



HAL
open science

Pathway analysis in metabolomics: pitfalls and best practice 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

► To cite this version:

Cecilia Wieder, Clément Frainay, Nathalie Poupin, Pablo Rodríguez-Mier, Florence Vinson, et al.. Pathway analysis in metabolomics: pitfalls and best practice for the use of over-representation analysis. 2021. hal-03240376

HAL Id: hal-03240376

<https://hal.science/hal-03240376>

Preprint submitted on 28 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Pathway analysis in metabolomics: pitfalls and best** 2 **practice for the use of over-representation analysis**

3 4 **Authors:**

5 Cecilia Wieder ¹, Clément Frainay ⁴, Nathalie Poupin ⁴, Pablo Rodríguez-Mier ⁴, Florence Vinson ⁴, Juliette
6 Cooke ⁴, Rachel PJ Lai ³, Jacob G Bundy ², Fabien Jourdan ^{4,5}, Timothy Ebbels* ¹

7
8 ¹ Section of Bioinformatics, Division of Systems Medicine, Department of Metabolism, Digestion, and
9 Reproduction, Faculty of Medicine, Imperial College London, London SW7 2AZ, UK

10 ² Section of Biomolecular Medicine, Division of Systems Medicine, Department of Metabolism, Digestion,
11 and Reproduction, Faculty of Medicine, Imperial College London, London SW7 2AZ, UK

12 ³ Department of Infectious Disease, Faculty of Medicine, Imperial College London, London SW7 2AZ, UK

13 ⁴ Toxalim (Research Centre in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS,
14 31300 Toulouse, France

15 ⁵ MetaToul-MetaboHUB, National Infrastructure of Metabolomics and Fluxomics, Toulouse, France.

16
17 Corresponding author email address: t.ebbels@imperial.ac.uk

18
19 **Key words:** pathway analysis, over-representation analysis, pathway enrichment, model interpretation,
20 bioinformatics

21

22

23 Abstract

24 Over-representation analysis (ORA) is one of the commonest pathway analysis approaches used
25 for the functional interpretation of metabolomics datasets. Despite the widespread use of ORA
26 in metabolomics, the community lacks guidelines detailing its best-practice use. Many factors
27 have a pronounced impact on the results, but to date their effects have received little systematic
28 attention in the field. We developed *in-silico* simulations using five publicly available datasets
29 and illustrated that changes in parameters, such as the background set, differential metabolite
30 selection methods, and pathway database choice, could all lead to profoundly different ORA
31 results. The use of a non-assay-specific background set, for example, resulted in large numbers
32 of false-positive pathways. Pathway database choice, evaluated using three of the most popular
33 metabolic pathway databases: KEGG, Reactome, and BioCyc, led to vastly different results in
34 both the number and function of significantly enriched pathways. Metabolomics data specific
35 factors, such as reliability of compound identification and assay chemical bias also impacted
36 ORA results. Simulated metabolite misidentification rates as low as 4% resulted in both gain of
37 false-positive pathways and loss of truly significant pathways across all datasets. Our results
38 have several practical implications for ORA users, as well as those using alternative pathway
39 analysis methods. We offer a set of recommendations for the use of ORA in metabolomics,
40 alongside a set of minimal reporting guidelines, as a first step towards the standardisation of
41 pathway analysis in metabolomics.

42

43

44 Author summary

45 Metabolomics is a rapidly growing field of study involving the profiling of small molecules
46 within an organism. It allows researchers to understand the effects of biological status (such as
47 health or disease) on cellular biochemistry, and has wide-ranging applications, from biomarker
48 discovery and personalised medicine in healthcare to crop protection and food security in
49 agriculture. Pathway analysis helps to understand which biological pathways, representing
50 collections of molecules performing a particular function, are involved in response to a disease
51 phenotype, or drug treatment, for example. Over-representation analysis (ORA) is perhaps the
52 most common pathway analysis method used in the metabolomics community. However, ORA
53 can give drastically different results depending on the input data and parameters used. In this
54 work, we have established the effects of these factors on ORA results using computational
55 simulations applied to five real-world datasets. Based on our results, we offer the research
56 community a set of best-practice recommendations applicable not only to ORA but also to other
57 pathway analysis methods to help ensure the reliability and reproducibility of results.

