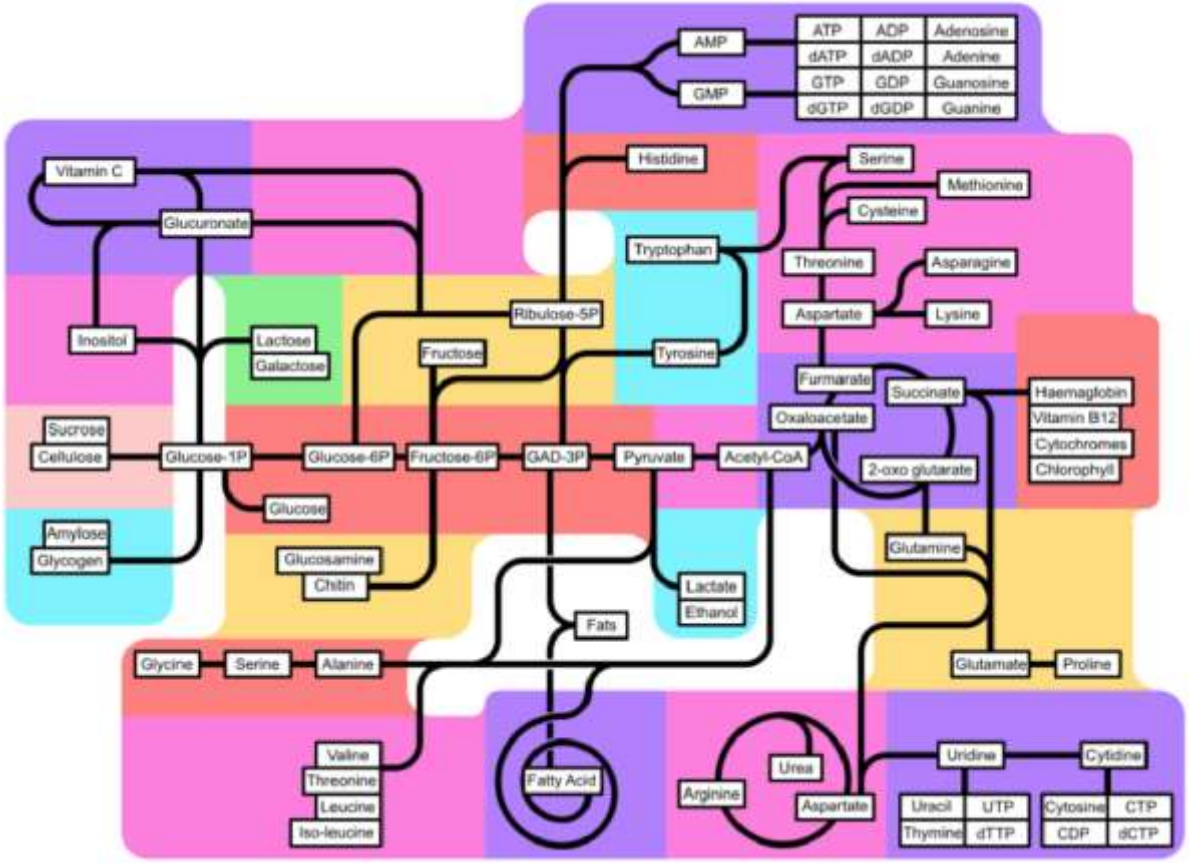
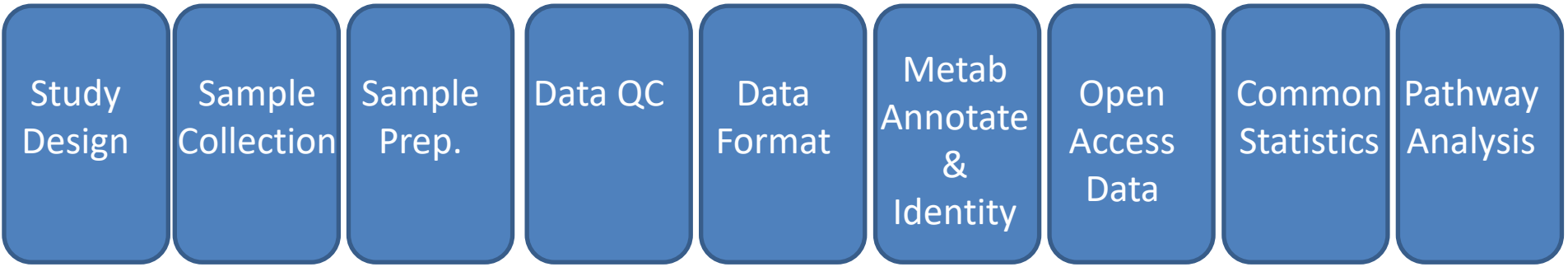
So you want to do ‘omics?



Marie Phelan

(Manager of ~~HighField-NMR~~ Expectations at University of Liverpool)

The omics Pipeline:



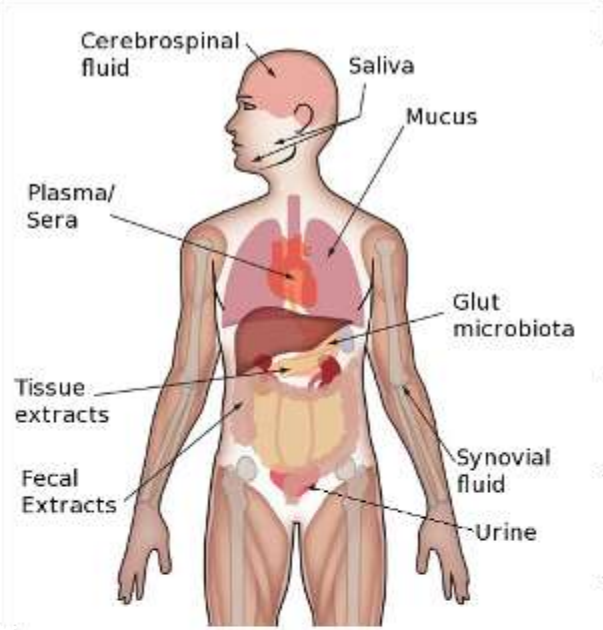
Study Design

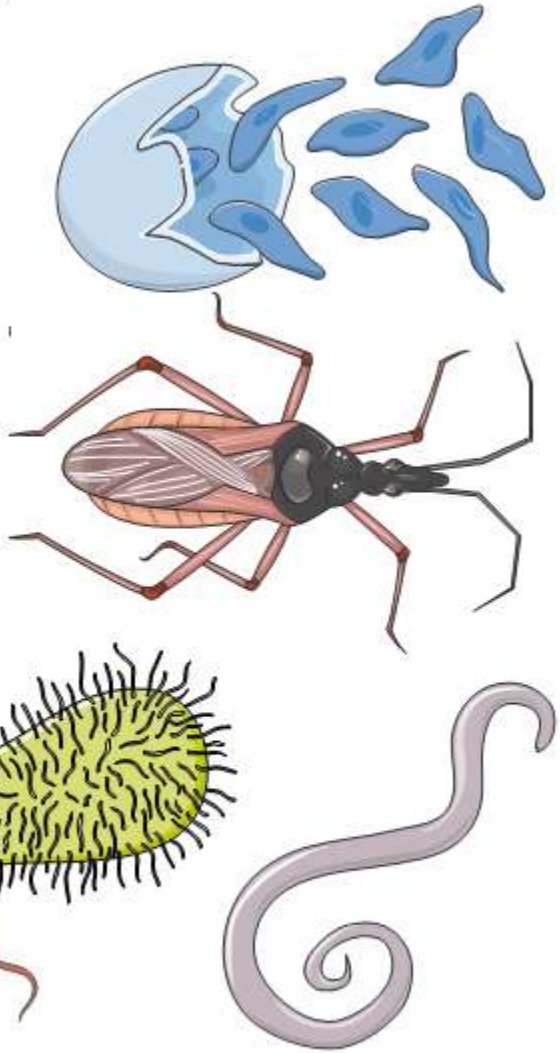
Clinical Samples
-Biofluids
-Tissue (Biopsy)

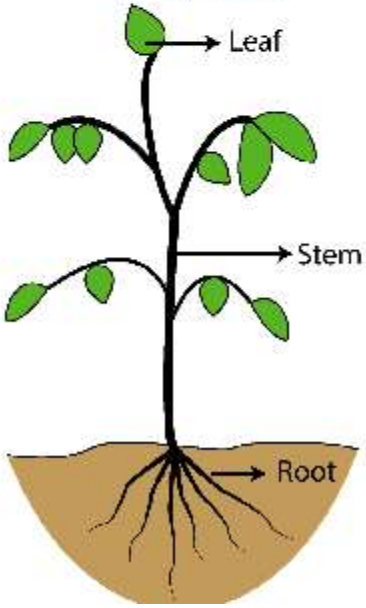
Clinical Models
-post-mortem material
-Primary cells & cell lines

Microbes
-Bacteria/fungi
-Parasites
-Whole model organisms

Technique Development
-Software development (CBF)
-lipids

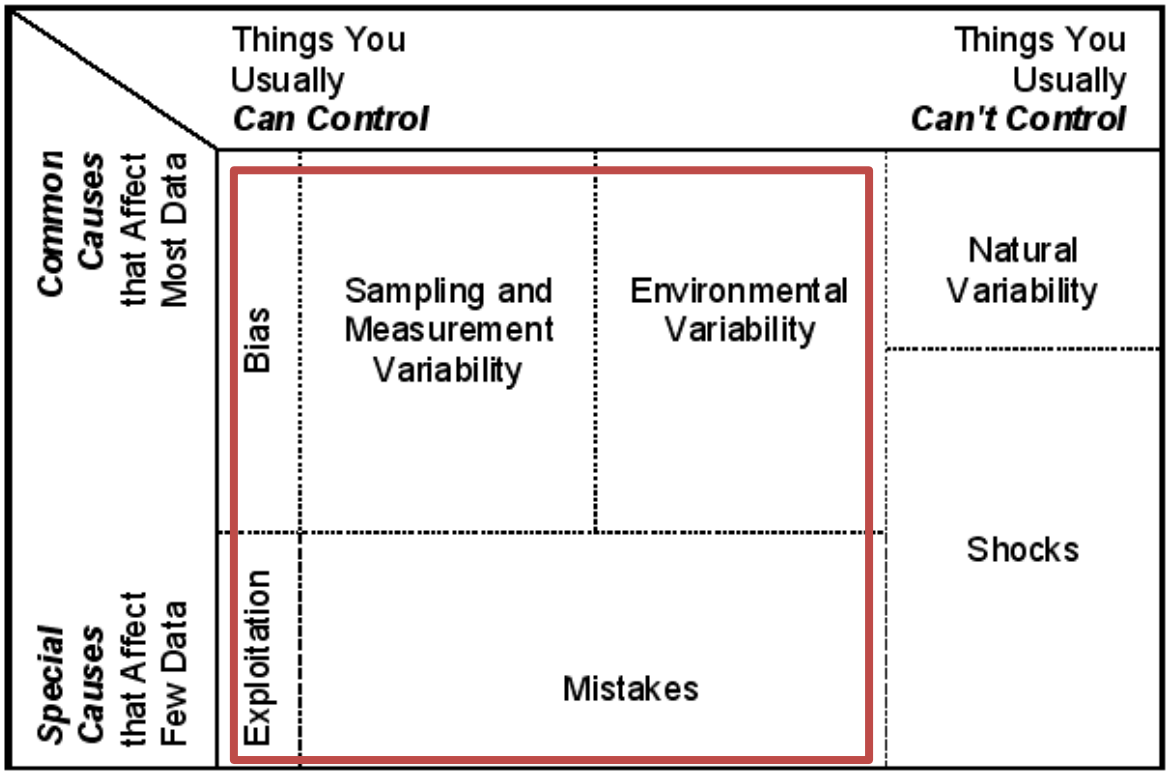






Data sources of variation – Experimental Design

Data = biological meaning + error



- Best way of controlling error is through a **robust experimental design** (e.g. minimise cohort variability, have controls in place, randomise your sampling, etc.).
- Remember the 3 Rs: **Reference**, **Replicate**, **Randomise**
- **Normalisation** and **scaling** steps can minimize undesirable variance from dilution effects etc. but are not a miracle

Get to your researchers before they design – avoid legacy samples as best you can!

Clinical Study?

Cohort Design

Essential to minimise the effects of(and record variation of) :

- Age Range
- gender
- drug use (statins, smoker status)
- metabolic disorders (diabetes, cardiovascular disease etc.)
- genetic heritage/race
- body mass

dependent on the research question other factors may also require consideration

Sample pre-collection

draw fasting blood (10-12 hours)

restrictions on diet (no alcohol, carb-rich food) & exercise 24 hours prior to draw

Sample storage

Separate plasma/serum from whole blood as directed prior to storage at -80^{o}C

Freeze thaw cycles effects sample integrity

Aliquot into either 500ul or 1ml fractions prior to storage

Data sources of variation – Confounding Factors

Care not to confuse correlation with causation

Minimise unwanted variation

-Gender & age matching
-Paired data?
-Appropriate controls and standards
-Beware of correlating conditions:
    -smoking and age/gender etc

Use appropriate analysis for study

-study numbers (per group)
-study groups
-genetic and phenotypic variation

Sample Origin

UNIVERSITY OF LIVERPOOL

NMR Centre for Structural Biology

Metabolomics Biomaterial Classification

Name:_____ **Date:**_____

Supervisor:_____ **Institute:**_____

Address:_____

_____

_____

_____

Duration of the Project: from: _____ to: _____

Sample derives from: Genus, Species, Strain :_____

Biomaterial(s):_____

Does the Biomaterial contains whole cells? Yes No

Sample Tracking

Requirement by law for Human-derived samples

Essential for **all** studies to have confidence that samples correctly transferred to NMR

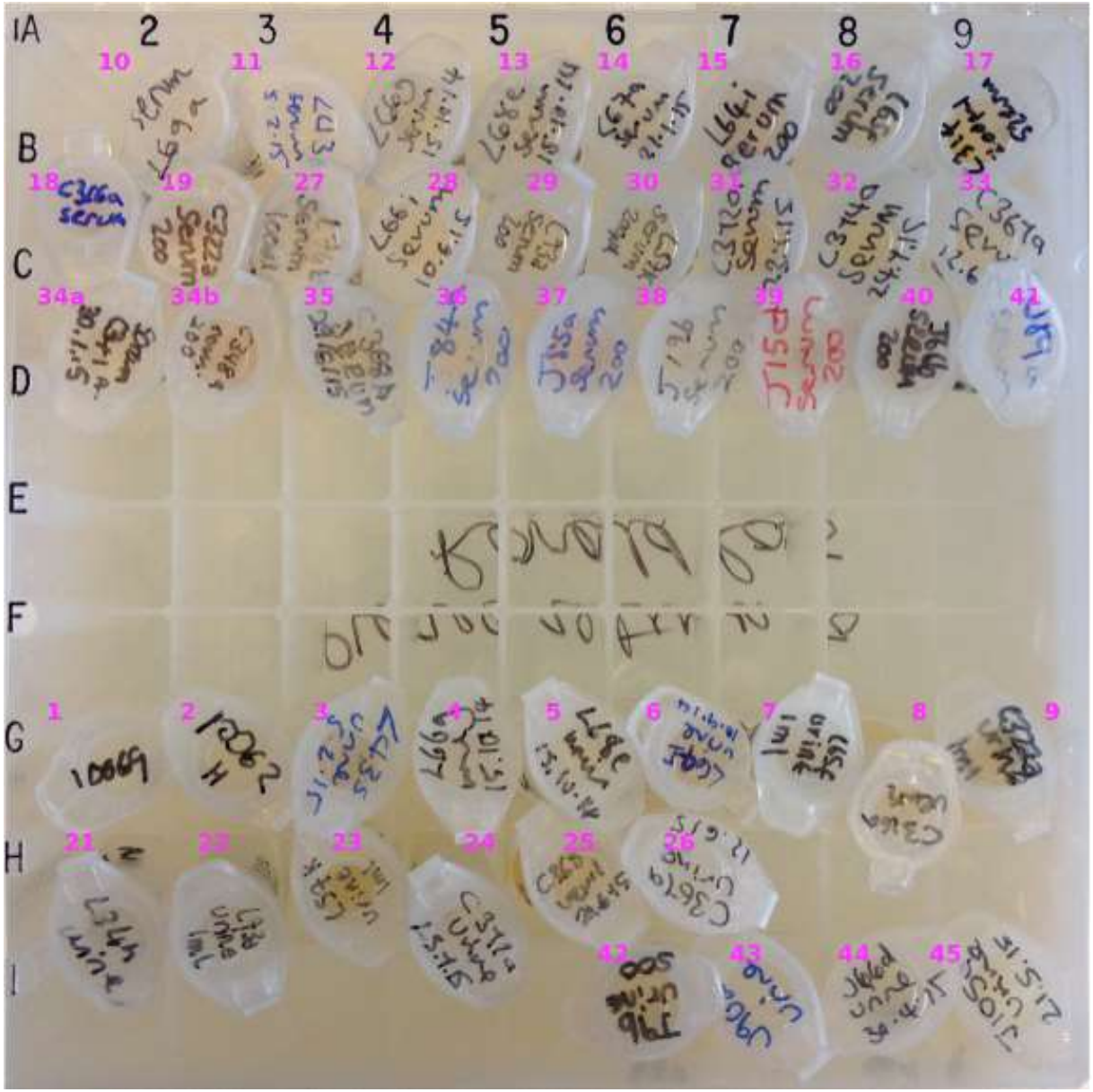
Best case scenario:

Samples individually labelled located in specific positions in a cryobox *AND* a spreadsheet of the annotation is provided alongside

Worst case scenario:

Samples provided labelled without associated spreadsheet or prescribed order

How can I guarantee that I have interpreted the order (and annotation) correctly?



SAMPLE NUMBER	URINE	SERUM
L69a	1	10
L62h	2	
L43s	3	11
L66g	4	12
L68e	5	13
L67c		14
L64i	6	15
L65f	7	16
C314a		17
C316a	8	18
C322a	9	19
L74b		27
L34h	21	
L66i		28
L73a	22	29
L57k	23	30
L30n		
C372a	24	31
C374a		32
C367a	26	33
C348a		34
C341a		
C368a	25	35
J84a		36
J85a		37
J19b		38
J15d		39
J66b		40
J89a		41
J9b	42	
J90a	43	
J66d	44	
J105a	45	

Annotated Spreadsheet – direct to IconNMR

Study Name and Date [COL_A]: File name in which the results will be saved. Please do NOT use spaces or special characters - only use text, number & underscore. For example Plasma_antibiotics_150114 is a plasma antibiotic study submitted on 15th October 2014.
Solvent [COL_B]: Which solvent is used in your sample? These depend on biomaterial and extraction procedure; for biofluids (blood, urine etc) state “H2O+D2O”, aqueous cell or tissue extracts “D2O”, lipophilic extracts either “MeOD” or “CDCl3” – if in doubt ask
Experiment [COL_C]: These are experiments set at the spectrometer – this can be left blank or set to the default is PROF_noesy
Position in NMR rack [COL_D]: These are the positions of the samples as prepared for NMR – if not yet prepared for NMR leave this blank.
Unique identifier [COL_E]: researchers own short-hand identity code for the sample.
Storage Tube ID [COL_F]: The annotation on the storage tube also should be recorded - for clarity we recommend numbering consecutively from 1 to n (n = total number of samples). If samples in 96-well block provide row and column position. A1, B1, C1, D1, E1, F1, G1, H1 then A2, B2, C2, D2, E2, F2, G2, H2 etc.
Cohort [COL_G] and replicate [COL_H]: identity for the cohort and biological/technical replicate number. If time course is provided this may be added as an additional column.
Study details [COL_I]: provide details of the study (investigator, institute, area of study)

NMR Sample Requirements

Sample number and Controls:

Sample groups need to be **consistent** and **reproducible.**

Study needs to be designed so that the variance **within** sample groups is less than the variance **between** sample groups ~1 ml - 100 µl fluid <100 mg tissue.

This depends on the type of study – human systemic (blood, urine etc.) samples vary dependent on diet, environment, ethnicity, age and gender.

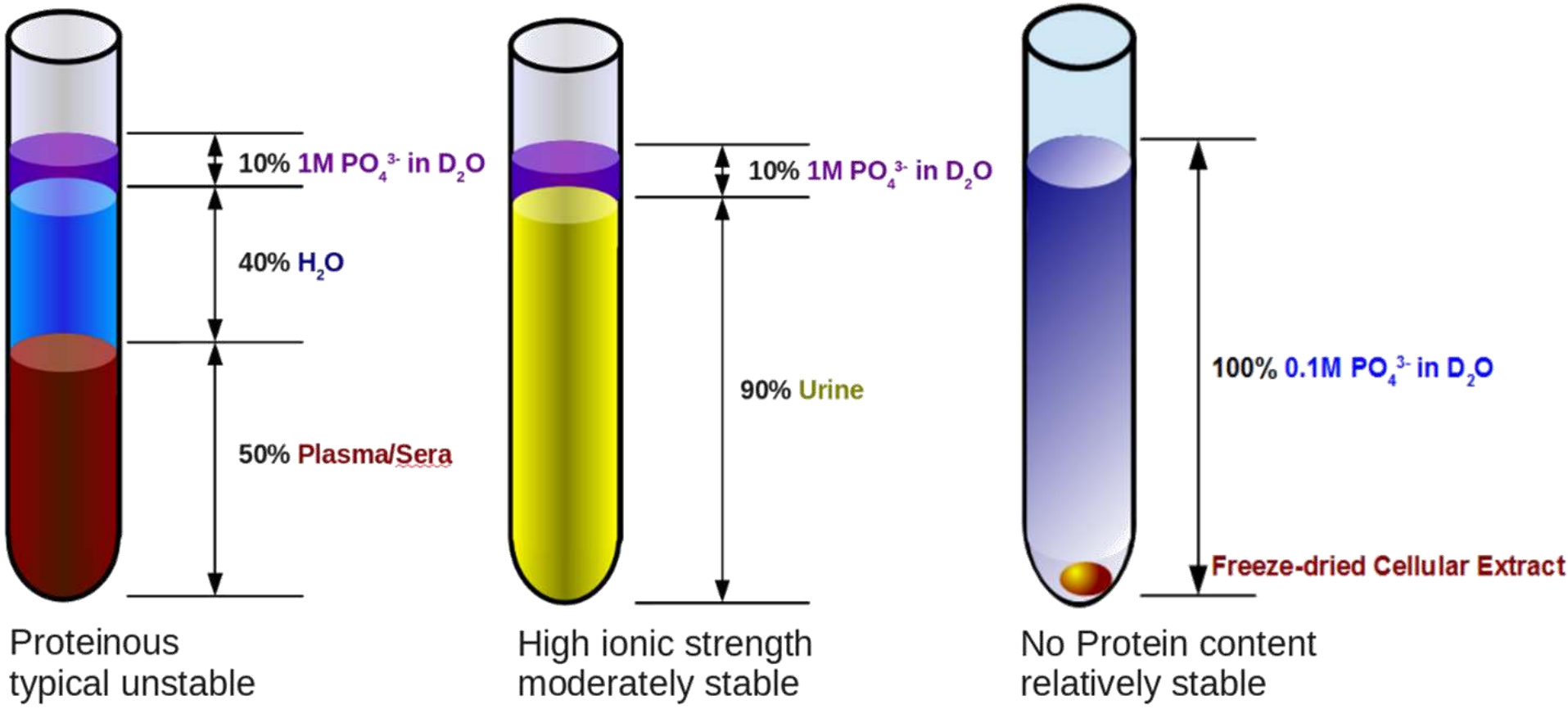
Samples cultured in the lab will have a lot less **intrinsic variance.**

Whole population studies typically require 1000+ samples per group.

Can reduce this by lowering the **intrinsic variance** per group with reductions in study criteria, i.e. age, gender and specific environmental requirements etc.

