


PATHWAY ANALYSIS FOR METABOLOMICS





Alex Sanchez

INTRODUCTION AND OBJECTIVES

INTRODUCING MYSELF








Statistics and Bioinformatics
Integrative analysis of omics data



SAMANTHA




Nutrition and Metabolomics



EIT Health is supported by the EIT,
a body of the European Union

DiGu Met Diet, Gut & Metabolomics

Software development



Alex Sánchez-Pla

Full Professor of Statistics.

Faculty of Biology Universitat de Barcelona

UB Director MSc of Statistics & Bioinformatics





UNIVERSITAT DE BARCELONA

Facultat de Biologia





Universitat Oberta de Catalunya

(UB) Coordinator of the Bioinformatics PhD program

(UB) Coordinator of the UOC-UB MSc Program in Bioinformatics & Biostatistics



Vall d'Hebron Institut de Recerca

Former Head of Statistics and Bioinformatics Unit
Former Head of the Data Science Projects Platform



EU-PEARL
EU PATIENT-CENTRIC CLINICAL TRIAL PLATFORM



EOSC-Life



IMPACT



UAB



Vall d'Hebron Institut de Recerca



Vall d'Hebron



INTRODUCING OUR GROUPS



Statistics & Bioinformatics and Nutrition & Metabolomics groups @ UB

SESSION OBJECTIVES

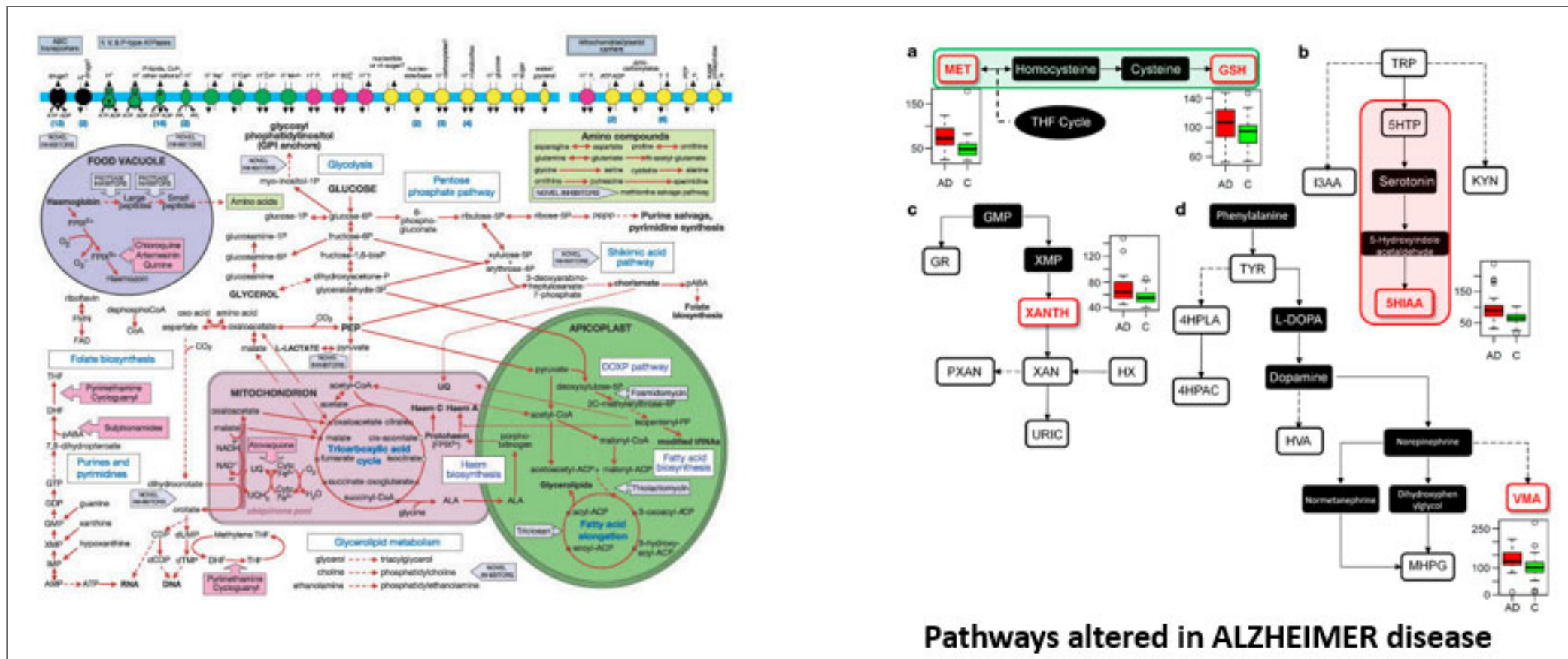
- Overview of Pathway Analysis for Metabolomics
- Introduce its components and
- Go through some methods with some detail
- Discuss some limitations and provide recommendations.
- Introduce some tools for Pathway Analysis
- Get a practical grasp of how to apply it.

SESSION OUTLINE

1. Introduction and objectives
2. Metabolite lists: What do they mean
3. Information sources to support interpretation
4. Methods and Tools to extract information
5. The limitations of PwA. Some recommendations
6. Software tools for PwA
7. Practical session

HEALTH, DISEASE AND PATHWAYS

- Metabolism is a complex network of chemical reactions within the confines of a cell that can be analyzed in self-contained parts called pathways.
- We often assume that “normal” metabolism is what happens in healthy state or, that disease can be associated with some type of alteration in metabolism.



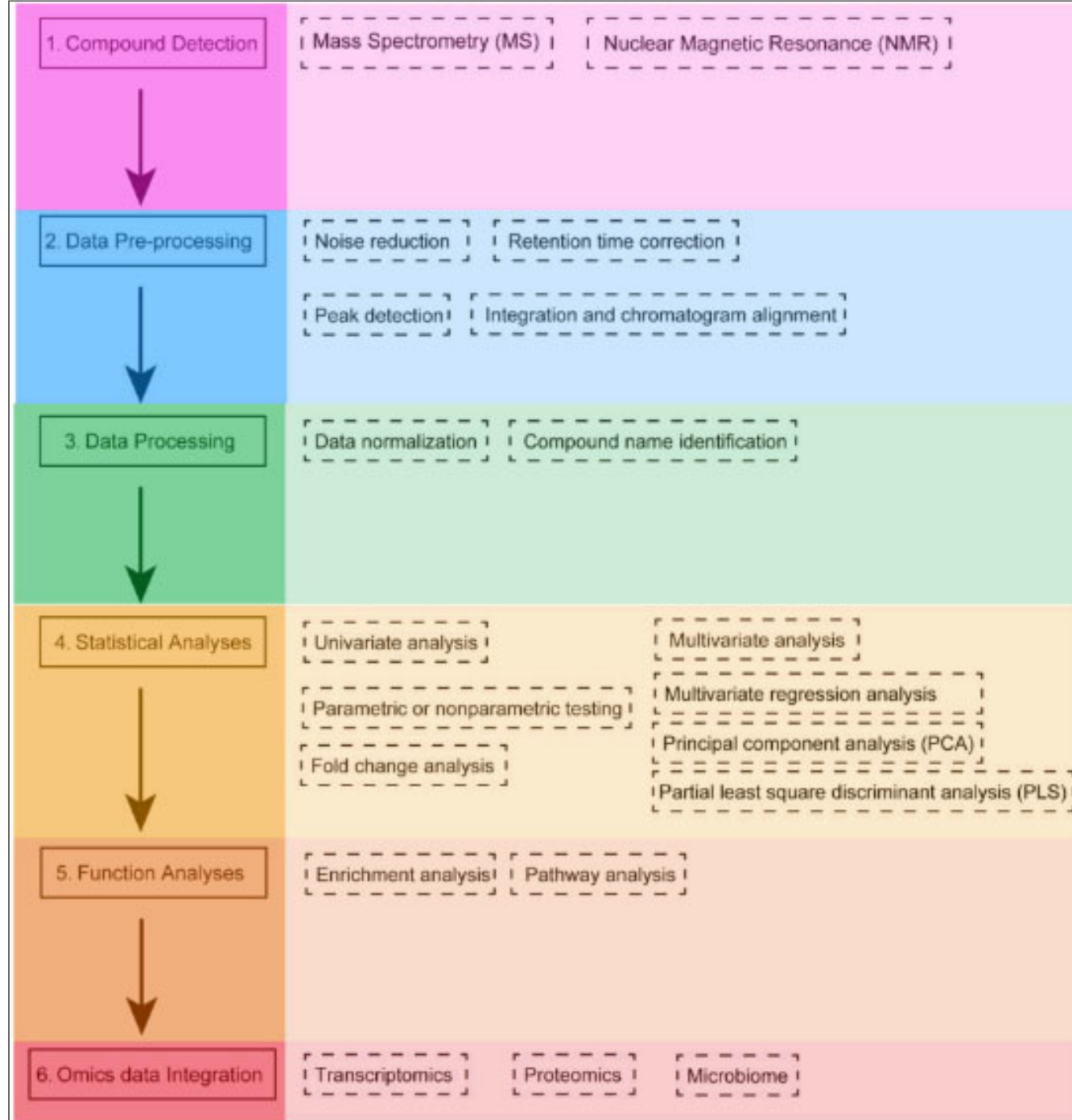
Characterization of disease attempted studying how ths disrupts pathways