58 Introduction

59 Pathway analysis (PA) plays a vital role in the interpretation of high-dimensional molecular
60 data. It is used to find associations between pathways, which represent collections of molecular
61 entities sharing a biological function, and a phenotype of interest [1]. Based on existing
62 knowledge of biological pathways, molecular entities such as genes, proteins, and metabolites
63 can be mapped onto curated pathway databases, which aim to represent how these entities
64 collectively function and interact in a biological context [2]. Originally developed for the
65 interpretation of transcriptomic data, PA has now become a popular method for analysing
66 metabolomics data [3,4]. There are several inherent differences between transcriptomic and
67 untargeted metabolomics data, however, which must be considered when performing PA with
68 metabolites. Firstly, metabolomics datasets tend to cover a much lower proportion of the total
69 metabolome than transcriptomic datasets do of the genome. Hence, metabolomics datasets tend
70 to contain far fewer metabolites than transcripts found in transcriptomic datasets. Secondly,
71 mapping compounds to pathways is not as straightforward as the equivalent mapping with
72 genes and proteins, and there is often a significant level of uncertainty surrounding metabolite
73 identification, both with respect to structures and database identifiers in any metabolomics
74 dataset.

75 There are several methods for PA, which can be classed into three broad categories:
76 over-representation analysis (ORA), functional class scoring (FCS), and topology-based methods
77 [5]. In this paper, we focus on ORA, one of the most mature and widely used methods of PA both
78 within the metabolomics [6,7] and transcriptomics [8] communities. ORA has found widespread
79 use in the identification of significantly impacted pathways in numerous metabolomics studies
80 [9–13]. It works by identifying pathways or metabolite sets that have a higher overlap with a set
81 of molecules of interest than expected by chance. The approach typically uses Fisher's exact test
82 to examine the null hypothesis that there is no association between the compounds in the
83 pathway and the outcome of interest [14].

84 To perform ORA, three essential inputs are required: a collection of pathways (or custom
85 metabolite sets), a list of metabolites of interest, and a background or reference set of
86 compounds. Pathway sets can be obtained freely from several databases, for example, the Kyoto
87 Encyclopaedia of Genes and Genomes (KEGG) [15], Reactome [16], BioCyc [17], or MetExplore
88 [18] databases, or commercial counterparts such as the Ingenuity PA (IPA) database [19]. The
89 list of metabolites of interest is generated by the user, most commonly obtained from
90 experimental data and by using a statistical test to find metabolites whose levels are associated
91 with an outcome (e.g. disease vs. control), and selecting a threshold (e.g. on the p -values) to
92 filter the list. The background set contains all molecules which can be detected in the
93 experiment. For example in transcriptomic studies, this consists of all genes or transcripts
94 which can be quantified. In targeted metabolomics, the background would contain all
95 metabolites detectable by the assay; in untargeted metabolomics, all annotatable metabolites. p -
96 values for each pathway are calculated using a right-tailed Fisher's exact test based on the
97 hypergeometric distribution. The probability of observing at least k metabolites of interest in a
98 pathway by chance is given by equation 1:

$$P(X \geq k) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}} \quad (1)$$

99
100 where N is the size of background set, n denotes the number of metabolites of interest, M is the
101 number of metabolites in the background set annotated to the i^{th} pathway, and k gives the
102 number of metabolites of interest which are annotated to the i^{th} pathway. A visual
103 representation of ORA is shown in Fig 1. Finally, multiple testing correction (to allow for the fact
104 that, typically, the calculation is made for multiple pathways, rather than just one pathway) can
105 be applied to obtain a final list of significantly enriched pathways (SEP).