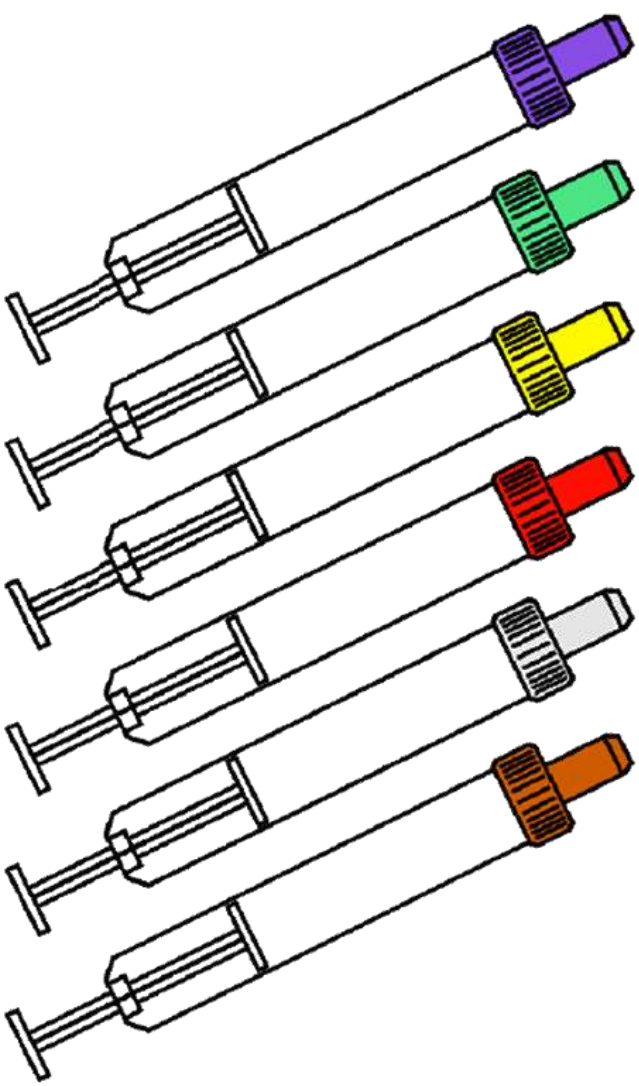
Power Calculations are increasingly required by MRC applications… estimate sample (n) required for a given effect size… not easy for multivariate data

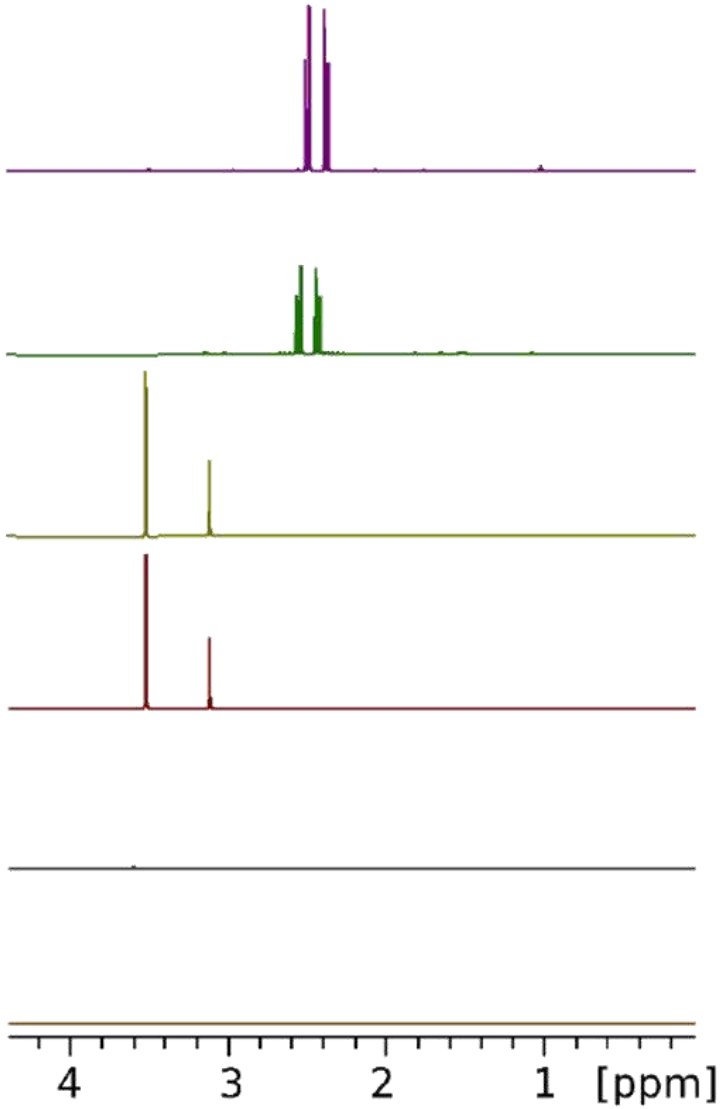
NMR Sample Preparation



Beckonert *et al* Sample Preparation: Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols* 2007 2(11):2692-703.

Blood – Special considerations





Phelan & Lian 'NMR Metabolomics: A Comparison of the Suitability of Various Commonly Used NHS Blood Collections tubes' *Current metabolomics*, **2016**, 4, 78-81

Manual Preparation or Liquid Handler

Advantages

- Often quicker
- Lower sample wastage

Limitations

- Danger of inconsistent pipetting – requires practise
- Danger of sample confusion – requires concentration (=> safer for smaller cohorts)
- Higher variance between researchers

Training and testing on small cohort or test set essential to hone technique prior to large sample preparation.

This also ensure adequate time dedicated to the task (must not rush)

Suitable for

- Samples without particulate (pre-spun)
- Large volumes
- Large number
- Removing samples from tubes for long-term storage

Unsuitable for

- Viscous samples (blood, SF)
- Volatile samples (lipidomics)
- Small volumes

At the spectrometer: Requirements

Instructions for Installation and Optimization of Metabonomics NMR Parameter Sets

Temporary version still under extensive revision!

13.10.2009

Author: Hartmut Schaefer

hartmut.schaefer@bruker-biospin.de

Important technical requirements for optimum results

Procedures and methods described in the instructions are based on availability of the following hard- and software

- Digital Receiver Unit (DRU) which needs an AVII or AVIII
- operation with inverse probes
- probes with Automated Tuning and Matching Unit ATM
- BTO 2000 for room temperature (RT) probes
- cooling unit BCU 05
- sample changer BACS or SampleJet
- TopSpin 2.0 or later

Requirements prior to use this publication

- System up and running
- properly installed and all compounds in Bruker specification
- properly routed
- probe properly registered via EDHEAD
- EDTE, flow properly set particularly for Cryo
- CORTAB available and correct
- PROSOL values for all standard solvents properly determined

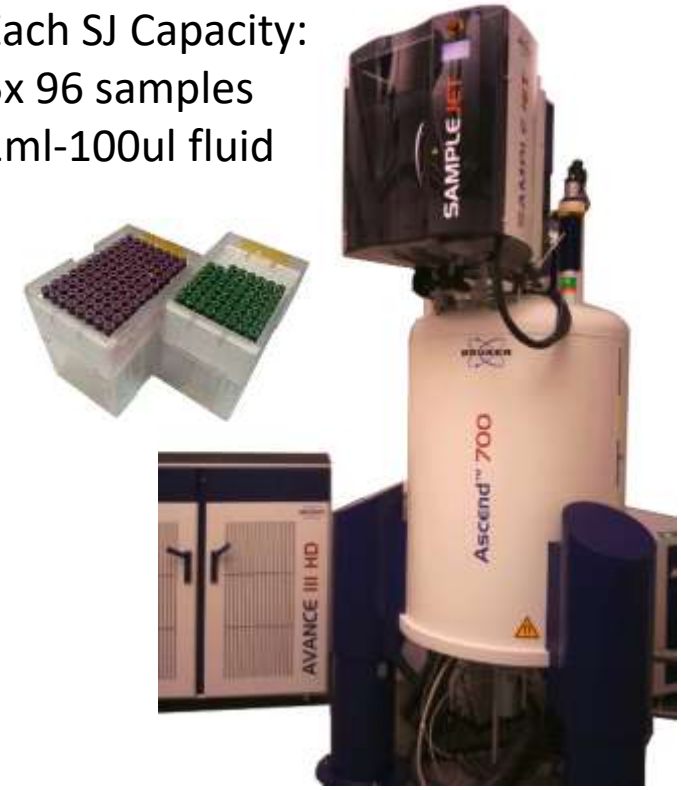
At the spectrometer: Hardware

- All spectrometers **SampleJets** and ^{1}H ^{13}C ^{15}N **triple resonance** probes for **solution** NMR
- 800MHz also has **solid-state** capabilities
- **Computational Suite** for analysis and training

Each SJ Capacity:
5x 96 samples
1ml-100ul fluid

600MHz Avance III
c.2007
^{1}H ^{13}C ^{15}N helium-cooled CryoProbe with atma
RT SampleJet

800MHz Neo
c.2018
^{1}H ^{13}C ^{15}N helium-cooled CryoProbe with atma
X ^{1}H BroadBand (X includes ^{11}B, ^{13}C, ^{31}P, ^{133}Cs, ^{195}Pt) Probe
X ^{1}H HR-MAS
Chilled (rack only) SampleJet



700MHz Avance IIIHD
c.2015
^{1}H ^{13}C ^{15}N helium-cooled CryoProbe
^{1}H ^{13}C ^{15}N RT Probe with atma
Chilled (rack and spinner) SampleJet

Take a (old) digital tour: www.tinyurl.com/LivNMRtour

At the spectrometer: Quality Assurance

Set of Procedures that are performed in advance of sample analysis

- Equipment within specification
- Consumables of a certain quality
- Standard Operating Procedures
- Standard Reference Materials

Investigation into Quality Assurance and Quality Control:
Dunn *et al* 2017 doi:10.1007/s11306-01-1188-9

In Practice prior to ***every*** batch/day/run:

- Run blanks
- Run temperature calibration
  (use 99.8% 2H-methanol standard from Bruker and au calctemp after zg30 1d)
- Run 3D shimming and check LWHH on DSS
  (use Sucrose water suppression standard from bruker and topshim gui to select 3D)

BONUS save shimmap to icon (for given lock solvent) to ensure every sample starts from standard shims

At the spectrometer: Initial Set-up

Courtesy of Pete Gierth/Hartmut Schaefer!

Optimise setup for long relaxing 1H signals with perfect baseline:

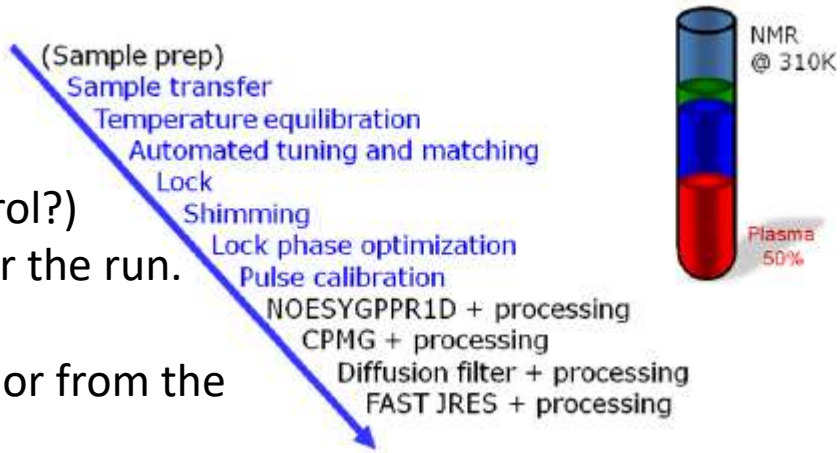
- de
- D1
- O1 (use zgpr to optimse o1 effectively – poor water suppression gives greater improvement when accurately set)

Default parameters from Bruker optimised for Avance II hardware in Liverpool available: www.ebi.ac.uk/MetaboLights

Really only want icon to handle samples of similar composition (a batch of serum for example have similar ionic strength, protein content viscosity etc.)

In Practice *every* batch/day/run:
Optimise a representative sample (pooled Or control?) for o1 and overwrite this in every parameter set for the run.

Keep all parameters as recommended in literature or from the initial set-up : (fixed RG, SW, NS, TD, etc. etc)



At the spectrometer: icon automations

Courtesy of Pete Gierth/Hartmut Schaefer!

au_plasma_noesy (1st expt):

```
// fix presat field strength 25Hz

   float PresatField = 25.0;
   float P1, PL1, PL9, F1;

// get current data

   GETCURDATA

// fix parameters

   STOREPAR ("PULPROG", "noesygppr1d");
   STOREPAR ("DIGMOD",  3);
   STOREPAR ("D 1",     4.0);
   STOREPAR ("D 8",     0.01);
   STOREPAR ("D 16",    0.0002);
   STOREPAR ("SWH",     18028.846);
   STOREPAR ("RG",      90.5);
   STOREPAR ("TD",      98304);
   STOREPAR ("NS",      32);
   STOREPAR ("DS",      4);
   STOREPAR ("GPZ 1",   50.0);
   STOREPAR ("GPZ 2",   -10.0);
   STOREPAR ("GPNAM1",  "SMSQ10.100");
   STOREPAR ("GPNAM2",  "SMSQ10.100");
   STOREPAR ("P 16",    1000.00);
   STOREPAR ("ZGOPTNS", "-DFLAG_BLK");

// optimize lockphase

   AUTOPHASE;

// determine 90deg pulse automatically, no display of results
// ATTENTION: pulse calibration starts with PROSOL values
// current values are IGNORED

   XCMD("pulsecal fast quiet");

// ensure PLdB9 consistency to 25.0 Hz RF-field

   FETCHPAR ("P 1",    &P1);
   FETCHPAR ("PLdB 1", &PL1);

   P1    = P1*1.0e-6;

   F1    = 1.0/4.0/P1;
   PL9   = PL1 + 20.0*log10 (F1/PresatField);

// set PLdB9

   STOREPAR ("PLdB 9", PL9);

// parameter migration to subsequent experiment

   XCMD("saveprofpars");

// run experiment

   ZG
```

au_plasma_cpmg:

```
// get current data

   GETCURDATA

// migrating parameters from CYPNO 00000

   XCMD("getprofpars")

// fix parameters

   STOREPAR ("PULPROG", "cpmgpr1d");
   STOREPAR ("DIGMOD",  3);
   STOREPAR ("SWH",     12019.230);
   STOREPAR ("D 20",    0.0003);
   STOREPAR ("L 4",     128);
   STOREPAR ("RG",      90.5);
   STOREPAR ("TD",      73728);
   STOREPAR ("DS",      4);

// run experiment

   ZG

   QUIT
```

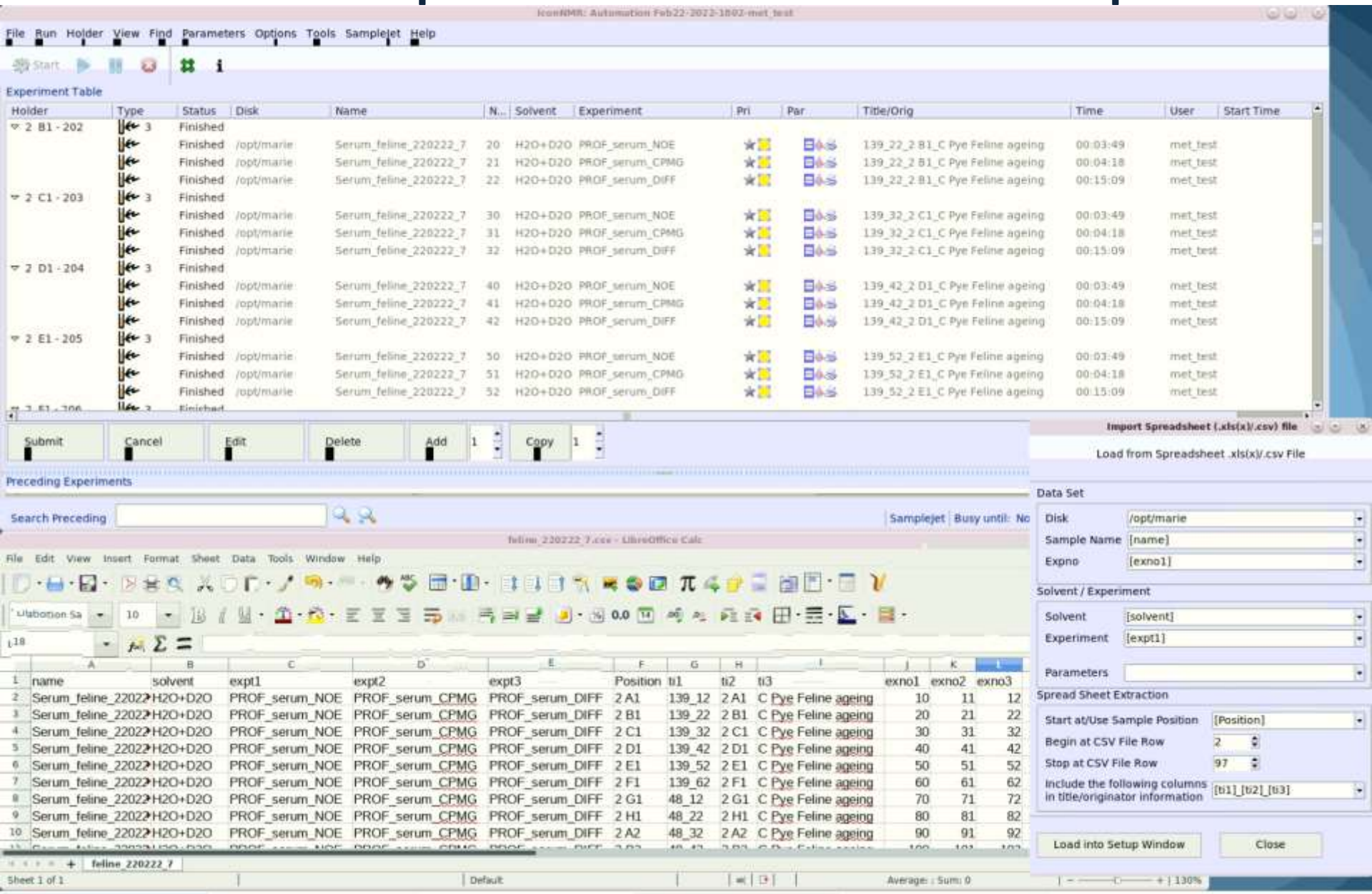
At the spectrometer: icon configurations

Courtesy of Pete Gierth/Hartmut Schaefer!

Set lock actions for each solvent
Can set different standard shim files
NB H2O+D2O tend to be blood serum/plasma @37^{o}C
D2O tend to be freeze dried extracts @25^{o}C

rsh 3D shimmap – established that day/batch/run at identical temperature
Then perform 1D topshim routine specific for solvent system
Atma tuning

At the spectrometer: Automation set-up



Automated processing – High-Throughput

Pros:

- Quick
- Consistent
- Independent of setup/day
- Good enough for most applications

Cons:

- May introduce systematic bias
- Hands-off approach
- Not always suitable (referencing)
- May lose subtle effects

Quality Control

Set of activities that are done during or immediately after analysis to demonstrate the quality of the data.

- Datasets meet certain criteria
  - Signal-to-noise
  - Temperature stability
  - Column/matrix stability

Published data should meet minimum level of reporting:

Sumner *et al* 2007
doi: 10.1007/s11306-007-0082-2
Salek *et al* 2013
doi: 10.1186/2047-217X-2-13

NMR QC Checklist

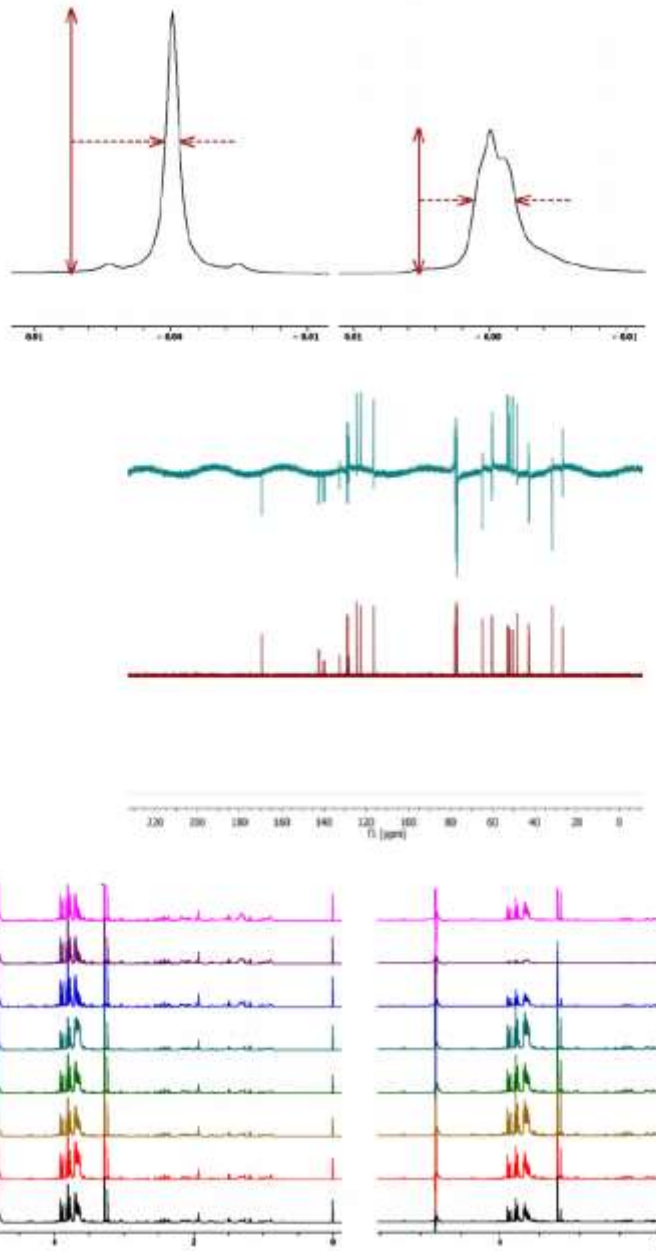
(i) **Referencing** Quality of spectrum with regards to a reference material – signal strength, shape and width and position important.

(ii) **Baseline** flat without curvature or sine wiggle.

(iii) **Signal to noise** as expected (check against representative spectrum)

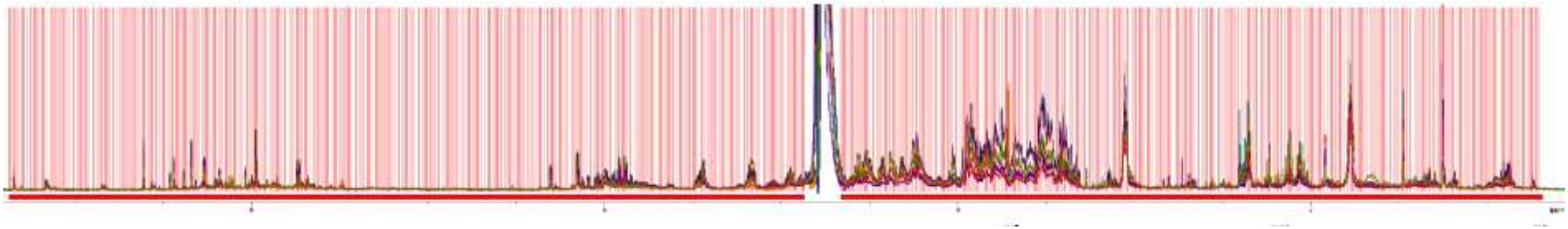
(iv) **Water suppression** good. Narrow water signal (between 0.2 and 0.4 ppm wide). No baseline distortion beyond that range.

