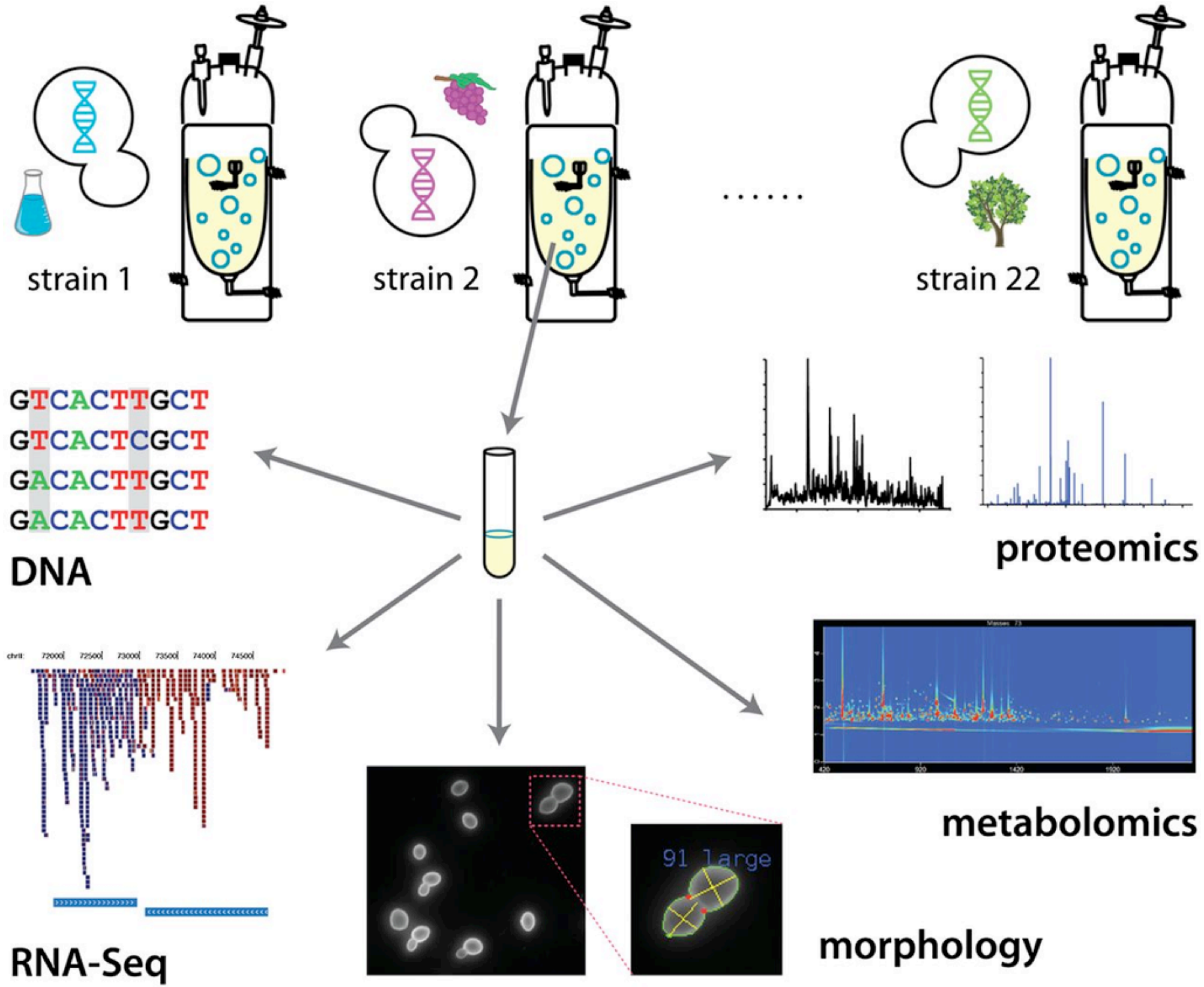
Phenotype

Part II: Using Metabolomics in biological research

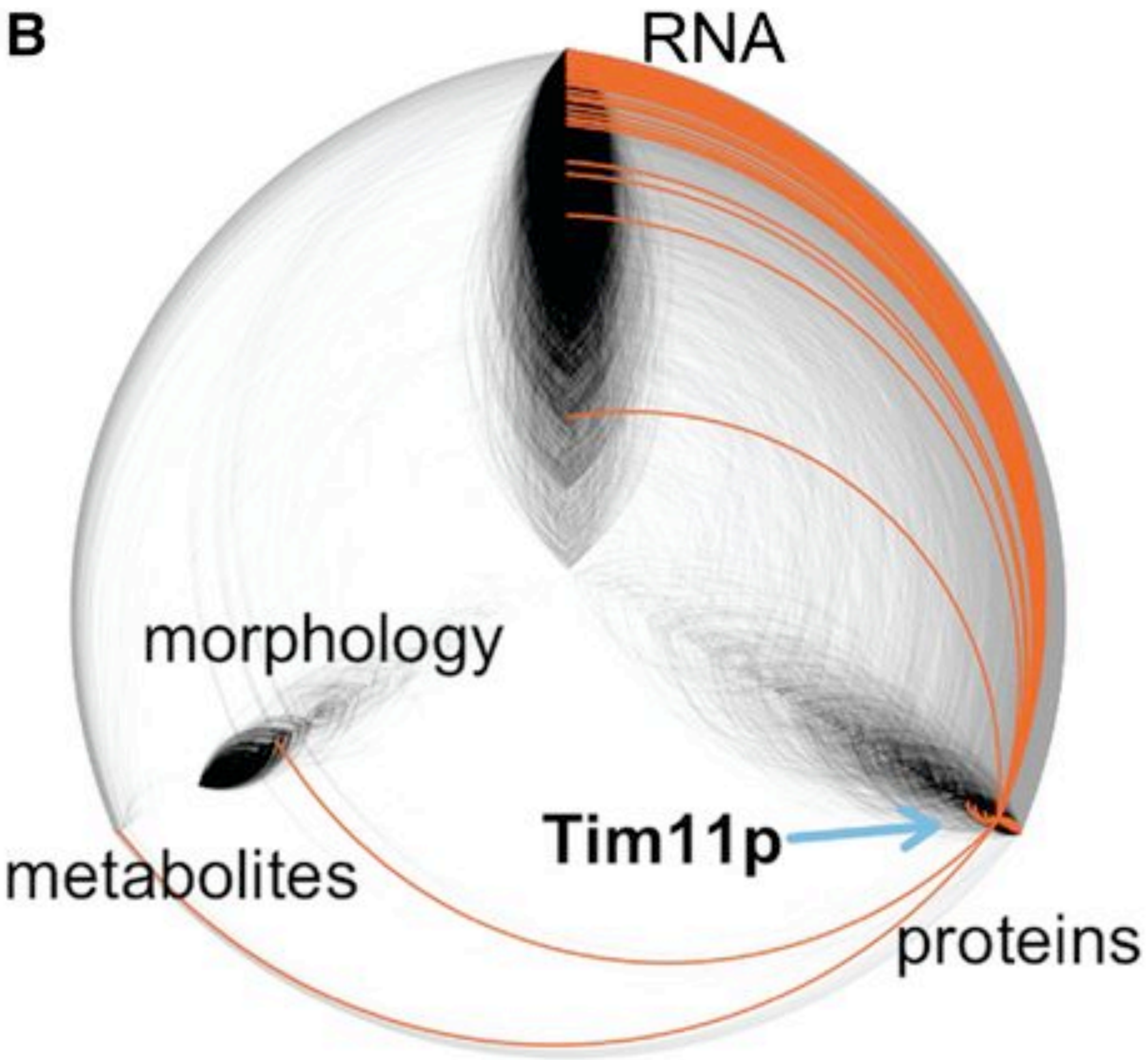
- Yeast Phenomics

- Pancreatic Cancer

Yeast phenomics

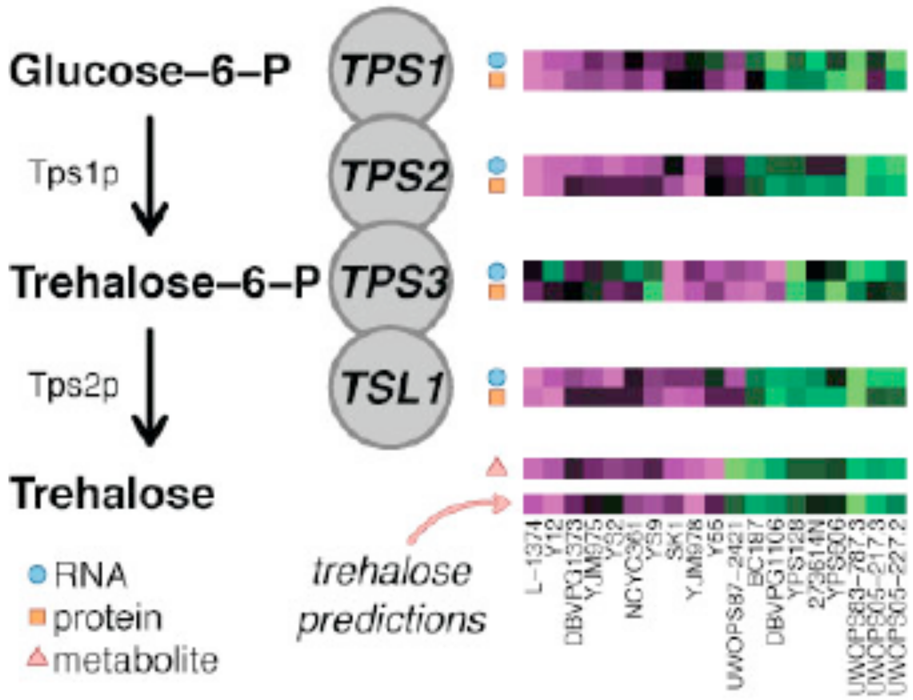


Integrating data

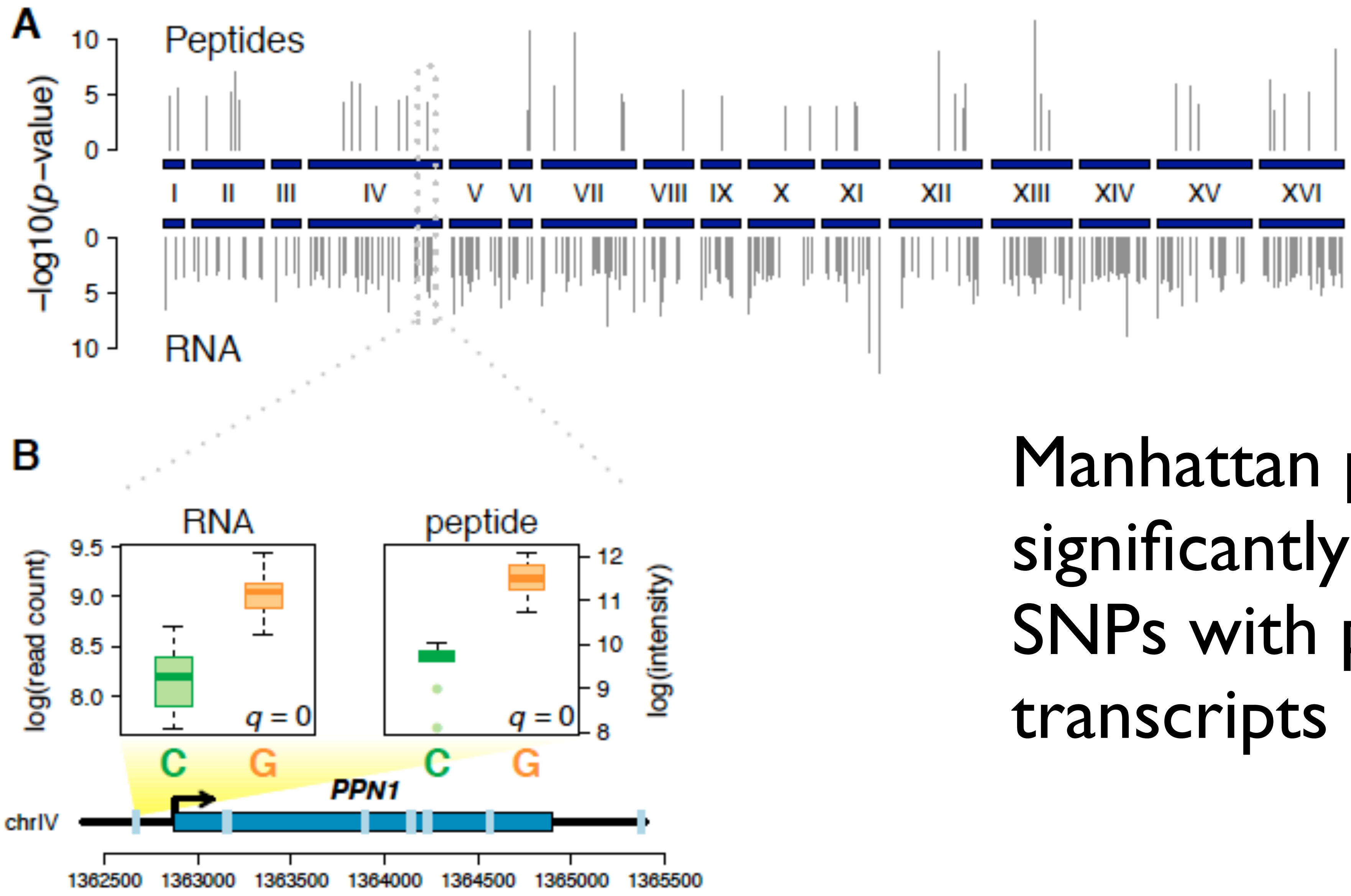


7078 phenotypes correlated to at least one other phenotype