SO WHAT IS PATHWAY ANALYSIS?

- ... any analytic technique that benefits from biological pathway or molecular network information to gain insight into a biological system. (Creixell et al., Nature Methods 2015 (12 (7)))
- Pathway Analysis methods rely on high throughput information provided by omics technologies to:
 - Contextualize findings to help understand biological processes
 - Identify features associated with a disease
 - Predict drug targets
 - Understand how to intervene in disease
 - Conduct target literature searches
 - Integrate diverse biological information

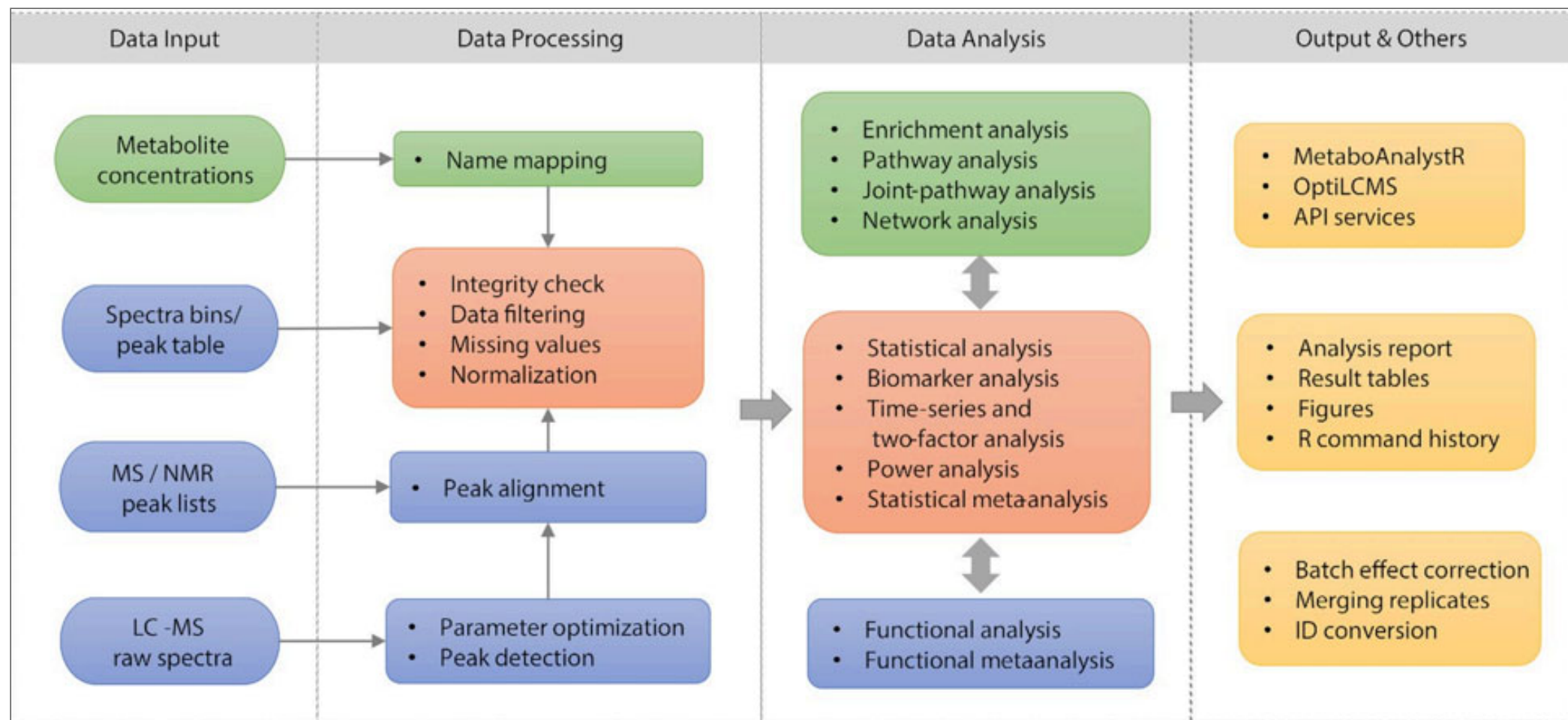
FROM SAMPLES TO *FEATURES* LISTS

BIOINFORMATICS WORKFLOWS



A Metabolomics Workflow Example

FROM SAMPLES TO *FEATURES* LISTS (2)



Metabolomics Workflows in MetaboAnalyst 5.0

ANALYSIS YIELDS METABOLITES LISTS

Metabolite
Amino acid
5-oxoproline (pyroglutamic acid)
7-Methylguanine
Creatinine
Histidine
Kynurenic acid
L-Tryptophan
N-(2-Furoyl)glycine
N-Acetylneuraminic acid
Spermidine
Organic compounds
(±)-Sulfobutanedioic acid
D-Tagatose
D-Xylulose
Glutaminy-Gamma-Glutamate
L-Galacto-2-heptulose
N-Acetylgalactosamine 6-sulfate
Phenol sulphate
Trigonellinamide
Tyrosine
Salicyluric acid
Carbohydrates
Gluconic acid
Sorbitol
Xenobiotics

An unordered list of metabolite IDs

Metabolite	Fold change	<i>p</i> -value	AUC	<i>p</i> -value
N-(2-Furoyl)glycine	13.83	0.001	0.902	0.001
Histidine	2.61	0.005	0.799	0.005
D-Tagatose	2.47	0.031	0.732	0.031
Gluconic acid	1.88	0.014	0.656	0.146
Sorbitol	1.60	0.038	0.763	0.014
(±)-Sulfobutanedioic acid	1.58	0.031	0.732	0.031
Phenol sulphate	1.58	0.042	0.719	0.042

Fold changes and AUC of metabolites whose concentrations were significantly increased in the patients with breast cancer compared to the healthy controls

- Metabolites lists are diverse:
 - Truncated vs All the features analyzed
 - Ordered vs unordered
 - Only IDs vs IDs with difference measures

AN OPEN PROBLEM: METABOLITES IDS

- To be able to do Pathway Analysis, metabolites need to be *mappable* to their sources of information.
 - Must be uniquely identifiable by names/IDs.
 - Must be possible to link/relate these names/IDs with the corresponding IDs in the source of information we wish to rely.
- This is *far from possible* for all metabolites.
- Uniquely and unambiguously naming all metabolites is, in the best of cases, “work in progress”.