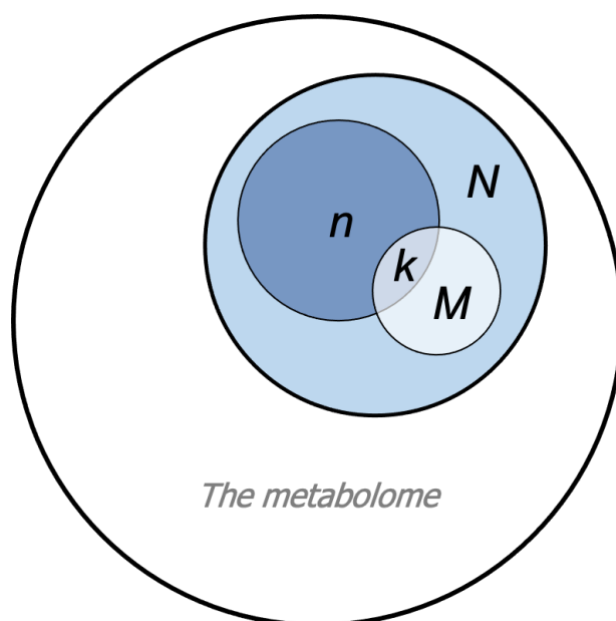


Fig 1: Over Representation Analysis (ORA) Venn diagram representing ORA parameters corresponding to Equation 1. N represents compounds forming the background set, which covers part of the full metabolome. M represents compounds in the pathway of interest. n represents compounds of interest (i.e. differentially abundant metabolites), and k represents the overlap between the list of compounds of interest and compounds in the pathway.

106

107 Despite the widespread use of ORA in metabolomics [4] the community lacks a set of
108 guidelines detailing its best use practices. Varying ORA inputs can result in large changes to
109 outputs, which raises the question of how such parameters should be chosen in order to obtain
110 the most reliable results. Moreover, as ORA was initially developed for use with transcriptomic
111 data and later adapted for use on metabolomic data, there are certain considerations
112 particularly important to metabolomics that may affect ORA results, such as the level of
113 compound identification. Our aim here, therefore, is to investigate the robustness of ORA in
114 typical metabolomics analysis, by examining the impact of varying the input data and
115 parameters. The factors examined are: the background set, selection of significant metabolites,
116 pathway database choice, organism-specific pathway sets, metabolite misidentification, and
117 chemical bias of the assay. Using five experimental datasets, we vary the inputs, each time
118 comparing to the original or standard settings, thus demonstrating the effect of these choices on
119 the output lists of significant pathways. Based on our modelling, we offer a set of
120 recommendations for ORA applied to metabolomics data, as well as a set of minimal reporting

121 recommendations which we hope can help contribute to future best-practice guidelines. It is
122 hoped that this research will promote a deeper understanding of the use ORA in metabolomics,
123 allowing researchers to better interpret their data in a pathway context.

124

125 Results

126 **Nonspecific background sets result in erroneously high levels of enriched pathways**

127 First, we examined several factors which are common to all ORA applications, beginning with
128 the background set. Five experimental datasets have been used throughout this work (Table 1,
129 see Methods), on which the following results are based.

130 The term background set (of size N , see Eqn. 1) is used to describe all the compounds
131 identifiable using a particular assay. For example, for a targeted approach, this corresponds to
132 the compounds assayed; for an untargeted approach, this corresponds to all annotatable
133 compounds. For mass-spectrometry (MS) studies, the background set would ideally refer to the
134 library of chemical standards used in metabolite annotation. Despite being a key parameter of
135 ORA, specifying the background set is an often-overlooked step. The use of a generic, non-assay-
136 specific background set implies that non-observed compounds are considered in the Fisher's
137 exact test formula, which, by definition, will always be absent from the list of metabolites of
138 interest (of size n , Eqn. 1). We investigated the effect of using a nonspecific background set,
139 consisting of all compounds annotated to at least one KEGG pathway, compared to an assay-
140 specific background set, consisting only of compounds identified and present in the abundance
141 matrix of each dataset. The nonspecific KEGG human background set contained considerably
142 more compounds (3373) than any of the example datasets.

143 A clear discrepancy was observed in many of the pathway p -values when using the
144 nonspecific vs. specific background set (Fig. 2a). A greater proportion of pathways had lower p -
145 values when using the nonspecific background set than the specific version. Interestingly, some
146 pathways were significant at $p \leq 0.1$ when using one background set but were not significant
147 using the other, as evident in the upper right and lower left quadrants of Fig 2a. We also

148 investigated the number of significantly enriched pathways (SEP) before and after multiple
149 testing correction (using Benjamini-Hochberg False Discovery Rate (BH FDR)) when using the
150 two different background sets (Fig. 2b). When using the specific background set, there were far
151 fewer SEPs at $p \leq 0.1$ (solid bars) and $q \leq 0.1$ (hatched bars) than there were using the
152 nonspecific background set. Surprisingly, when using the specific background set (lighter
153 coloured bars), two datasets contained no pathways which remained significant after multiple-
154 testing correction (no hashed bars). Since our further analyses require several pathways to be
155 enriched in the original datasets, we decided to use a significance threshold corresponding to an
156 uncorrected p-value of ≤ 0.1 . While we do not recommend this threshold in practice as it is
157 relatively liberal, this approach allowed us to demonstrate the characteristic behaviour of ORA
158 across a wide range of datasets.