Predicting phenotypes

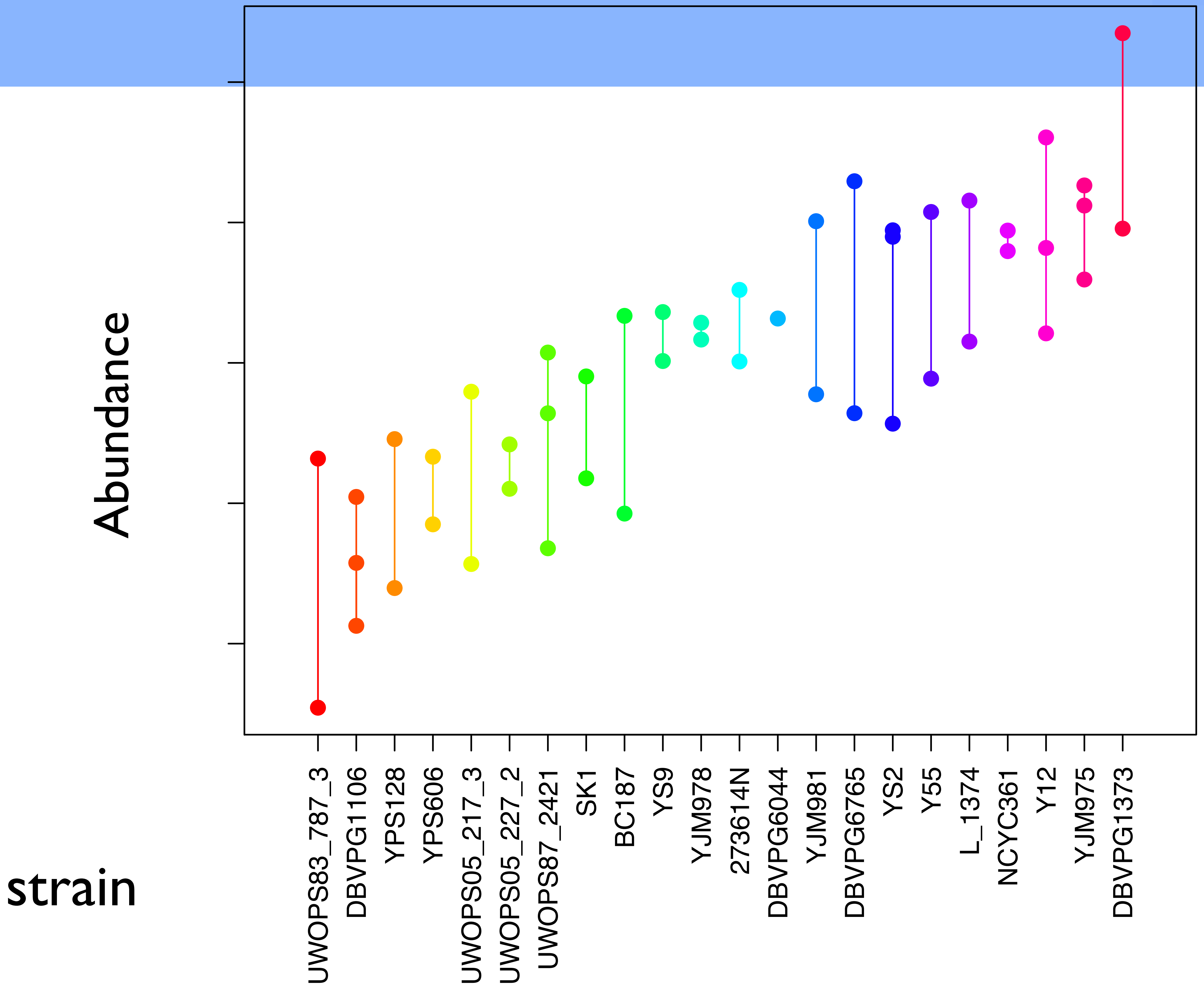


Genetic associations



Manhattan plot of significantly associated SNPs with peptides and transcripts

A metabolite with heritable variation: Ribose

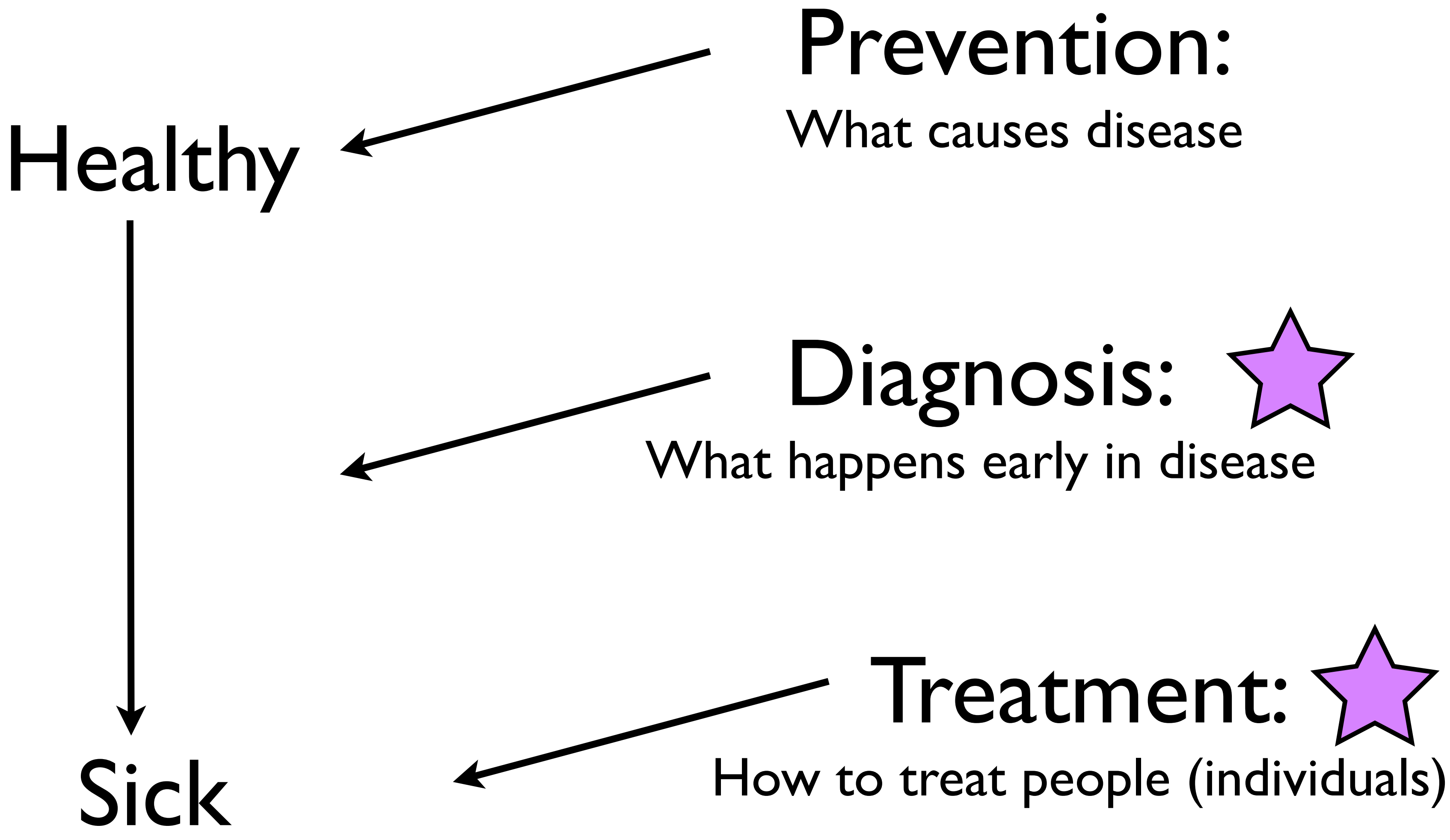


Summary

- 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

Sick

Prevention:

What causes disease

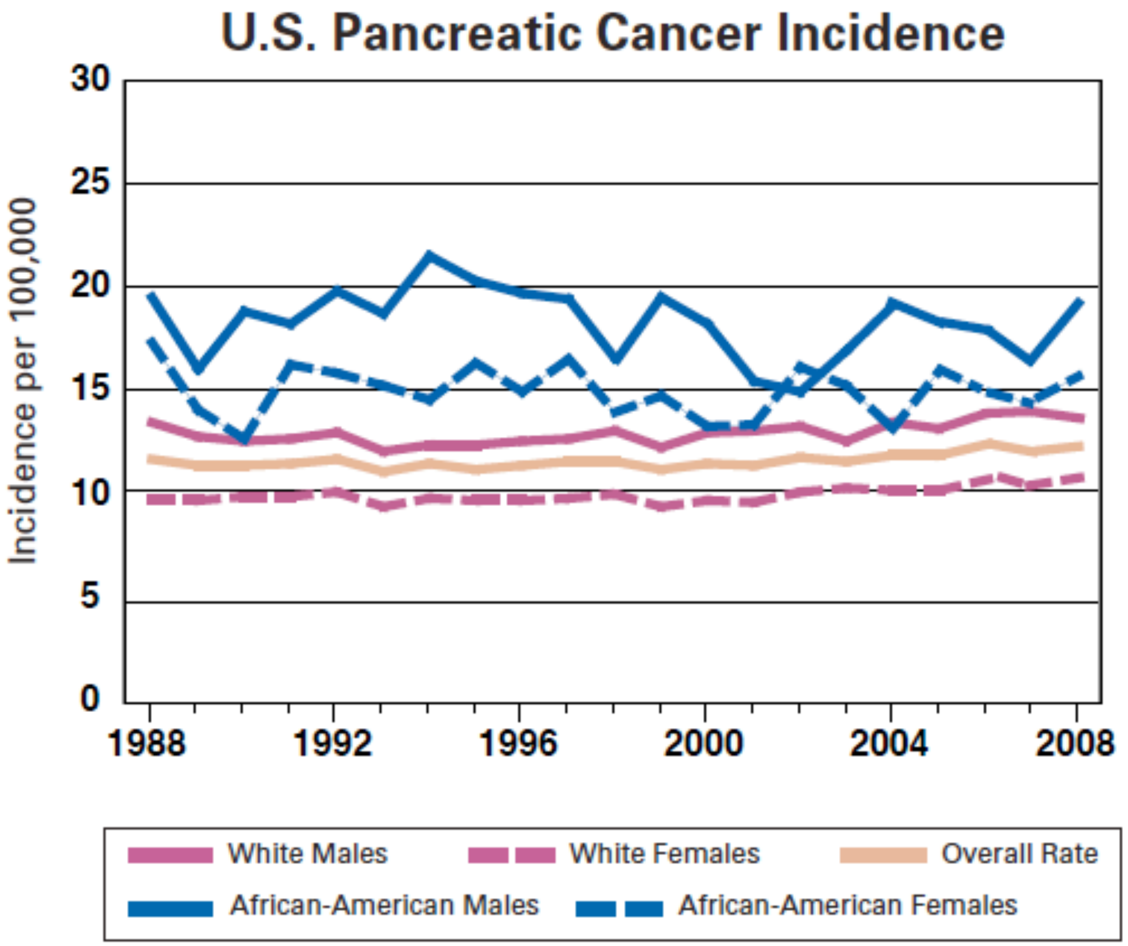
Diagnosis: ★

What happens early in disease

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

Sally Ride

Steve Jobs

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

Use genomic technologies to improve Pancreatic Cancer patient outcomes

Solutions:

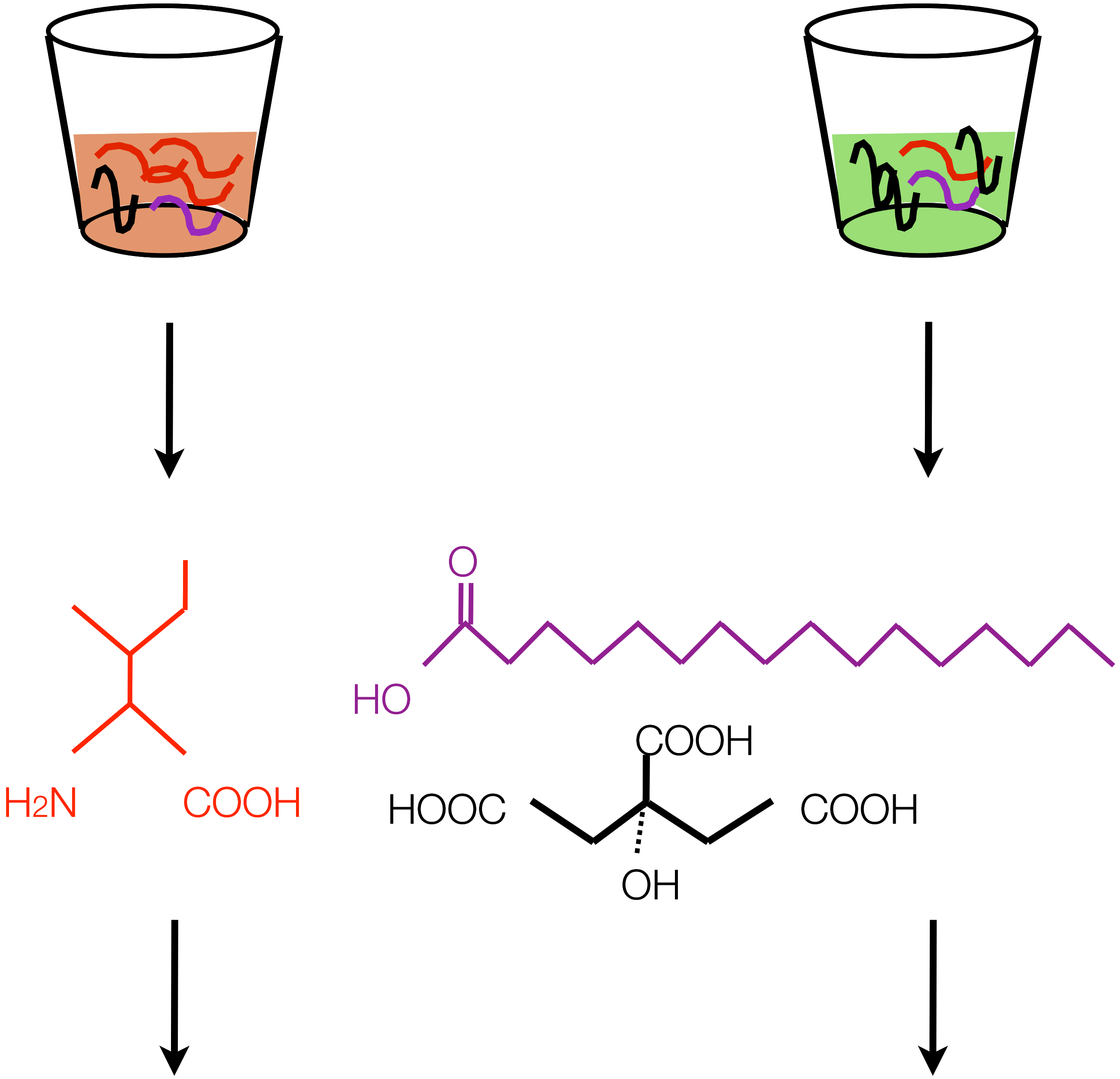
1) Better diagnostic markers

2) Improved and/or personalized treatment options

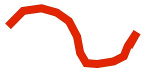
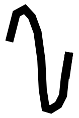
A role for metabolism in pancreatic cancer