DIFFERENT ANNOTATION LEVELS

1. **Exact structure**, including stereochemistry and bond geometry
2. **Regiochemistry level** (stereochemistry and bond geometry unknown)
3. **Molecular species level** (regiochemistry unknown)
4. **Species level** (no information on structural features)

MANY NAMES AND DESCRIPTORS

- **Computed descriptors**

- IUPAC name
- InChI, InChIKey
- SMILES (canonical or isomeric)

IUPAC name

(3S,8S,9S,10R,13R,14S,17R)-10,13-dimethyl-17-[(2R)-6-methylheptan-2-yl]-2,3,4,7,8,9,11,12,14,15,16,17-dodecahydro-1H-cyclopenta[a]phenanthren-3-ol

InChI

1S/C27H46O/c1-18(2)7-6-8-19(3)23-11-12-24-22-10-9-20-17-21(28)13-15-26(20,4)25(22)14-16-27(23,24)5/h9,18-19,21-25,28H,6-8,10-17H2,1-5H3/t19-,21+,22+,23-,24+,25+,26+,27-/m1/s1

InChIkey

HVYWMOMLDIMFJA-DPAQBDIFSA-N

SMILES

CC(C)CCCC(C)C1CCC2C1(CCC3C2CC=C4C3(CCC(C4)O)C)C

Isomeric SMILES

C[C@H](CCCC(C)C)[C@H]1CC[C@@H]2[C@@]1(CC[C@H]3[C@H]2CC=C4[C@@]3(CC[C@@H](C4)O)C)C

Computed descriptors for Cholesterol

MANY NAMES AND DESCRIPTORS

- **Non-systematic identifiers**

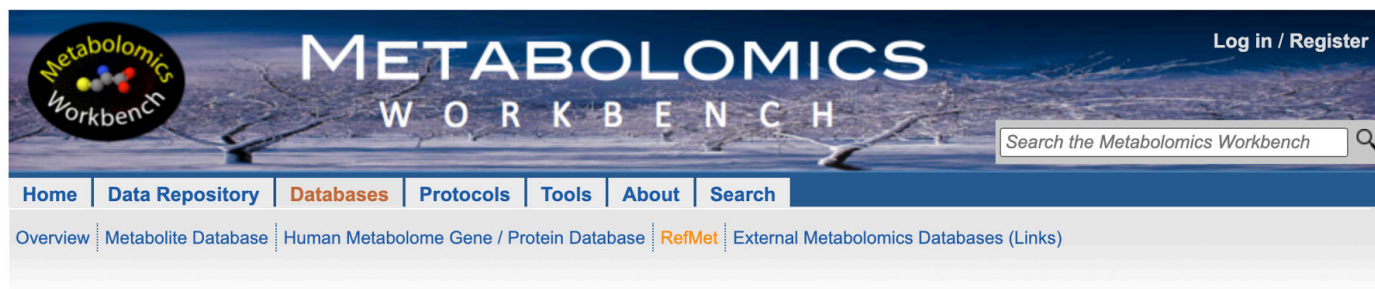
- Common name
- RefMet Name
- PubChem ID
- HMDB ID
- ChEBI ID
- KEGG ID
- LipidMaps ID
- Drug Bank ID
- Metabolomics Workbench ID
- CAS
- Deprecated CAS
- ...

MANY SYNONYMS

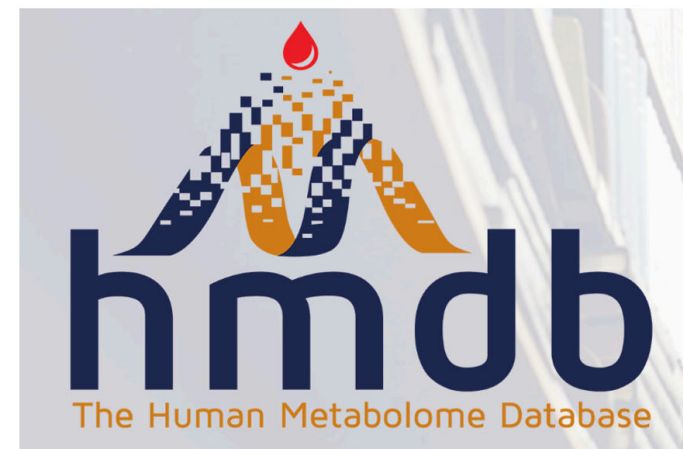
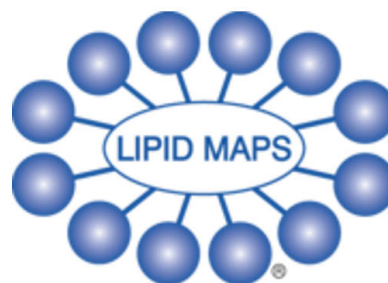
- Cholest-5-en-3-ol (3 β)-
- Cholesterol
- (3 β)-Cholest-5-en-3-ol
- Cholest-5-en-3 β -ol
- Cholesterin
- 5:6-Cholesten-3 β -ol
- Dythol
- 3 β -Hydroxycholest-5-ene
- Provitamin D
- Cholesteryl alcohol
- (-)-Cholesterol
- Δ^5 -Cholesten-3 β -ol
- Lidinit
- Lidinite
- NSC 8798
- SyntheChol
- Marine Cholesterol
- C 8667
- 137: PN: WO2023069707 SEQID: 165 claimed sequence
- MeSH ID: D002784