(v) **Phase** Peaks are uniform and symmetrical

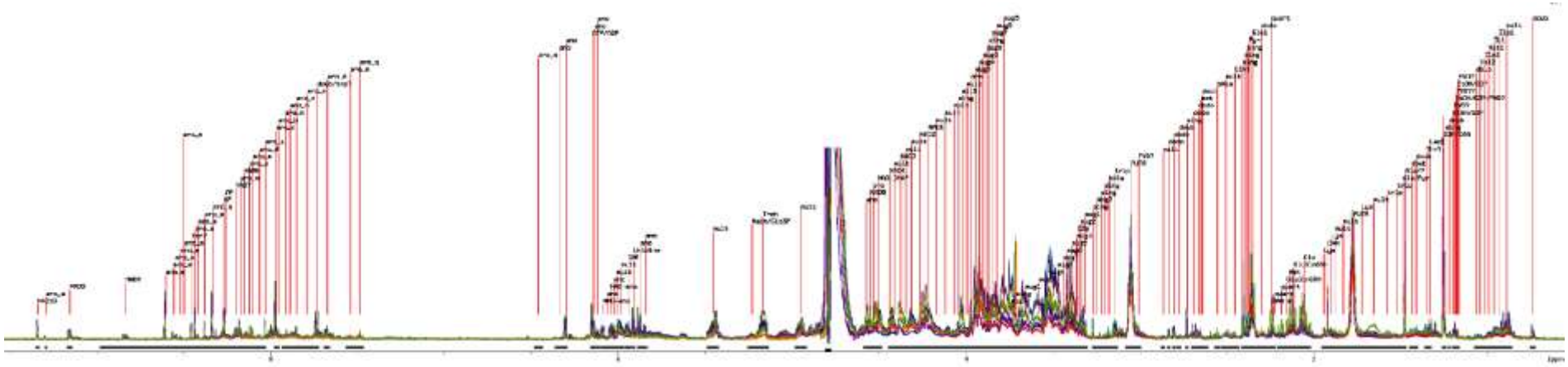


Data Analysis – Spectral Integration

Either



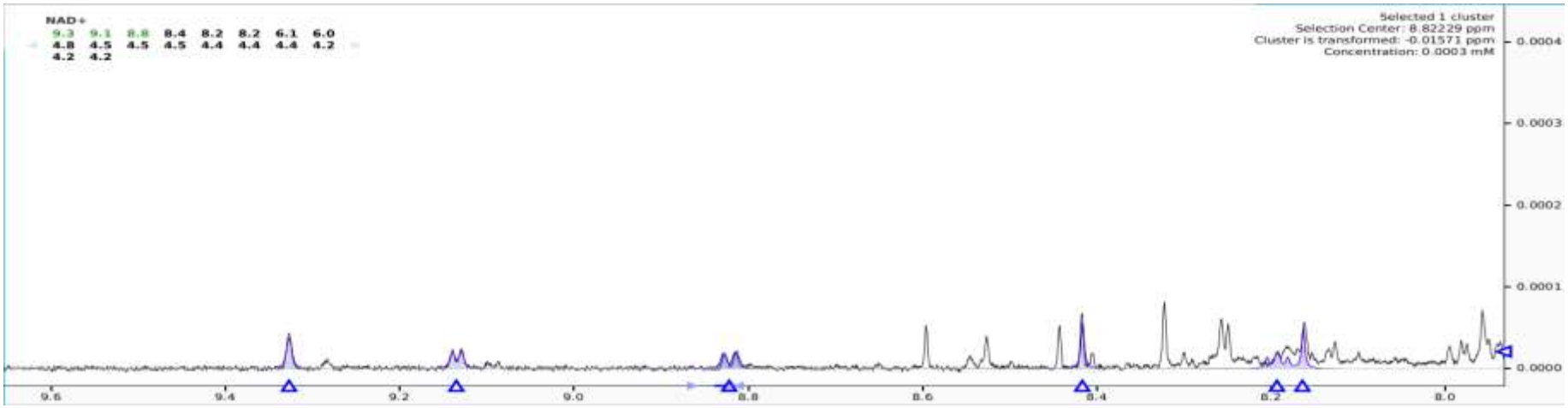
1. divide spectrum into equal increments and integrate intensity for each increment



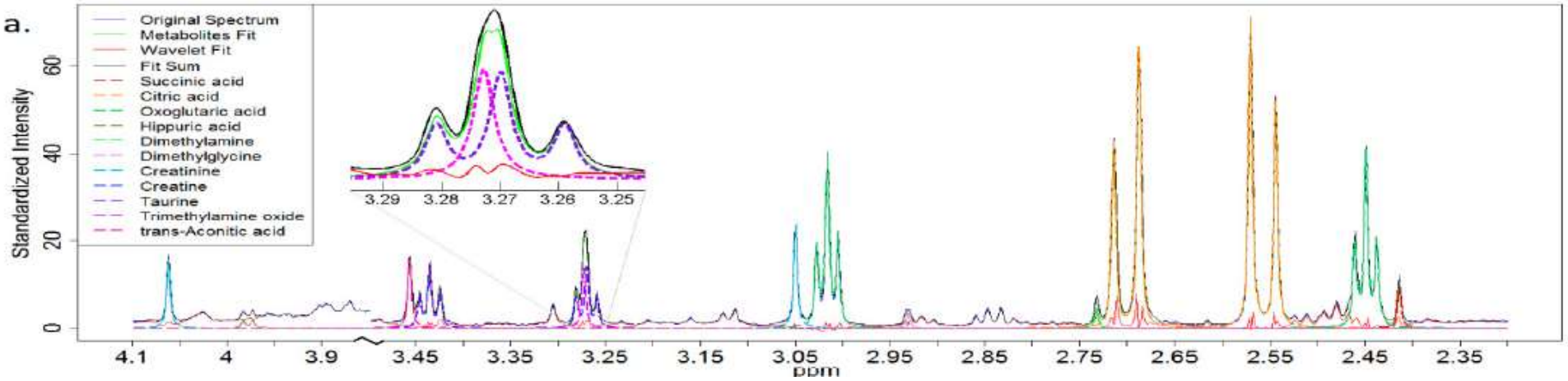
2. divide spectrum up into individual peaks and integrate intensity of each peak/group of peaks (requires a pattern file)

Data Analysis – Spectral Deconvolution

Either

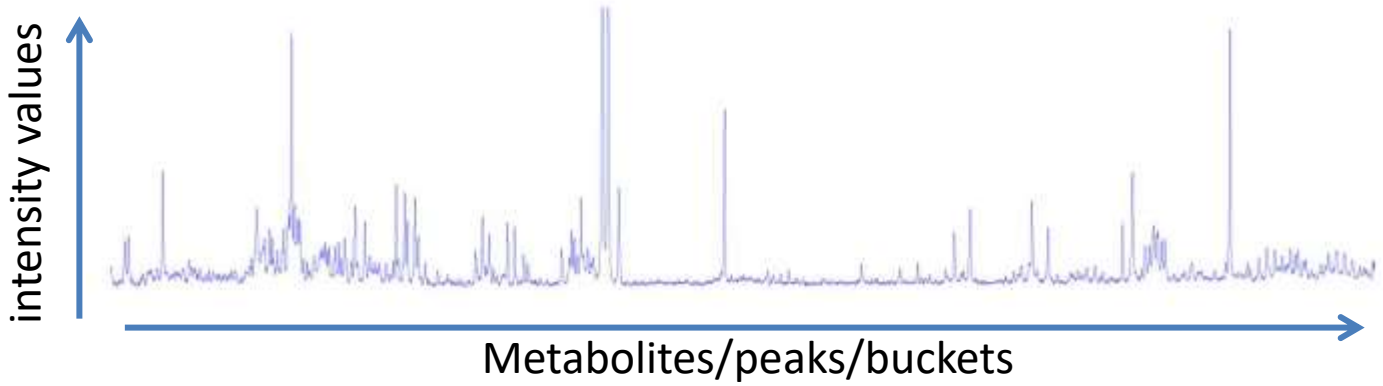


1. Use proprietary software (chenomx)



2. Use wave fitting programme

Spectral Formatting



Transformation of a Spectrum

Only useful to NMR Analysts using specific software

samples

Metabolites/peaks/buckets

intensity values

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		Pseudo_6	Pseudo_6	Pseudo_6	Pseudo_6	Pseudo_6	Pseudo_6	Pseudo_6	Pseudo_6	Pseudo_6	Pseudo_6	Pseudo_6	Pseudo_6	Pseudo_6	Pseudo_6
2	9.99975	3.86E-05	8.23E-06	1.36E-05	3.38E-05	2E-07	2.6E-05	-5.7E-05	-3.6E-05	2.35E-05	4.48E-05	-3E-06	-4.4E-06	-5.6E-06	0.000026
3	9.99925	3.49E-05	-5.1E-07	1.4E-05	1.67E-05	1.43E-05	8.16E-06	-6.5E-05	2.02E-05	0.000026	2.89E-05	-2.1E-05	6.82E-05	-2E-06	2.15E-05
4	9.99875	-6.5E-06	6.67E-06	3.38E-05	2.69E-05	1.18E-				5	3.48E-05	-2.1E-05	4.57E-05	2.76E-06	3.77E-05
5	9.99825	-9E-06	-5.5E-06	5.06E-05	1.35E-05	-2.3E-				5	1.83E-05	7.01E-06	1.96E-05	1.47E-05	4.65E-05
6	9.99775	-2.1E-05	-1.2E-05	5.25E-05	-3.6E-05	-3.6E-				5	1.22E-06	2.21E-05	4.1E-05	1.11E-05	4.06E-05
7	9.99725	-1.7E-05	-1.1E-05	5.91E-05	-5E-05	-2.6E-05	1.06E-05	3.35E-05	4.13E-05	3.09E-06	7.63E-06	2.58E-05	3.46E-05	2.37E-05	2.01E-05
8	9.99675	-2.5E-06	-2.1E-05	9.55E-06	-1.9E-05	-2.4E-06	1.4E-05	8.02E-06	-3.9E-07	-2.1E-05	5.63E-06	9.77E-06	2.61E-05	4.96E-05	2.65E-06
9	9.99625	5.85E-06	-1.2E-05	-2.5E-05	6.78E-06	-4.6E-06	-1.1E-06	2.29E-05	-1.5E-05	1.68E-05	1.01E-05	1.42E-05	2.21E-05	4.51E-05	1.67E-05
10	9.99575	-4.6E-07	-2.5E-05	-6.9E-06	-9E-06	-1.7E-05	1.49E-05	7.39E-05	-3.3E-05	3.42E-05	0.000017	-5.4E-06	2.76E-05	2.59E-05	6.22E-06
11	9.99525	-3E-06	-8.2E-06	-1.2E-05	-1.8E-05	-2.2E-05	1.72E-05	7.91E-05	-2.7E-05	3.5E-05	3.24E-05	-1E-05	1.79E-05	1.15E-05	7.5E-06
12	9.99475	-3.8E-06	3.99E-05	4.84E-05	-5.5E-05	-3.7E-06	2.85E-05	3.54E-05	-1.6E-06	1.9E-05	4.84E-05	-1.9E-05	6.39E-05	2.51E-05	1.82E-06

Into a Matrix of numbers.

Interpretation by multiple analysts and tools

Why Worry About Statistics for NMR?

- NMR is Multivariate
  i.e. each peak is a separate variable

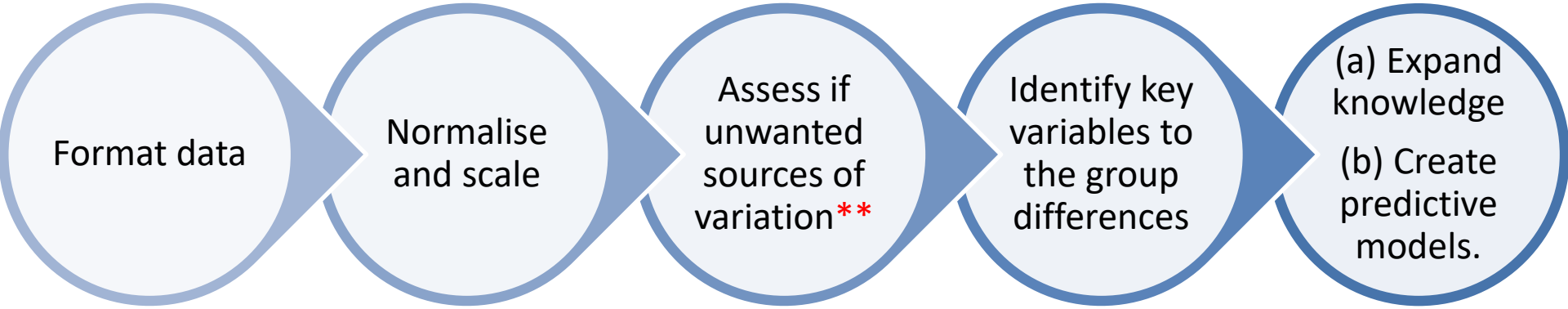
- NMR is used to build models
  how do we validate / assign value to these models?

- Multiple sources of information must be combined
  how do we combine information in a balanced and accurate manner?

Statistics is employed in:

- Ligand binding/screening
- Structure Calculations and Validations
- Metabolomics
- Concentration Calculations

Data Analysis



Statistical Data analysis

Usual software used for statistical analysis:

- SPSS
- Minitab
- Stata
- OriginPro
- Graph Pad
- Simca
- ...

R:

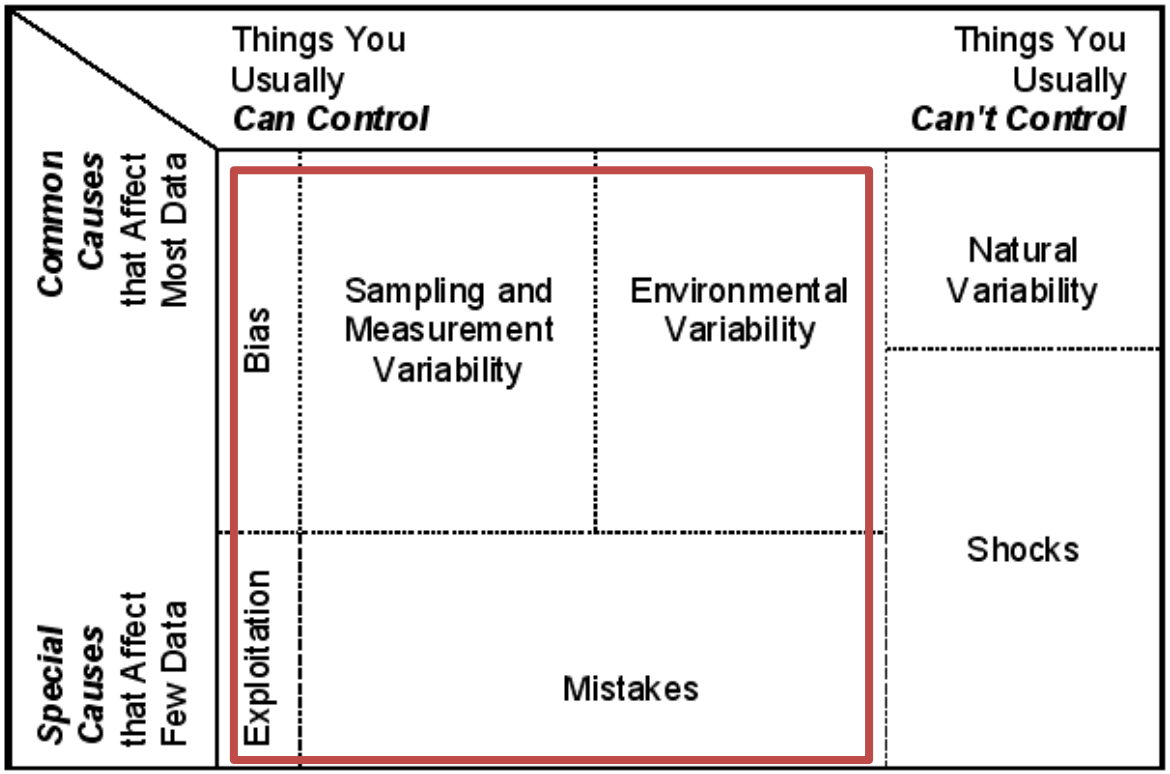
- Open source
- Powerful and flexible (it is much more than an statistical analysis software)
- Operates at command line and it is also a **programming language** therefore can be extra powerful

https://www.r-project.org/

For easier visualisation and use:
https://www.rstudio.com/

Statistical Analysis - Data sources of variation

Data = biological meaning + error



- Best way of controlling error is through a **robust experimental design** (e.g. minimise cohort variability, have controls in place, randomise your sampling, etc.).
- Remember the 3 Rs: **Reference**, **Replicate**, **Randomise**

- correct use of **normalisation** and **scaling** steps can minimize undesirable variance from dilution effects etc.

Refresher – the language of statistics

•Reporting and understanding statistical treatments requires understanding of key terms such as significance, variable, class, scaling, average etc.

•Unfortunately not only are some of these terms used in everyday language but even within statistics certain terms are taken to mean different things.

•Where there are multiple uses for a given term we are working with the most widely used definition.

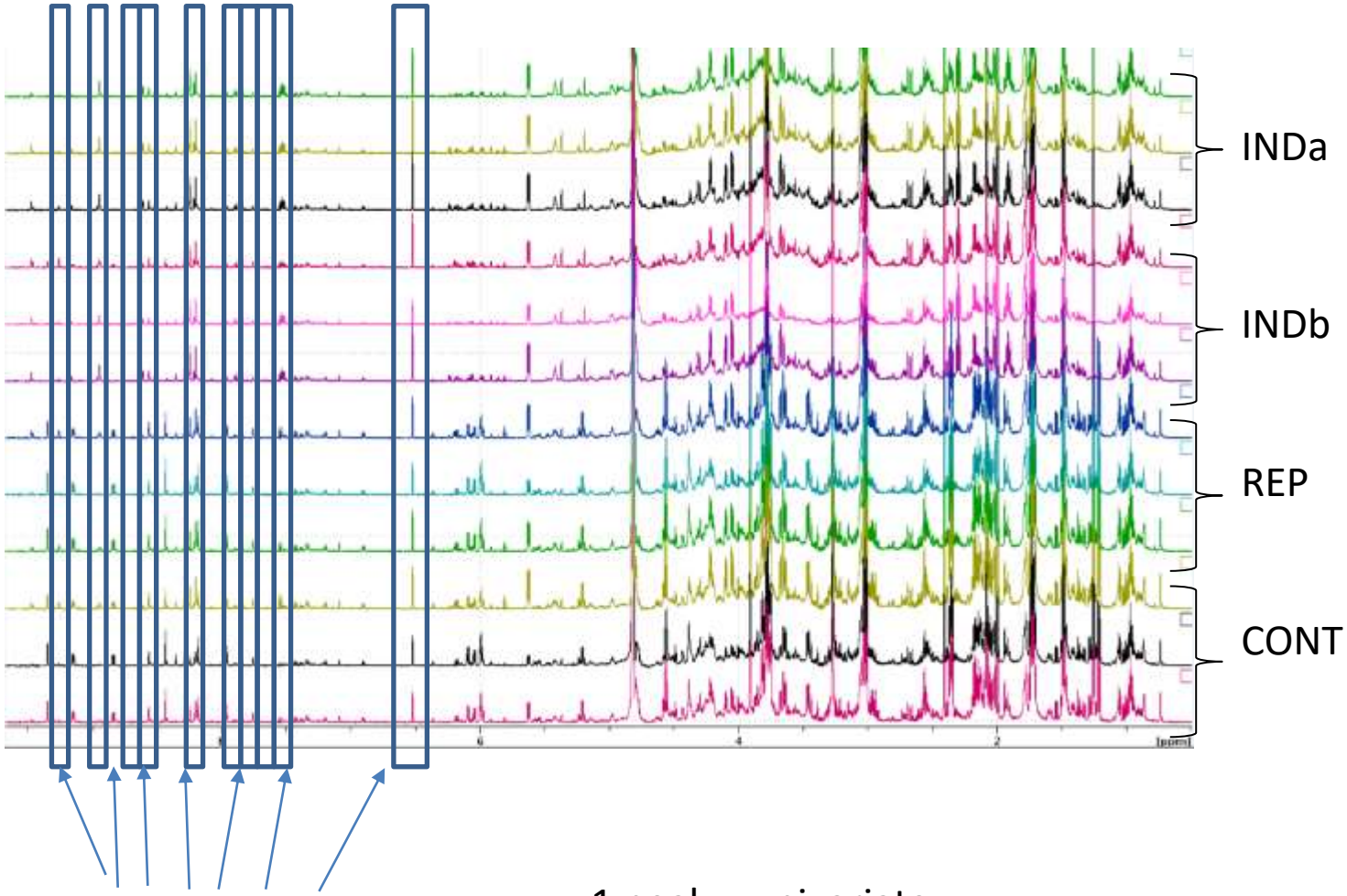
•Be aware that certain journals/articles/books may use less common definitions.

•Also avoid using the common language definitions in scientific reporting; i.e. only say something is **significant** if you can back it up with statistical analysis.

group statistic	population parameter	description
n	N	number of members of sample or population
$\bar{x}$ "x-bar"	μ "mu" or μ_x	mean
M or Med	(none)	median
s (TIs say Sx)	σ "sigma" or σ_x	standard deviation For variance, apply a squared symbol (s^2 or σ^2).
r	ρ "rho"	coefficient of linear correlation
$\hat{p}$ "p-hat"	p	proportion
z t χ^2	(n/a)	calculated test statistic

Significance tests

Are these metabolite/s signals significantly different?



How many peaks/metabolites / variables?

1 peak = univariate
Multiple peaks = multivariate

How many groups/cohorts/classes of sample?

1 group
One-sample t-test

2 groups
Two-sample t-test

3+ groups
ANOVA

Data preparation before performing Hypothesis testing

- Objectives:
  (a) reduce variance between experiments (batch effect)
  (b) make variables within an experiment comparable, independently of their absolute value in order to assess the changes more accurately.

- Some of these include
  (a) data transformations such as log, square-root
  (b) data scaling such as mean centering or *Pareto* scaling or
  (c) normalisation by a reference variable or sample.

Why do we need to scale & normalise biofluids?

- Intensities are relative to the largest signal (and the detector)
- Some biofluids may be diluted at different levels by the biosystem (urine)
- Metabolites of interest may be degrees of magnitude lower than other metabolites – need statistical test that will consider all metabolites equally

Data preparation - Is it appropriate to normalise?

Dilution effects in biofluids such as urine require normalisation as intensities **between** spectra will be artificially different.

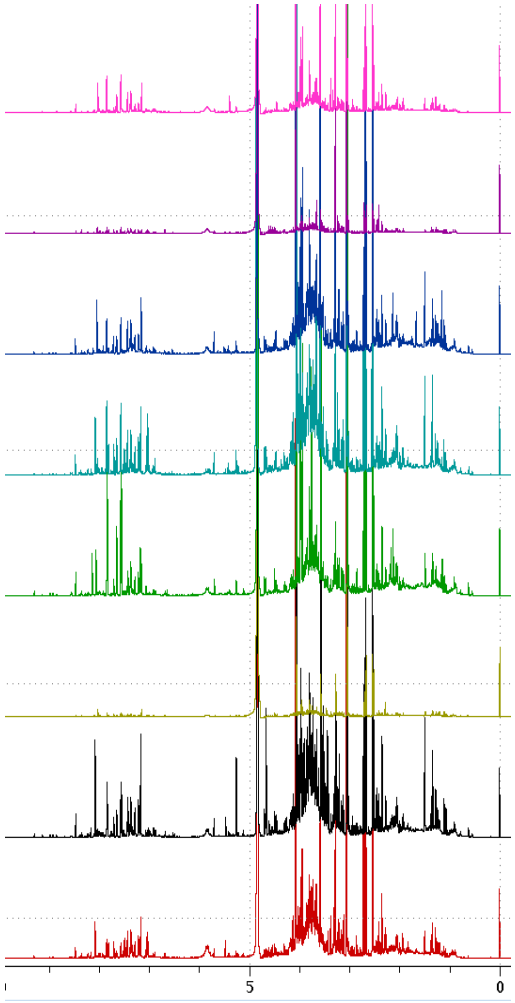
Tissue and **Cells** need optimised extraction:

- ensure **consistent** levels of metabolites.
- extract should either be **constant** or
- **normalised** by use of a reference material (typically TSP)
- **Add** Reference pre-extraction: at a ratio [TSP]:[biomaterial]

Systemic fluids are **homeostatic**:

peritoneal
plasma
synovial fluid etc.

thus normalisation may not be appropriate (or make a difference).



Data Preparation – Does the data need to be transformed?