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

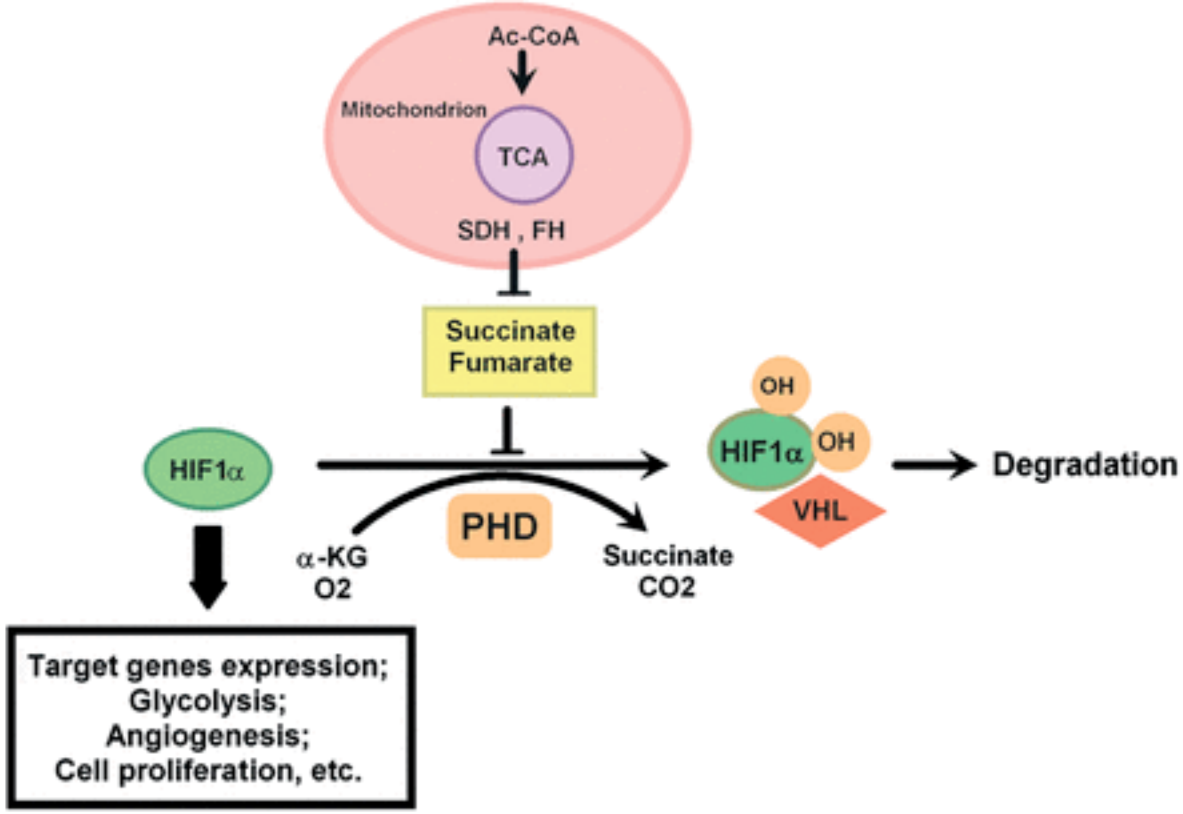
	Sample 1	Sample 2
	3	1
	1	1
	1	3

Analyses

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

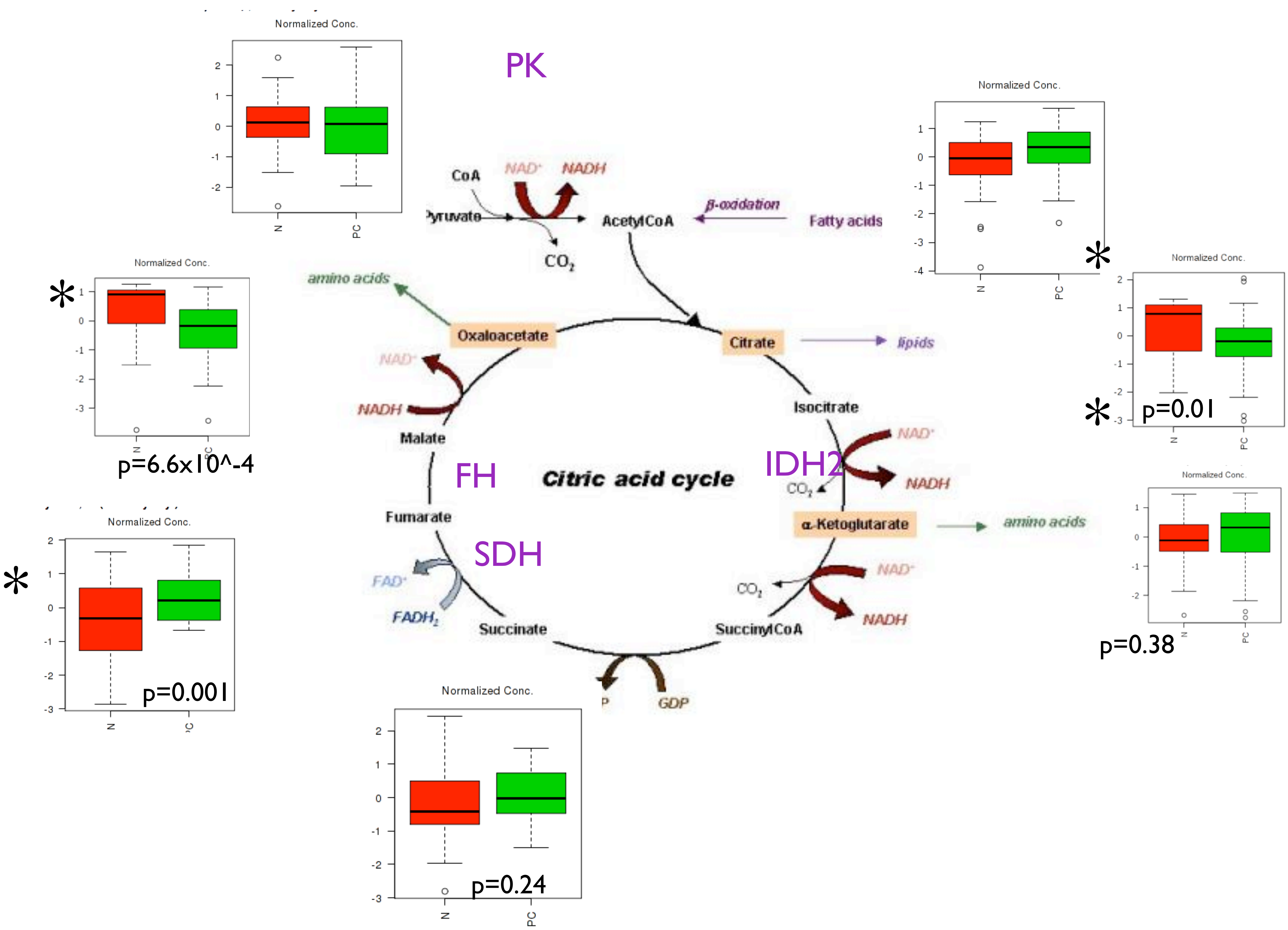
TCA cycle

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



Review: Wu and Zhao 2012

Urine-TCA cycle



p=6.6x10^-4

p=0.001

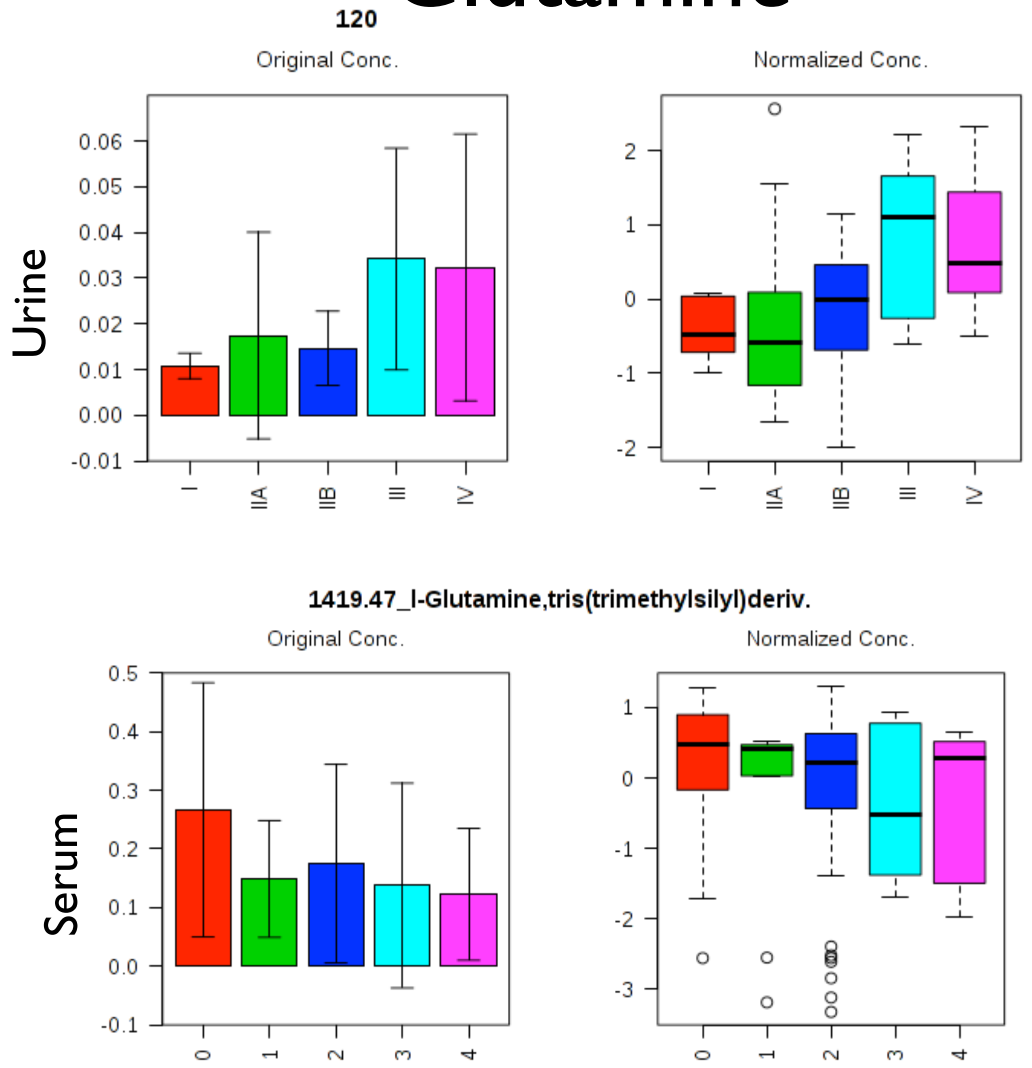
p=0.24

p=0.01

p=0.38

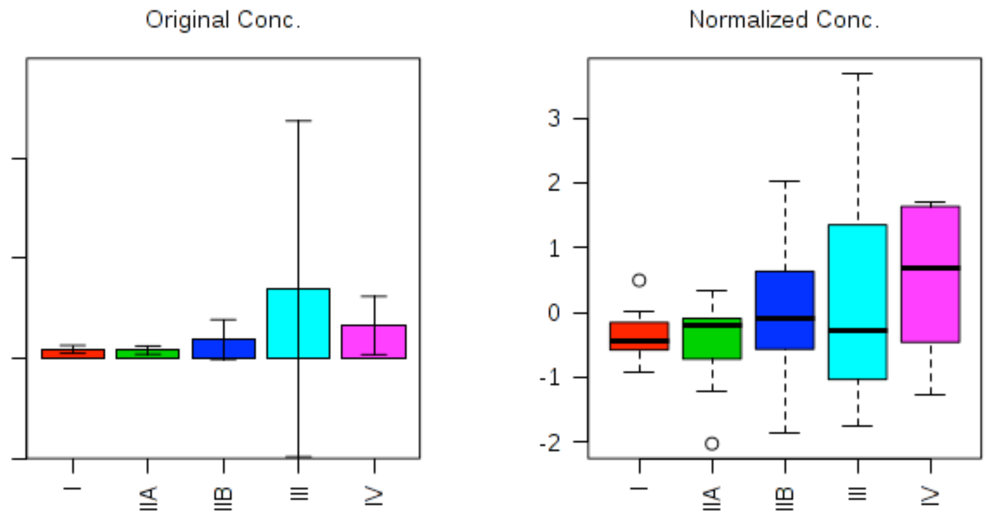
Most significant effects

Glutamine



Pancreatic Cancer cells are characterized by their "glutamine addiction"

