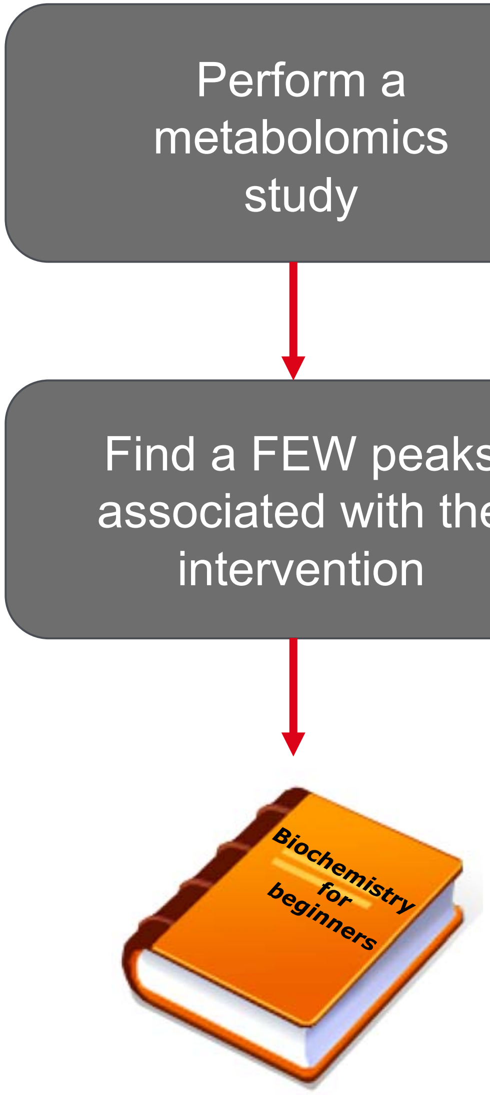
Pathway analysis in metabolomics

Tim Ebbels

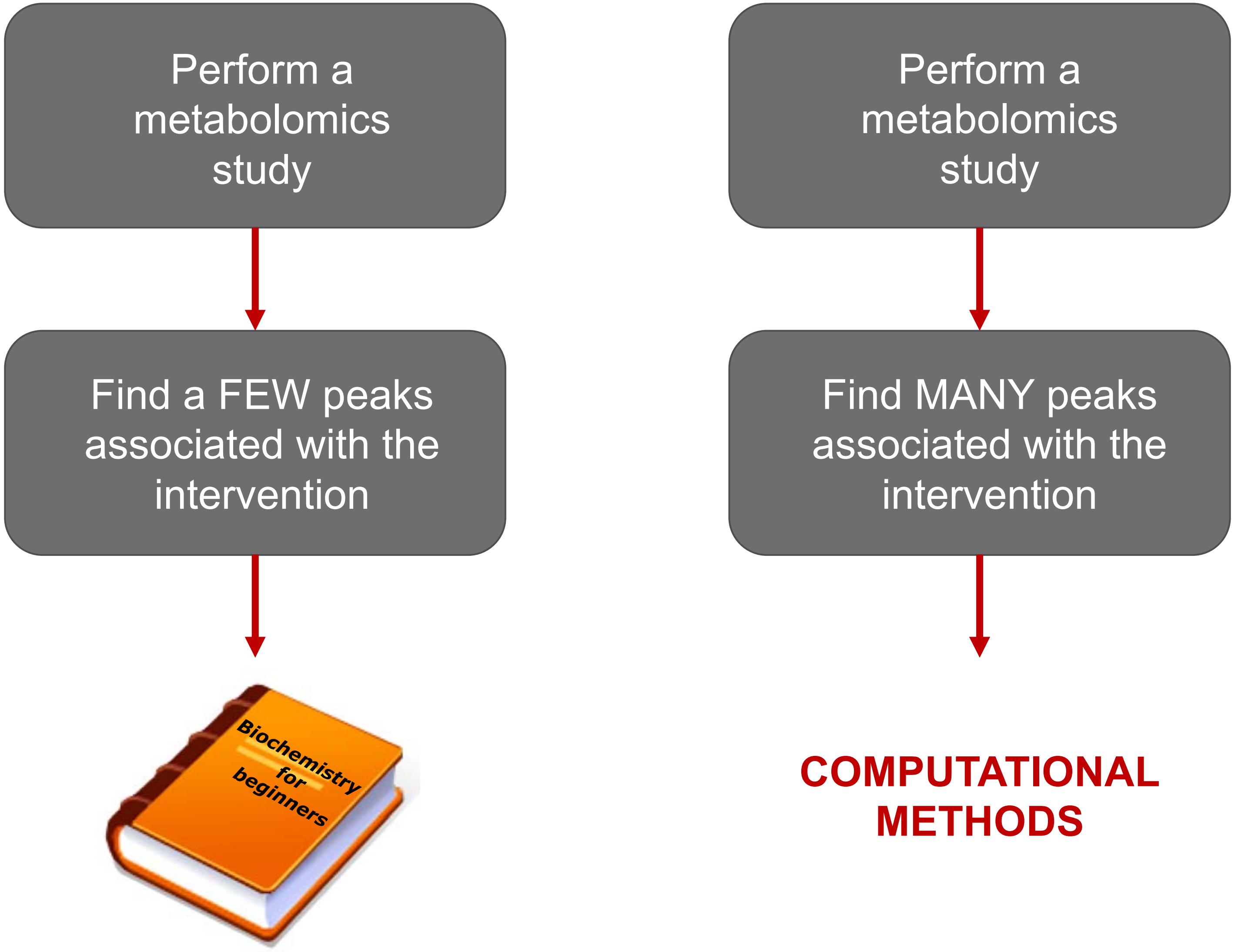
Associating metabolites with pathways



Perform a metabolomics study

Find a FEW peaks associated with the intervention

Associating metabolites with pathways



Perform a metabolomics study

Find a FEW peaks associated with the intervention

Perform a metabolomics study

Find MANY peaks associated with the intervention

COMPUTATIONAL METHODS

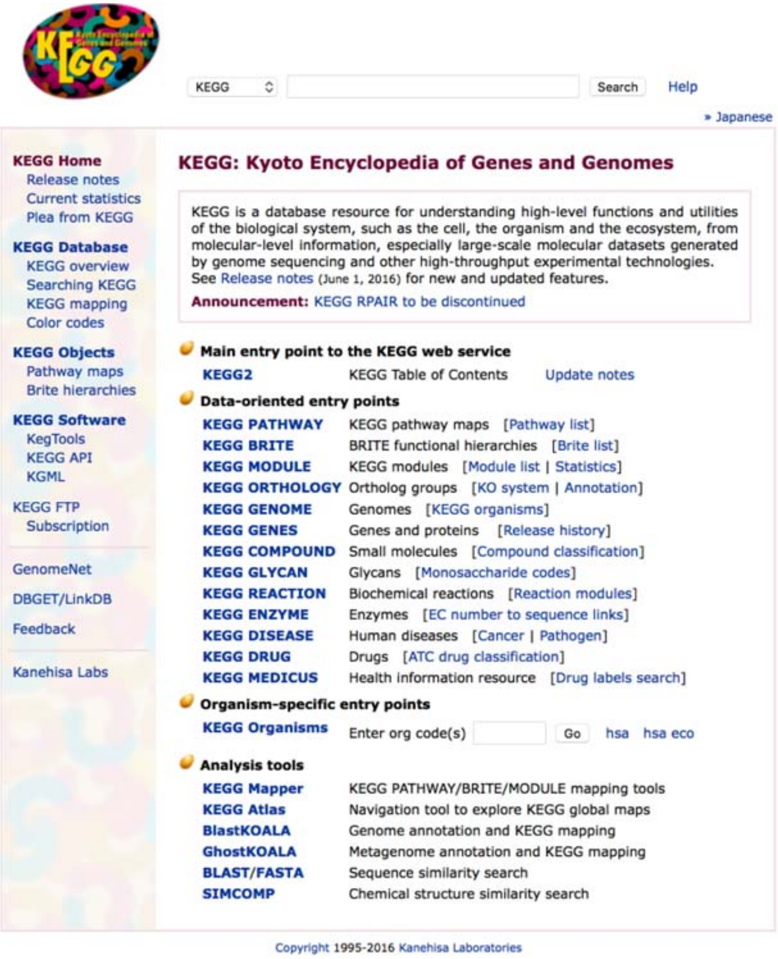
KEGG – Kyoto Encyclopedia of Genes and Genomes

- http://www.genome.jp/kegg/
- KEGG compound
  - Currently lists 17685 compounds

Can download all compound information

using the R package KEGGREST

‘KEGG_Compounds_May_2015.txt’



KEGG Home
Release notes
Current statistics
Plea from KEGG

KEGG Database
KEGG overview
Searching KEGG
KEGG mapping
Color codes

KEGG Objects
Pathway maps
Brite hierarchies

KEGG Software
KegTools
KEGG API
KGML

KEGG FTP
Subscription

GenomeNet
DBGET/LinkDB
Feedback

Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See Release notes (June 1, 2016) for new and updated features.