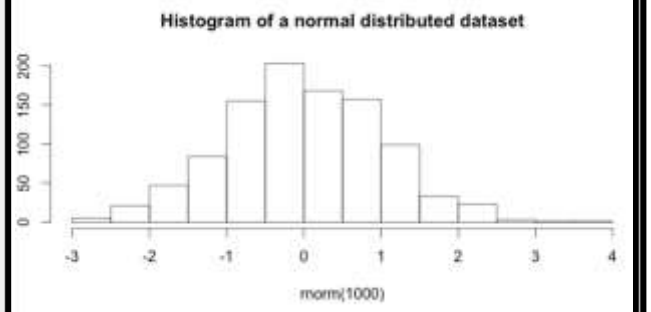
First need to ask – does the data follow a normal distribution?

a) It needs to have the **shape** of a normal distribution→ we can plot an histogram to see how it looks like.

b) We can also do an Statistical test called the **Shapiro-Wilk** test that would indicate if data is normal or not.

c) However for big datasets with many many replicates it is not very reliable and so we can do a **Q-Q-plot**.

a



b

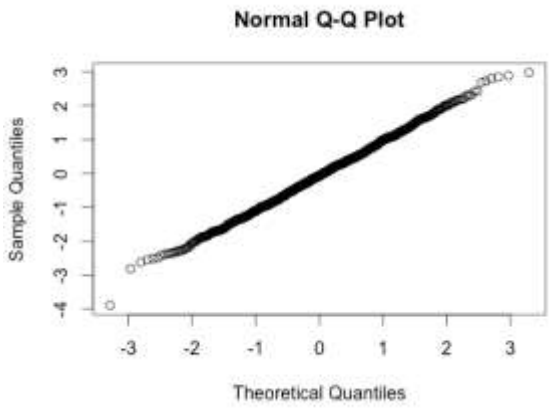
```
> shapiro.test(data)

        Shapiro-Wilk normality test

data:  data
W = 0.97118, p-value = 0.2583
```

Usually p<0.05 or p<0.01 is significant; not this case

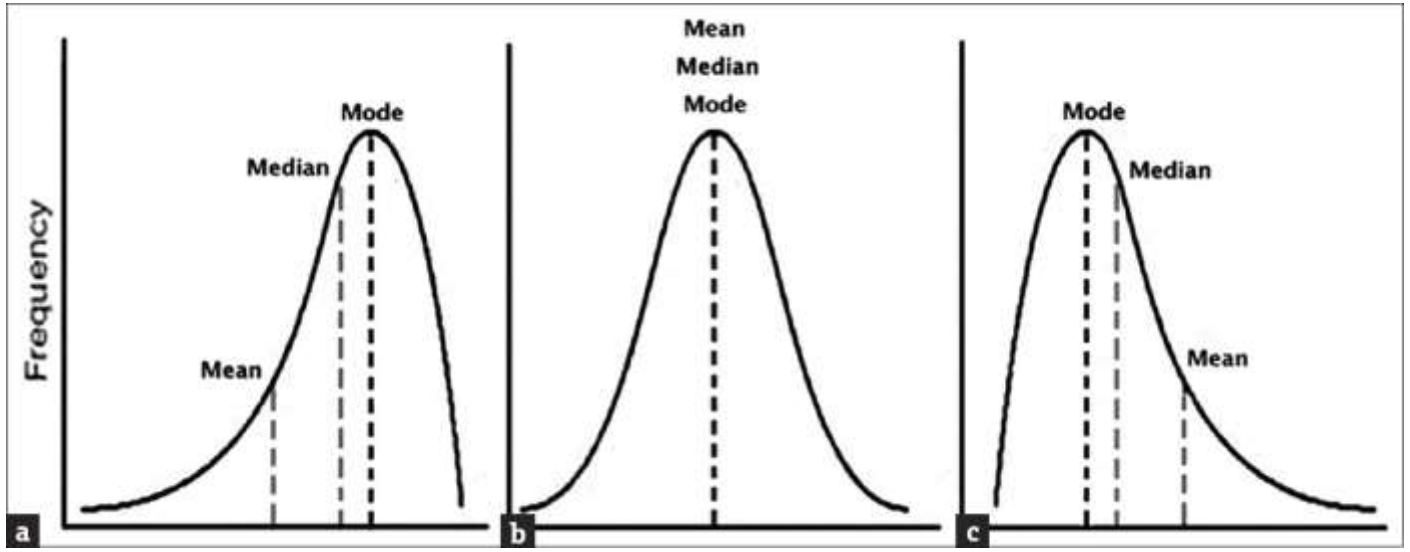
c



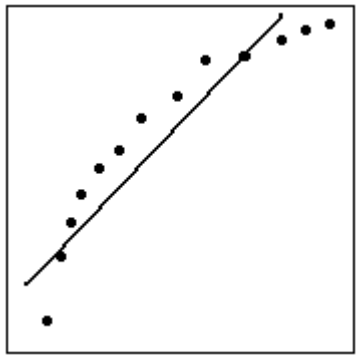
When is it appropriate to transform the data?

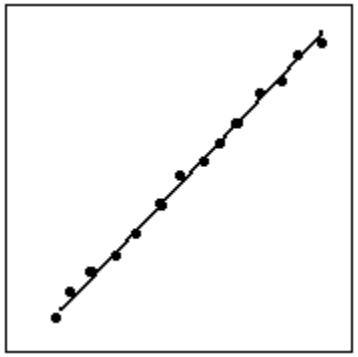
Data that exhibits a **skewed** distribution
A log transformation will return a more **normal distribution**

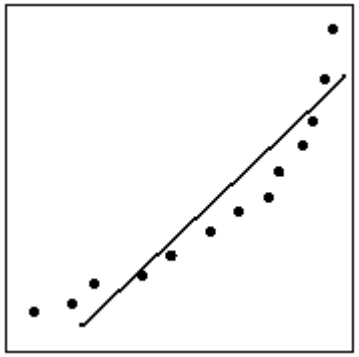
Negative Skew | Normal (no skew) | Positive skew



(perfect symmetry)



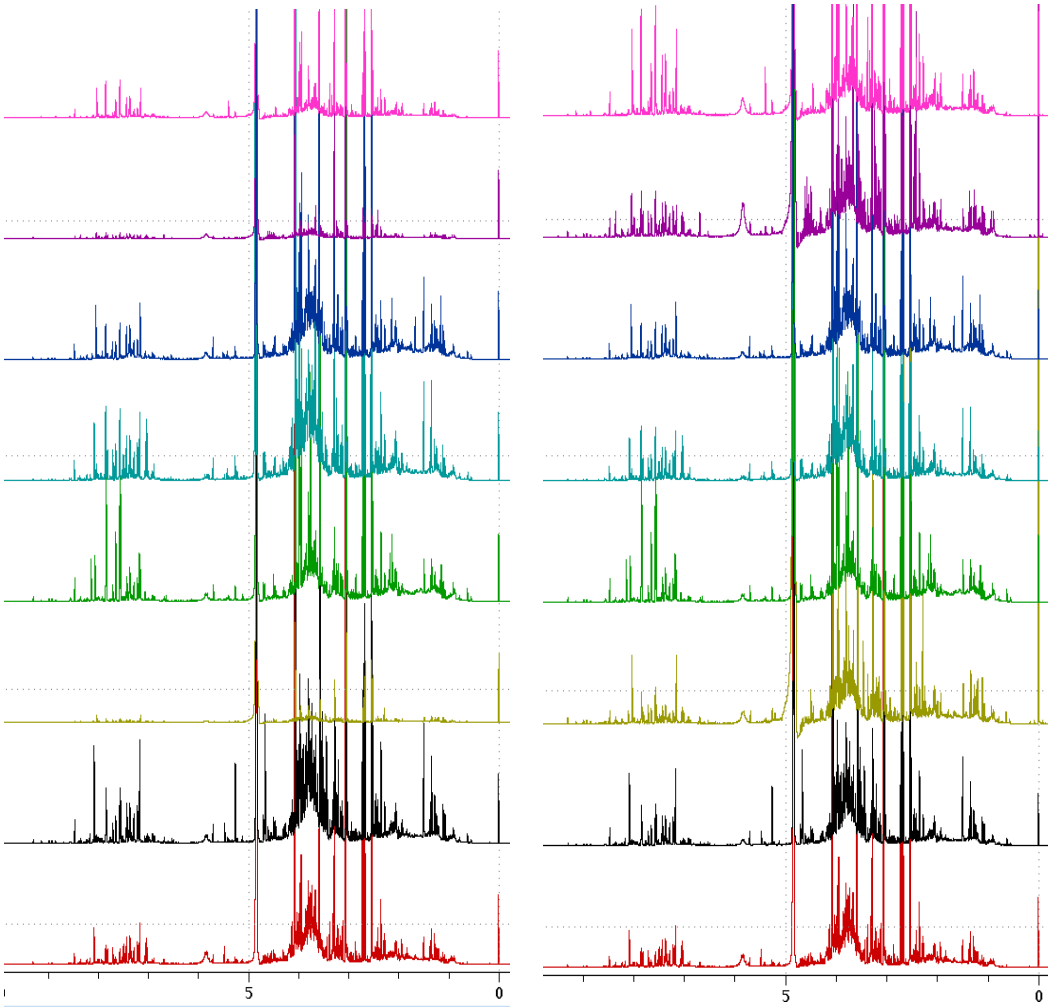




Examples of Data preparation - Normalisation and transformation

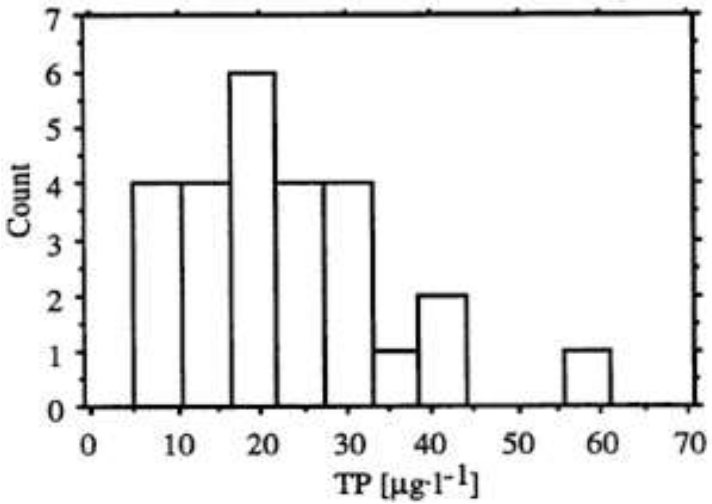
Before normalisation

After normalisation

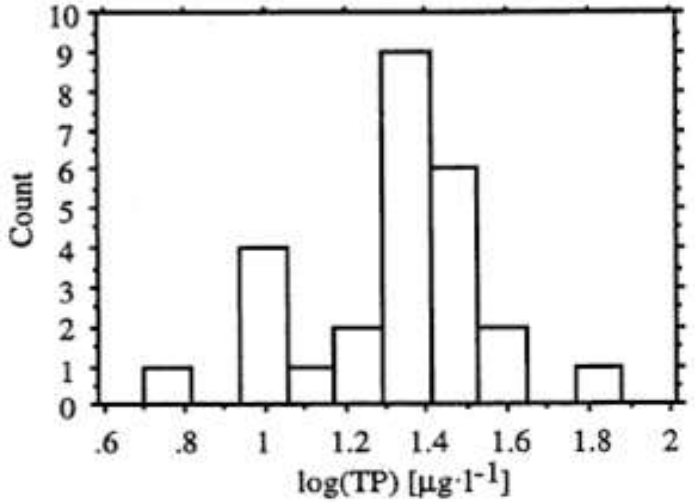


Urine Spectra

Before transformation

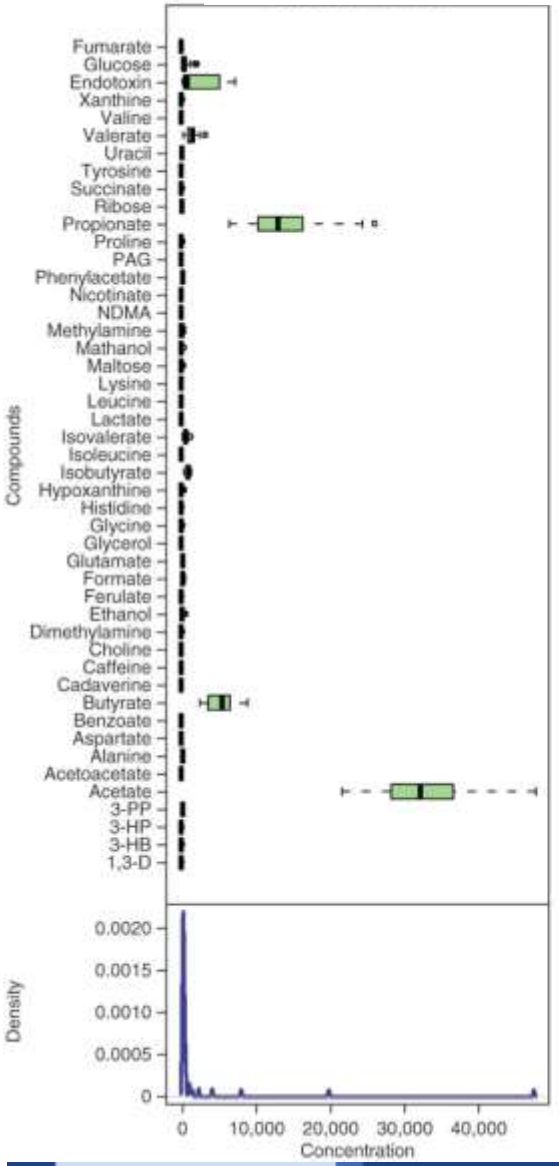


After transformation

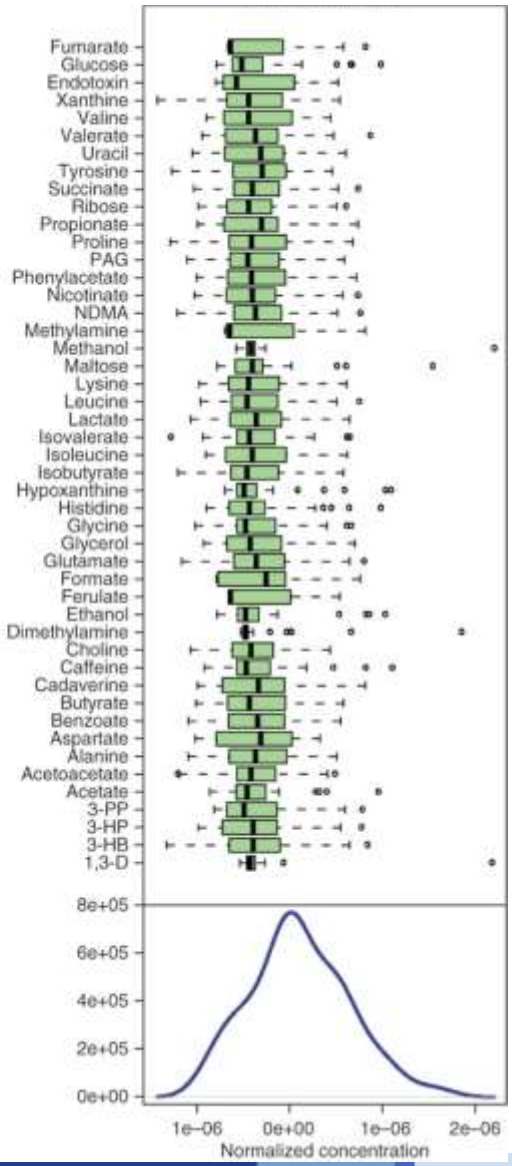


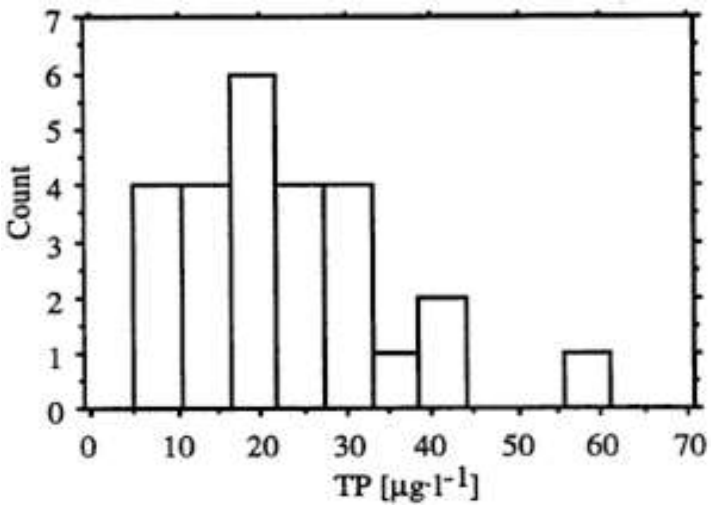
Normalisation and Scaling

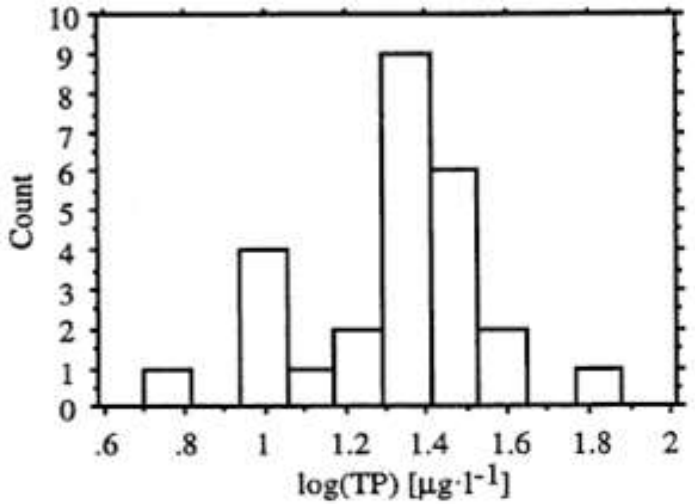
Before scaling



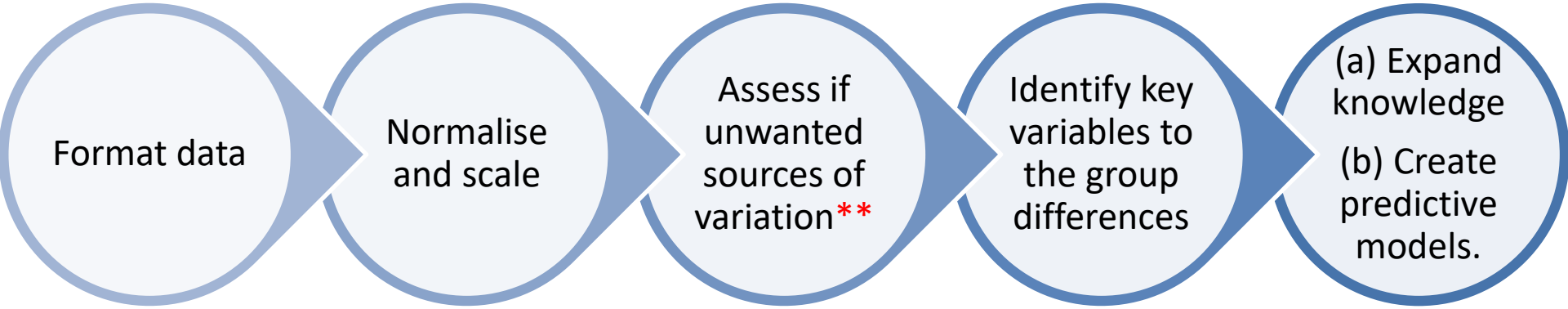
After scaling





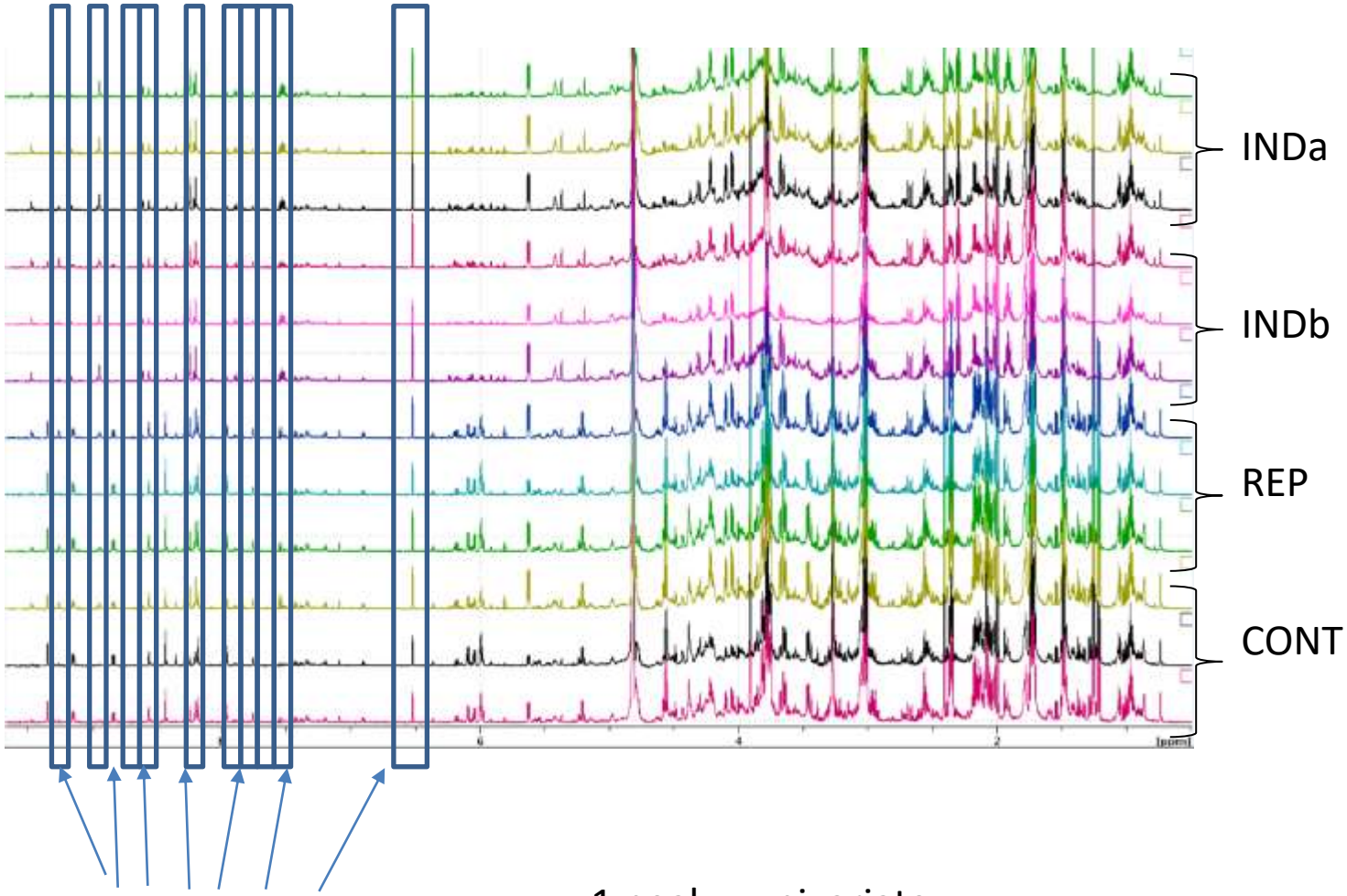


Data Analysis



Significance tests

Are these metabolite/s signals significantly different?



How many peaks/metabolites / variables?

1 peak = univariate
Multiple peaks = multivariate

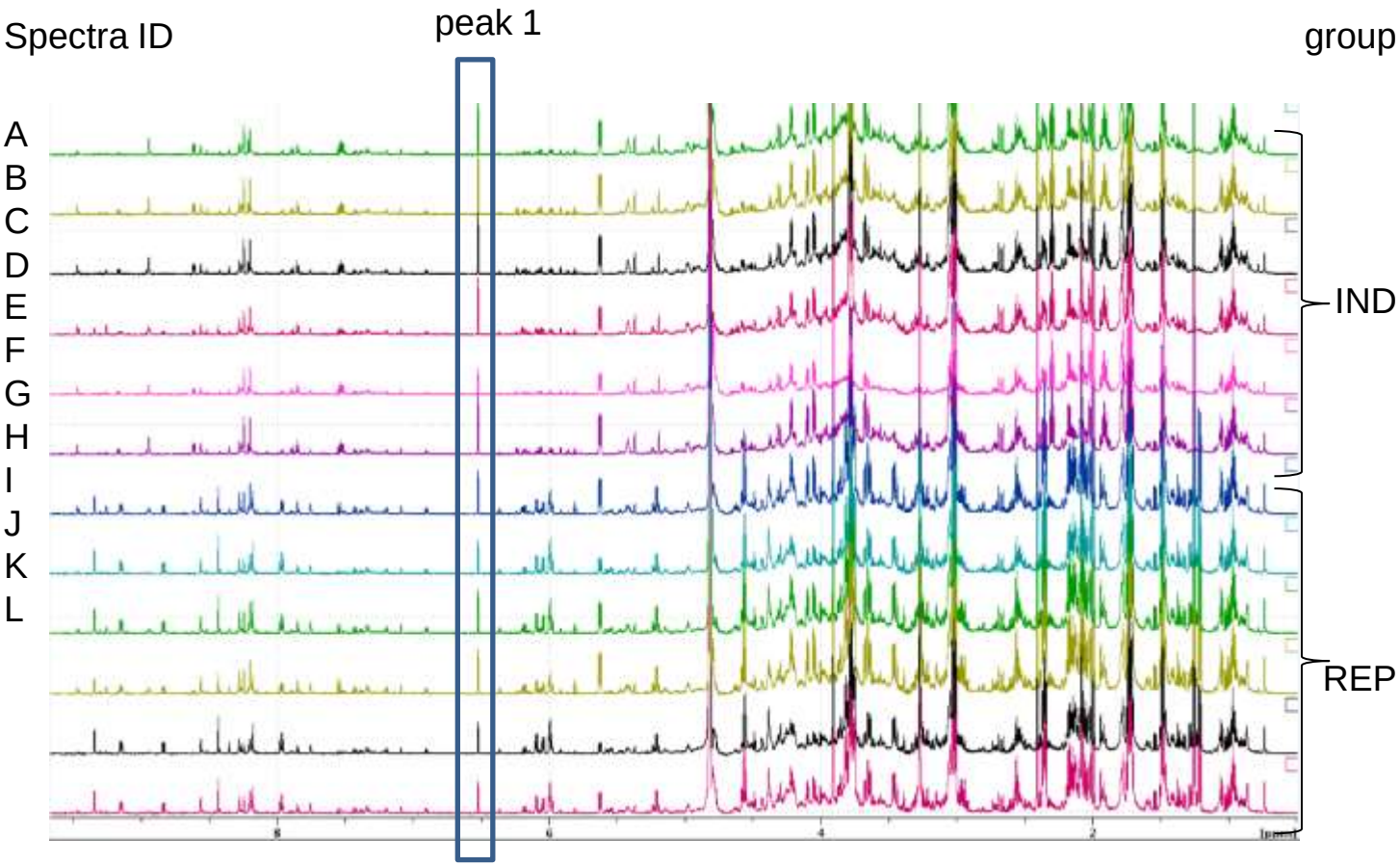
How many groups/cohorts/classes of sample?