Other names for Cholesterol

MANY SOLUTIONS



RefMet: A Reference list of Metabolite names



KEGG COMPOUND Database

Chemical substances integrated with genomics

Some compound databases

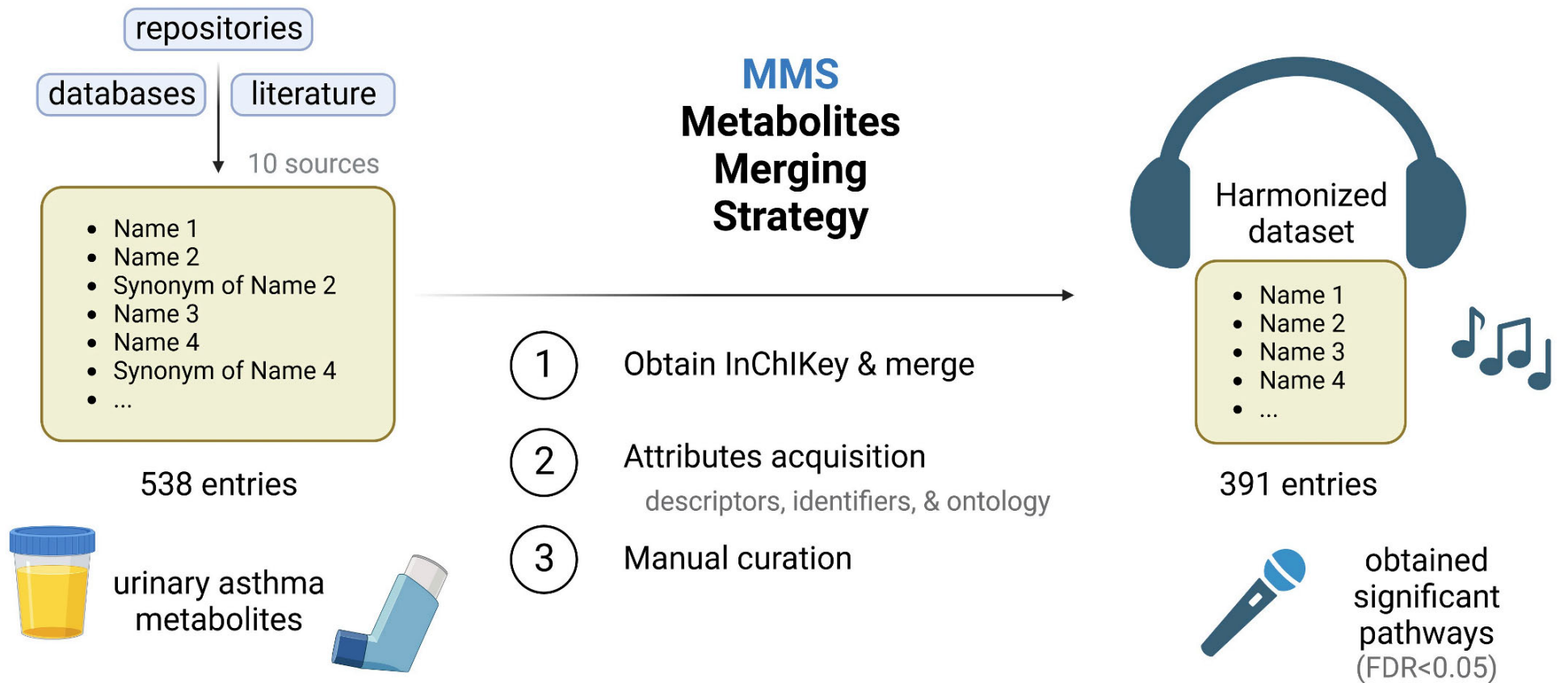
MANY SOLUTIONS

A Metabolites Merging Strategy (MMS): Harmonization to Enable Studies' Intercomparison

Héctor Villalba¹, Maria Llambrich^{2 3}, Josep Gumà^{1 4}, Jesús Brezmes^{2 3}, Raquel Cumeras^{1 2}

Affiliations + expand

PMID: 38132849 PMCID: [PMC10744506](#) DOI: [10.3390/metabo13121167](#)



This study highlights the need for standardized and unified metabolite datasets to enhance the reproducibility and comparability of metabolomics

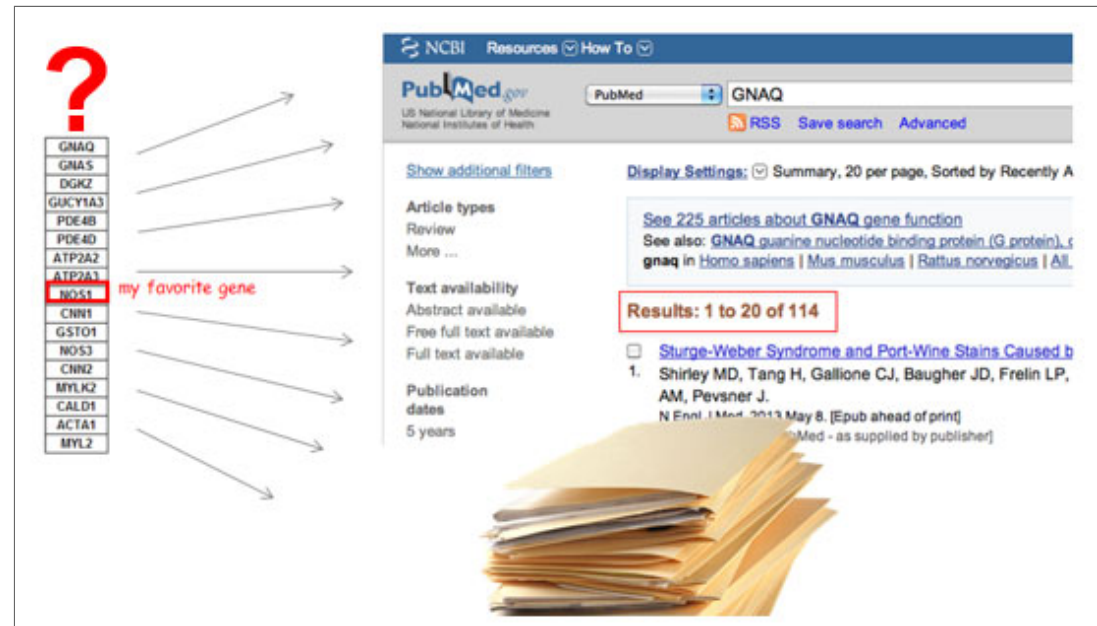
studies.

<https://pubmed.ncbi.nlm.nih.gov/38132849/>

THE *WHERE TO, NOW?* QUESTION

Once a list of feature is obtained it can be studied on a one-by-one basis

- Select some features for biochemical validation,
- Map individual features to specific pathways,
- Perform functional assays,
- Do a literature search ...



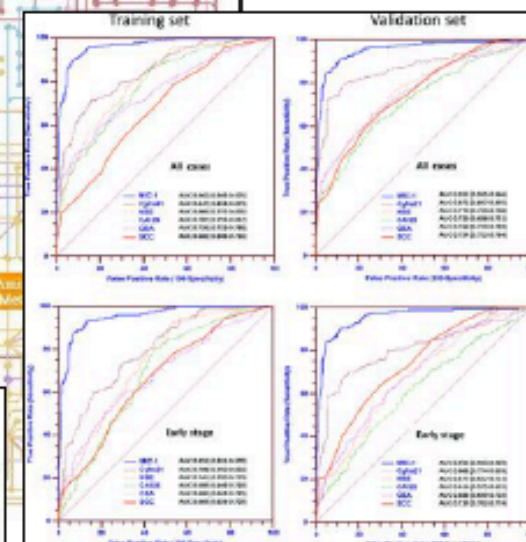
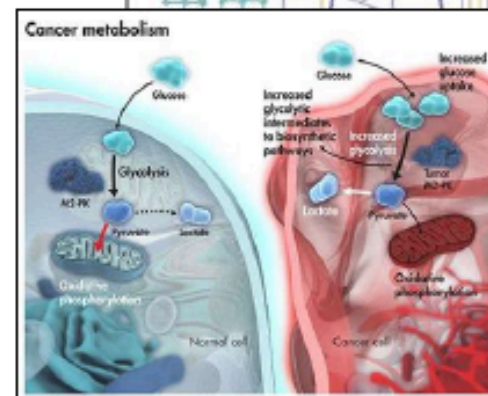
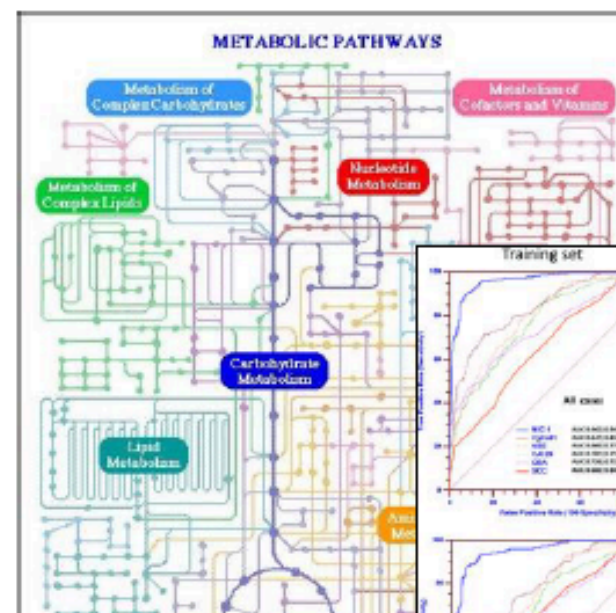
- This will yield useful information, but
 - It may be slow and resource-consuming
 - It does not account for **interaction** between features.