Glycine



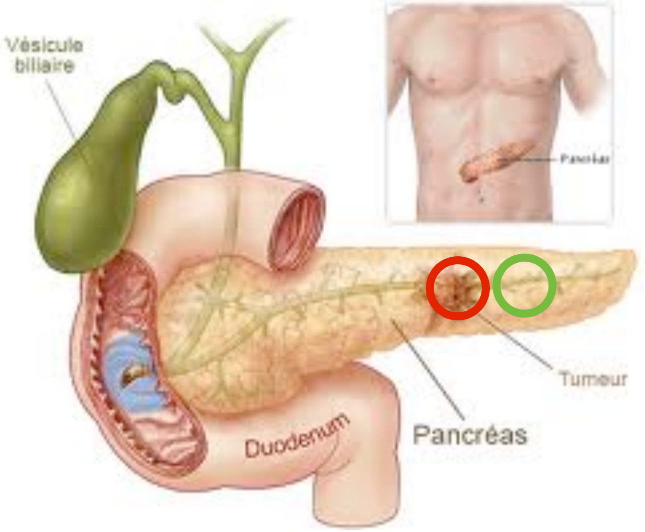
Text

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

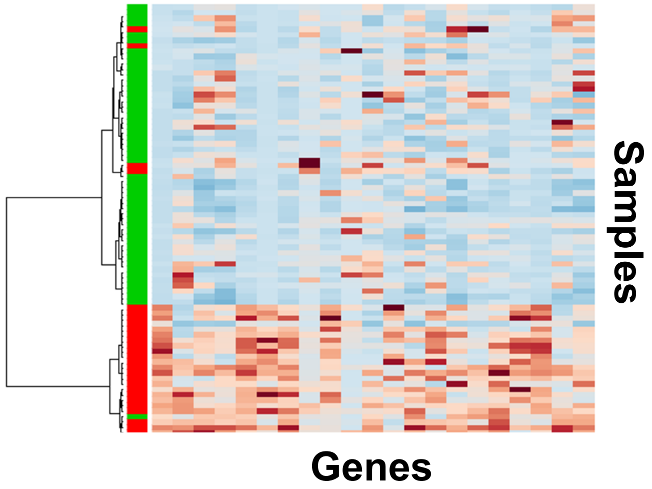
Multi- “Omics” approach

- 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



Tumor tissue
Normal tissue

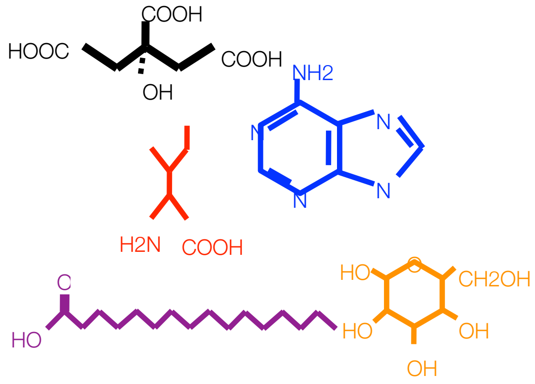


Gene expression changes in pancreatic cancer

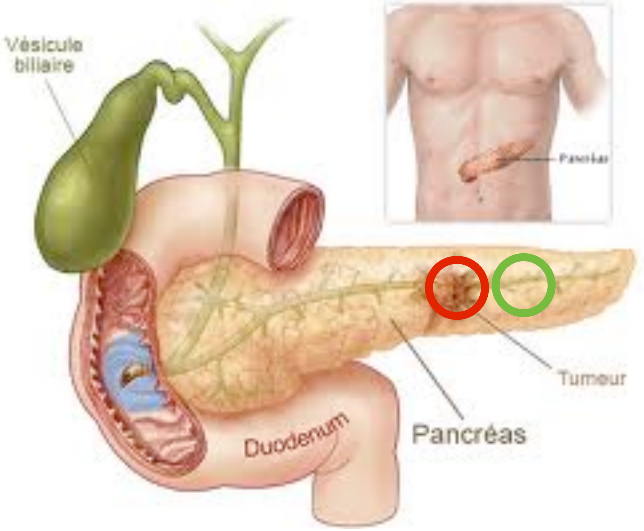
- Is there evidence of altered metabolic pathways in gene expression data?
- Are the same pathways we identified in blood and urine changing in tumor samples?
- What do we learn by intersecting these data?

Pancreatic Cancer-
Integrating Metabolomics and Genomics

Serum from pancreatic cancer patients

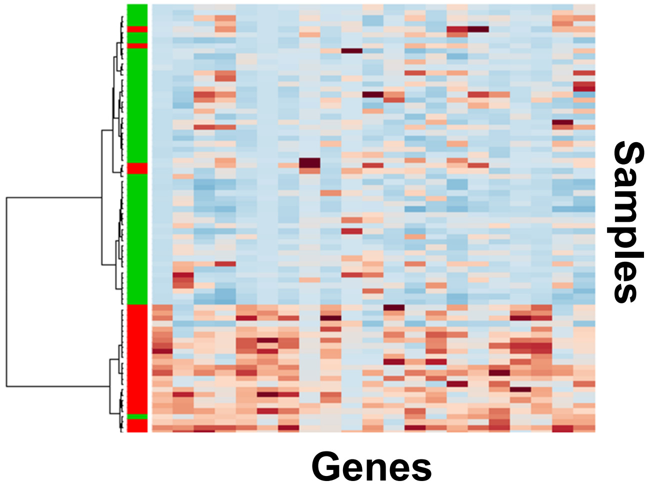


Metabolic changes in pancreatic cancer



Tumor tissue

Normal tissue

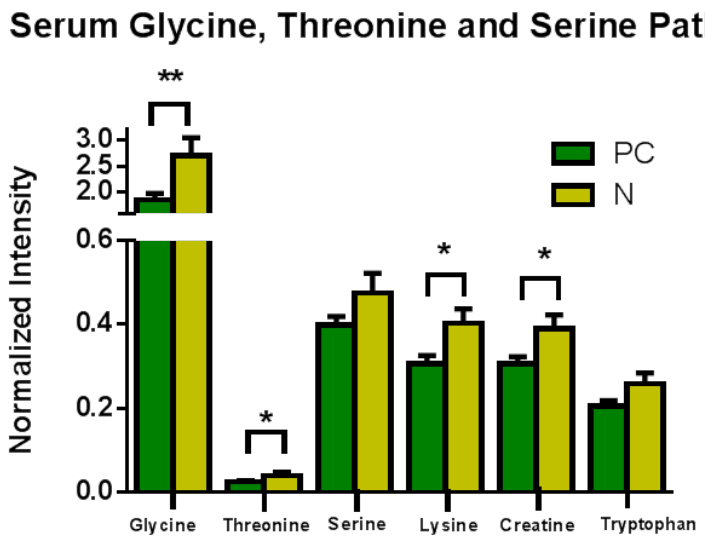


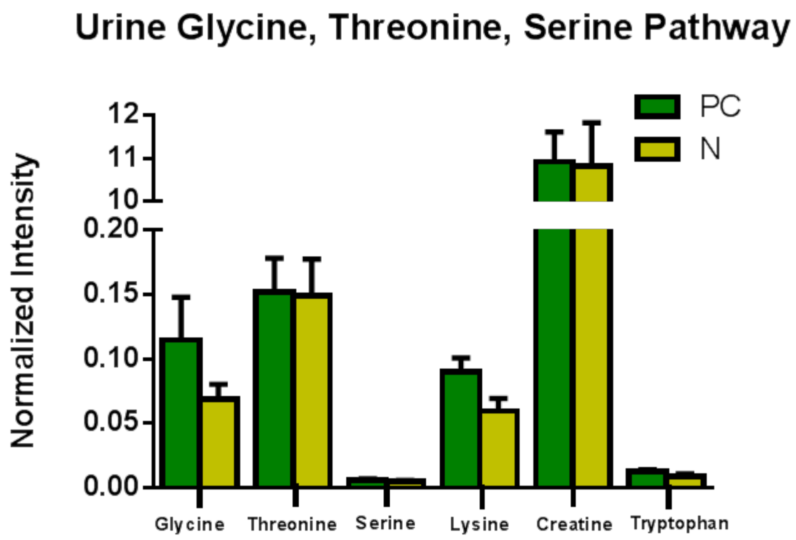
Gene expression changes in pancreatic cancer