1 group
One-sample t-test

2 groups
Two-sample t-test

3+ groups
ANOVA

Simple Scenario – what are the parameters?



N = number of individuals in dataset = 12

n = Number of individuals per group = 6

Groups = 2

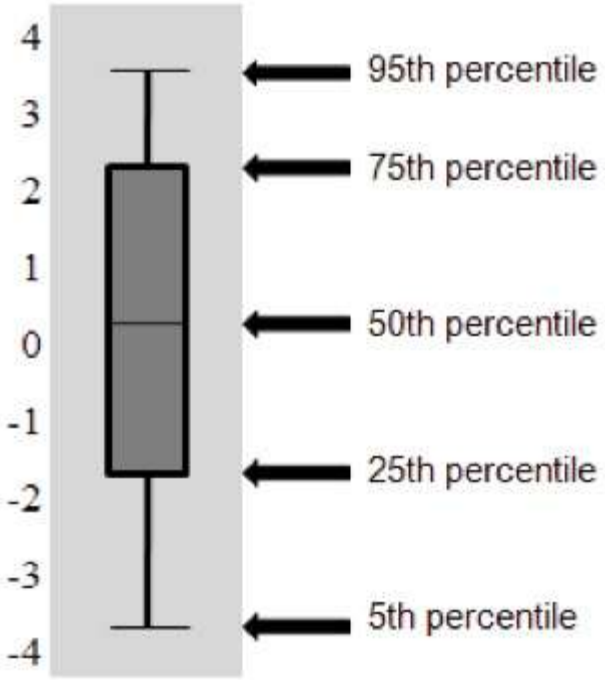
Variables = peak 1 intensity and groups

- Each NMR spectrum is chopped into peak intensities.
- Multiple spectra make up a sample **group**.
- Multiple **groups** make up a **dataset**
- 1 spectrum is one patient or experiment with multiple peaks.
- To begin with consider only one peak (peak 1)

How can we visualise the distribution of peak intensities

Box plots

- Box-plot (or box-and-whisker plot)
- **box** covers the inter-quartile range
- **whiskers** (typically) indicate 5-95 %
- Outliers are indicated separately
- Central bar = position of the median
- Useful for comparing the characteristic of different samples.

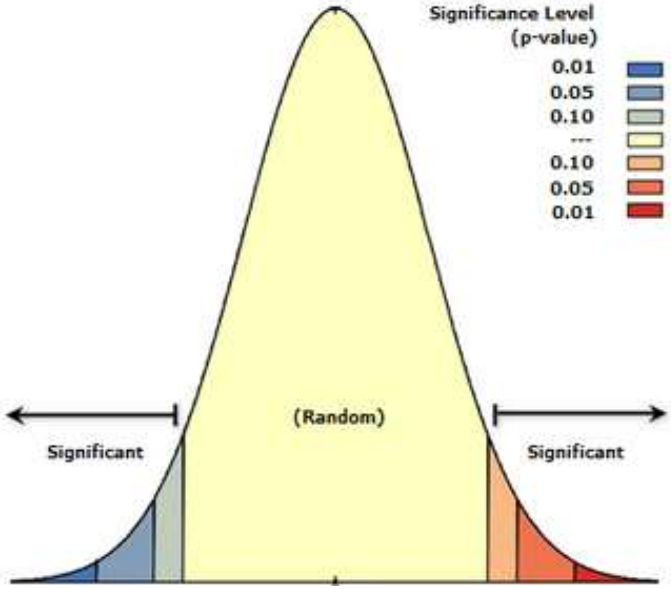


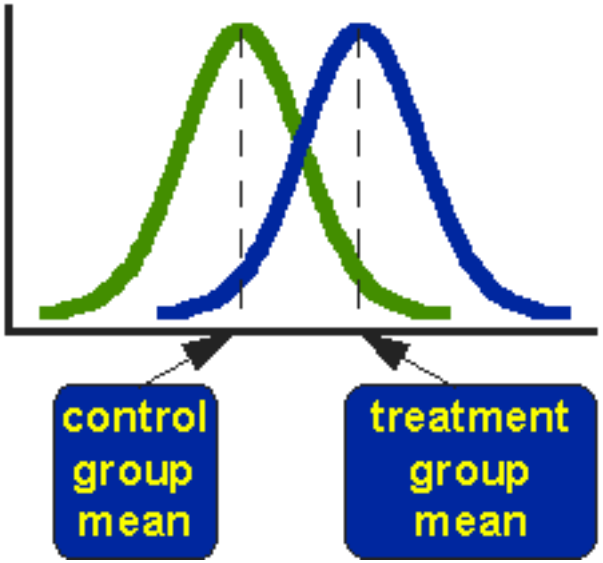
- All peak intensities are **different** between individual datapoints
- How can we tell what differences are **Statistically Significant**?

Significance Testing

Testing a null Hypothesis: Between two groups of spectra the metabolite concentrations are not different:

$mean_{group1}=mean_{group2}$





- Distribution of concentrations of metabolites observed by NMR and MS tend to follow **Normal Distribution**
- Two-group analysis can be performed on any metabolite peak using **student's t-test** (or a suitable variation)
- For identification of significant differences comparing multiple **t-tests adjustment** must be made to the ***P value***

Sources of errors in hypothesis tests

Errors

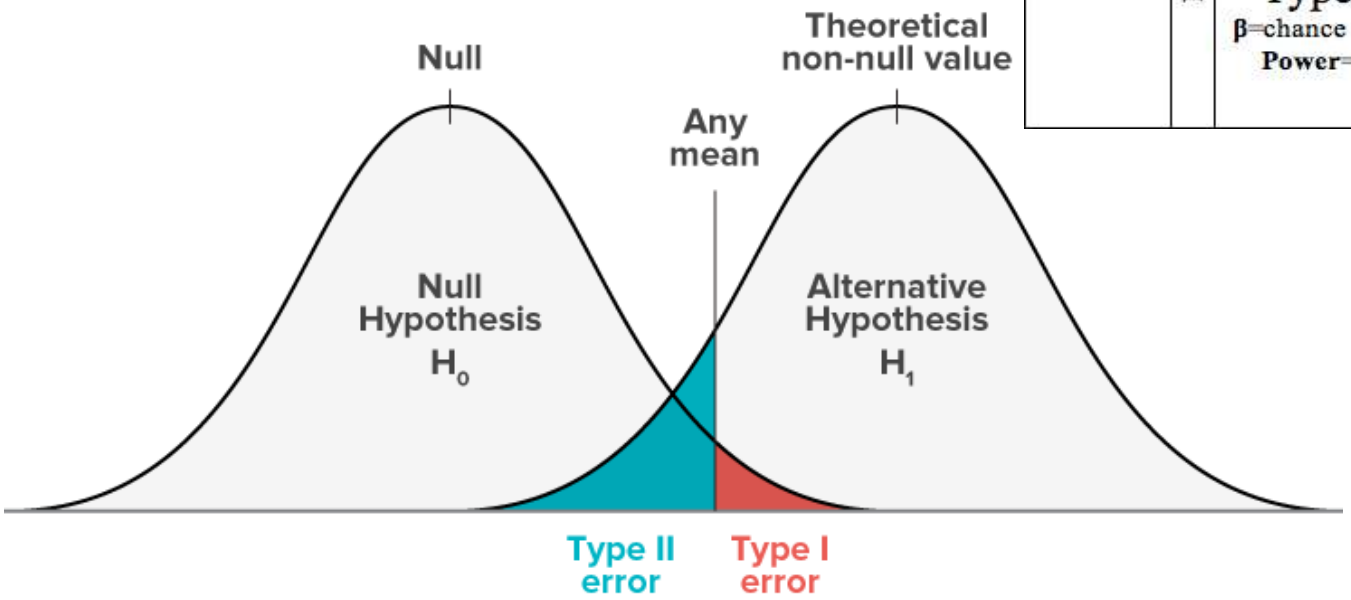
2 types of error

Type I error rate (α) is usually fixed at 0.05.

This means that in the long run we accept that 5% of tests will be false.

In reality this is a gross underestimation

		The "Truth" (Real Difference): Yes	The "Truth" (Real Difference): No
Study Findings (Significant difference)	Yes	(A) √	(B) Type I Error **p**=likelihood result occurred by chance **α**=p-value cutoff for significance
	No	(C) Type II Error **β**=chance of type II error **Power**=(1−β)•100%	(D) √



Paired data

Paired data

Often found in samples collected **before** and **after** an event such as:

- treatment
- delay (or incubation period)

Could also be **matched individuals** with (near) identical:

- age
- genetic background
- environment
- disease severity

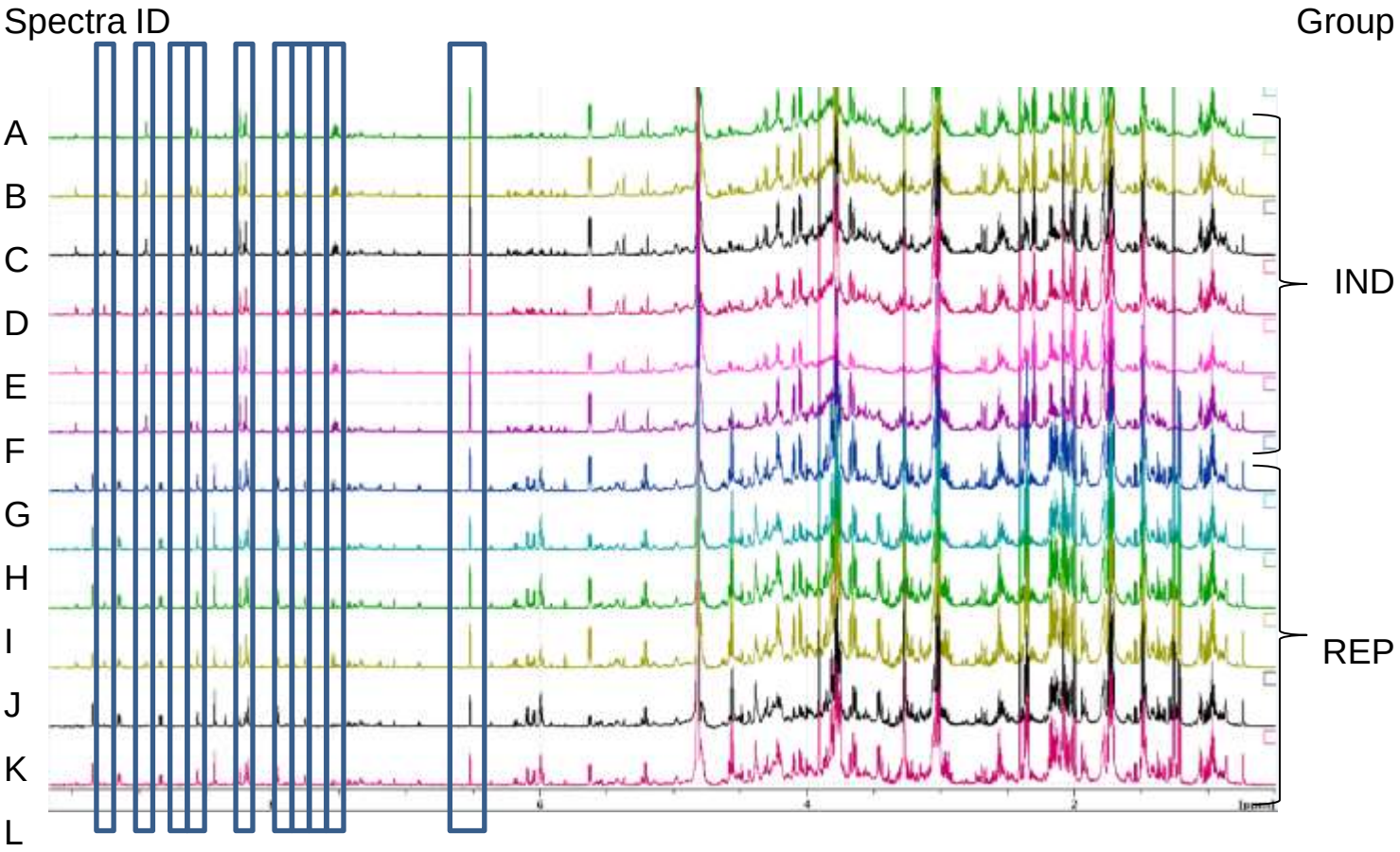
Why use paired samples?

Reduces variability caused by effects incidental to the study

Therefore reduces the signal-to-noise

Effectively enhances statistical power in small dataset

More realistic scenario – what are the parameters?



N = number of samples in dataset = 12
n = Number of samples per group = 6
Groups = 2
Variables = multiple peaks (which we shall consider individually) and group

- Each NMR spectrum is chopped into peak intensities.
- Multiple spectra make up a sample group.
- Multiple groups make up a dataset.
- 1 spectrum is one patient or experiment with multiple peaks.
- Multiple peaks?

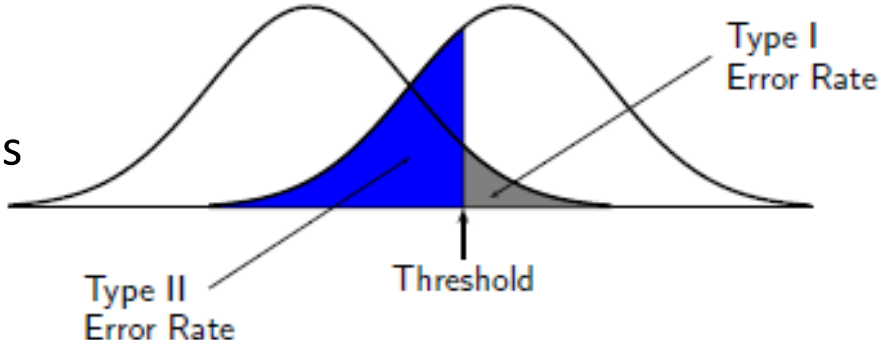
For multiple tests we need to adjust the P value

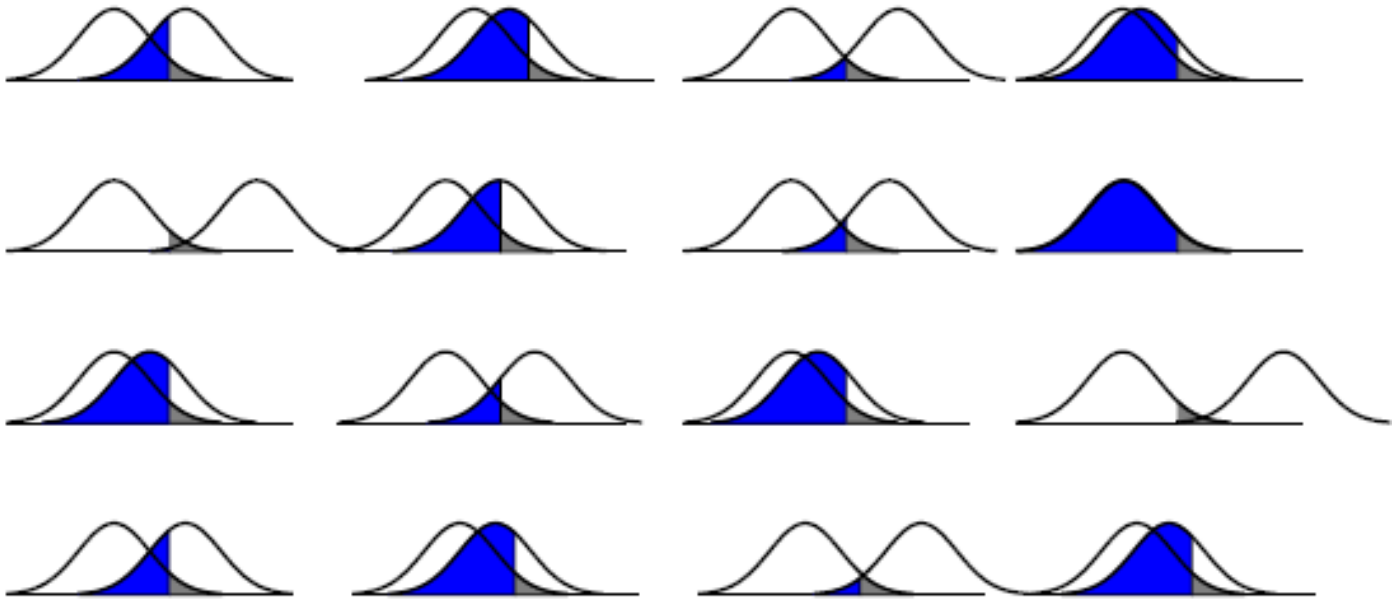
P value is essentially a rejection threshold

this prevents **false positives**

i.e. an insignificant variance being marked as significant

This is known as a Type I error





Multiple T-tests the probability of a **type I** (false positive) increases with the number of tests! When performing multiple tests (e.g. 16) with a fixed threshold per test of 0.05 the probability for *at least* one of the tests to be a type I error:

$$1 - (1 - 0.05)^{16} = 0.56$$

P adjustment methods – which to choose?

Choose a procedure that balances the competing demands of sensitivity and specificity.

Bonferroni

Control Family Wise Type I Error (FWER)

FWER the probability of at least one type I error

Benjamini-Hochberg (BH) 1995

Control false discovery rate (FDR)

FDR the expected proportion of type I errors among H0 rej

Bonferroni

gives fewer Type I errors - performs well in sparse cases (T0 ~ m)

However

Bonferroni over controls FDR and will not in general minimise FNR

in non-sparse cases power can be improved by other methods

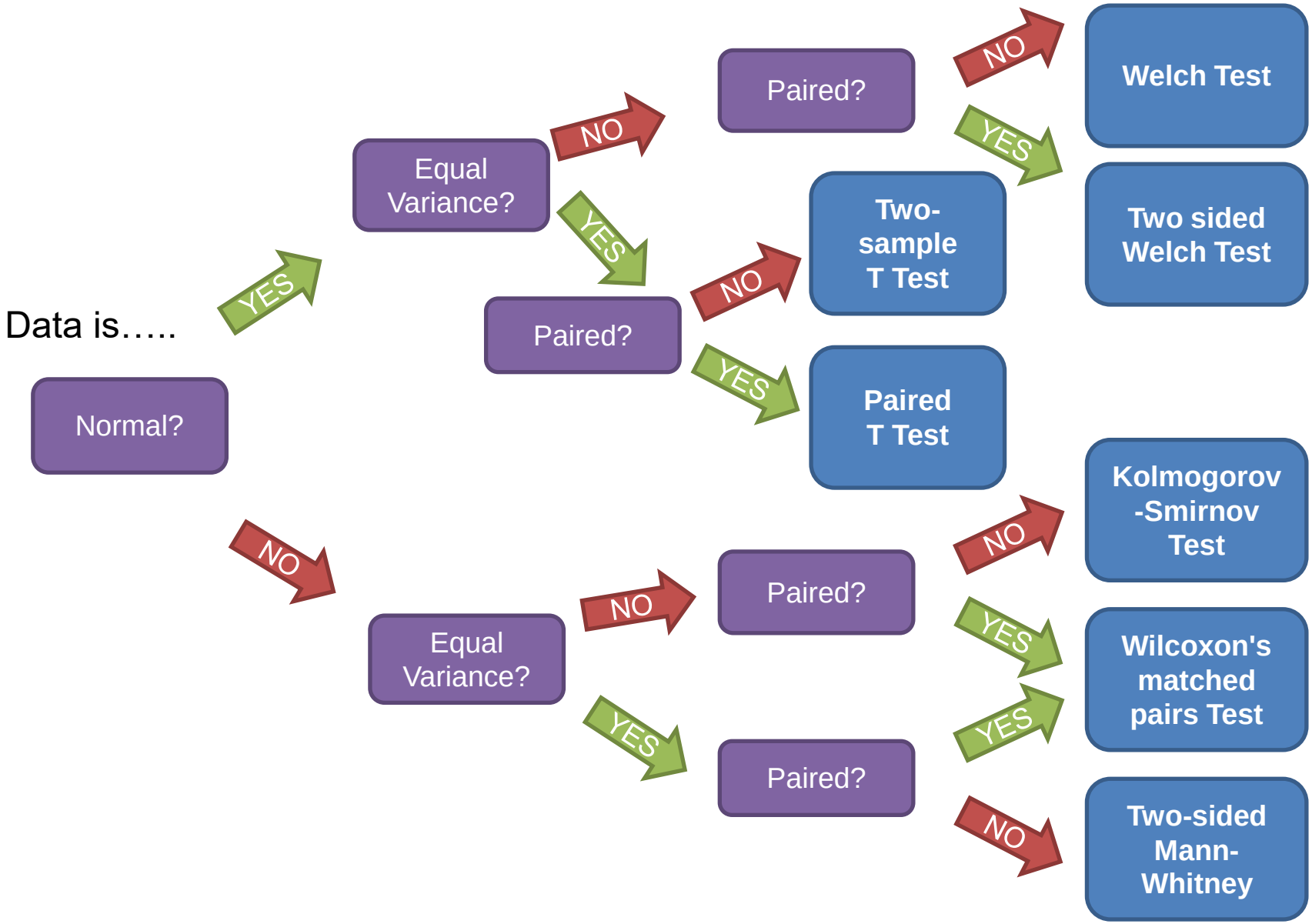
	H0 ret	H0 rej	Total
H0 true	TN	FD	T0
H0 false	FN	TD	T1
Total	N	D	m

T/F = True/False

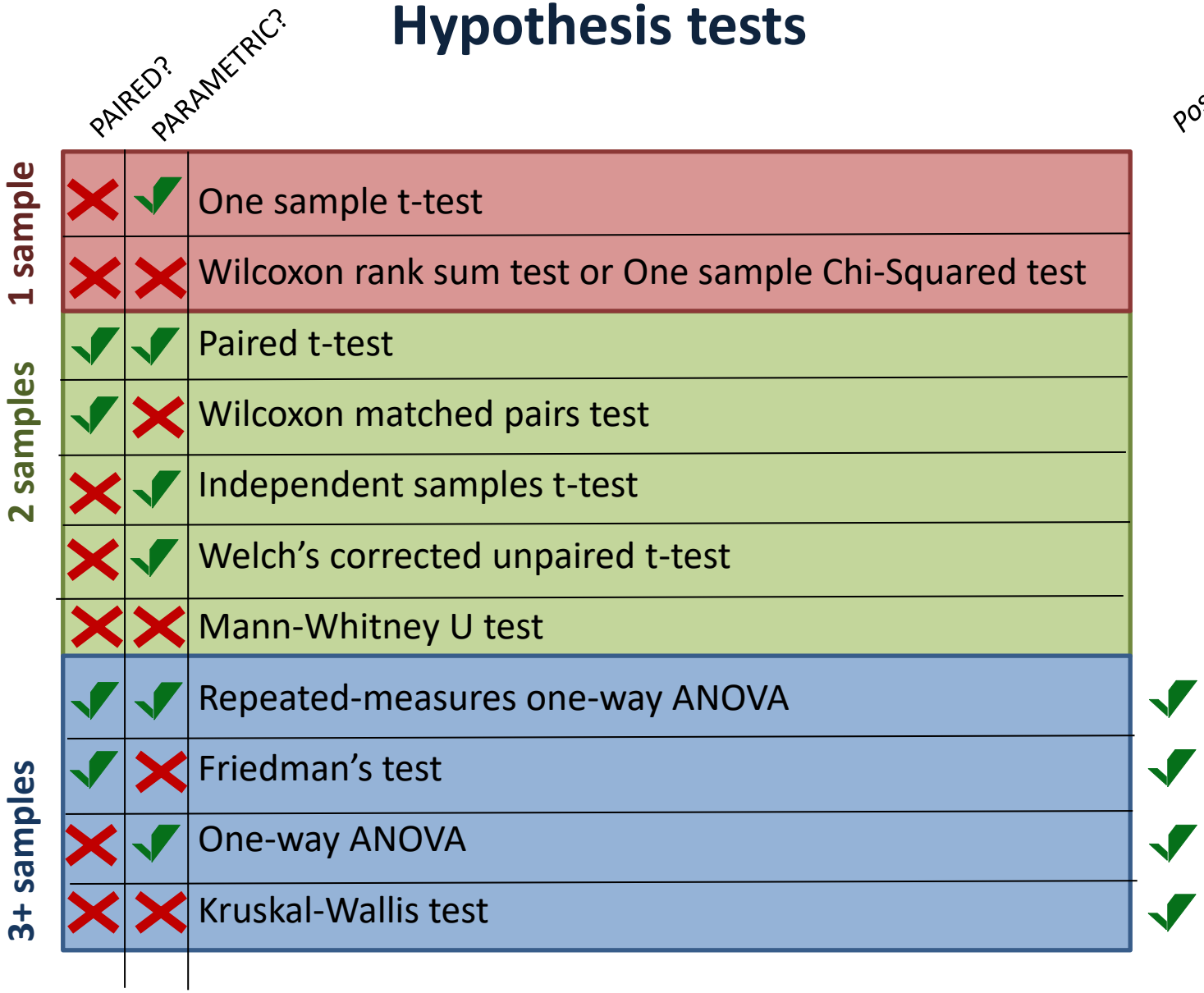
D/N = Discovery/Nondiscovery

Retained/rejected

Which hypothesis test?