Announcement: KEGG RPAIR to be discontinued

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents	Update notes

Data-oriented entry points

KEGG PATHWAY	KEGG pathway maps [Pathway list]
KEGG BRITE	BRITE functional hierarchies [Brite list]
KEGG MODULE	KEGG modules [Module list | Statistics]
KEGG ORTHOLOGY	Ortholog groups [KO system | Annotation]
KEGG GENOME	Genomes [KEGG organisms]
KEGG GENES	Genes and proteins [Release history]
KEGG COMPOUND	Small molecules [Compound classification]
KEGG GLYCAN	Glycans [Monosaccharide codes]
KEGG REACTION	Biochemical reactions [Reaction modules]
KEGG ENZYME	Enzymes [EC number to sequence links]
KEGG DISEASE	Human diseases [Cancer | Pathogen]
KEGG DRUG	Drugs [ATC drug classification]
KEGG MEDICUS	Health information resource [Drug labels search]

Organism-specific entry points

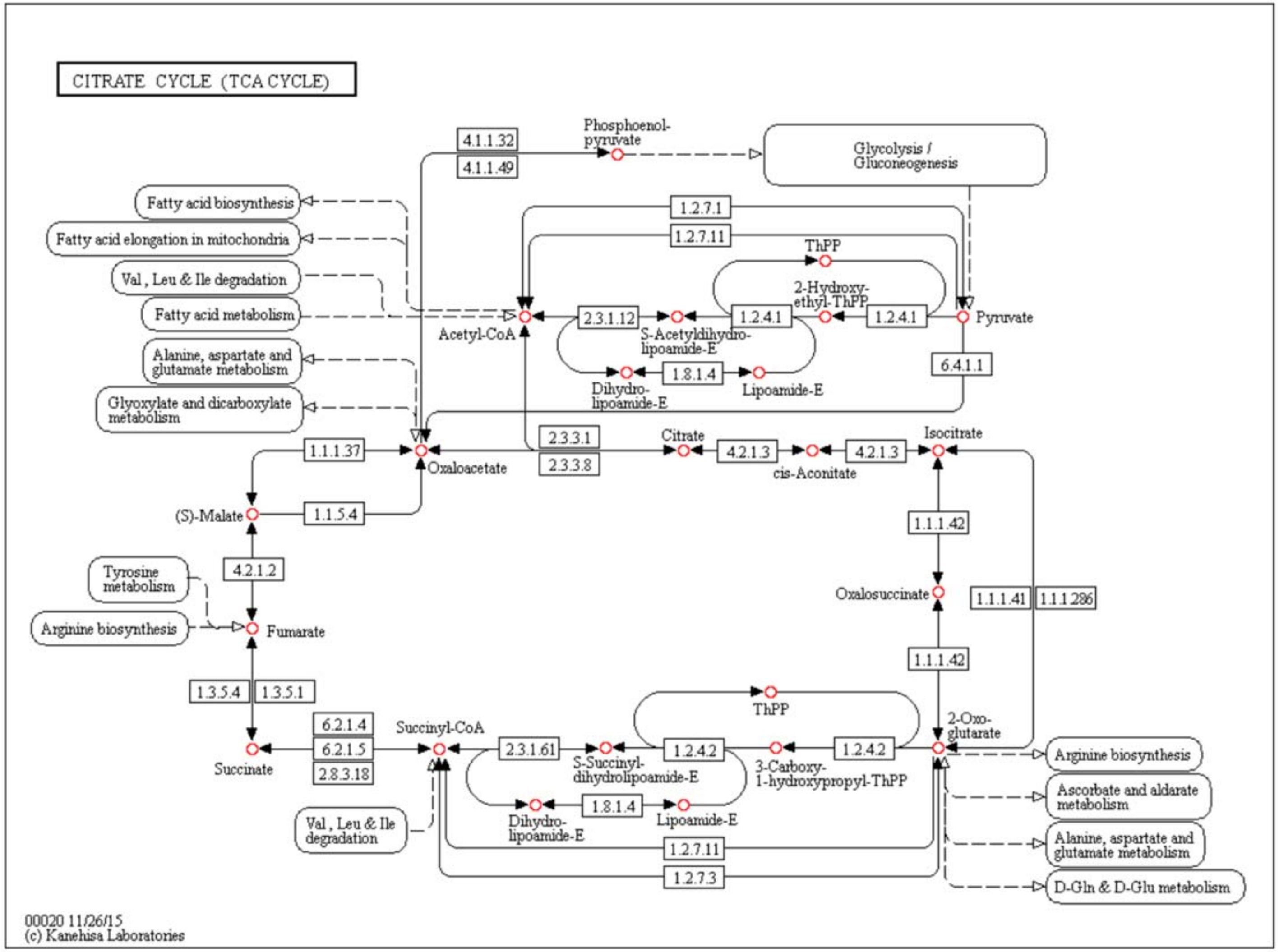
KEGG Organisms	Enter org code(s) Go hsa hsa eco

Analysis tools

KEGG Mapper	KEGG PATHWAY/BRITE/MODULE mapping tools
KEGG Atlas	Navigation tool to explore KEGG global maps
BlastKOALA	Genome annotation and KEGG mapping
GhostKOALA	Metagenome annotation and KEGG mapping
BLAST/FASTA	Sequence similarity search
SIMCOMP	Chemical structure similarity search

Copyright 1995-2016 Kanehisa Laboratories

KEGG PATHWAY



Problems with pathway definitions?

Do metabolic pathways exist?

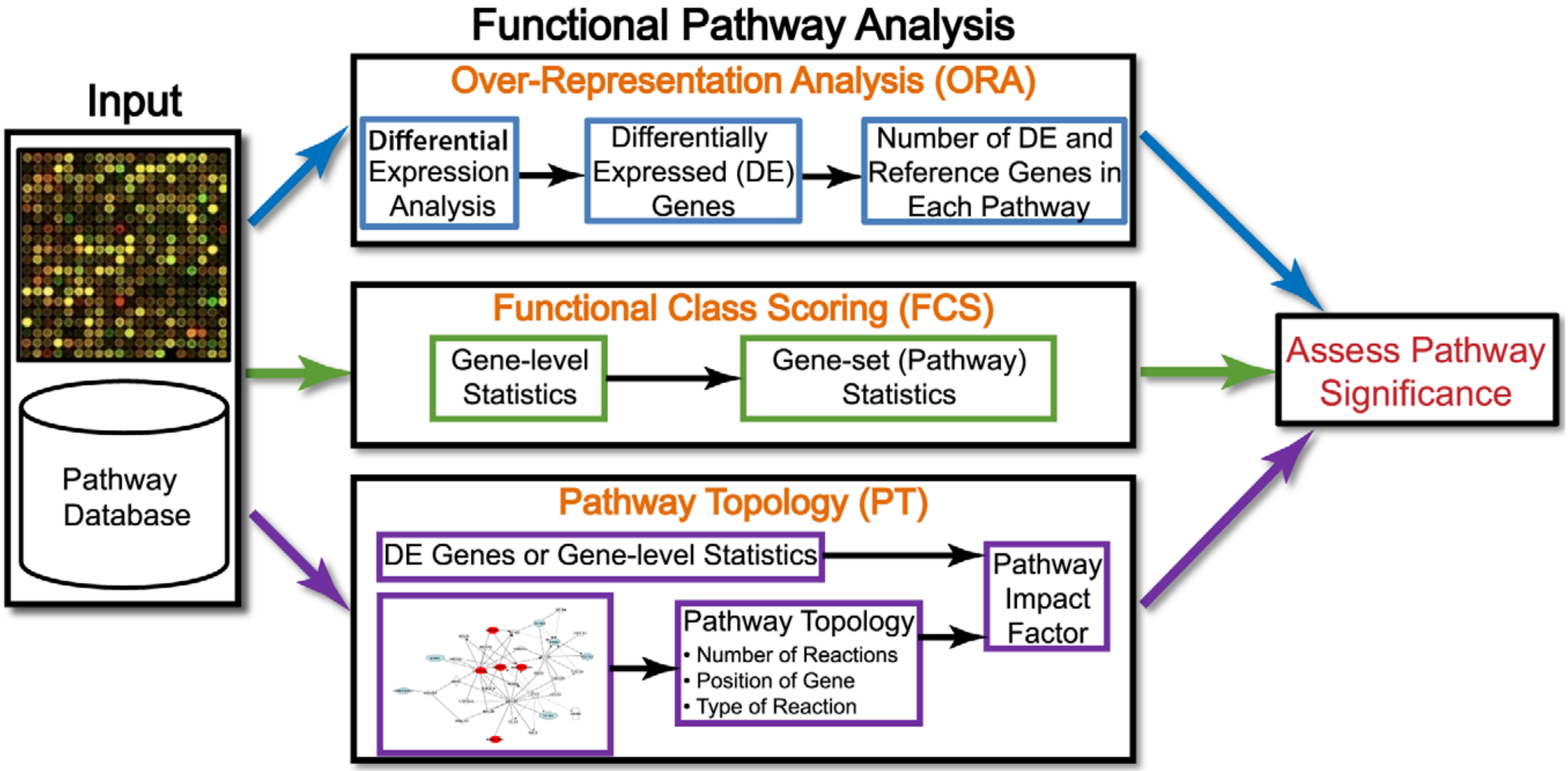
Problems with pathway definitions?