159 A key difference between the specific and nonspecific background sets used in the
160 simulations in Fig. 2 is the number of compounds they each contain. For the human datasets
161 (Yachida, Stevens, and Quirós) for example, the nonspecific background set contained a total of
162 3373 unique compounds, whereas the specified background sets for these datasets ranged in
163 size from 286 to 1110 compounds. It is therefore reasonable to ask whether the changes seen in
164 Fig 2a and b could be due to the size of the background sets. Accordingly, we investigated how
165 the size of the background set affects ORA results. In Fig 2c, we simulated a reduction in the
166 number of compounds identified in the experiment and identify differentially abundant (DA)
167 metabolites based on the compounds in the reduced background set. This could also reflect the
168 differences in the number of metabolites identifiable on different platforms, for example, MS
169 and NMR assays. In Fig 2d, we aimed to demonstrate how changing the number of compounds
170 in the background set but keeping the number of DA metabolites static affects the number of
171 SEP (hence changing the ratio of DA compounds to background set compounds). Both removal
172 of compounds at random and non-DA compounds from the background set resulted in a
173 decrease in the proportion of SEP ($p \leq 0.1$) as compared to using 100% of the compounds in the
174 background set. Reduction of the background set at random (Fig. 2c) resulted in a steady

175 decrease in the number of significant pathways, as DA or non-DA compounds may be removed
176 and the new list of DA metabolites is calculated based on the reduced background set.
177 Reduction of the background set without removal of the original DA metabolites resulted in a
178 much more variable decline in the number of significant pathways (Fig. 2d). Datasets that had
179 larger background sets to begin with, such as Fuhrer et al., appeared to be the least affected by
180 the background set reduction. This is likely attributed to the fact that even when the reduced
181 background set contained just 10% of the original compounds, it still contained over 240
182 metabolites. The trends observed in Fig. 2d also imply that a higher ratio of background set
183 compounds to DA compounds provides more power in detecting SEPs.