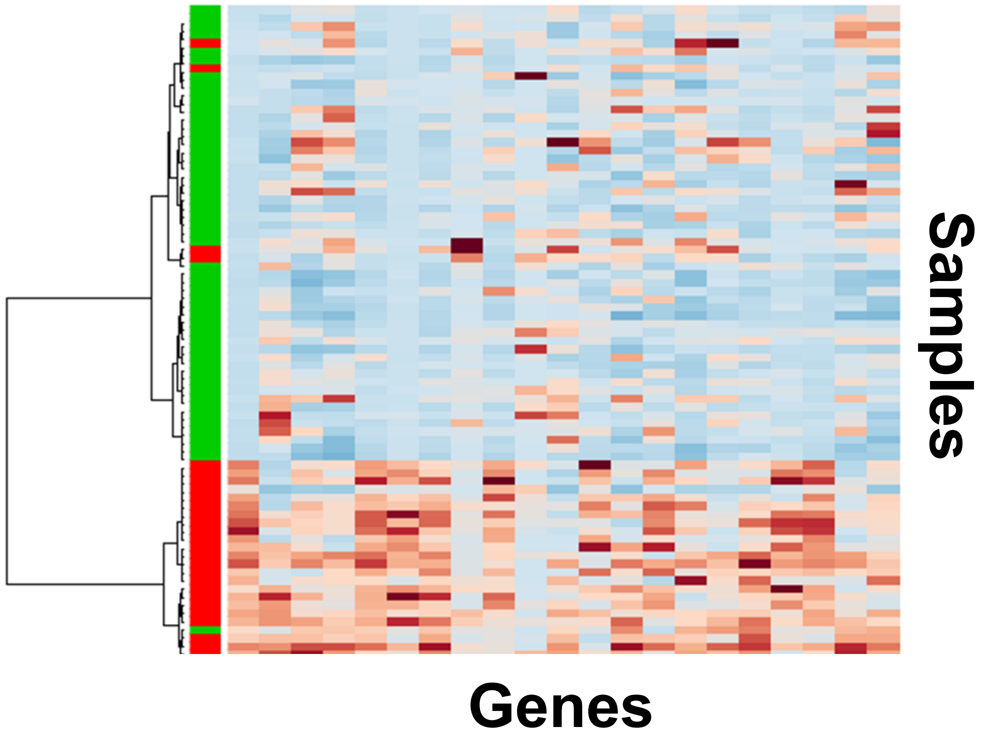
Common Pathways

Identification of pathways important to tumor growth and patient survival

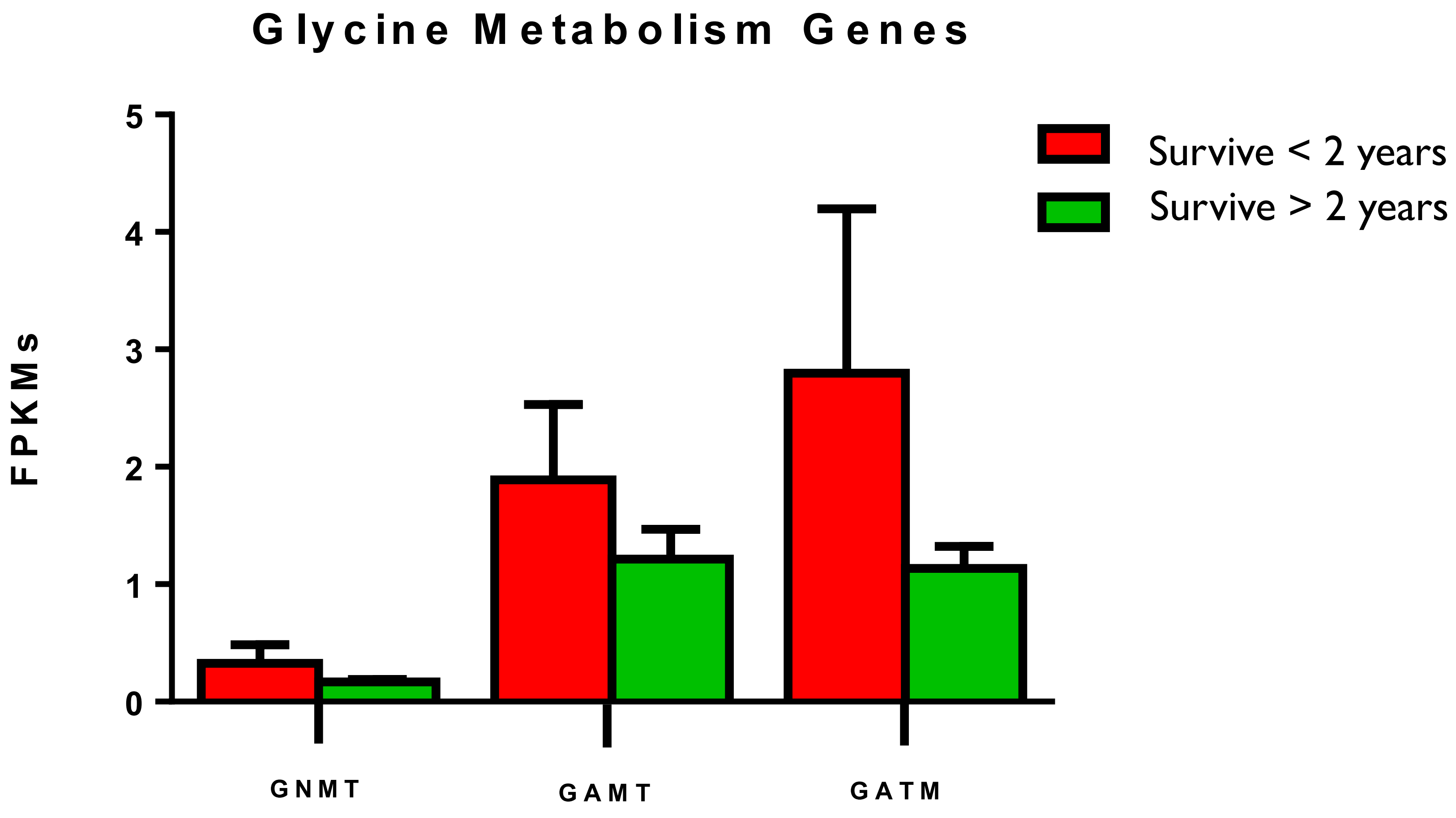
Glycine, Threonine, Serine Synthesis

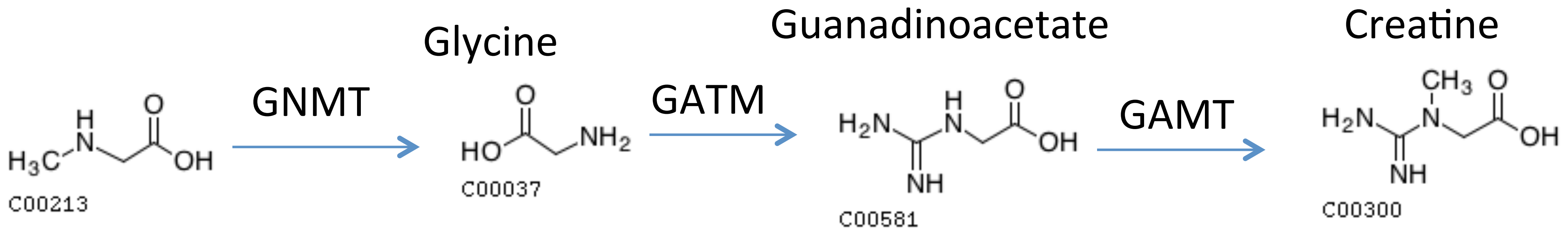




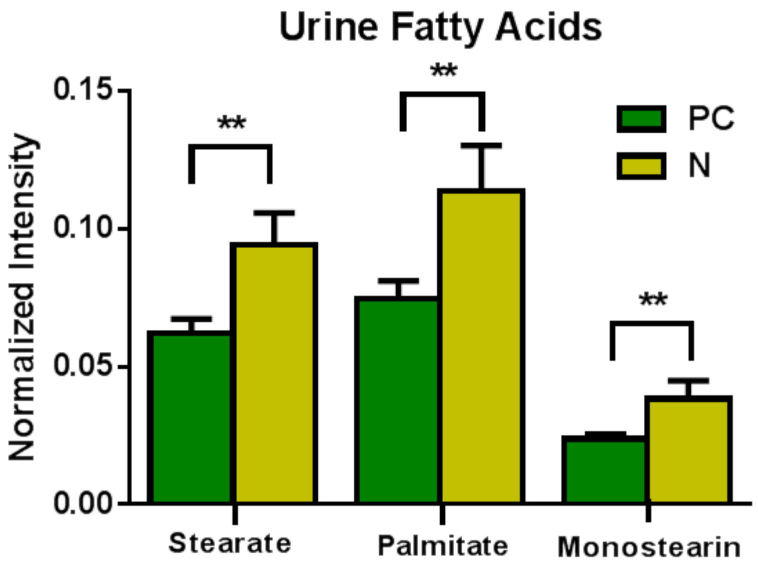


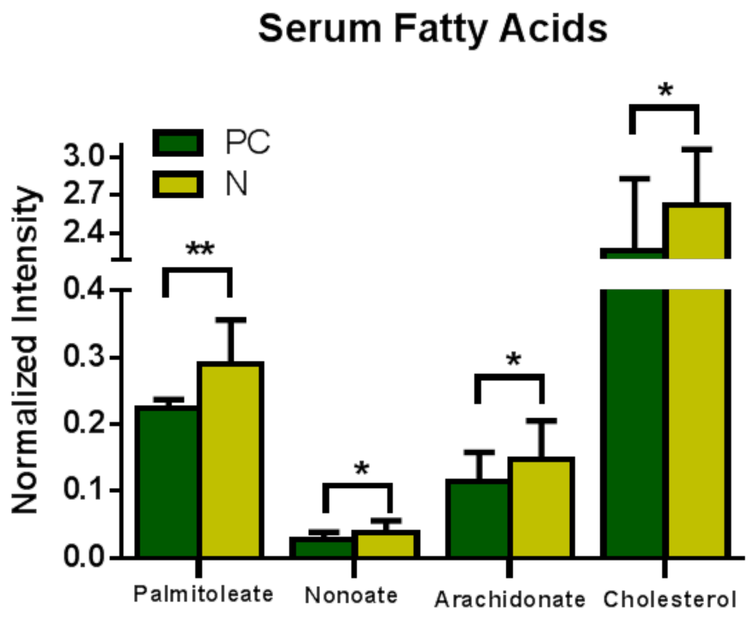
Glycine pathway gene expression associated with poor prognosis