Hypothesis tests



M Marusteri, V Bacarea. Comparing groups for statistical differences: how to choose the right statistical test? Biochemia Medica 2010;20(1):15-32.

Significance tests

Multiple comparisons: ANalysis Of VAriance (ANOVA):

1. Compares one signal/bucket/metabolite across treatments/conditions
2. Hypothesis:
   H0 hypothesis: no difference between groups / all groups are from the same population i.e. treatment has no effect
   H1 hypothesis: at least one group is different from the rest
3. Requirements: Replicates, independent observations, normal data.

Significance tests are used to:

- Check data integrity - Quality control (when performed between replicates)
- To identify buckets that change across treatments. A significant ANOVA result on a bucket indicates **at least one** of the treatments are affecting that bucket. To determine **which** treatment(s) is significantly different *post-hoc* analysis can reveal this information.
- If your data can be classified into subsets (different strains, different sex, different age, etc.) then ANOVA can also be used to test whether there is an overall difference between these blocks for a given metabolite.

Post-hoc analyses are needed to adjust the p-values to reduce the false discovery rate, some also test within significant groups which comparisons are responsible for the significance

Multivariate Analysis

Unsupervised:

- Unsupervised techniques use **no information** about the **groupings** in order to transform the data
- The information is effectively ***compressed***, reducing the number of variables in the data without losing much information.
- Popular example of an **unsupervised multivariate** analysis is **Principal Component Analysis (PCA)**

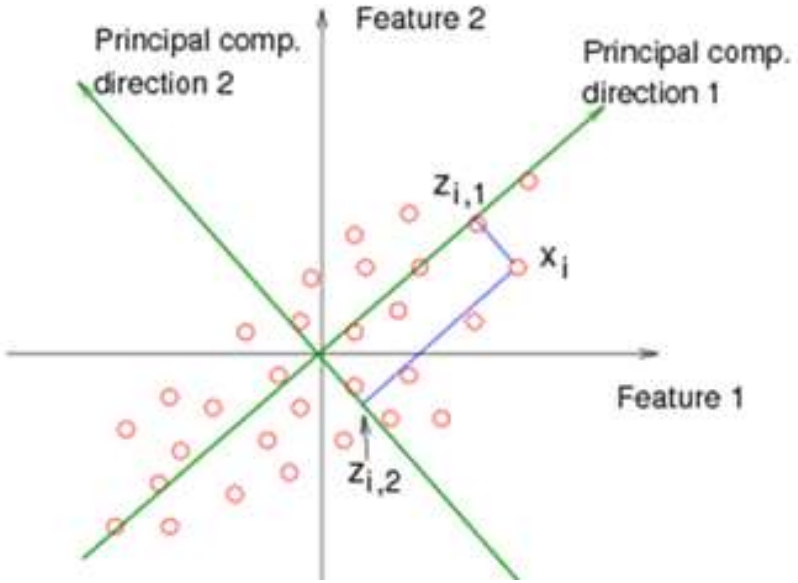
Discriminant/Supervised:

- Supervised techniques **do** use **information** about the **groupings** in order to transform the data.
- Essentially discriminant analysis **suppress** variance **within group** and **enhance** variance **between groups.**
- Popular example of a **supervised multivariate** analysis is **Partial Least Squares Discriminant Analysis (PLS-DA)**

Principal Component Analysis - PCA

PCA:

- PCA is an unsupervised data transformation that produces a set of uncorrelated variables called **principal components** (PCs)
- Unsupervised techniques use **no information** about the **groupings** in order to transform the data
- The first PC captures the maximum amount of variance in the data
- The second - the maximum possible amount of the remaining variance, and so on.
- The information is effectively ***compressed***, reducing the number of variables in the data without losing much information.
- **Score plots** are used to assess the data structure of the PCs
- The data is transformed from a coordinate system of **metabolites/buckets** into a new coordinate system of **PCs**



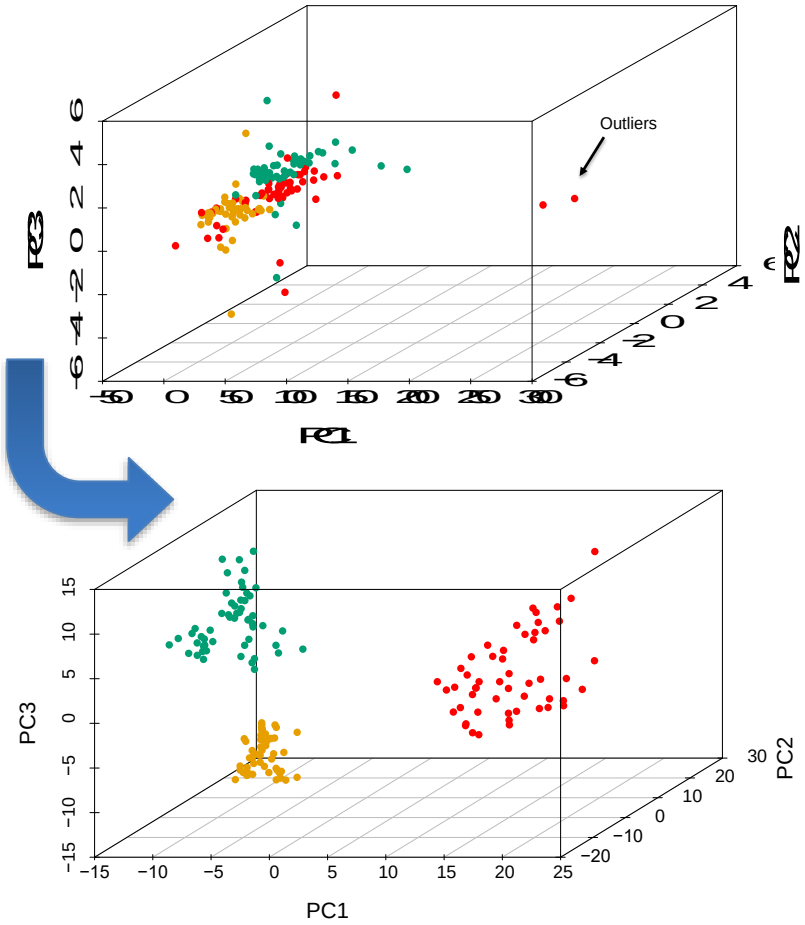
Used for:

- Visualization of structure within data
- Reducing number of variables for building more robust models

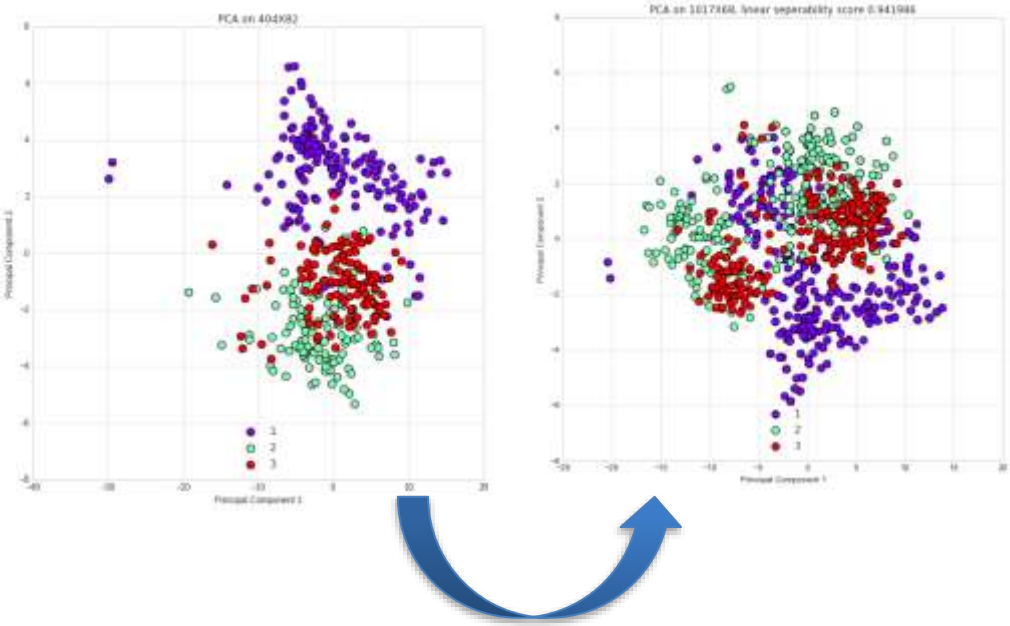
Principal Component Analysis - PCA

Can be used to detect and correct for:

Outliers

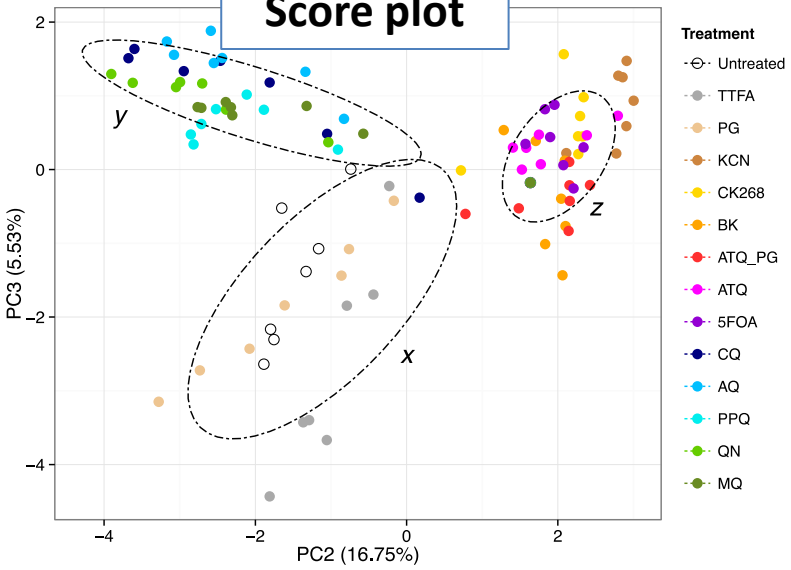


Batch effect

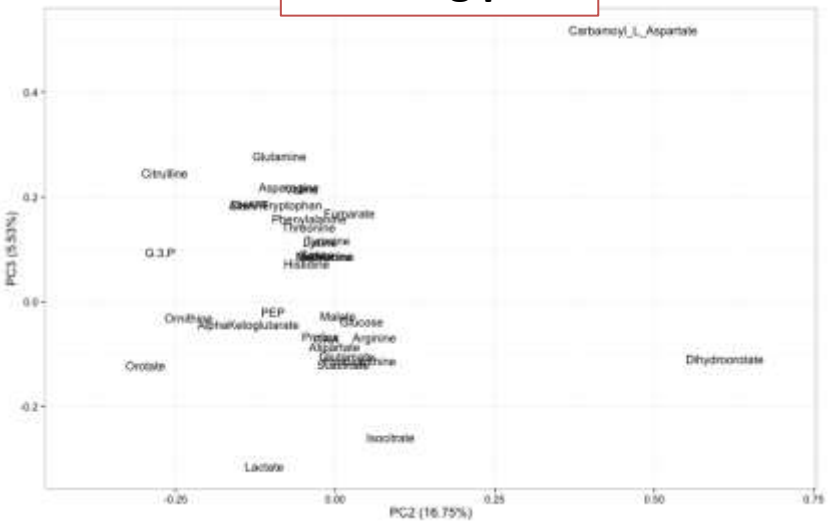


Output of Principal Component Analysis - PCA

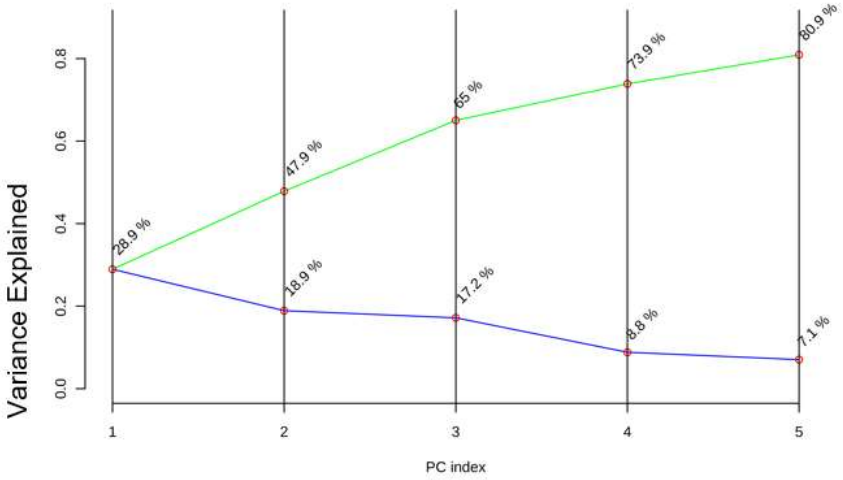
Score plot



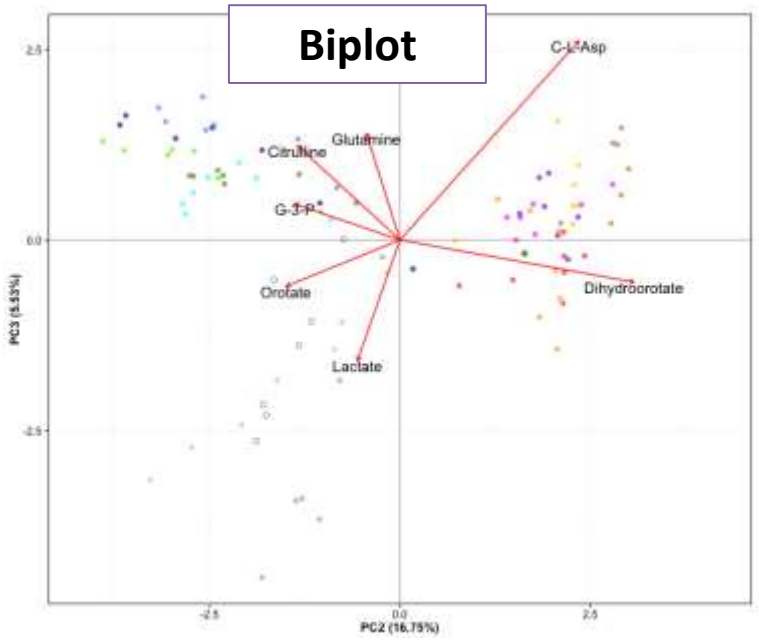
Loading plot



Scree

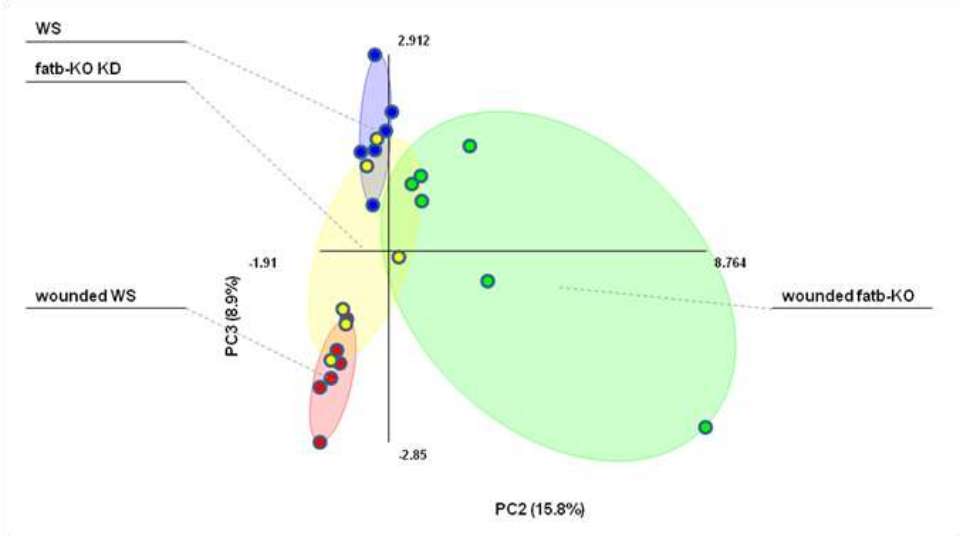


Biplot

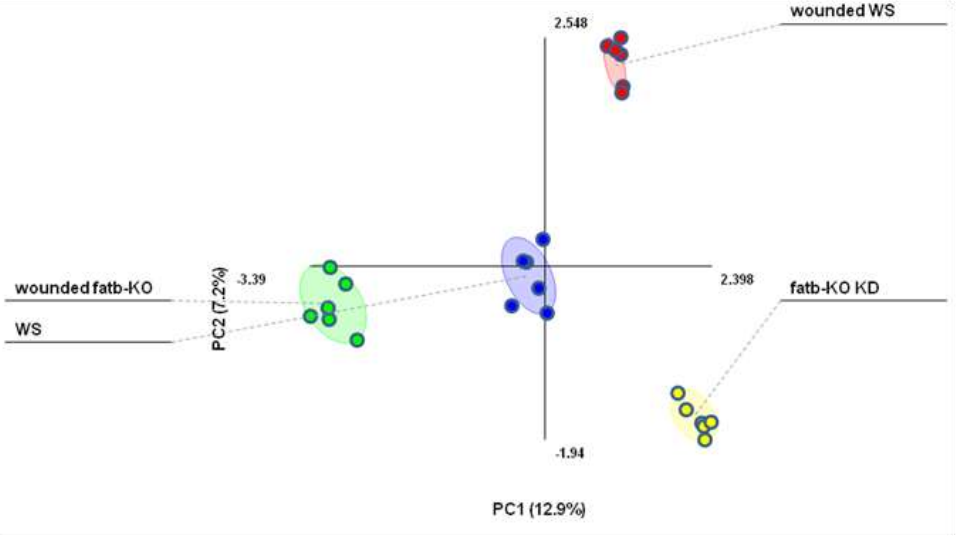


Partial Least Squares – Discriminant Analysis (PLS-DA)

- **Discriminant Analysis** techniques are supervised models
- Supervised classification requires grouping information prior to model building
- The resulting model **maximizes** the effects of metabolites giving **variance *between*** the **groups**…
- ….and **reduces** the **variation** found ***within*** each **group**.
- The output is a model with predictive capability.
- With all supervised models it is possible to **overfit** the data.
- Fitting the model to too many components will lead to over-fitting and consequently meaningless results.

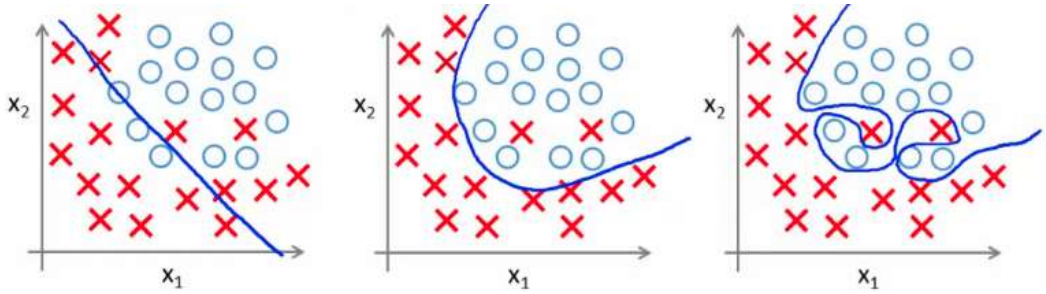


PCA

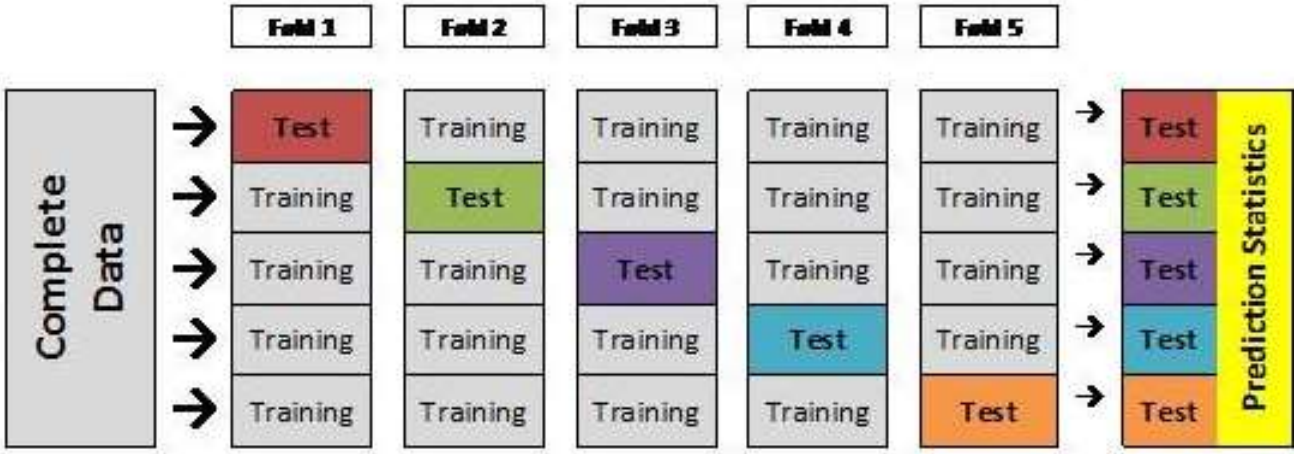


PLS-DA

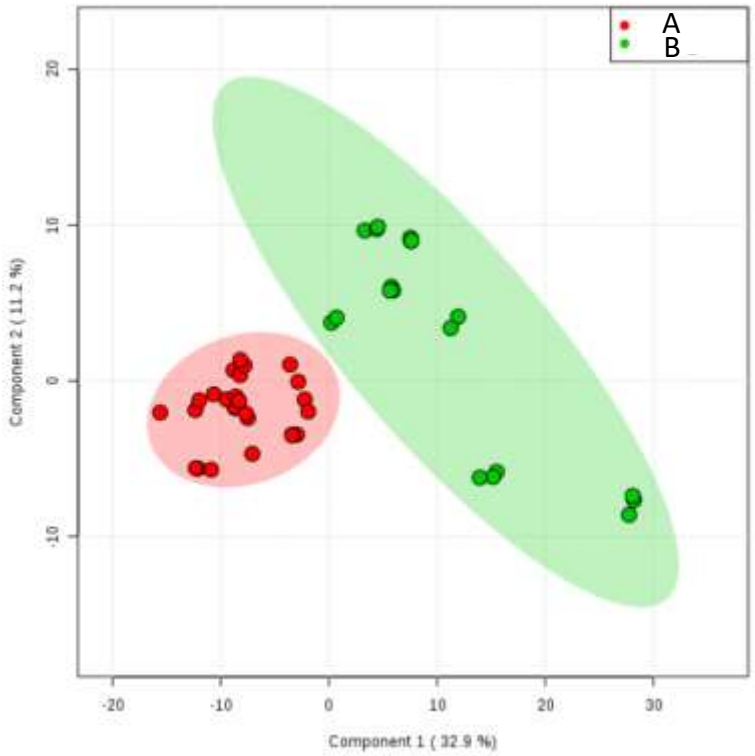
Over-fitting:

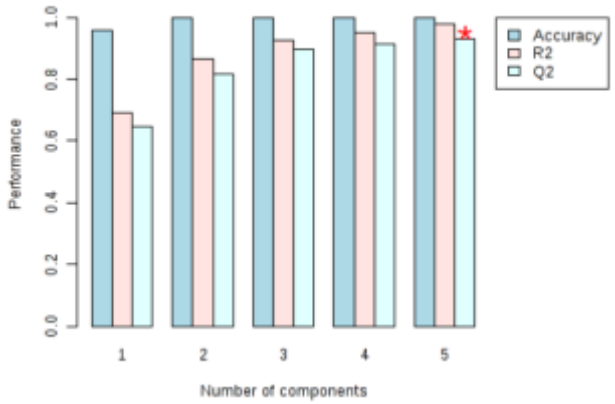


- **Cross-validation** to avoid over-fitting is usually performed by splitting the data into *n* subsets and building a model on the data **leaving one subset out**.
- That subset is used to **test the model**.
- The process is **repeated** until each subset have served as the test set.
- The test results are used to assess the model **accuracy** and **robustness**.
- Cross validation is tried using varying number of **components**
- An **appropriate** number of components for the model is selected based on the model accuracy across the subsets.