Do metabolic pathways exist?

- Yes, with some complications
  - May be condition-dependent
  - KEGG pathways (or other textbook/database pathways) may represent an arbitrary way of dividing up a metabolic network

Imperial College London

Types of Pathway Analysis

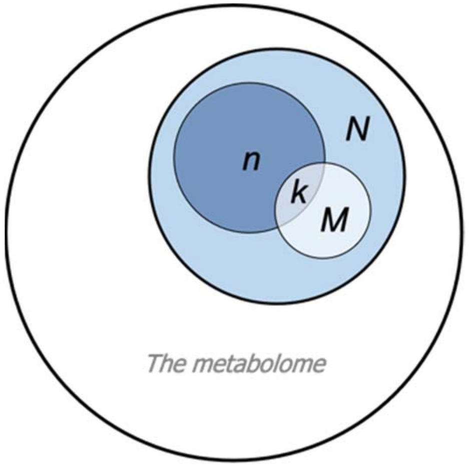

Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Pathway Impact Factor
Assess Pathway Significance

*Khatri, P., M. Sirota and A. J. Butte (2012). "Ten years of pathway analysis: current approaches and outstanding challenges." <u>PLoS Comput Biol</u> **8**(2): e1002375.*

Over-representation analyses

- INPUT
  - List of significant metabolites
- CALCULATION
  - For each pathway, calculate the probability that the list has more metabolites from the pathway than would be expected by chance
- OUTPUT
  - List of significantly associated pathways, with *P* values adjusted for multiple correction testing

Over-representation analysis (ORA)

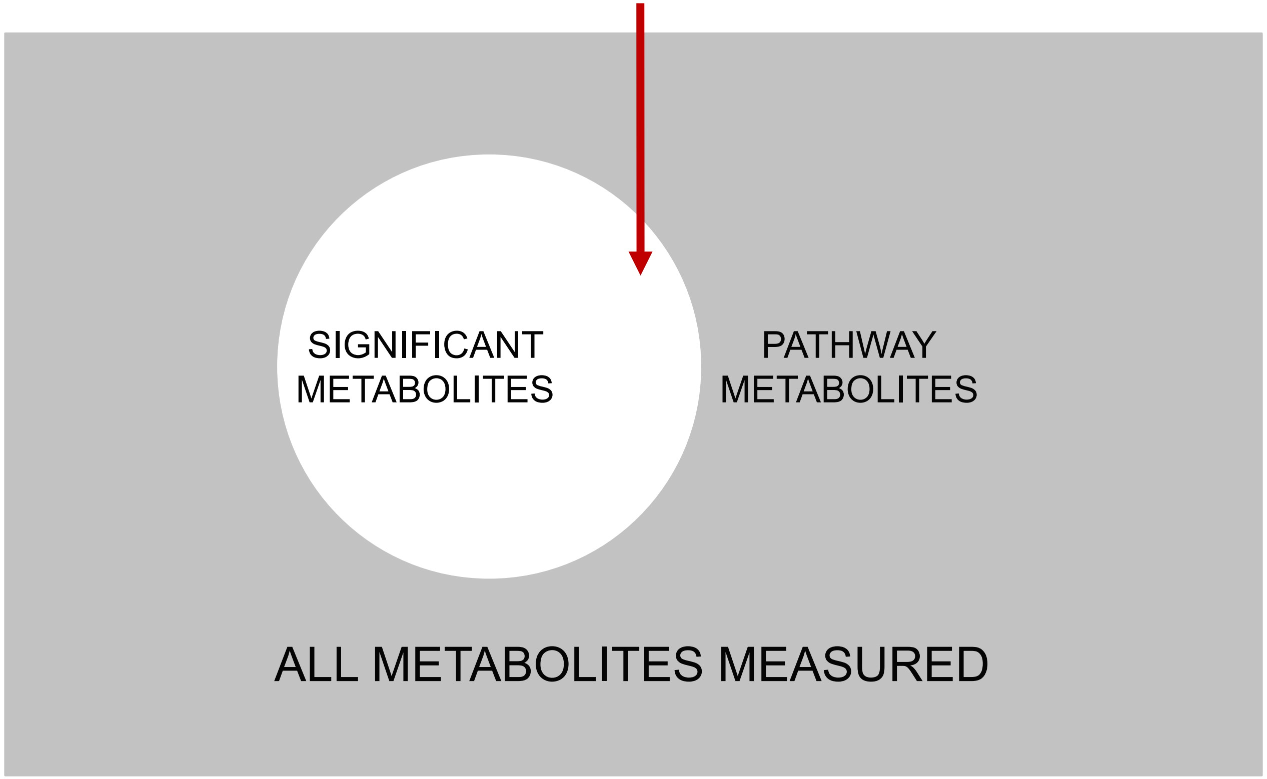


	Iq#sdwkz d|	Q rw#lq#sdwkz d|
G liihuhqwldo|# dexqgdqw	6	78
Q rw#gliihuhqwldo|# dexqgdqw#	;	435

- *N represents compounds forming the background set, which covers part of the full metabolome.*
- *M represents compounds in the pathway of interest.*
- *n represents compounds of interest (i.e. differentially abundant metabolites)*
- *k represents the overlap between the list of compounds of interest and compounds in the pathway.*

Over-representation analysis to identify pathways

HOW SIGNIFICANT IS THE INTERSECTION?



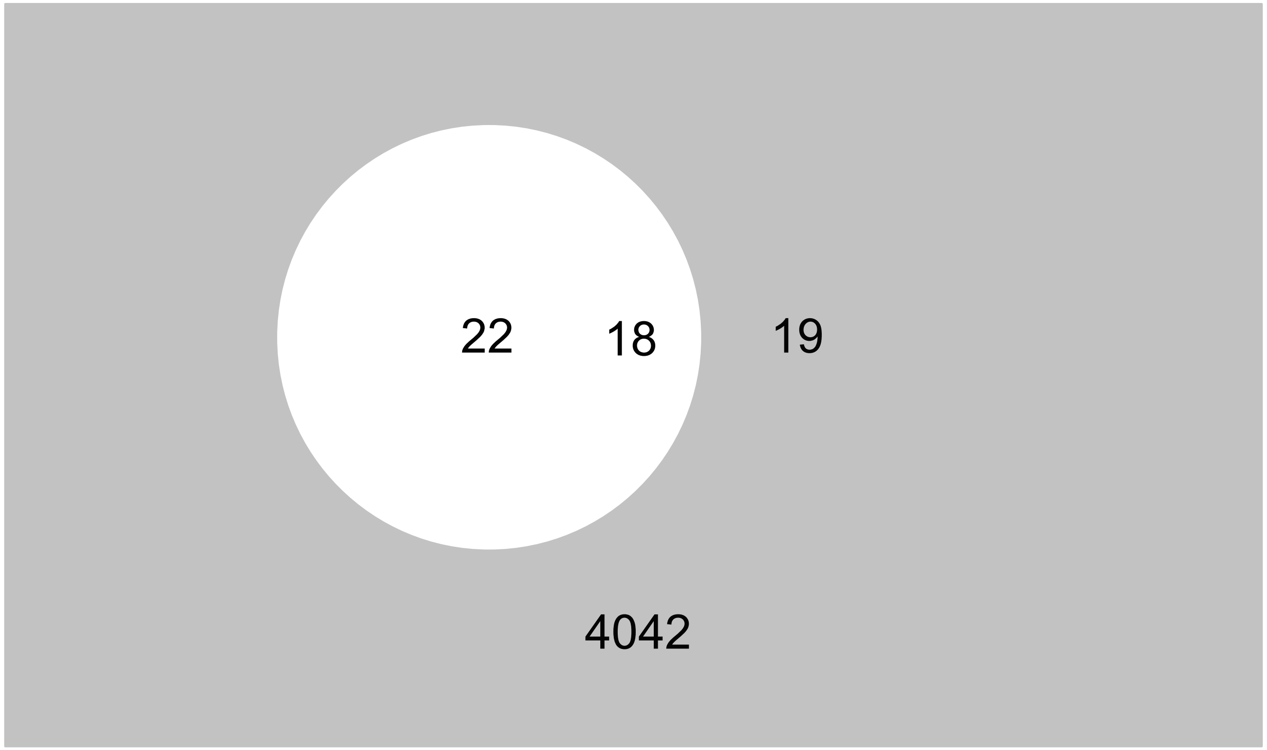
SIGNIFICANT METABOLITES

PATHWAY METABOLITES

ALL METABOLITES MEASURED

Over-representation analysis to identify pathways

What is the likelihood of seeing this exact arrangement of the data?