AND HERE COMES PATHWAY ANALYSIS

- Pathway Analysis studies the list as a whole.
- With this aim it combines:
 - The list of features, with
 - Pre-existing sources of information related to them
- And, after some processing, it yields
 - some type of scores about
 - groups of features appearing to be significantly related with the process being studied.

HOW CAN WE INTERPRET THESE LISTS?

Compound	Retention Time (min)	Conc. in Urine (μM)	Compound	Retention Time (min)	Conc. in Urine (μM)
Dns-α-phospho-L-lysine	0.92	<0 L	Dns-Ile	6.35	25
Dns-α-phospho-L-tyrosine	0.95	<0 L	Dns-3-aminosalicylic acid	6.44	0.5
Dns-adenosine monophosphate	0.99	<0 L	Dns-pipecolic acid	6.50	0.5
Dns-α-phosphoethanolamine	1.05	16	Dns-Leu	6.54	54
Dns-glucosamine	1.08	22	Dns-cystathionine	6.54	0.3
Dns-α-phospho-L-threonine	1.09	<0 L	Dns-Leu-Pro	6.60	0.4
Dns-8-dimethylamine purine	1.20	<0 L	Dns-5-hydroxylysine	6.65	1.6
Dns-3-methyl-histidine	1.22	80	Dns-Cystine	6.73	160
Dns-aurine	1.25	634	Dns-N-norleusine	6.81	0.1
Dns-samosine	1.34	28	Dns-5-hydroxydopamine	7.17	<0 L
Dns-Arg	1.53	36	Dns-dimethylamine	7.33	299
Dns-Asn	1.55	133	Dns-5-HIAA	7.46	19
Dns-hypoxanthine	1.58	10	Dns-umbelliferone	7.47	1.9
Dns-homocarnosine	1.61	3.9	Dns-2,3-diaminopropionic acid	7.53	<0 L
Dns-guanidine	1.62	<0 L	Dns-L-serine	7.70	15
Dns-Gln	1.72	633	Dns-4-acetyamidophenol	7.73	58
Dns-allantoin	1.83	3.6	Dns-prosine	7.73	6.9
Dns-L-citrulline	1.87	2.9	Dns-homocystine	7.76	3.3
Dns-1 (or 3)-methylhistamine	1.94	1.9	Dns-acetaminophen	7.97	82
Dns-adenosine	2.06	2.6	Dns-Phe-Phe	8.03	0.4
Dns-methylguanidine	2.20	<0 L	Dns-5-methoxy tryptophan	8.04	2.1
Dns-Ser	2.24	641	Dns-Lys	8.86	164
Dns-aspartic acid amide	2.44	26	Dns-aniline	8.17	<0 L
Dns-4-hydroxy-proline	2.59	2.3	Dns-Ileu-Phe	8.22	0.3
Dns-Glu	2.57	21	Dns-His	8.35	1550
Dns-Asp	2.60	80	Dns-4-fidylsine	8.37	<0 L
Dns-Thr	3.03	157	Dns-benzylamine	8.38	<0 L
Dns-epinephrine	3.05	<0 L	Dns-1-ephedrine	8.50	0.6
Dns-ethanolamine	3.11	471	Dns-tryptamine	8.63	0.4
Dns-aminoadipic acid	3.17	70	Dns-pyridoxamine	8.84	<0 L
Dns-Gly	3.43	2510	Dns-2-methyl-benzylamine	9.24	<0 L
Dns-Ala	3.66	593	Dns-5-hydroxytryptophan	9.25	0.12
Dns-aminolevulinic acid	3.97	30	Dns-1,3-diaminopropane	9.44	0.23
Dns-D-amino-butyric acid	3.98	4.6	Dns-putrescine	9.80	0.5
Dns-D-amino-hippuric acid	3.98	2.9	Dns-1,2-diaminopropane	9.86	0.1
Dns-5-hydroxymethyluracil	4.58	1.9	Dns-tyrosinamide	9.79	29
Dns-tryptophanamide	4.70	5.5	Dns-dopamine	10.08	140
Dns-isoguanine	4.75	<0 L	Dns-castaverine	10.08	0.08
Dns-6-aminopentanoic acid	4.79	1.0	Dns-histamine	10.19	0.4
Dns-sarcosine	4.81	7.2	Dns-3-methoxy-tyramine	10.19	9.2
Dns-3-amino-isobutyrate	4.81	65	Dns-Tyr	10.28	321
Dns-2-aminobutyric acid	4.81	17	Dns-cysteamine	10.42	<0 L



From Lists to Biology

ONTOLOGIES, DATABASES AND METABOLITE SETS

THE ELEMENTS OF PATHWAYS ANALYSIS

- Loosely speaking, to do Pathway Analysis one needs:
 - A list of features, characterizing a process.
 - A source of information about these features.
 - An algorithm to highlight relevant information by linking *list* and *source*.
 - A tool implementing the algorithm.
- In this section, we focus on *sources of information* and on *how to provide it to the algorithms*.

SOURCES OF INFORMATION FOR PWA

Some common databases in Metabolomics

ONTOLOGIES, DATABASES ET ALT.

Although incomplete s.o.i are multiple and diverse.

- **Ontologies:** Structured vocabularies for categorizing and describing relationships within a domain. **GO**, **ChEBI**
- **Pathway Databases:** Detailed information about biological pathways and their the biological context. **KEGG**, **Reactome**, **SMPDB**.
- **Compound Databases:** Information on small molecules for identification and characterization of metabolites. **HMDB**, **PubChem**, **LipidMaps**, and **MassBank**
- And many more: **Networks DBs**, **Spectral DBs**, ...