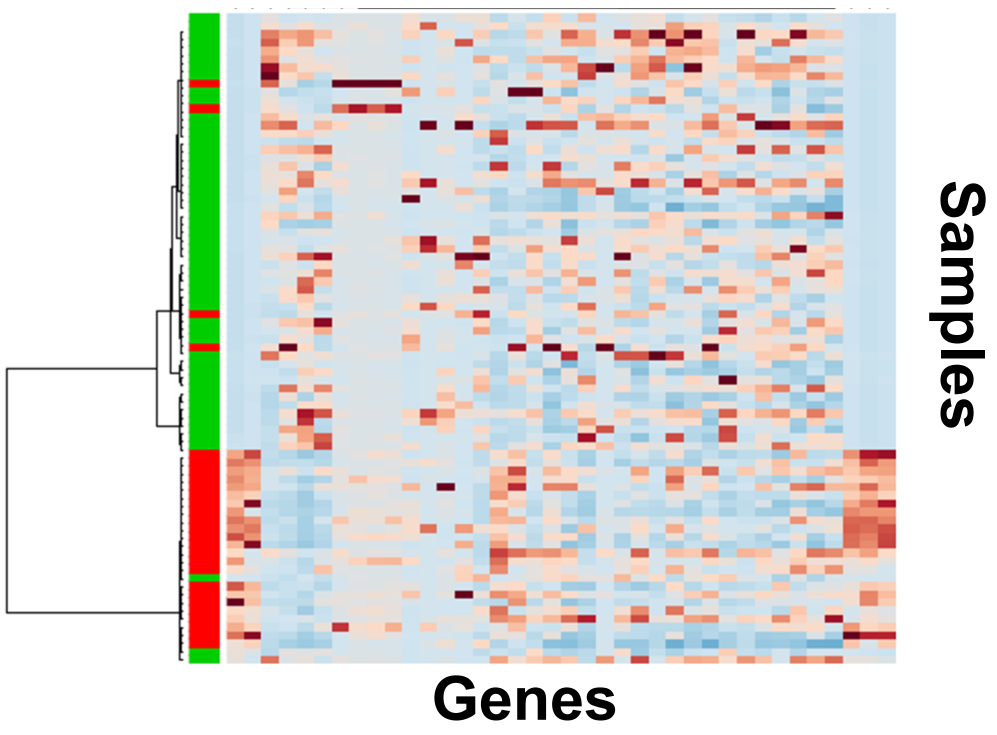




Fatty Acid Biosynthesis

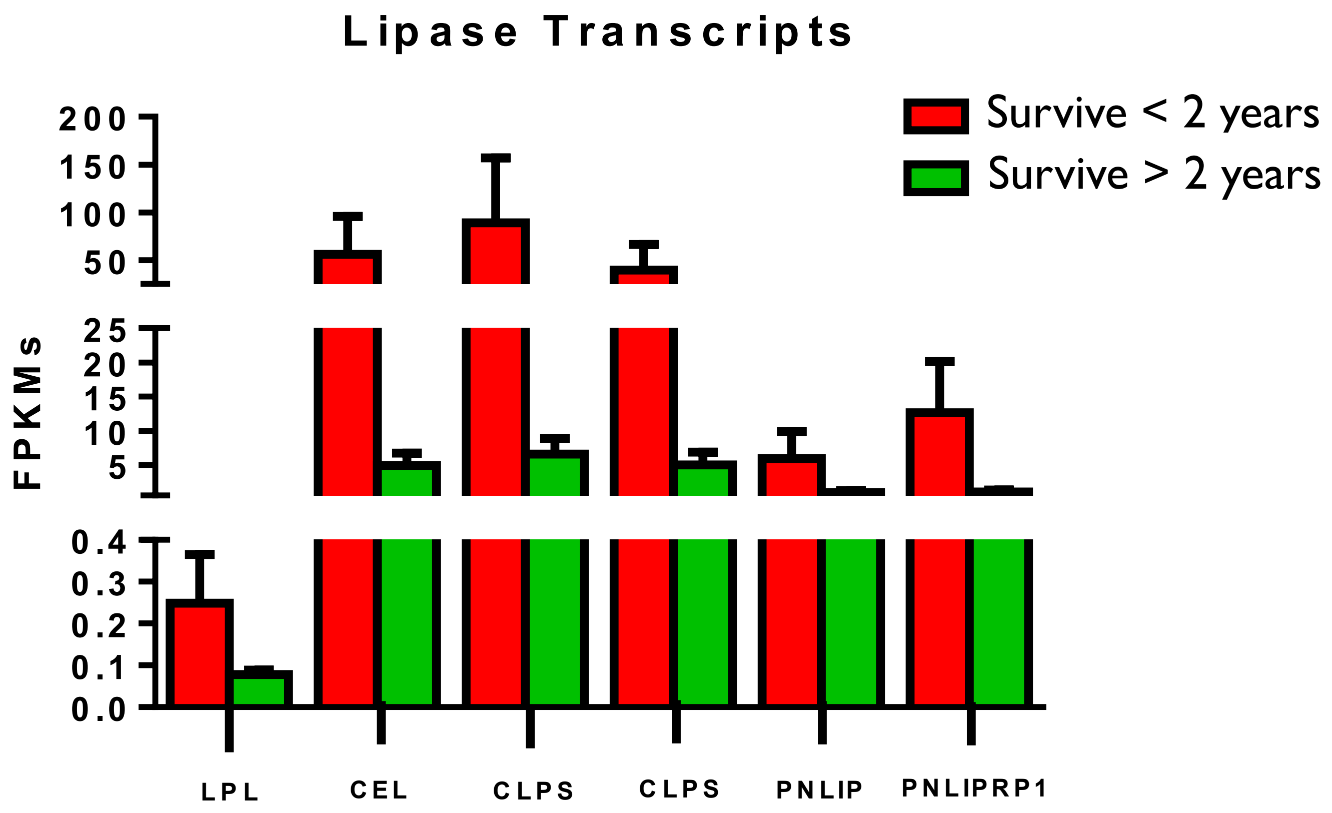




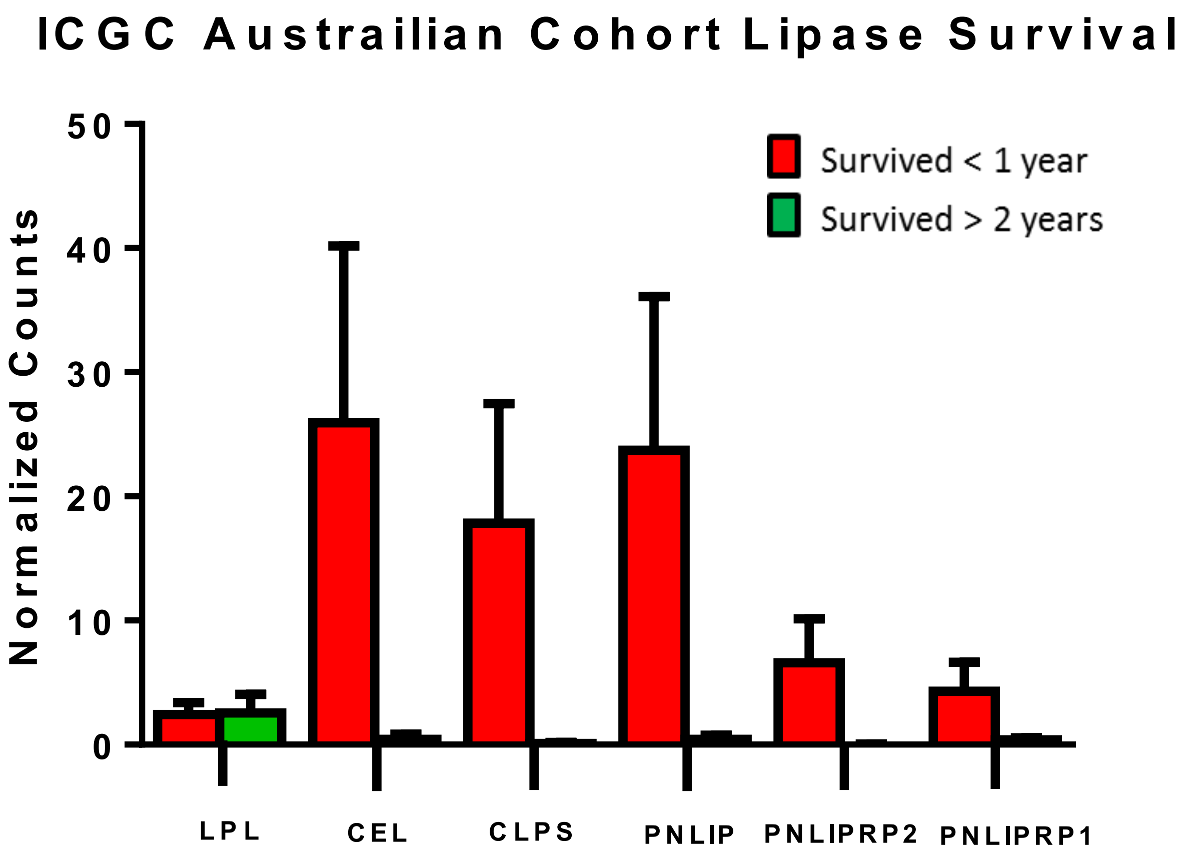


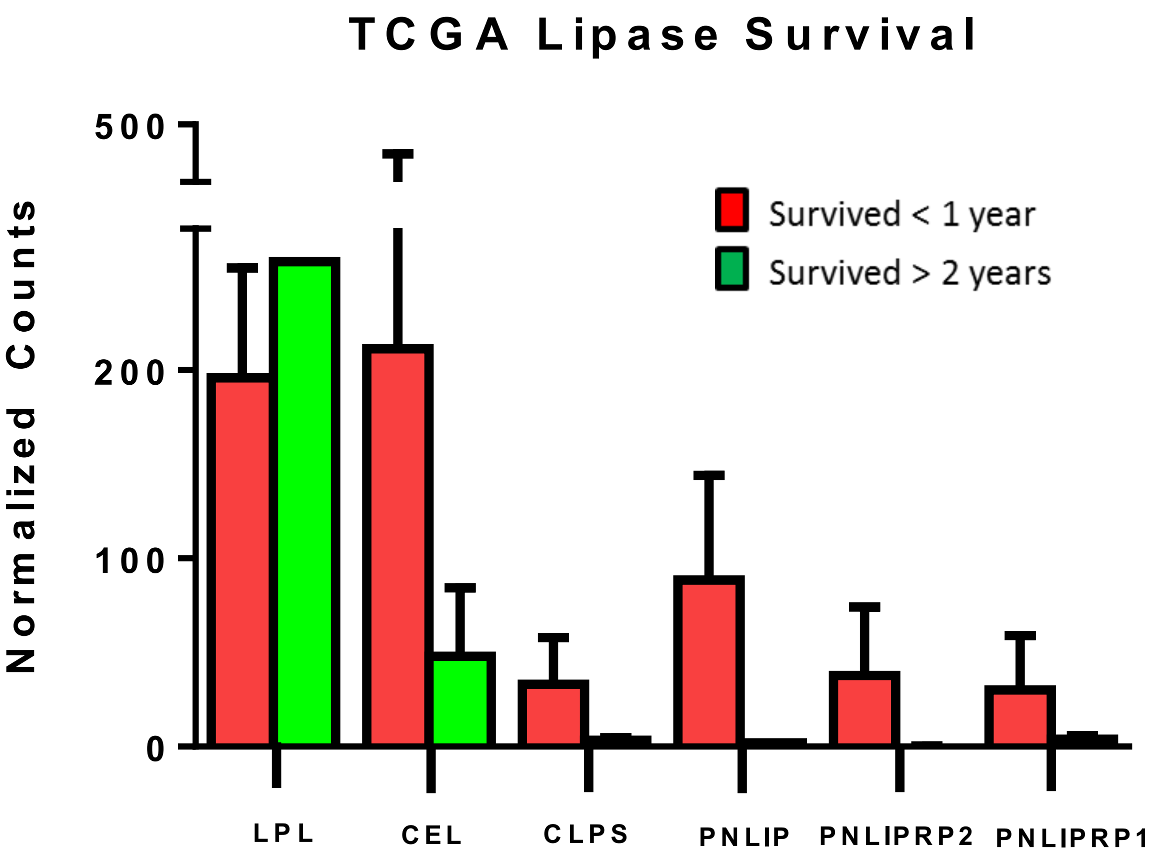
A correlation between survival and lipase expression