Calculating *P* values

- One-sided Fisher's exact test
  - 40 out of 4042 measured metabolites are significant
  - 18 out of 40 of the significant metabolites belong to pathway A
  - 37 out of 4042 metabolites are known to be associated with pathway A

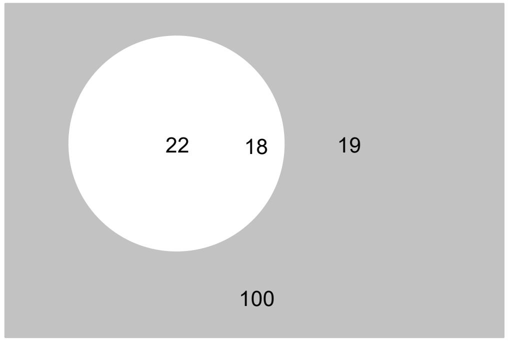
	In pathway	**Not in pathway**	
Significant metabolites	18	22	40
Non-significant metabolites	19	3983	4002
	37	4005	4042

P value = 1.45537e-28

- Hypergeometric function can also be used to calculate *P* values

Background lists

- Curated list of metabolites known to be found in tissue/cells/fluid of interest with the analytical method you're using
- *P* values will be higher and more realistic
- Ideally
  - Assign every metabolite in your sample
- Possible proxy
  - Use literature to find lists of metabolites people have assigned in similar cells/tissues/fluids, using a similar analytical platform
  - The Human Metabolome Database – http://www.hmdb.ca
    - » Filter by saliva, blood, urine, CSF, other fluids
    - » Can also get KEGG, CHeBI, BioCyc, … identifiers from HMDB



Imperial College London

Problems with mappings – lactate example

Molecule	KEGG ID	KEGG pathway(s)
S-Lactate	C00186	Glycolysis/Gluconeogensis; Pyruvate metabolism; Propanoate metabolism; Styrene metabolism; Metabolic pathways; Biosynthesis of secondary metabolites; Microbial metabolism in diverse environments; HIF-1 signalling pathway
R-Lactate	C00256	Pyruvate metabolism
Lactate	C01432	None

OPTIONS

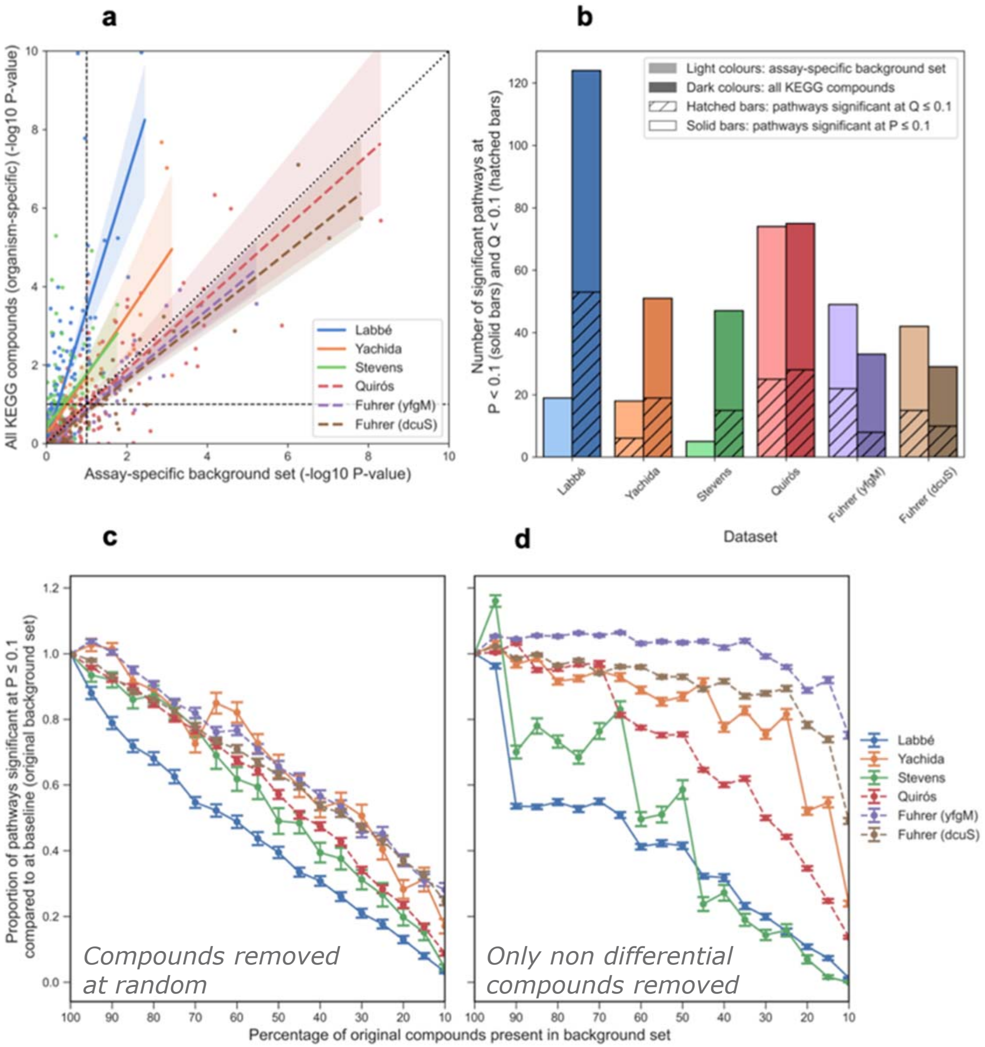
Map to lactate only	This is in no pathway
Map to R-lacate and S-lactate	Pyruvate metabolism will be biased towards being significant – as it will have 2 significant metabolites
Map to S-lactate	In this case this is the best option, but if R-lactate was in pathways without S-lactate, then this would cause problems

Non-specific background sets result in erroneously high levels of enriched pathways

***Non-specific background set**: e.g. all compounds in KEGG, often default option*

***Assay-specific background set:** all compounds annotatable in the dataset*

- *Lower p-values observed when using non-specific background set*
- *Many likely to be false-positives as highly unlikely to detect all compounds with the assay used!*
- *Larger background sets provide higher power for detection of significant pathways*
- *Ratio of differentially abundant (DA): non-DA compounds impacts the power of ORA*



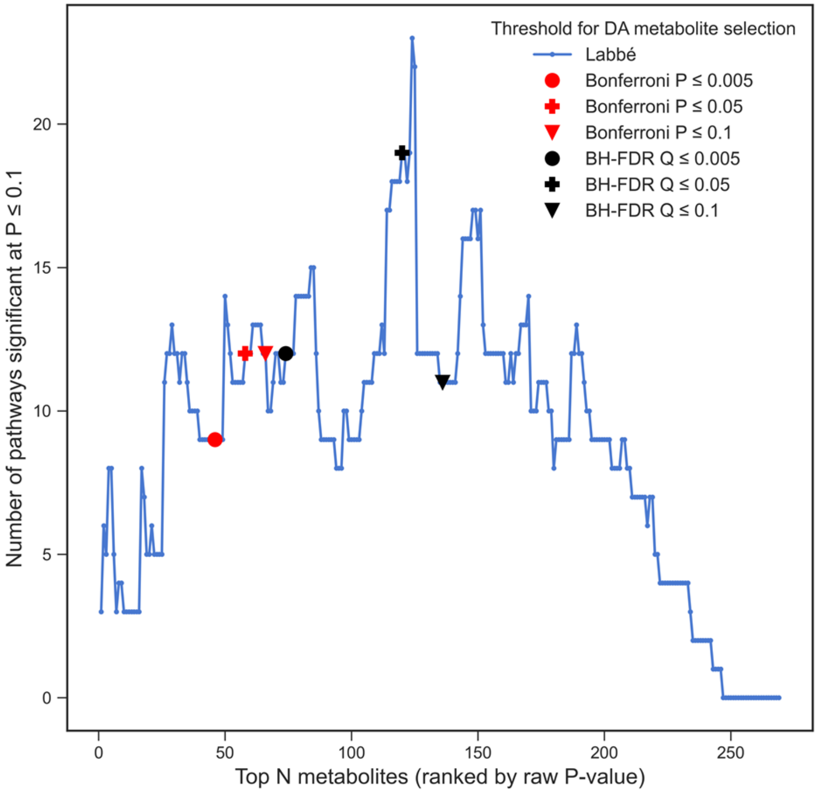
Compounds removed at random

Only non differential compounds removed

Increasing the number of differential metabolites can result in higher or lower numbers of significant pathways

Should we always use a threshold of P ≤ 0.05 to select differentially abundant metabolites?