THE HUMAN METABOLOME DB

The Human Metabolome Database

- Detailed information about human metabolites, their structures, pathways, origins, concentrations, functions and reference spectra
- HMDB has 248,855 metabolites, 132,335 pathways, 3.1 million MS and NMR spectra, metabolite biomarker data on >600 diseases
- A resource established to provide reference metabolite values for human disease, human exposures & population health
- Captures both targeted and untargeted metabolomics (and exposomics) data

THE FOOD CONSTITUENT DATABASE

The Food Constituent Database

- Database of 70,000+ compounds found in 727 foods and their effects on flavour, aroma, colour and human health
- Comprehensive concentration information to ID foods that are rich in particular micronutrients
- Links chemistry to food types (biological species) to flavour, aroma, colour and human health
- Supports sequence, spectral, structure and text searches

THE KEGG DB

·
Kyoto Encyclopedia of Genes and
Genomes

- The “Go-to” Metabolic Pathway Database
- Has 535 “canonical” pathway diagrams or maps covering 5994 organisms for a total of 604,808 pathways
- ~170 metabolic pathways covering 18,553 compounds, includes many disease pathways (80), protein signaling (70) pathways, and biological process pathways (70)
- Metabolic pathways are highly schematized and mostly limited to catabolic and anabolic processes

SMALL MOLECULE PATHWAY DATABASE

The Small Molecule Pathway Database
(SMPDB)

- Nearly 48,900 hand-drawn small molecule pathways – 404 drug action pathways – 20,251 metabolic disease pathways – 27,876 metabolic pathways – 160+ signaling and other pathways
- Depicts organs, cell compartments, organelles, protein locations, and protein quaternary structures
- Maps gene chip & metabolomic data
- Converts gene, protein or chemical lists to pathways or disease diagnoses

OBTAINING METABOLITE SETS

- As described, PwA matches lists of metabolites with previously defined metabolite sets that characterize a process, a disease or a group.
- Some sources of information (Ontologies, Pathways DBs) directly provide metabolite sets.
- For compound DBs, Metabolite sets have to be built
 - By manual curation
 - Automatically (some type of clustering)

METABOLITES SET LIBRARIES

Overview of MSEA's metabolite set libraries

METAMAP CLUSTERS

.

CHEMICAL SIMILARITY CLUSTERS

.

CHEMICAL ONTOLOGIES

.

ANALYSIS METHODS

TYPES OF PATHWAY ANALYSIS

Khatri et al. 10 years of Pathway Analysis

OVER-REPRESENTATION ANALYSIS

- Given
 - A feature (metabolites) list (from some study).
 - A collection of feature (metabolites) sets (...)
- The goal is finding out if any of the feature sets *surprisingly enriched* in the feature list?
 - Need to define “surprisingly” (statistics)
 - Need to deal with test multiplicity?

OBTAINING FEATURE LISTS

.

ASSESSING “SURPRISINGLY”

Given a feature list, “fl”, and a feature set, “FS”, check if the % of genes in “fl” annotated in “FS” the same as the % of genes globally annotated in “FS”?

- If both percentages are similar → *No Enrichment.*
- If the % of features in “FS” is greater in “fl” than in the rest of genes → *“fl” is enriched in “FS”*

EXAMPLE

.

ASSESS SIGNIFICANCE: FISHER TEST

- The example shows two cases
 - One where percentages are quite different
 - Another where percentages are similar.
- How can we set a threshold to decide that the difference is “big enough” to call it “Enriched”
 - Use Fisher Test or, equivalently,
 - a test to compare proportions or
 - a hypergeometric test.

EXAMPLE 1: SURPRISINGLY ENRICHED

P-value small, odds-ratio high: List is surprisingly enriched in Feature Set

EXAMPLE 2: NON-ENRICHED

P-value high, odds-ratio around 1: List is not enriched in Feature Set

SUMMARY: RECIPE FOR ORA

1. Define feature list (e.g. thresholding analyzed list) and background list,
2. Select feature sets to test for enrichment,
3. Run enrichment tests and adjust for multiple testing
4. Interpret your enrichments
5. Publish! ;)

POSSIBLE PROBLEMS WITH ORA

- No “natural” value for the threshold
- Possible loss of statistical power due to thresholding
- No resolution between significant signals with different strengths
- Weak signals neglected
- Different results at different threshold settings
- Based on the wrong assumption of independent feature (or feature group) sampling, which increases false positive predictions.

FUNCTIONAL CLASS SCORING

- Also known as:
 - *Analysis of ranked lists*
 - *Metabolite Set Enrichment Analysis*
 - Rooted in the *Gene Set Enrichment Analysis* (GSEA) method developed to overcome ORA limitations.
-

THE GSEA METHOD (1)

- GSEA method compares, for each feature set, the distribution of the test statistic within the set with the overall distribution of those statistics, i.e. the calculated for all genes.
- To do this, test statistics are ranked (from biggest to smallest) and for gene set a running sum is computed such that
 - If a feature is in the set *add* a certain quantity ($\sqrt{(N - N_s)/N_s}$)
 - If a feature is not in the set, *subtract* a (small) quantity ($\sqrt{N_s/(N - N_s)}$)

THE GSEA TESTS

- If the distribution of the *running sum* doesn't differ from a *random walk* then the list can be declared significantly enriched in that set.
- Original test was a Kolmogorov-Smirnov test (K-S test) statistic with P-values computed by randomization.

GSEA EXTENSIONS/ALTERNATIVES

- **Wilcoxon test:**

It uses rank-based methods to assess whether the feature sets are distributed differently across the groups.

- **Globaltest:**

- It evaluates the association between a predefined set of features and a clinical outcome of interest.
- Instead of testing individual features, it assesses the global effect of the gene set on the outcome.
- This method is beneficial in identifying pathways or feature sets that have a combined influence on a phenotype, rather than relying on individual feature-level analysis.

PWA FOR UNTARGETED STUDIES

- What to do when you don't know what the metabolites ions are?
- Most popular option is Mummichog (Li et al. 2013).

MUMMICHOG PATHWAY MAPPING

- Ions are divided into significant and non-significant groups.
 - E.g 1000 ions, 150 with $p\text{-val} < 0.05$
- Repeat many times
 - Randomly take 150 of the remaining non-significant ions and mapped onto known pathways.
 - This provides an estimate of how likely it is to observe random association of non-significant ions with pathways.
- The significant ions are now mapped to the pathways and evidence is sought for enhanced associations (Fisher exact test)

MUMMICHOG CHANGE OF APPROACH

Mummichog redefines the work flow of untargeted metabolomics

MULTIPLE TESTING PROBLEM AND ADJUSTMENTS

MULTIPLE TESTING

- Whatever approach we use for pathway Analysis there is a common characteristic: *Every test is applied for every feature set in a long collection of sets*
- This leads to a *multiple testing problem*: the Type I error probability of falsely rejecting the null hypothesis increases with the number of tests.
- In order to avoid an artificial inflation of *False positive discoveries* some adjustments are recommended.

HYPOTHESIS TESTS DECISION TABLE

		Actual Situation “Truth”	
		H₀ True	H₀ False
Decision	Do Not Reject H₀	Correct Decision $1 - \alpha$	Incorrect Decision Type II Error β
	Rejct H₀	Incorrect Decision Type I Error α	Correct Decision $1 - \beta$

$\alpha = P(\text{Type I Error}) \quad \beta = P(\text{Type II Error})$

In a test with a null and an alternative hypothesis there are 2 possible right decisions and two possible incorrect ones (Type I and Type II errors)