184

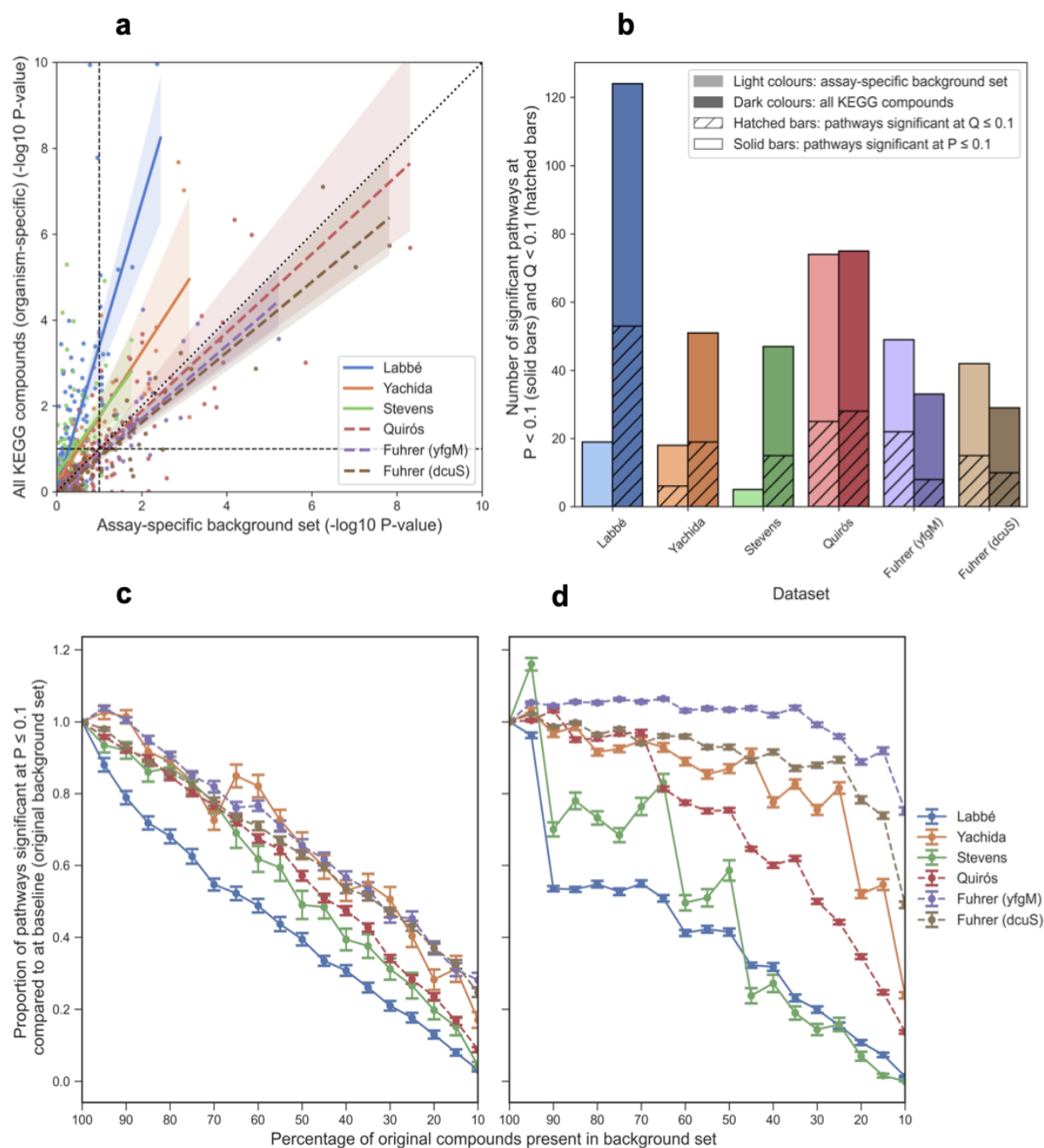


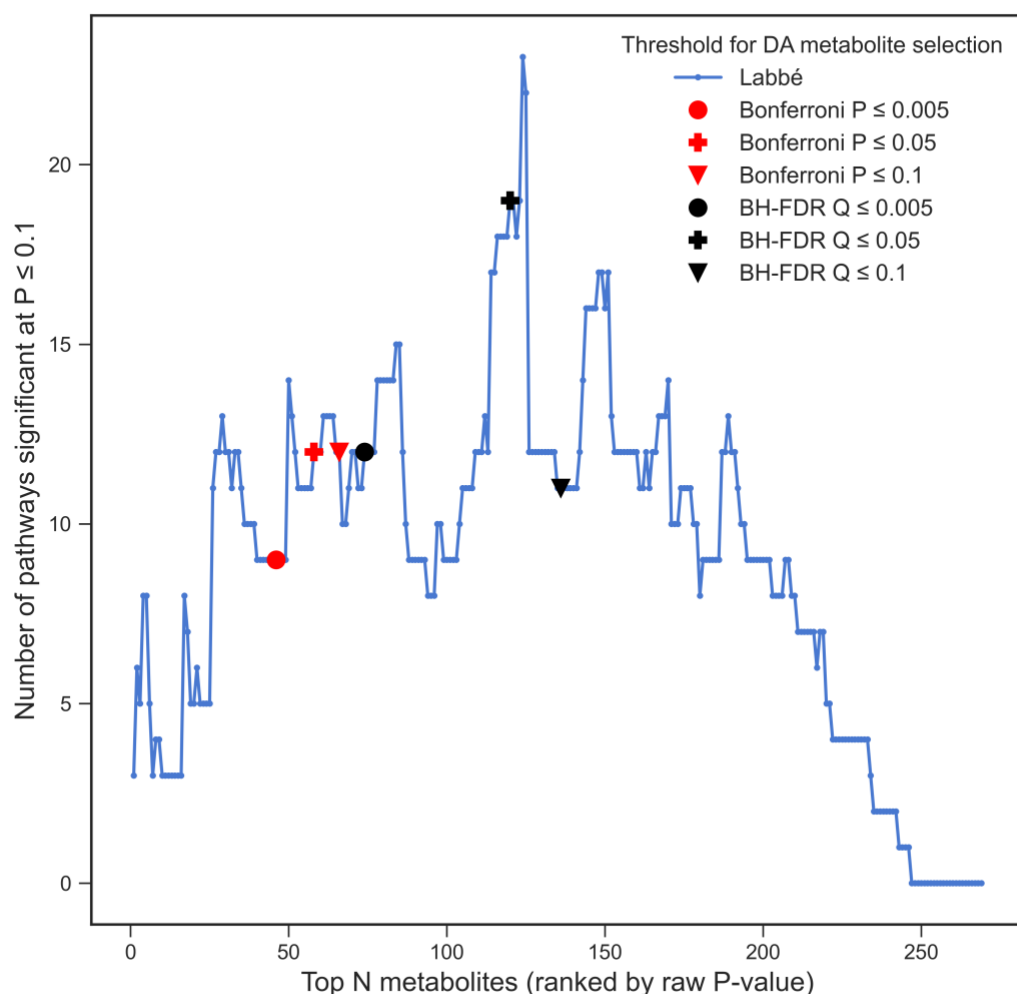
Fig 2: Effect of background set. **a** Scatter plot of $-\log_{10}$ p-values of pathways when using an assay-specific background set consisting of all measurable compounds in each dataset (x-axis) compared to using a non-specific background set containing of all compounds annotated to at least one KEGG pathway (y-axis). Dashed black lines represent a p-value threshold equivalent to $p = 0.1$. Regression lines are shown with shading representing the 95% confidence interval. **b** Number of pathways significant at $p \leq 0.1$ (solid bars) and the number of pathways significant at $q < 0.1$ (hatched bars, BH FDR correction). Datasets are ordered by number of compounds mapping to KEGG pathways. **c and d** The effect of reducing the size of the background set. **c** Compounds were removed from the background set at random and DA metabolites were identified based on the modified background set. **d** Only non-DA compounds were removed from the background set at random. In all panels a, c & d, dashed lines represent datasets where no chromatography/electrophoresis was used. Error bars represent standard error of the mean.

185

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

189 The list of compounds of interest is a key parameter of ORA, as any compound falling below the
190 significance threshold will not be able to contribute to the enrichment of a pathway. Methods
191 used to select DA metabolites typically rely on p -values or q -values derived from a statistical
192 test, for example when comparing metabolite abundances between study groups, or regression-
193 based approaches for continuous outcomes. An threshold such as $q \leq 0.05$ is often used to select
194 DA metabolites, however, as with all hypothesis testing this is an arbitrary choice. Furthermore,
195 in untargeted metabolomics, hundreds or thousands of metabolites are often profiled and
196 therefore multiple testing correction is essential. We therefore investigated the effect of using