- *Used t-tests to determine DA level of each metabolite*
- *Ranked metabolites by unadjusted P-value and added one by one to list of DA metabolites*
- *Addition of just one metabolite can result in large fluctuations in number of significant pathways*
- *Each dataset has a number of DA metabolites which yields the highest number of significant pathways (global maximum)*



Pathway database choice is key

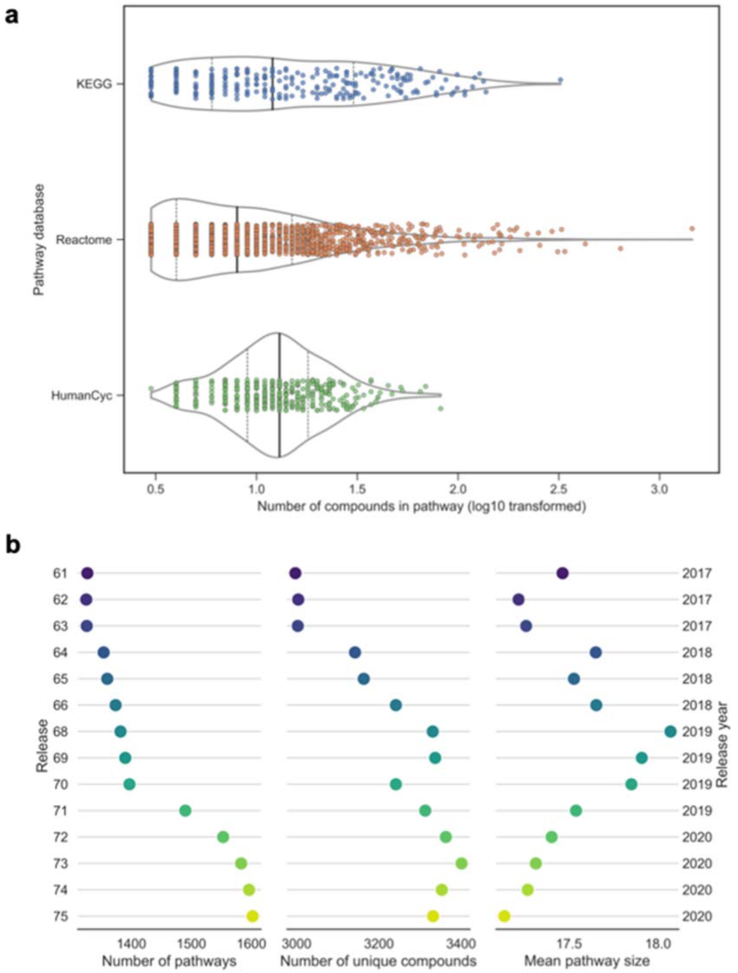
Inherent differences between pathway databases (KEGG, Reactome, BioCyc):

- *Pathway size*
- *Compounds in pathway*
- *Pathway boundaries*
- *Areas of metabolism covered*

ORA results rarely agree!

ORA results are short-lived

- *With each pathway database update, new pathways and compounds are added*
- *Existing pathways may be modified*
- *Results should be continuously updated using the latest database version*



Metabolite misidentification results in both gain and loss of truly significant pathways

- *Some level of metabolite misidentification expected in all experimental datasets*
- *Simulated misidentifications by mass and compound formula*
- *All datasets had pathway loss rate (false negative rate) and pathway gain rate (false positive rate) > 0*

Therefore, some pathways are significant purely due to misidentification, while others lost

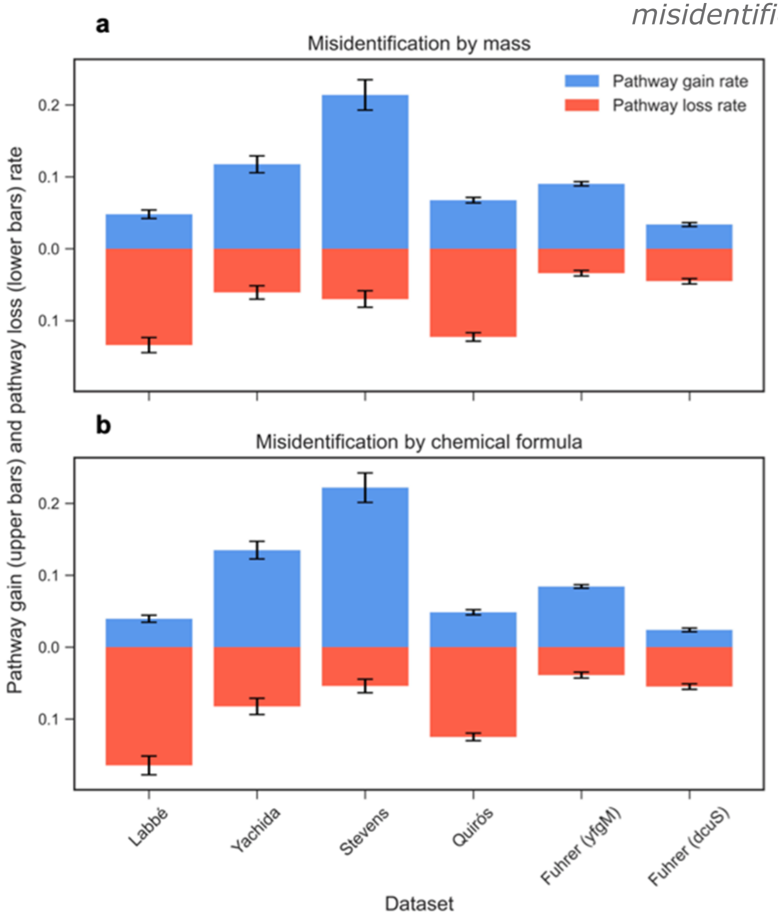
The pathway loss rate and pathway gain rate at f% metabolite misidentification are then defined as:

$$\mathit{Pathway\ loss\ rate}(A, B_f) = 1 - \frac{|A \cap B_f|}{|A|}$$

$$\mathit{Pathway\ gain\ rate}(A, B_f) = \frac{|B_f - A|}{|A|}$$

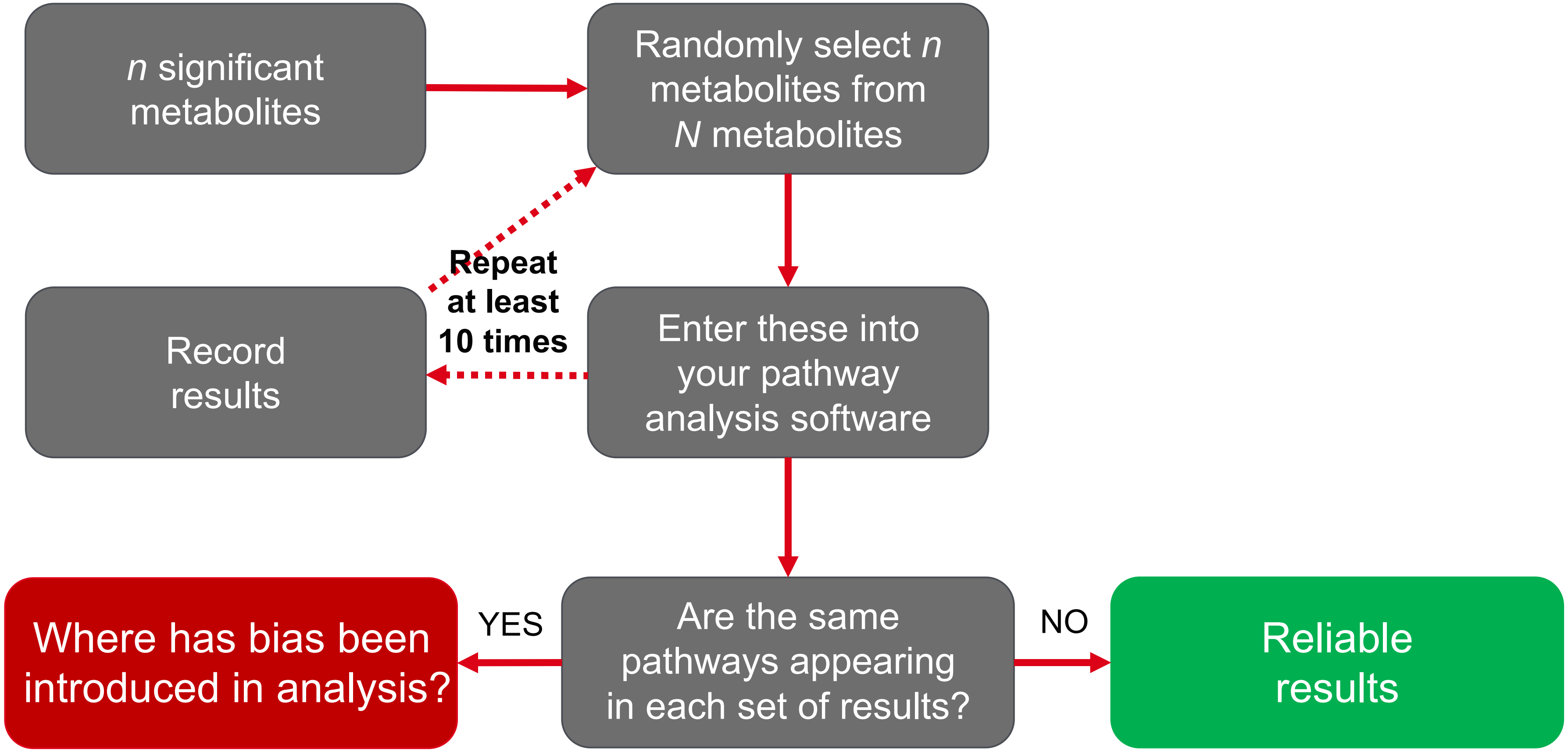
where |A| indicates the cardinality (number of elements) in the set A, and |B-A| indicates the set formed by those members of B which are not members of A.