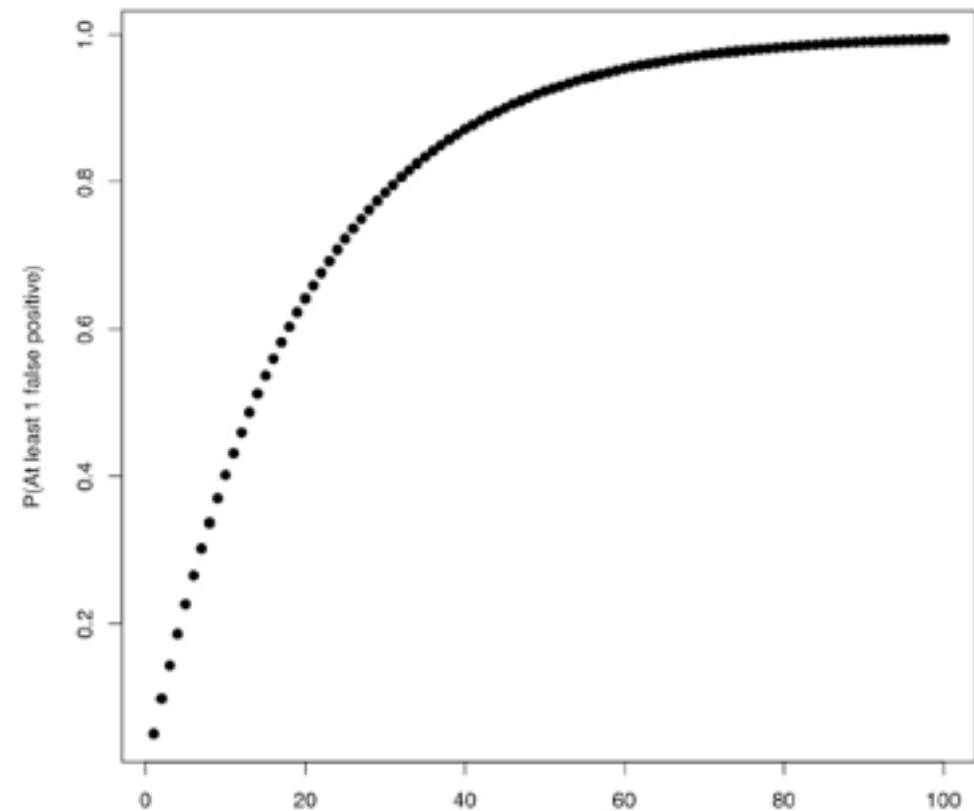
WHY MULTIPLE TESTING MATTERS

- The probability of observing one false positive if testing once is:
 - $P(\text{Making a type I error}) = \alpha$
 - $P(\text{not making a type I error}) = 1 - \alpha$
- Now imagine we perform m tests independently
 - $P(\text{not making a type I error in } m \text{ tests}) = (1 - \alpha)^m$
 - $P(\text{making at least a type I error in } m \text{ tests}) = 1 - (1 - \alpha)^m$
- As m increases the probability of having at least one type error tends to increase

TYPE I ERROR NOT USEFUL HERE

Probability of At Least 1 False Positive

Number of tests: m	P(making at least a type I error) = $1-(1-a)^m$
1	0.050000
2	0.097500
3	0.142625
4	0.185494
5	0.226219
6	0.264908
7	0.301663
8	0.336580



HOW TO DEAL WITH THIS ISSUE?

- Controlling for type I error is not feasible if many tests.
 - Idea: Modify α (or alternatively the p-value) so the error probability is ***controlled overall***
 - This may mean different things:
 1. The probability of *at least one error* in m tests is $< \alpha$
 2. The expected number of false positives is below a fixed threshold.
- ...

FAMILY WISE ERROR RATE

- Let M be the number of annotations tested.
- Given p-value, p compute $p_{adj} = p \times M$, or
- Given significance level α compute $\alpha_{adj} = \alpha/M$.
- The adjusted P-value, p_{adj} is greater than or equal to the probability that **one or more** of the observed enrichments are due to random draws.
- This adjustment is said to *controlling for the Family-Wise Error Rate* (FWER).
- Bonferroni method controls FWER.

BONFERRONI CAVEATS

- This adjustment is very stringent and can “wash away” real enrichments leading to false negatives,
- Often one is willing to accept a less stringent condition, that is accepting some false positives to avoid too many false negatives.
- This is may be done using the “*false discovery rate*” (FDR), which leads to a gentler correction when there are real enrichments.

FALSE DISCOVERY RATE

- FDR is the expected proportion of “False Positives” that is of the observed enrichments due to chance.
- Less restrictive than Bonferroni adjustment which is a bound on the probability that **any one** of the observed enrichments could be due to random chance.
- Typically, FDR adjustments are calculated using the Benjamini-Hochberg procedure.
- FDR threshold is often called the “q-value”

AN EXAMPLE

	raw	Bonferroni	FDR
Quinolate	0.000003	0.000218	0.000218
Glucose	0.000016	0.001036	0.000276
3-Hydroxyisovalerate	0.000019	0.001187	0.000276
Leucine	0.000020	0.001232	0.000276
Succinate	0.000029	0.001802	0.000276
Valine	0.000031	0.001922	0.000276
N,N-Dimethylglycine	0.000034	0.002125	0.000276
Adipate	0.000035	0.002206	0.000276
myo-Inositol	0.000040	0.002508	0.000279
Acetate	0.000069	0.004376	0.000415
Glutamine	0.000073	0.004616	0.000415
Creatine	0.000079	0.004978	0.000415
Alanine	0.000104	0.006570	0.000505

	raw	Bonferroni	FDR
Betaine	0.000115	0.007265	0.000519
Methylamine	0.000127	0.008002	0.000533
Pyroglutamate	0.000172	0.010811	0.000616
3-Hydroxybutyrate	0.000175	0.010994	0.000616
cis-Aconitate	0.000183	0.011547	0.000616
Formate	0.000186	0.011730	0.000616
Tryptophan	0.000196	0.012323	0.000616
Dimethylamine	0.000282	0.017772	0.000846
Creatinine	0.000327	0.020605	0.000937
Tyrosine	0.000525	0.033090	0.001439
Sucrose	0.000710	0.044700	0.001862
3-Indoxylsulfate	0.000924	0.058182	0.002327
Lactate	0.000978	0.061634	0.002371
Threonine	0.001134	0.071410	0.002645
Asparagine	0.001204	0.075839	0.002709
Histidine	0.001272	0.080105	0.002762

	raw	Bonferroni	FDR
trans-Aconitate	0.001349	0.084962	0.002832
Xylose	0.001445	0.091016	0.002915
Serine	0.001486	0.093637	0.002915
Pyruvate	0.001527	0.096207	0.002915
2-Hydroxyisobutyrate	0.001952	0.122970	0.003581
Lysine	0.001989	0.125320	0.003581
Fumarate	0.002326	0.146544	0.004071
2-Aminobutyrate	0.002924	0.184225	0.004979
Fucose	0.003358	0.211567	0.005568
Citrate	0.004126	0.259970	0.006666
tau-Methylhistidine	0.004324	0.272399	0.006810
Trigonelline	0.005797	0.365230	0.008816
Hippurate	0.005877	0.370276	0.008816
Trimethylamine N-oxide	0.006344	0.399666	0.009295
O-Acetylcarnitine	0.007151	0.450507	0.010239
Ethanolamine	0.008639	0.544251	0.012094

	raw	Bonferroni	FDR
Glycine	0.014320	0.902160	0.019612
Taurine	0.019209	1.000000	0.025748
1,6-Anhydro-beta-D-glucose	0.026248	1.000000	0.034230
pi-Methylhistidine	0.026623	1.000000	0.034230
Guanidoacetate	0.027876	1.000000	0.035124
Glycolate	0.028844	1.000000	0.035631
4-Hydroxyphenylacetate	0.031695	1.000000	0.038400
Carnitine	0.035584	1.000000	0.042298
2-Oxoglutarate	0.044770	1.000000	0.052232
Isoleucine	0.051845	1.000000	0.059386
1-Methylnicotinamide	0.063494	1.000000	0.071431
Hypoxanthine	0.093111	1.000000	0.102912
3-Aminoisobutyrate	0.181820	1.000000	0.197494
Tartrate	0.188030	1.000000	0.200778
Pantothenate	0.223280	1.000000	0.234444
Methylguanidine	0.241610	1.000000	0.249532



	raw	Bonferroni	FDR
Uracil	0.295780	1.000000	0.300551
Acetone	0.425500	1.000000	0.425500