197 varying significance levels and different multiple correction testing approaches to select
198 metabolites of interest on ORA results. To this end, DA compound lists of increasing length were
199 constructed by adding compounds, from lowest p -value to highest, one at a time. ORA was
200 performed following the addition of each compound to the DA list. The number of SEPs detected
201 using a DA list corresponding to Bonferroni adjusted p -values and BH FDR q -values at
202 thresholds of 0.005, 0.05, and 0.1 was also determined. Note that here, we are discussing the
203 significance level relating to selection of DA metabolites (the first step of ORA), not pathways
204 (second step of ORA). Fig 3 shows an example of this procedure on the Labbé et al. dataset. Plots
205 for all datasets are shown in Fig S1. With the addition of each metabolite to the DA list, the
206 number of SEPs tended to increase to a global maximum, followed by a decrease to zero where
207 the DA list consisted of the entire background set. Several fluctuations can be observed as local
208 minima and maxima in Fig. 3, demonstrating that the addition of just a single compound can
209 have a pronounced effect on the number of SEP. As expected, the list of DA metabolites
210 determined by Bonferroni correction at varying alpha thresholds resulted in fewer significant
211 pathways than using BH FDR correction. Generally, higher alpha thresholds resulted in more DA
212 metabolites and hence more significant pathways. In the case of selecting metabolites based on
213 BH FDR q -values however, more significant pathways were obtained using $\alpha \leq 0.05$ than $\alpha \leq$

214 0.005 or ≤ 0.1 . In summary, the addition of DA metabolites in order of significance will always
215 result in an increase, followed by a decrease in the number of significant pathways. Thus, it is
216 critical for practitioners to understand where their chosen significance threshold lies in this
217 overarching trend.