4% simulated misidentification



Imperial College London

Checking for bias


n significant metabolites
Randomly select *n* metabolites from *N* metabolites
Enter these into your pathway analysis software
Record results
Repeat at least 10 times
Are the same pathways appearing in each set of results?
YES
Where has bias been introduced in analysis?
NO
Reliable results

Drawbacks to overrepresentation approaches

- Works from a list of significant metabolites
  - Correlation with phenotype/genotype
  - *t* tests between classes
  - Fold change (LIMMA)
  - ANOVA

Continuous measures

- We then apply a cut-off (e.g FDR < 0.05)
- All significant genes are treated equally once included in the list

Enrichment analysis

- Enrichment approaches take a value for each element measured
- No need for background list
  - All measured elements are listed with their value
- Different variants exist
  - Set enrichment analysis (similar to GSEA for genes)
  - Wilcoxon enrichment
  - Quantitative Enrichment Analysis (e.g. MSEA available via MetaboAnalyst 3.0)
- Details of the calculation used depend on the variant used

Set Enrichment Analysis

- Uses *a priori* metabolite sets that have been grouped together by their involvement in the same biological pathway
- Analyzes whether the majority of the metabolites you have identified fall in the extremes of this list
  - the top and bottom of the list correspond to the largest differences in metabolite levels between your groups
  - If the metabolite set falls at either the top (over-represented) or bottom (under-represented), it is thought to be related to group differences
- Methods assumes metabolite independence
- *P* values are calculated through repeated permutations or *t* tests

Wilcoxon Enrichment

- Used by ConsensusPathwayDB – http://consensuspathdb.org
- Takes two values per metabolite (e.g. mean control value and mean treated value)
- Tests for pathways where the difference between these are significantly different from 0, indicating pathways which are positively or negatively enriched

Quantitative Enrichment Analysis

- An adaptation of the globaltest algorithm for gene enrichment
  - “*global test is meant for data sets in which many covariates (or features) have been measured for the same subjects, together with a response variable, e.g. a class label, a survival time or a continuous measurement. The global test can be used on a group (or subset) of the covariates, testing whether that group of covariates is associated with the response variable.*”
- For each metabolite a *Q* value is calculated based on the average of the squared covariance between the metabolite values and the outcome value
- For each pathway the *Q*-stat is the average *Q* value of genes in this pathway
- *P* values are calculated based on the asymptotic distribution expected of the *Q*-stats

Over-representation vs Enrichment analysis

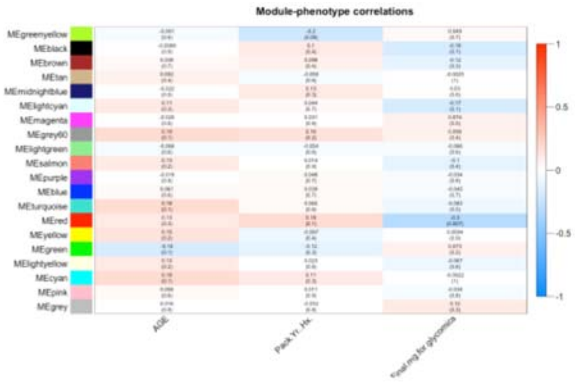
	Enrichment analysis	Over-representation
Can examine any pathways / processes / sets of genes or metabolites of interest	✓	✓
Takes a list of metabolites or genes	✗	✓
Uses a continuous value for each metabolite of gene	✓	✗
Can combine metabolite and transcript data	✓	✓
P values generated by…	Repeated permutations or exact mathematical calculations	Exact mathematical calculations

Web resources for metabolic pathway analysis

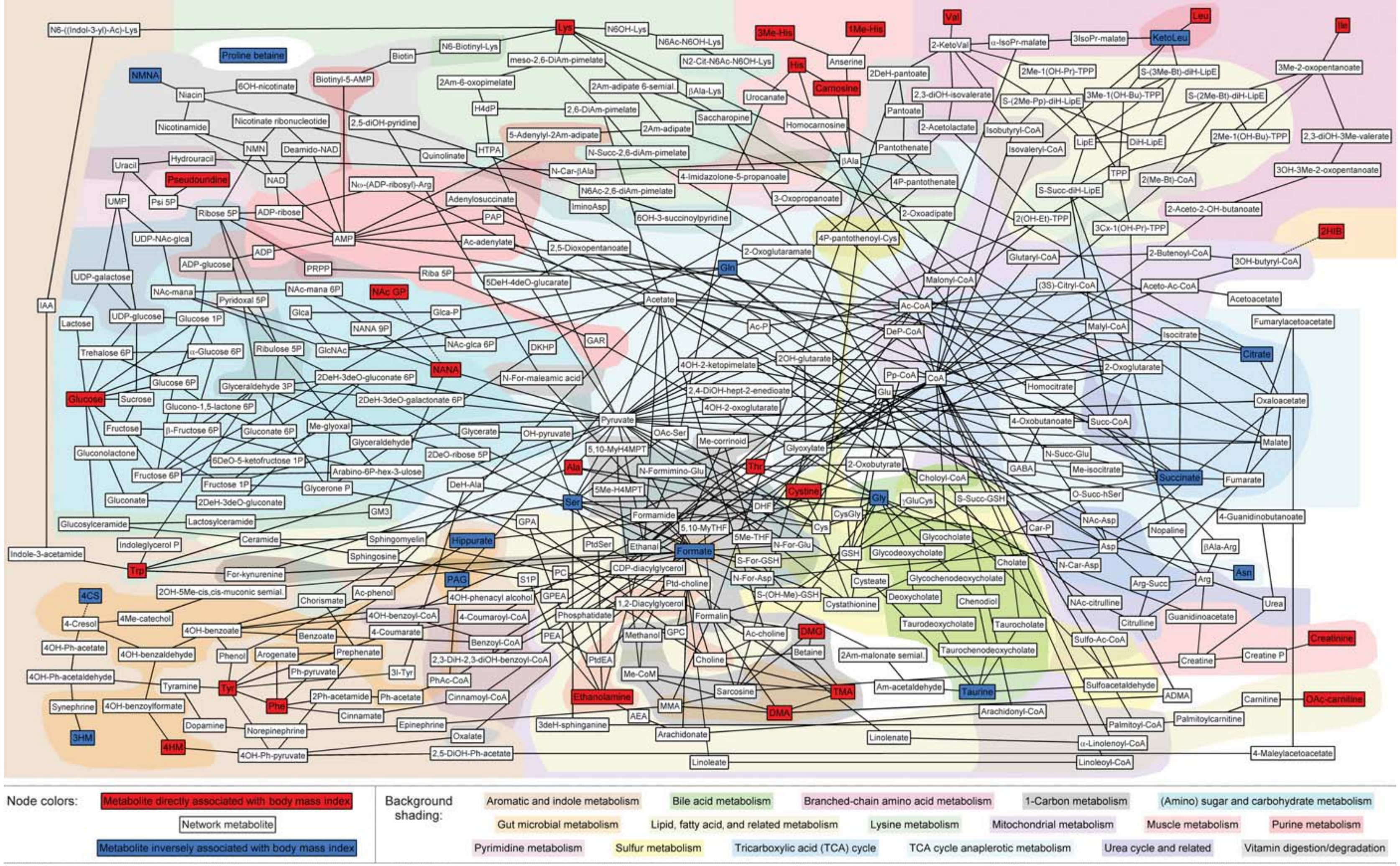
- MetaboAnalyst 3.0 (McGill University)
  - http://www.metaboanalyst.ca/faces/home.xhtml
  - Enrichment analyses
- BINChE
  - http://www.ebi.ac.uk/chebi/tools/binche/
  - Enrichment analysis of small molecules (uses ChEBI IDs – MetaboAnalyst has a tool to convert IDs)
- Functional annotation of a metabolite list
  - http://cpdb.molgen.mpg.de/CPDB/mfct_annot
- Other resources
  - Booth *et al.* (2013). Computational tools for the secondary analysis of metabolomics experiments. *Comput Struct Biotechnol* **4**, e201301003. doi:10.5936/csbj.201301003.