LIMITATIONS AND RECOMMENDATIONS

SOME LIMITATIONS


- Incomplete Pathway Databases
- Metabolite Misidentification
- Chemical Bias of Assays
- Background Set Selection
- Selection of Compounds of Interest
- Multiple testing issues

PLOS COMPUTATIONAL BIOLOGY

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis

Cecilia Wieder, Clément Frainay, Nathalie Poupin, Pablo Rodríguez-Mier, Florence Vinson, Juliette Cooke, Rachel PJ Lai, Jacob G. Bundy, Fabien Jourdan, Timothy Ebbels 

PATHWAY ANALYSIS TOOLS

PATHWAY ANALYSIS TOOLS

Tool	Identifiers for metabolite input	Pathway databases used	Methods available	Format
ConsensusPathDB	KEGG, Chebi, Pubchem, CAS, HMDB	ChEMBL, Drugbank, Biocarta, EHMN, HumanCyc, INOH, KEGG, Netpath, PID, Reactome, SMPDB, TTD, Wikipathways	Overrepresentation, Wilcoxon	Website
IMPaLA	KEGG, Chebi, Pubchem, CAS, HMDB	Biocarta, EHMN, BioCyc, INOH, KEGG, Netpath, PID, PharmGKB, Signalink, Reactome, SMPDB, Wikipathways	Overrepresentation, Wilcoxon	Website
MetaboAnalyst	Names, HMDB, KEGG, Pubchem, ChEBI, METLIN	SMPDB, KEGG, Biocarta, SNP-associated metabolite sets, User provided sets	Overrepresentation, MSEA, topological analysis	Website, R package
PaintOmics	KEGG	KEGG	Overrepresentation	Website
MPINet	Pubchem	ChEMBL, Drugbank, Biocarta, EHMN, HumanCyc, INOH, KEGG, Netpath, PID, Reactome, SMPDB, TTD, Wikipathways	Overrepresentation	R package
3omics	Pubchem	KEGG, HumanCyc	Overrepresentation	Website
MarVis	IDs, Names, Accurate masses	Kegg, BioCyc	Overrepresentation, K-S test, Wilcoxon	Website
LIPEA	KEGG, HMDB, abbreviations, swissLipids, lipidMaps, ChEBI	KEGG	Overrepresentation	Website
Lipid mini-on	Names	LipidMaps	Overrepresentation	website

Common pathway analysis tools for metabolomics data.

A COMPARISON OF TOOLS

Marco-Ramell *et al. BMC Bioinformatics* (2018) 19:1
DOI 10.1186/s12859-017-2006-0

BMC Bioinformatics

RESEARCH ARTICLE

Open Access

Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data



Anna Marco-Ramell^{1,2}, Magali Palau-Rodriguez^{1,2}, Ania Alay³, Sara Tulipani¹, Mireia Urpi-Sarda^{1,2}, Alex Sanchez-Pla^{3,4} and Cristina Andres-Lacueva^{1,2*}

Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data

THE SPACE OF TOOLS (IN 2017)

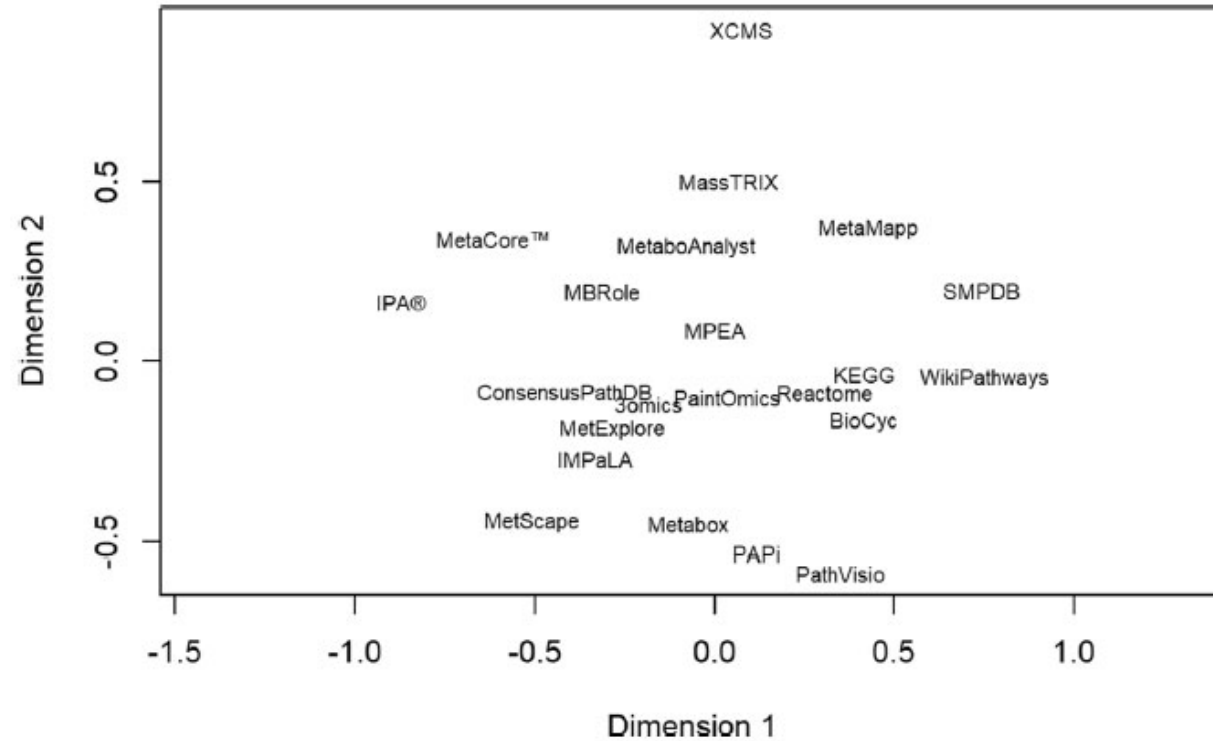


Fig. 1 Non-metric multidimensional scaling (NMDS) plot of the most used tools for metabolomics data enrichment based on Jaccard's distances. Additional file 3: Table S3 shows the main features of each tool

NOT THE SAME, NOT THAT DIFFERENT

- ORA tools provided consistent results among tools revealing that these analyses are robust and reproducible regardless of their analytic approach.
- Redundancy of identifiers, Use of chemical class identifiers and Incompleteness of databases sets limit the extent of the analyses and reduce their accuracy.
- More work in the completeness of metabolite/pathway databases is required to get more accurate and global insights of the metabolome.

SUMMARY

- Pathway Analysis is a useful approach to help gain biological understanding from omics-based studies.
- There are many ways, many methods, many tools
- Guide the choice by a combination of *meaning, availability, ease of use* and *usefulness*.
- Usually obtained from a good understanding of what it does and how it is done.
- Different methods may yield different results.
Worth checking!