Early vs. Late Survivors

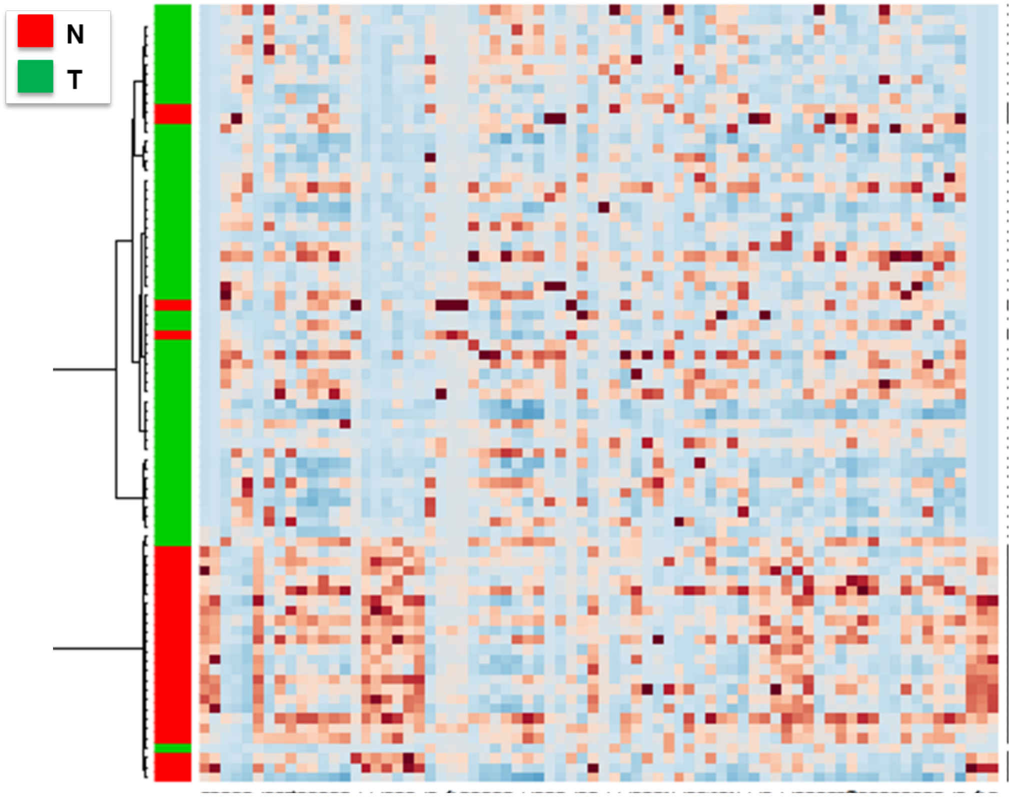


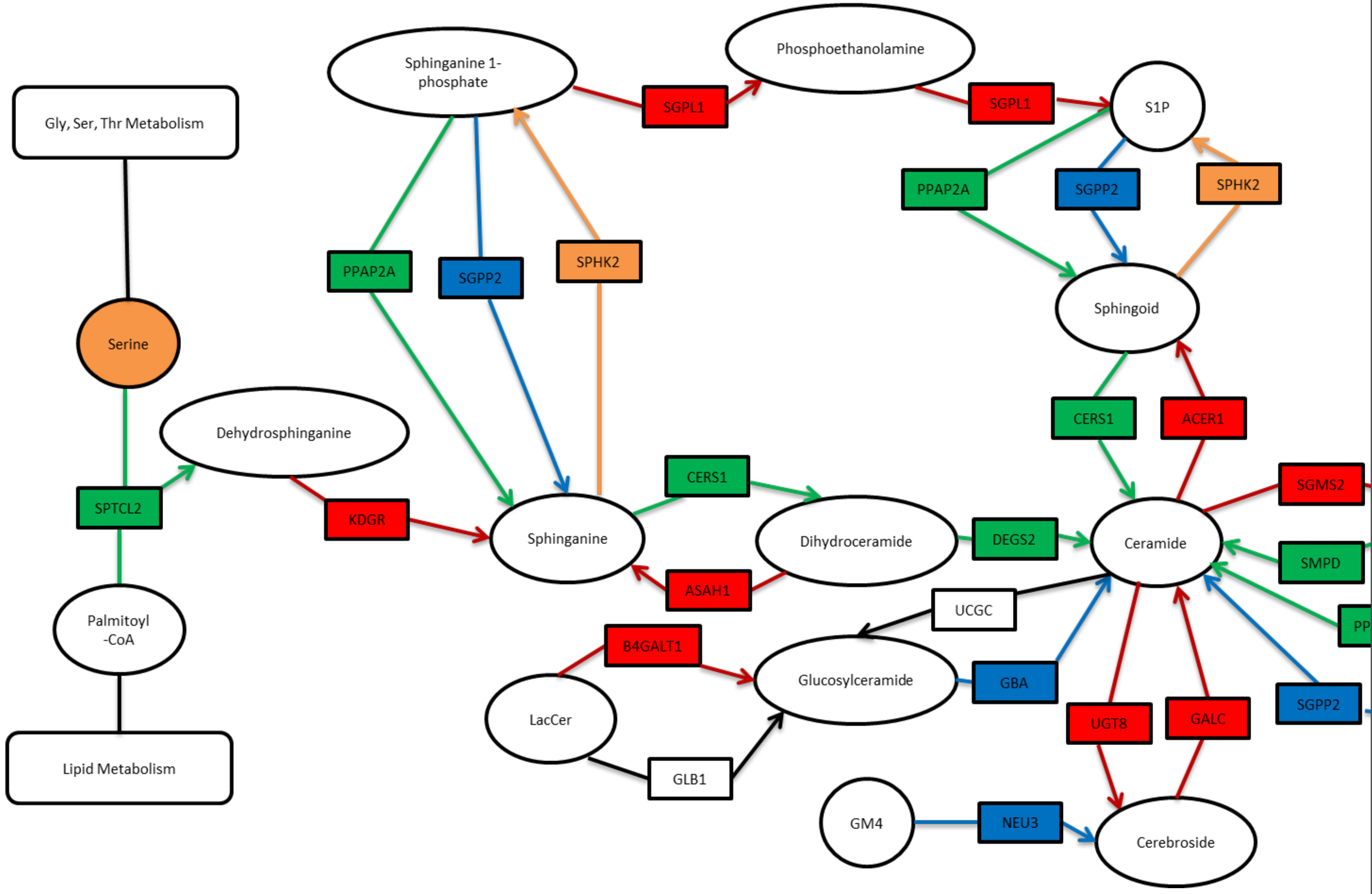
Replication in independent samples





A pathway to link these pathways: Sphingolipid Biosynthesis





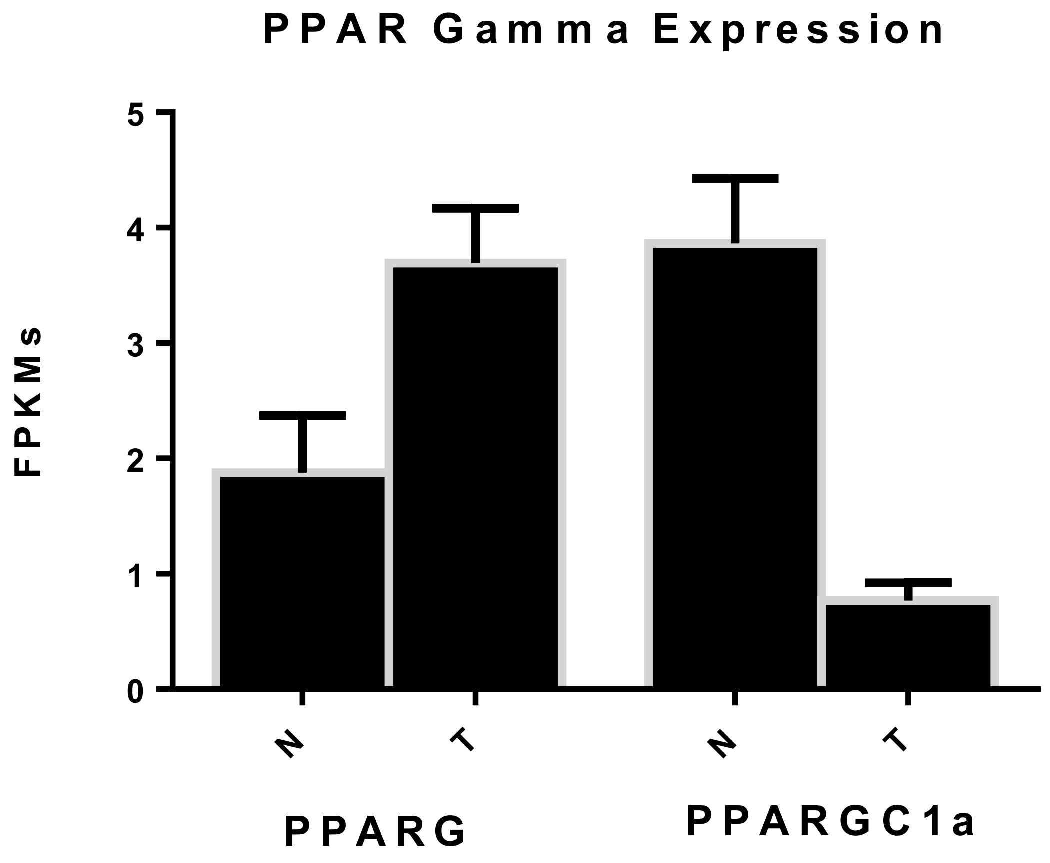
Is the link important to disease?

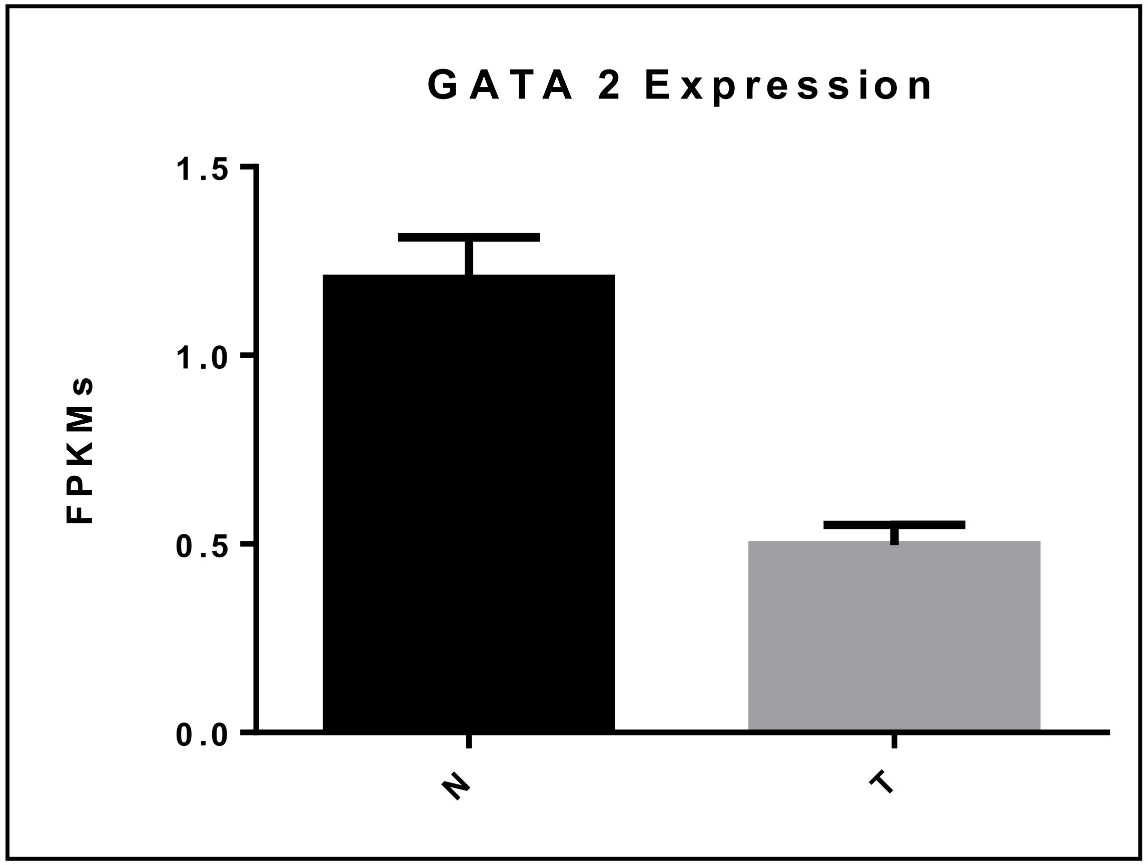
- If ceramide/sphingolipid biosynthesis is essential for apoptosis in cancer cells, they should reduce ceramide production, perhaps through downregulation of lipase genes.
- We can test in vitro whether apoptosis in cancer cell lines is sensitive to fatty acid concentration, and whether apoptosis requires 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





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