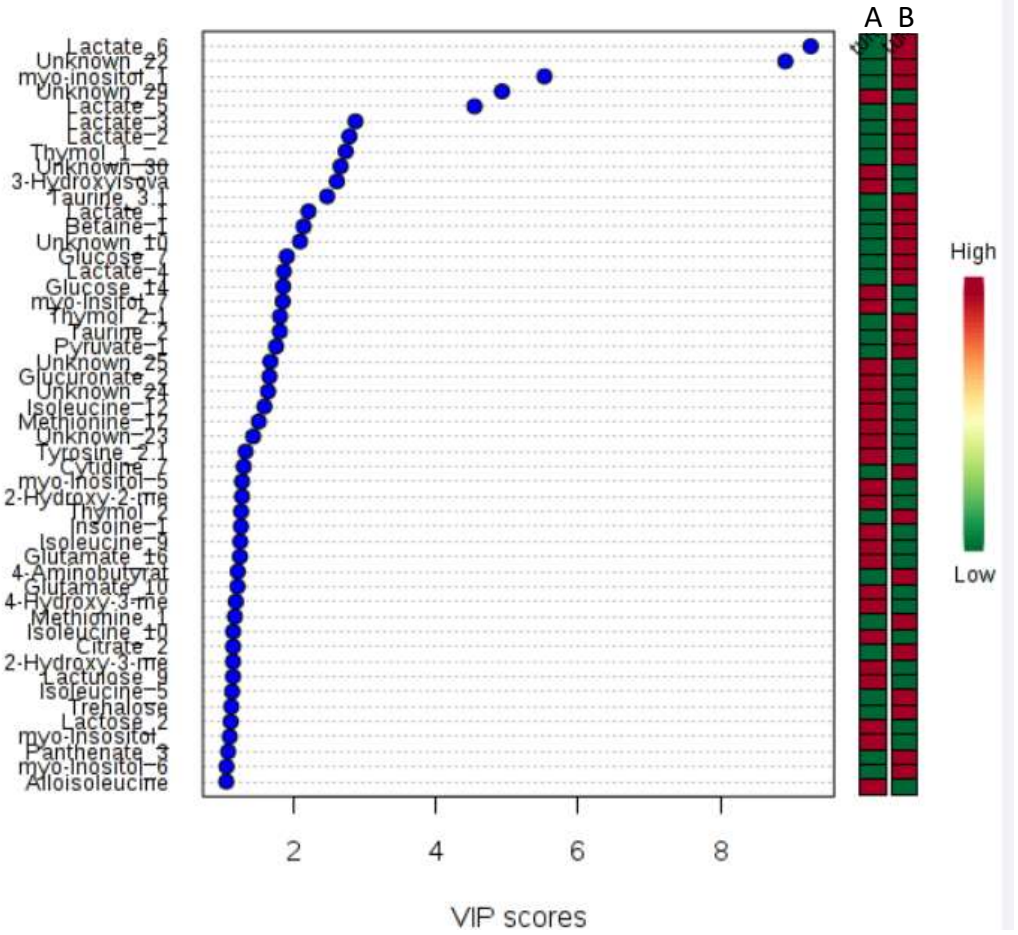


Output of the PLS-DA





Variable Importance in Projection (VIP)



R^2 corresponds to the sum of squares captured by the model

Q^2 is the cross-validated R^2

Multivariate Analysis – which to use?

Limitations:

Advantages:

Unsupervised:

- Does not necessarily report about groupings.
- Will highlight contamination or batch effects.
- Not considered a true statistical test – no true performance values.

- Unbiased.
- Reports on greatest variance between all samples.
- Useful for identifying batch effects.
- Useful for appraising technique.
- No lower limit on samples.

Discriminant/Supervised:

- Biased.
- Will model different groups – even if there is no 'real' difference.
- Requires many samples to test properly (cross-validation etc.).

- Reports variance that defines groupings.
- Will ignore/reduce 'unwanted' variance contamination or batch effects.
- Measures true performance values.

Can you identify significant Spectral differences?

No:

- Check:

  - experimental design

  - Statistical analysis

  - Number of samples?

  - Methods of metabolite measurement?

  - Amount of material? (signal/noise)

Yes:

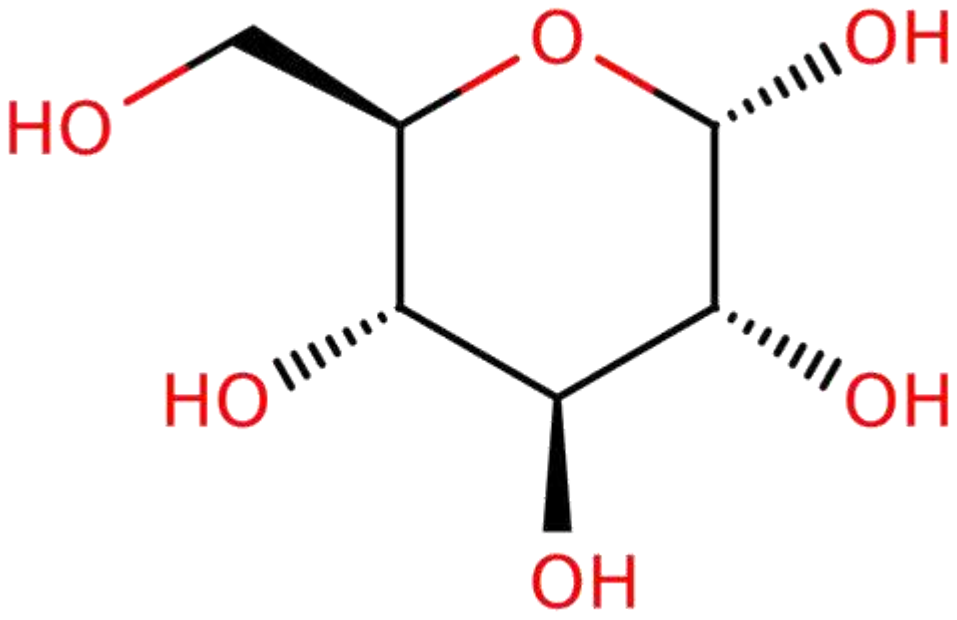
- Are these **effects** you are looking for?

- Do they relate to **sample prep** or **conditions tested**

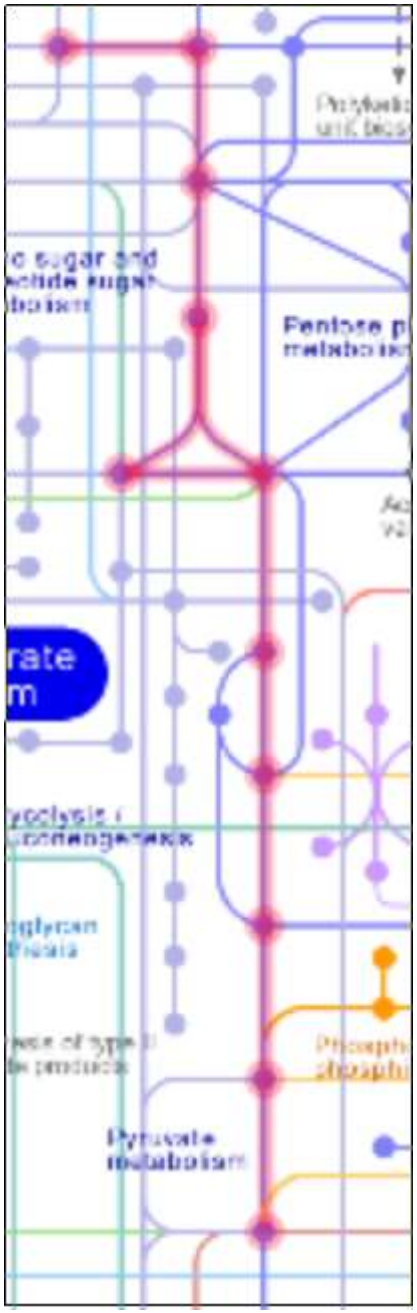
- Can they be **biologically contextualised**?

From Metabolite to Biological Pathways

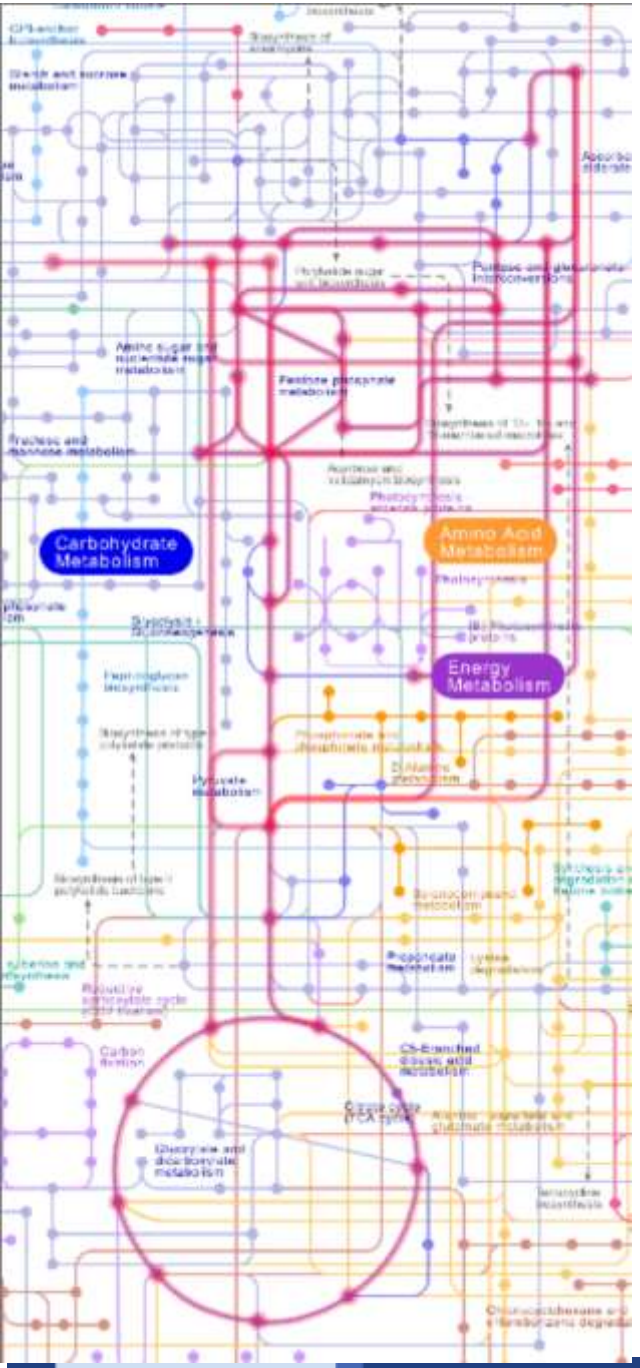
First let us consider the complexity of Metabolic Pathways:



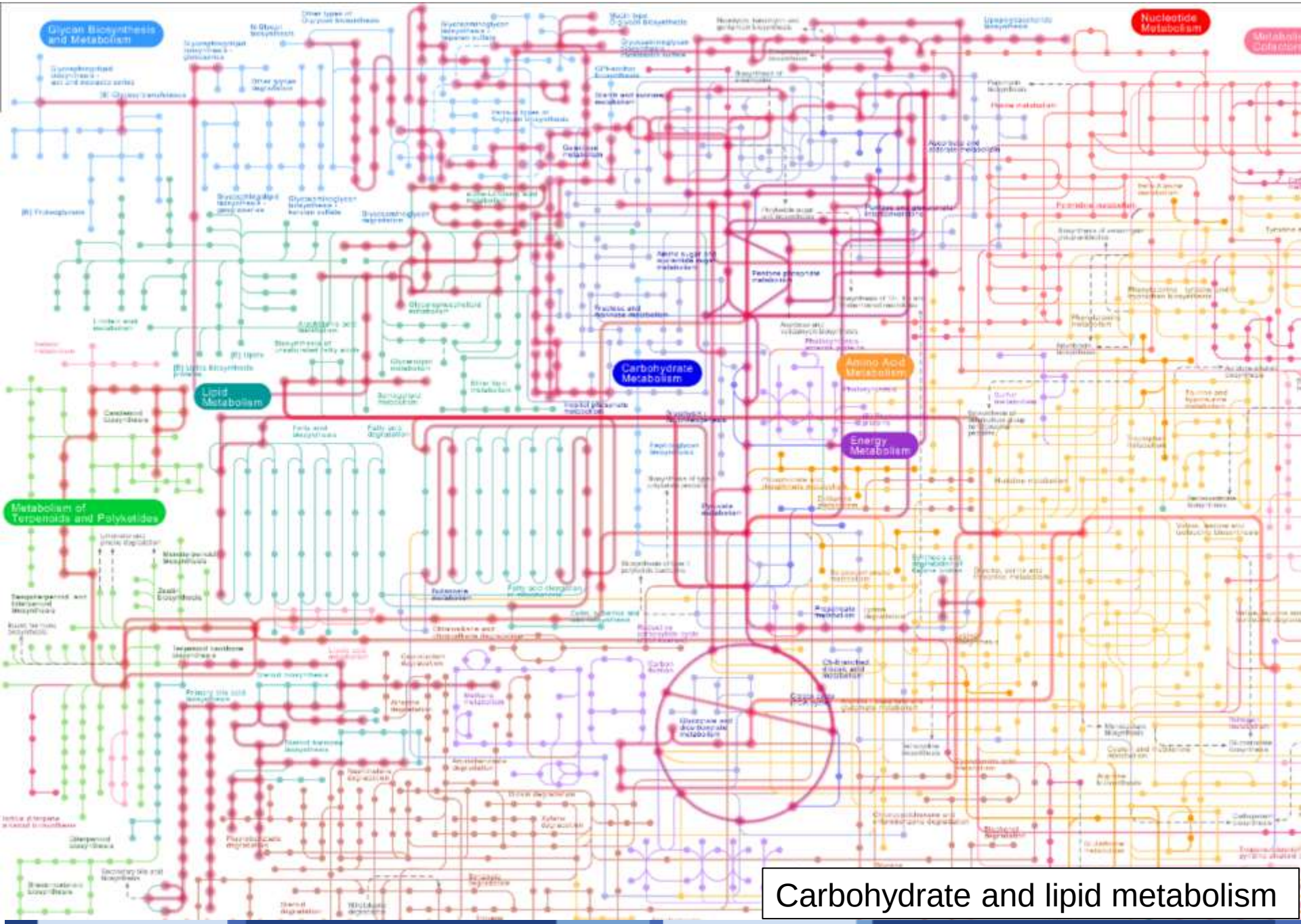
α-D-Glucose



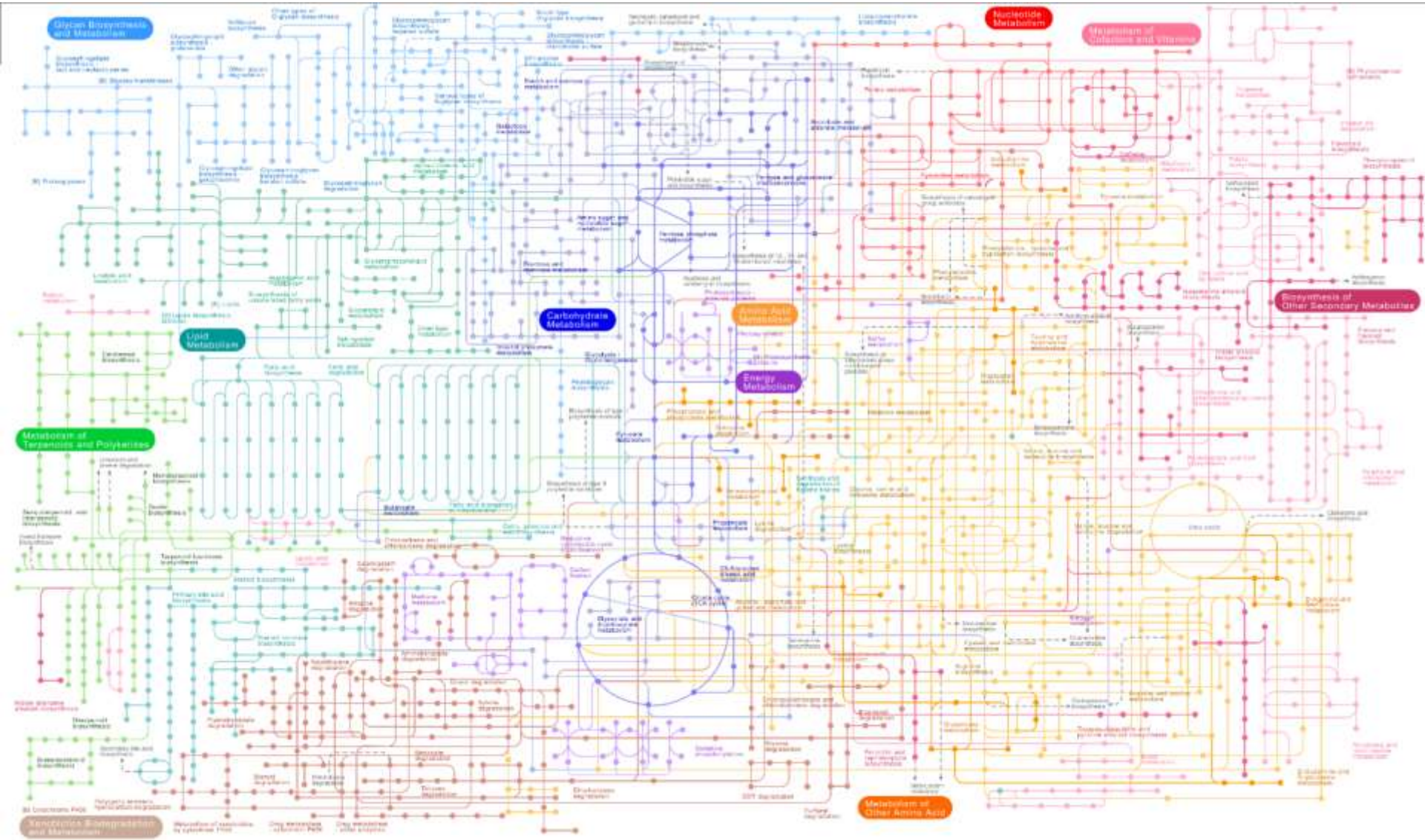
Glycolysis (Embden-Myerhof Pathway) Module



Central carbohydrate metabolism



Carbohydrate and lipid metabolism



All pathways available on KEGG

Software & Databases

Software:

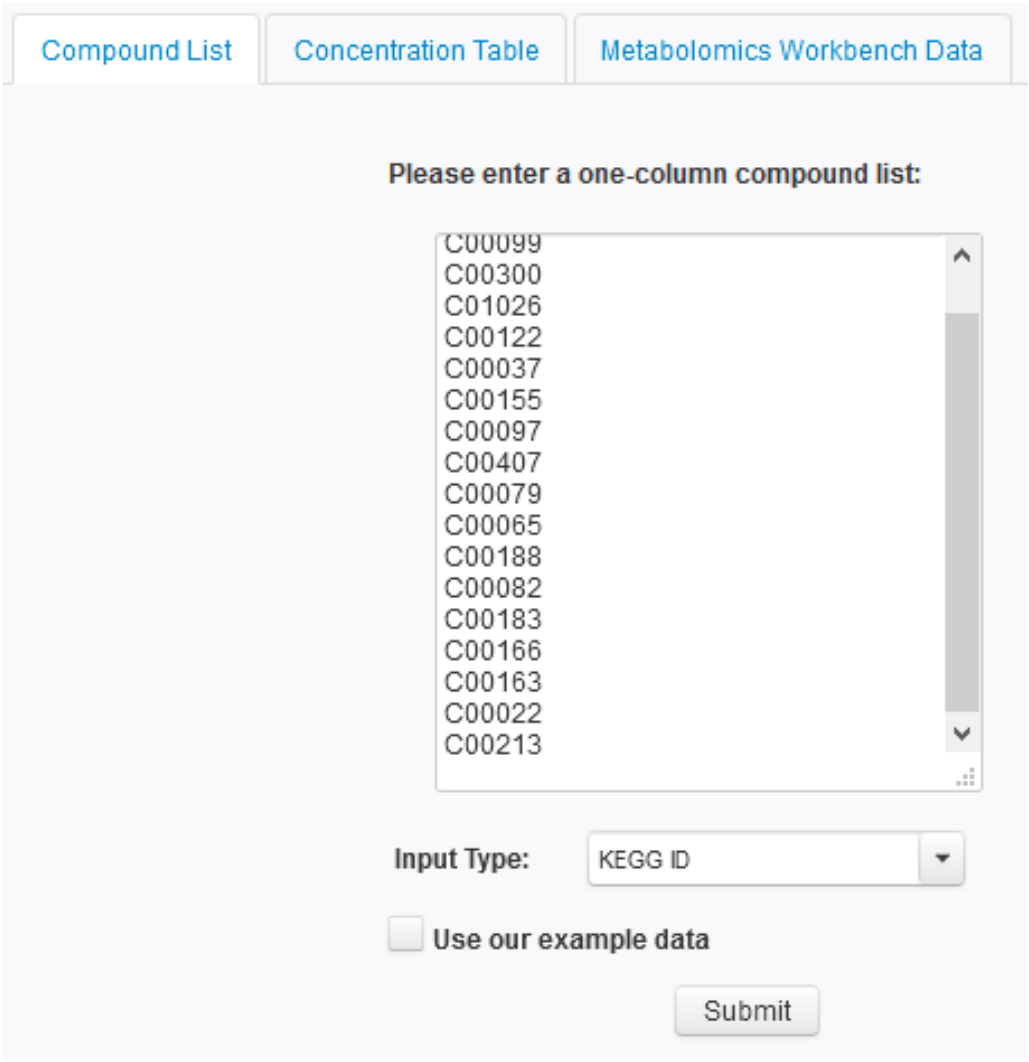
- MetaboAnalyst
  - Freeware
  - Common pathways
  - Common organisms
  - Metabolite only
  - Not always up-to-date

Databases:

- KEGG
  - Metabolite, protein and genetic data
  - API access
  - Required license for FTP access
- HMDB
  - Mammalian Metabolites
  - Synonyms datasets and chemical information
  - Curated literature - associated diseases

MetaboAnalyst

Input unique Identifiers:



Analysis (targeted)

pathway analysis integrates:

- enrichment analysis (MSEA)
- pathway topology analysis (impact)

Visualization for 26 model organisms:

- Human,
- Mouse,
- Rat,
- Cow,
- Chicken,
- Zebrafish,
- *Arabidopsis thaliana*,
- Rice,
- Drosophila,
- Malaria,
- *S. cerevisae*,
- *E.coli*,
- other species…

MetaboAnalyst

Pathways covered by:
Pathway Analysis (targeted)

1. Select pathway analysis parameters

Specify pathway analysis parameters:

Visualization method	(•) Scatter plot (testing significant features) ( ) Heatmaps (testing your selected features)
Enrichment method	(•) Hypergeometric Test ( ) Fisher's Exact Test
Topology analysis	(•) Relative-betweeness Centrality ( ) Out-degree Centrality
Reference metabolome	(•) Use all compounds in the selected pathway library ( ) Upload your own reference metabolome

2. Select pathway library
(Limited number of Organisms)

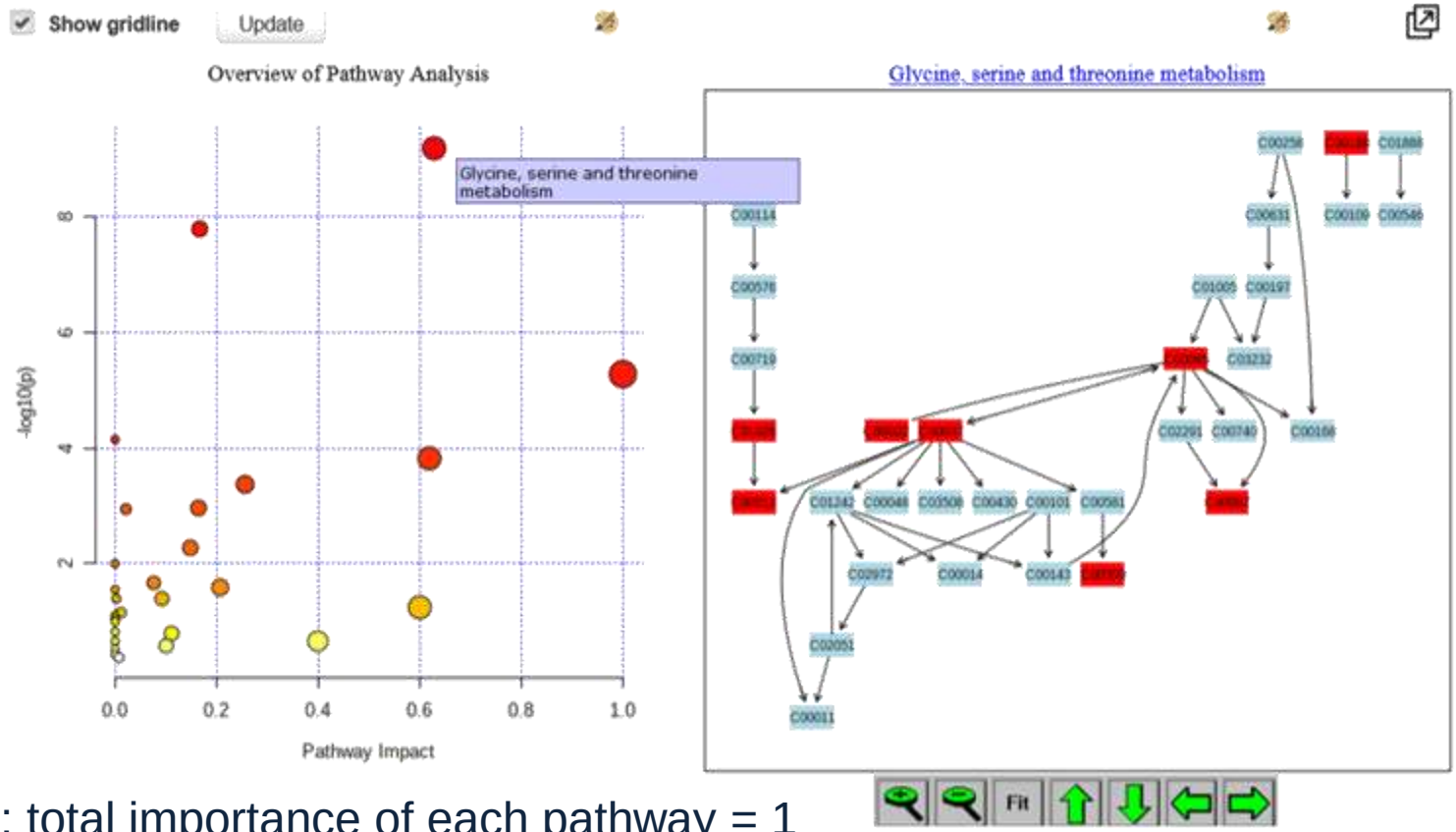
Select a pathway library: *(KEGG pathway info were obtained in Oct. 2019)*

Mammals	(•) Homo sapiens (KEGG) ( ) Homo sapiens (SMPDB) ( ) Mus musculus (KEGG) ( ) Mus musculus (SMPDB) ( ) Rattus norvegicus (rat) (KEGG) ( ) Bos taurus (cow) (KEGG)
Birds	( ) Gallus gallus (chicken) (KEGG)
Fish	( ) Danio rerio (zebrafish) (KEGG)
Insects	( ) Drosophila melanogaster (fruit fly) (KEGG)
Nematodes	( ) Caenorhabditis elegans (nematode) (KEGG)
Fungi	( ) Saccharomyces cerevisiae (yeast) (KEGG)
Plants	( ) Oryza sativa japonica (Japanese rice) (KEGG) ( ) Arabidopsis thaliana (thale cress) (KEGG) ( ) Chlorella variabilis (green alga) (KEGG)
Parasites	( ) Schistosoma mansoni (KEGG) ( ) Plasmodium falciparum 3D7 (Malaria) (KEGG) ( ) Plasmodium vivax (Malaria) (KEGG) ( ) Trypanosoma brucei (KEGG)
Prokaryotes	( ) Escherichia coli K-12 MG1655 (KEGG) ( ) Bacillus subtilis (KEGG) ( ) Pseudomonas putida KT2440 (KEGG) ( ) Staphylococcus aureus N315 (MRSA/VSSA) (KEGG) ( ) Thermotoga maritima (KEGG) ( ) Synechococcus elongatus PCC7942 (KEGG) ( ) Mesorhizobium japonicum MAFF 303099 (KEGG) ( ) Klebsiella pneumoniae MGH 78578 (serotype K52) (KEGG) ( ) Klebsiella varicola At-22 (KEGG) ( ) Streptococcus pyogenes M1 476 (serotype M1) (KEGG)

Submit

MetaboAnalyst

Output offers ‘impact’ and p-value (presented as –log10) :

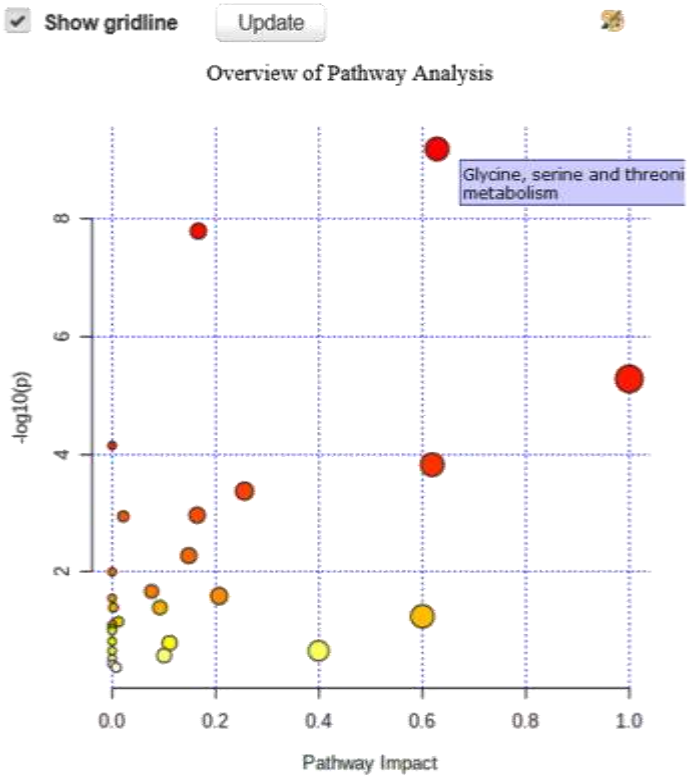


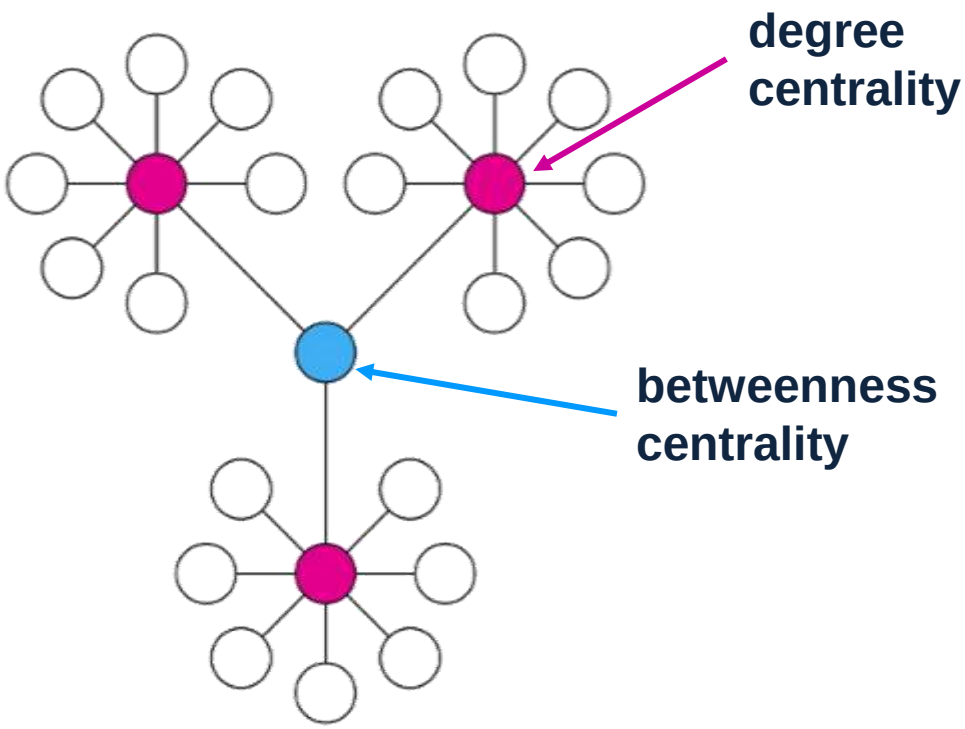
Impact: total importance of each pathway = 1
Each metabolite node is the **%** with respect to the total pathway importance
pathway impact is the cumulative **%** from the matched metabolite nodes.

Node importance estimation: **betweenness centrality** & **degree centrality**

MetaboAnalyst

Output offers ‘impact’ and p-value (presented as –log10) :





Impact: total importance of each pathway = 1
Each metabolite **node** is the **%** with respect to the total pathway importance
pathway impact is the cumulative **%** from the matched metabolite **nodes**.

Node importance estimation: **betweenness centrality** & **degree centrality**

MetaboAnalyst

Result is a list of metabolic pathways and associated p-value and impact:

Click the corresponding **Pathway Name** to view its graphical presentation; click **Match Status** to view the pathway compounds (with matched ones highlighted).

1 2

Pathway Name	Match Status	p	-log(p)	Holm p	FDR	Impact	Details
Glycine, serine and threonine metabolism	8/33	6.4736E-10	9.1889	5.4378E-8	5.4378E-8	0.62837	KEGG SMP
Aminoacyl-tRNA biosynthesis	8/48	1.611E-8	7.7929	1.3371E-6	6.766E-7	0.16667	KEGG
Phenylalanine, tyrosine and tryptophan biosynthesis	3/4	5.2309E-6	5.2814	4.2893E-4	1.4647E-4	1.0	KEGG SMP
Valine, leucine and isoleucine biosynthesis	3/8	7.1126E-5	4.148	0.0057612	0.0014936	0.0	KEGG SMP
Phenylalanine metabolism	3/10	1.502E-4	3.8233	0.012016	0.0025234	0.61904	KEGG SMP
Cysteine and methionine metabolism	4/33	4.2302E-4	3.3736	0.033418	0.0059222	0.25594	KEGG SMP SMP
Tyrosine metabolism	4/42	0.0010833	2.9653	0.084496	0.011925	0.16435	KEGG SMP SMP
Pantothenate and CoA biosynthesis	3/19	0.0011358	2.9447	0.087454	0.011925	0.02143	KEGG SMP
Glyoxylate and dicarboxylate metabolism	3/32	0.0052872	2.2768	0.40183	0.049347	0.14815	KEGG
Valine, leucine and isoleucine degradation	3/40	0.0099344	2.0029	0.74508	0.083449	0.0	KEGG SMP
Citrate cycle (TCA cycle)	2/20	0.021391	1.6698	1.0	0.16335	0.07615	KEGG SMP
Pyruvate metabolism	2/22	0.025652	1.5909	1.0	0.17956	0.20684	KEGG SMP
Propanoate metabolism	2/23	0.027903	1.5543	1.0	0.1803	0.0	KEGG SMP
Alanine, aspartate and glutamate metabolism	2/28	0.040286	1.3948	1.0	0.2256	0.0024	KEGG SMP SMP SMP
Glutathione metabolism	2/28	0.040286	1.3948	1.0	0.2256	0.09216	KEGG SMP
Synthesis and degradation of ketone bodies	1/5	0.056803	1.2456	1.0	0.29822	0.6	KEGG SMP
Arginine and proline metabolism	2/38	0.069981	1.155	1.0	0.34579	0.01212	KEGG SMP
Thiamine metabolism	1/7	0.07866	1.1042	1.0	0.36708	0.0	KEGG SMP
Taurine and hypotaurine metabolism	1/8	0.089408	1.0486	1.0	0.39528	0.0	KEGG SMP
Ubiquinone and other terpenoid-quinone biosynthesis	1/9	0.10004	0.99984	1.0	0.42016	0.0	KEGG SMP

1 2

Impact – based on number of nodes the metabolites have in the pathways
P-value – based on Metabolite Set Enrichment Analysis

MetaboAnalyst

Result is a list of metabolic pathways and

Click the corresponding **Pathway Name** to view its graphical presentation; click **Match Status** to view t

Pathway Name	Match Status						
Glycine, serine and threonine metabolism	8/33	6.4					
Aminoacyl-tRNA biosynthesis	8/48	1.6					
Phenylalanine, tyrosine and tryptophan biosynthesis	3/4	5.2					
Valine, leucine and isoleucine biosynthesis	3/8	7.1					
Phenylalanine metabolism	3/10	1.5					
Cysteine and methionine metabolism	4/33	4.2					
Tyrosine metabolism	4/42	0.0					
Pantothenate and CoA biosynthesis	3/19	0.0011358	2.9447	0.087454	0.011925	0.02143	KEGG SMP
Glyoxylate and dicarboxylate metabolism	3/32	0.0052872	2.2768	0.40183	0.049347	0.14815	KEGG
Valine, leucine and isoleucine degradation	3/40	0.0099344	2.0029	0.74508	0.083449	0.0	KEGG SMP
Citrate cycle (TCA cycle)	2/20	0.021391	1.6698	1.0	0.16335	0.07615	KEGG SMP
Pyruvate metabolism	2/22	0.025652	1.5909	1.0	0.17956	0.20684	KEGG SMP
Propanoate metabolism	2/23	0.027903	1.5543	1.0	0.1803	0.0	KEGG SMP
Alanine, aspartate and glutamate metabolism	2/28	0.040286	1.3948	1.0	0.2256	0.0024	KEGG SMP SMP SMP
Glutathione metabolism	2/28	0.040286	1.3948	1.0	0.2256	0.09216	KEGG SMP
Synthesis and degradation of ketone bodies	1/5	0.056803	1.2456	1.0	0.29822	0.6	KEGG SMP
Arginine and proline metabolism	2/38	0.069981	1.155	1.0	0.34579	0.01212	KEGG SMP
Thiamine metabolism	1/7	0.07866	1.1042	1.0	0.36708	0.0	KEGG SMP
Taurine and hypotaurine metabolism	1/8	0.089408	1.0486	1.0	0.39528	0.0	KEGG SMP
Ubiquinone and other terpenoid-quinone biosynthesis	1/9	0.10004	0.99984	1.0	0.42016	0.0	KEGG SMP

1 2

Matched metabolites:

Pathway	Metabolites
Glycine, serine and threonine metabolism	**L-Serine**; Choline; Betaine aldehyde; Betaine; Guanidinoacetate; 3-Phospho-D-glycerate; **N,N-Dimethylglycine**; L-Cystathionine; **Glycine**; O-Phospho-L-serine; **Sarcosine**; 5,10-Methylenetetrahydrofolate; **L-Threonine**; Lipoylprotein; Aminoacetone; D-Glycerate; [Protein]-S8-aminomethyldihydrolipoyllysine; Tetrahydrofolate; Dihydrolipoylprotein; 2-Phospho-D-glycerate; D-Serine; Hydroxypyruvate; **Creatine**; 3-Phosphonooxypyruvate; **L-Cysteine**; 2-Oxobutanoate; Glyoxylate; L-2-Amino-3-oxobutanoic acid; **Pyruvate**; CO2; 5-Aminolevulinate; Methylglyoxal; Ammonia

Impact – based on number of nodes the metabolites have in the pathways
P-value – based on Metabolite Set Enrichment Analysis

Metabolite Set Enrichment Analysis (MSEA)

MSEA is a method to test the probability that the metabolites identified represent metabolic pathways.

Utilises Fishers Exact test to assign a P value to each pathway based on the metabolites observed.

Requires a database containing all pathways and associated metabolites in the organism of study (can be accessed from KEGG).

Calculation cycles through each individual pathway in turn to determine a probability (P-value) for the likelihood of each pathway being represented:

	Query Metabolite	Query Pathway
In Pathway	A	B
Not in Pathway	C-A	D-B

A = Number of query metabolites matched with query pathway
B = Number of metabolite instances for the query pathway
C = total number of query metabolites
D = Total number of metabolite instances for all pathways in the organism

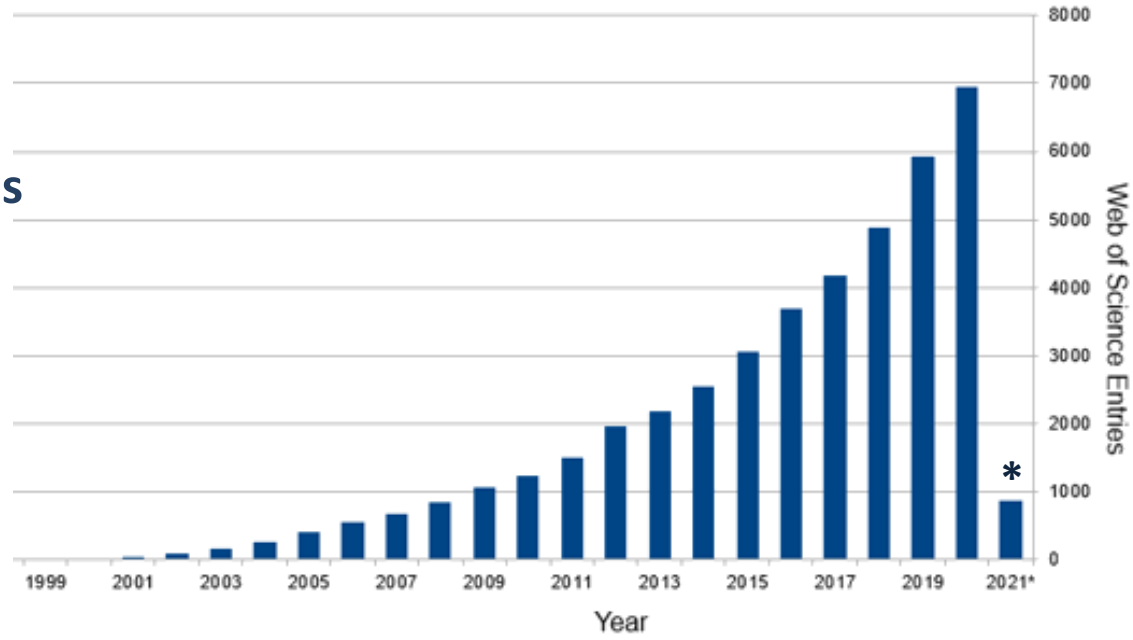
J Xia & DS Wishart *Nuc Acids Res*, 38(1) 2010, W71–W77, doi.org/10.1093/nar/gkq329

Data Deposition

Why Open Access?

- Increase Understanding
- Share best practise
- Comparative studies
- Publication in top-tier journals
- Improve Integration with
- other ‘omics

Metabolomics Research Publications:



*Jan-Feb 2021 only

MetaboLights

i – investigation details

Ontology source list (databases used)
Author list (role, address, email, affiliations)
Protocols
- (Sample collection,
- Extraction
- NMR sample
- NMR spectroscopy
- NMR assay
- Data transformation
- Metabolite identification,
- Statistical test)

Publication (DOI, abstract, title, authors)
Study factors

a – assay information

Most of the information will be **consistent** between samples

Details regarding technique

s – study information

m – metabolite profiles

Scripts available to convert HMDB ID to other formats

ebi.ac.uk/metabolights

Haug et al Nucl. Acids Res. (2013) doi: 10.1093/nar/gks1004 :

Training Workshops and Networking

Look out for coming events:
Liverpool Training Centre for Metabolomics
@LivUniTCM

1 day metabolomics pipeline workshop (Autumn 2022)
Step by step talks on metabolomics pipeline

Lipids working group
From sample collection to pathways & biomarker determination

Hands on R statistics for NMR metabolomics (January 2023)
Computer led statistical analysis

1 day metabolomics symposium (tbc 2023)
Showcase of metabolomics in research at Liverpool

Liverpool Workshop online:
www.tinyurl.com/NMRmetab

International Networks

National Phenome Centre (Imperial & St Marys, London):
Learn.nihr.ac.uk

Birmingham Metabolomics Training Centre
Birmingham.ac.uk/facilities/metabolomics-training-centre

Metabolomics Society:
Metabolomicssociety.org

European Bioinformatics Society:
Ebi.ac.uk/training/handson

Metabolomics Quality Assurance and Quality Control consortium
mQACC.org

Helpful Literature

Blood plasma

-Soininen et al Analyst. 2009 Sep;134(9):1781-5. doi: 10.1039/b910205a

Recommendations of the Metabolomics Society

-Sumner et al. Metabolomics (2007) 3:211 doi:10.1007/s11306-007-0082-2

-Salek et al GigaScience 2013 2:13 doi: 10.1186/2047-217X-2-13

Nature Protocols

-Beckonert et al Nature Protocols 2, - 2692 - 2703 (2007) doi:10.1038/nprot.2007.376

-Want et al Nature Protocols 8, 17–32 (2013) doi:10.1038/nprot.2012.135

-Want et al Nature Protocols 5, - 1005 - 1018 (2010) doi:10.1038/nprot.2010.50

Power Calculations

-general software overview:

McCrum-Gardner Int J Therapy & Rehab 2010, 17(1) doi:10.12968/ijtr.2010.17.1.45988 (single variable)

Drive full of resources (including further background reading):

www.tinyurl.com/NMRmetab-docs

Thanks for Listening

Involved In NMR teaching?
Take my teaching survey
(simplified after issues with submissions)
www.tinyurl.com/NMRteach

@LivUniCMR
@LivUniTCM

HF-NMR
NMR metabolomics

@LivUniNMR

mphelan@Liverpool.ac.uk
sites.google.com/view/nmrliverpool

Liverpool HF-NMR:
Dr Rudi Grosman

Liverpool CBF:
Dr Eva Caamano
Dr Arturas Grauslys

Steering Group
Dr Igor Barsukov
Dr Konstantin Luzyanin
Prof Jon Iggo
Prof Fred Blanc

Thank you!

LIV-SRF:
Prof Ian Prior
Dr Victoria Harman
Ben Mollitt
Julie Boileau
@livuniLivSRF

DOSY DISCOVERY DAY

12-09-2022 **Dr Juan Aguilar-Malavia** **University of Durham**

09:45-10:00. Welcome and overview (housekeeping, datasets used).
10:01-11:00. So many pulse sequences. Which one should I use?

11:31-12:30 Setting up DOSY, including questions

13:31-14:15. How not to interpret DOSY. Gareth.
14:16 – 14:45. How to process DOSY data using VNMRJ. Juan.
14:46 – 15:00. How to process DOSY data using TopSpin. Juan.
15:01 – 15:15. How to process DOSY data using Dynamics centre. Geoff.
15:16 – 15:45. How to process DOSY data using Mestrenova. Marie.
15:46 – 16:05. How to process DOSY data using GNAT. Mathias Nilsson.

16:06-16:30 troubleshooting/analysis Q&A. 30 min.

Preliminary Program & (free) Registration: **www.tinyurl.com/DOSY2022**
For more details: **j.a.aguilar@durham.ac.uk**

A caveat - Seeing is believing!

Identical R does not inform on the distribution of Errors