R packages for metabolite mapping/analyses

- KEGGREST
  - https://bioconductor.org/packages/release/bioc/html/KEGGREST.html
- Pathview
  - http://pathview.r-forge.r-project.org
  - Allows you to overlay your data onto KEGG pathways
  - Can also be used for gene expression data
- PAPi
  - https://www.bioconductor.org/packages/release/bioc/manuals/PAPi/man/PAPi.pdf
  - Predict metabolic pathway activity based on metabolomics data
- Grinn
  - http://kwanjeeraw.github.io/grinn/
  - Network analysis
  - Outputs can be imported into Cytoscape
  - Incorporates data from KEGG, SMPDB, HMDB, REACTOME, CheBI, UniProt and ENSEMBL



MetaboNetworks: Matlab resource



Posma *et al.* (2014). *Bioinformatics* **30**, 893–895; Elliott *et al.* (2015). *Sci Transl Med* **7**, 285ra62.

Imperial College London

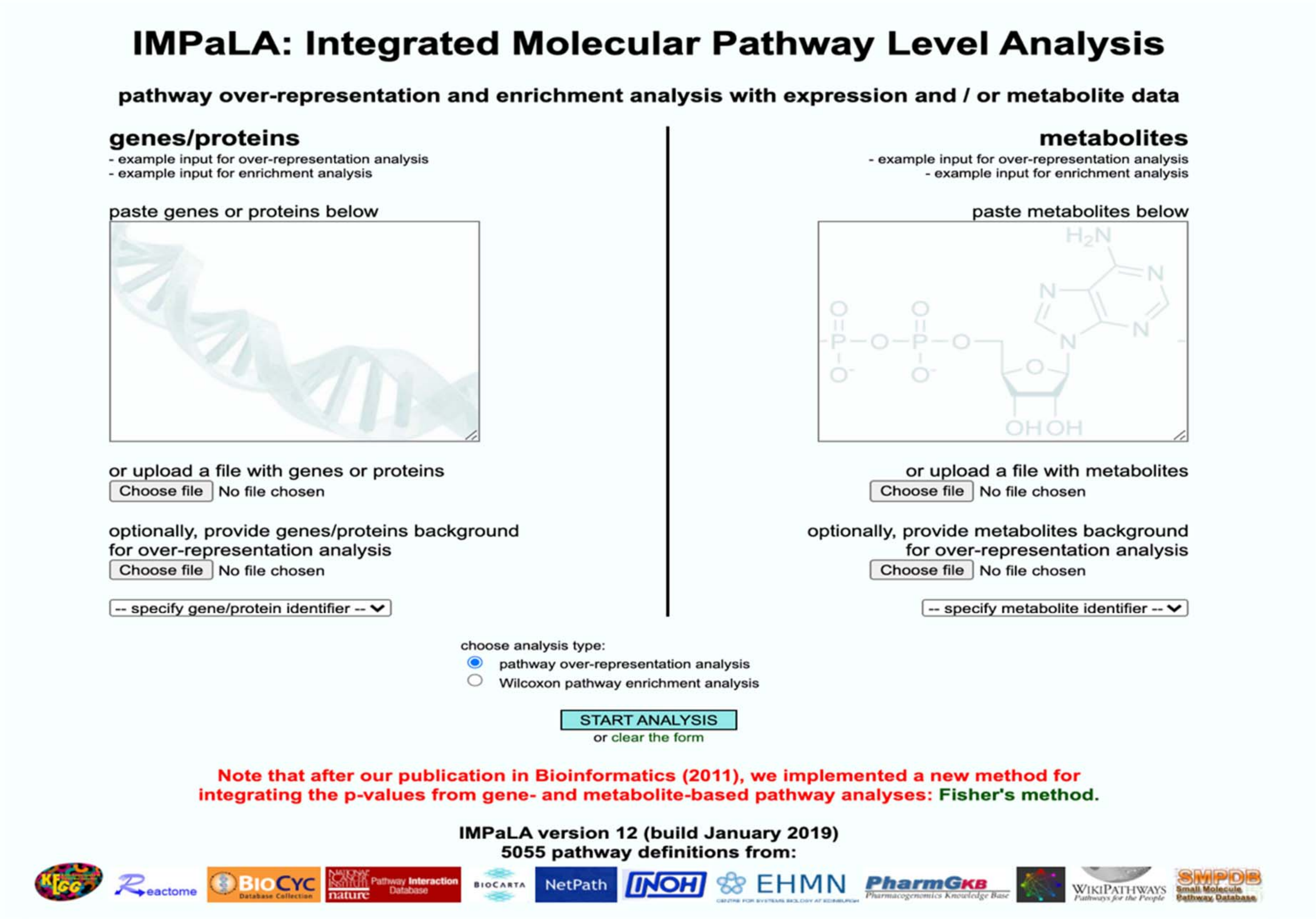
Integration of transcriptomic/proteomic and metabolite data

- IMPaLA: Integrated Molecular Pathway Level Analysis
  - Pathway over-representation and enrichment analysis with expression and/or metabolite data
  - http://impala.molgen.mpg.de
- 3Omics: a web based systems biology visualization tool for integrating human transcriptomic, proteomic and metabolomic data
  - http://3omics.cmdm.tw
- Ingenuity Pathway Analysis
  - Commercial software
- Visualization tools
  - https://omictools.com/transcriptomic-and-metabolomic-data-integration-category

Imperial College London

http://impala.molgen.mpg.de/

IMPaLA Pathway Analysis Tool



IMPaLA: Integrated Molecular Pathway Level Analysis

pathway over-representation and enrichment analysis with expression and / or metabolite data

genes/proteins

- example input for over-representation analysis
- example input for enrichment analysis

paste genes or proteins below

or upload a file with genes or proteins

Choose file No file chosen

optionally, provide genes/proteins background for over-representation analysis

Choose file No file chosen

-- specify gene/protein identifier --

metabolites

- example input for over-representation analysis
- example input for enrichment analysis

paste metabolites below

or upload a file with metabolites

Choose file No file chosen

optionally, provide metabolites background for over-representation analysis

Choose file No file chosen

-- specify metabolite identifier --

choose analysis type:

● pathway over-representation analysis

○ Wilcoxon pathway enrichment analysis

START ANALYSIS

or clear the form

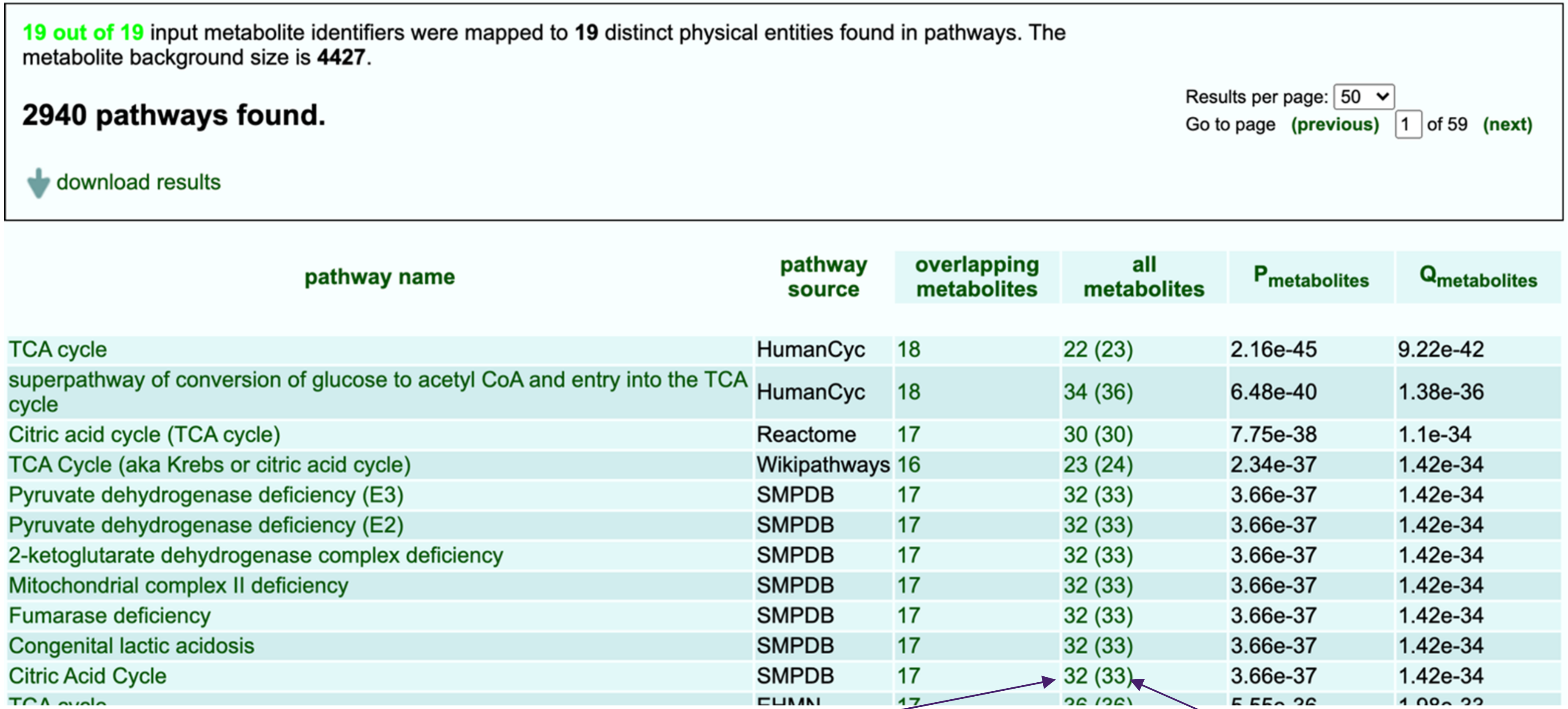
Note that after our publication in Bioinformatics (2011), we implemented a new method for integrating the p-values from gene- and metabolite-based pathway analyses: Fisher's method.

IMPaLA version 12 (build January 2019)
5055 pathway definitions from:

KEGG
Reactome
BioCyc Database Collection
National Cancer Institute Pathway Interaction Database nature
BIOCARTA
NetPath
INOH
EHMN
PharmGKB Pharmacogenomics Knowledge Base
WikiPathways Pathways for the People
SMPDB Small Molecule Pathway Database

IMPALA - Results

Mapping results



19 out of 19 input metabolite identifiers were mapped to **19** distinct physical entities found in pathways. The metabolite background size is **4427**.

2940 pathways found.

Results per page: 50
Go to page **(previous)** 1 of 59 **(next)**

download results

pathway name	pathway source	overlapping metabolites	all metabolites	$P_{metabolites}$	$Q_{metabolites}$
TCA cycle	HumanCyc	18	22 (23)	2.16e-45	9.22e-42
superpathway of conversion of glucose to acetyl CoA and entry into the TCA cycle	HumanCyc	18	34 (36)	6.48e-40	1.38e-36
Citric acid cycle (TCA cycle)	Reactome	17	30 (30)	7.75e-38	1.1e-34
TCA Cycle (aka Krebs or citric acid cycle)	Wikipathways	16	23 (24)	2.34e-37	1.42e-34
Pyruvate dehydrogenase deficiency (E3)	SMPDB	17	32 (33)	3.66e-37	1.42e-34
Pyruvate dehydrogenase deficiency (E2)	SMPDB	17	32 (33)	3.66e-37	1.42e-34
2-ketoglutarate dehydrogenase complex deficiency	SMPDB	17	32 (33)	3.66e-37	1.42e-34
Mitochondrial complex II deficiency	SMPDB	17	32 (33)	3.66e-37	1.42e-34
Fumarase deficiency	SMPDB	17	32 (33)	3.66e-37	1.42e-34
Congenital lactic acidosis	SMPDB	17	32 (33)	3.66e-37	1.42e-34
Citric Acid Cycle	SMPDB	17	32 (33)	3.66e-37	1.42e-34
TCA cycle	EHMN	17	36 (36)	5.55e-36	1.98e-33

No. background list metabolites in pathway

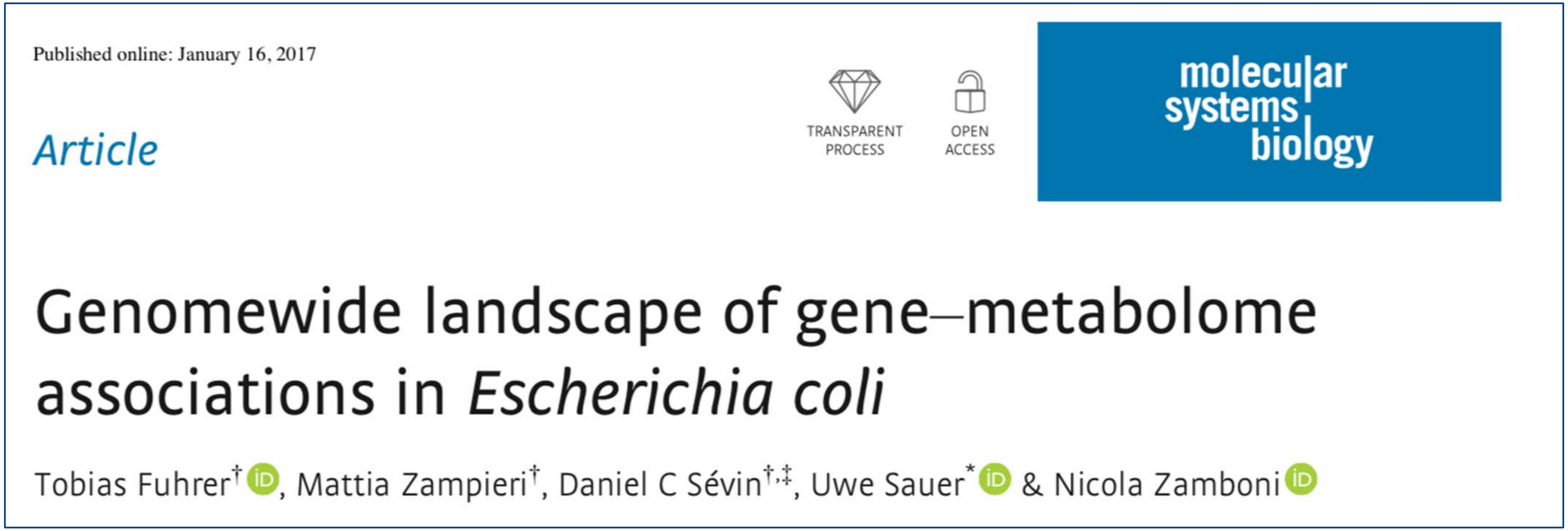
No. metabolites in pathway

Limitations of pathway analysis approaches

1. Do they work?

Limitations of pathway analysis approaches

1. Do they work?



Published online: January 16, 2017

Article

TRANSPARENT PROCESS

OPEN ACCESS

molecular systems biology

Genomewide landscape of gene–metabolome associations in *Escherichia coli*

Tobias Fuhrer[†], Mattia Zampieri[†], Daniel C Sévin[†,‡], Uwe Sauer[*] & Nicola Zamboni

“Beyond expected metabolic changes in the proximity to abolished enzyme activities, the association map reveals a largely unknown landscape of gene–metabolite interactions that are not represented in metabolic models.”

Limitations of pathway analysis approaches

1. Do they work?

2. Not always straightforward to interpret
   - Can go from a list of metabolites to a list of pathways
   - Can feel like just promoting ignorance to a new level

Limitations of pathway analysis approaches

1. Do they work?

2. Not always straightforward to interpret
   - Can go from a list of metabolites to a list of pathways
   - Can feel like just promoting ignorance to a new level

3. Still a lot of incompleteness in our understanding of factors that can lead to inaccurate or misleading results using pathway approaches
   - Effects of metabolite misassignment
   - Correct metabolic network may not be available
   - Effects of different pathway definitions

Limitations of pathway analysis approaches

1. Do they work?

2. Not always straightforward to interpret
   - Can go from a list of metabolites to a list of pathways
   - Can feel like just promoting ignorance to a new level

3. Still a lot of incompleteness in our understanding of factors that can lead to inaccurate or misleading results using pathway approaches
   - Effects of metabolite misassignment
   - Correct metabolic network may not be available
   - Effects of different pathway definitions

4. In general, **should not be used** as a tool for saying “these are the pathways that are up / down regulated”!
   - Guide to further analysis and/or experimentation

Conclusions and best practice for ORA

Suggested best practice guidelines

- Specify a realistic background set i.e., all the compounds which were detectable using the analytical platform used in the experiment.
- Use an organism-specific pathway set if the organism is supported by the pathway database.
- Perform ORA using multiple pathway databases and derive a consensus pathway signature using the results
- Use multiple-testing correction to select both DA metabolites and, where feasible, significant pathways.

Suggested minimum reporting criteria

- The statistical test/approach used for pathway analysis (e.g. Fisher’s exact test)
- The tool (and version) used to perform ORA.
- The pathway database, the corresponding compound identifier type (e.g. KEGG, ChEBI, BioCyc, etc.), its release number and which organism-specific pathway set was used (if any).
- Which compounds form the background set.

Wieder, C., C. Frainay, N. Poupin, P. Rodríguez-Mier, F. Vinson, J. Cooke, R. P. Lai, J. G. Bundy, F. Jourdan and T. Ebbels (2021). "Pathway analysis in metabolomics: pitfalls and best practice for the use of over-representation analysis." bioRxiv: 2021.2005.2024.445406.

Acknowledgements

Rachel Cavill

Lesley Hoyles

Jake Bundy

Cecilia Wieder