218
219 **Fig 3: Number of DA metabolites.** The effect of the number of DA metabolites in the list of metabolites of interest on the
220 number of significant pathways ($p \leq 0.1$) in the Labbé et al. dataset. Results corresponding to Bonferroni thresholds are
221 denoted by red markers while those corresponding to BH FDR thresholds are denoted by black markers. Marker shape
222 (circle, cross, or triangle) represents the adjusted p-value threshold for DA metabolite selection (0.005, 0.05, and 0.1
223 respectively).

224

225 **Pathway database choice is key**

226 An important consideration when conducting any type of pathway analysis is the nature of the
227 pathway sets used. Pathway sets can differ between databases in many ways, including the
228 number of pathways present, the size of pathways, how pathways are curated (either manually
229 or computationally, or a combination of both), and the organisms supported. We compared
230 several properties of three pathway databases: KEGG, Reactome, and BioCyc. As this work
231 focuses on metabolomics, only pathways which contain at least three metabolites were
232 considered for the purposes of this paper, and genes and proteins were excluded from the
233 pathway definition. Using human pathways as an example, as of December 2020, Reactome
234 contained the highest number of pathways (1631), followed by HumanCyc (390) (part of the
235 BioCyc collection) and KEGG, containing 261 pathways. A comparison of pathway sizes across
236 the three databases can be seen in Fig 4a, in which HumanCyc pathways are the largest across
237 the three databases, followed by KEGG and Reactome, based on median pathway size.

238 We next investigated the similarity of metabolite composition for KEGG and Reactome
239 pathways. Identifiers for metabolites in each pathway were first converted to KEGG IDs and the
240 ComPath [20] resource was used to find equivalent pathway mappings, linking KEGG and
241 Reactome pathways with the same metabolic functions. We calculated the Jaccard index (JI) for
242 each of the 23 pairs of equivalent pathways. The JI values were low (median = 0.08,
243 interquartile range = 0.01-0.16), suggesting a low level of similarity in metabolite composition
244 despite apparent equivalence of function. The same calculation was performed considering only
245 genes in equivalent KEGG and Reactome pathways. 55 pathways were comparable, and while
246 the JI values were slightly larger than those derived from comparison of metabolite-only
247 pathways (median = 0.19, interquartile range = 0.11-0.26), these also suggest low levels of
248 similarity in the gene composition of pathways from different databases. To explore whether
249 similar biological functions could be inferred from an ORA using different databases, we
250 compared the SEPs obtained using the Yachida *et al.* dataset based on KEGG, Reactome, and
251 HumanCyc pathways (Table S1). By manual inspection of pathway names, there appeared to be

252 low concordance between the results of the three databases in terms of biological function.

253 Similar observations were also made in the other datasets.

254 In addition to selecting a pathway database, many pathway databases offer both
255 reference and organism-specific pathway sets. Reference pathway sets are not associated with
256 any organism and can be useful where the organism under study does not have an associated
257 pathway set. We compared basic properties of the KEGG human and KEGG reference pathways
258 sets. The KEGG reference pathway set contained both more (377 vs. 261 pathways) and larger
259 pathways (mean pathway size 45 vs. 30 compounds). The two pathway sets had a median JI of
260 0.8 (IQR = 0.57-1.0) for pathways with a common ID (e.g. Glycolysis: HSA00010/MAP00010),
261 indicating a high level of similarity between pathways but that not all common pathways are
262 identical. We performed ORA for each example dataset using both the organism-specific and
263 reference pathway sets and compared the SEPs obtained (Table 2). While there was a large
264 overlap, many more pathways were significantly enriched in the reference pathway set alone as
265 opposed to in the organism-specific pathway set alone. This is likely due to the fact that the
266 reference set contains more pathways, although not all of these may be of biological relevance
267 to the organism in question.

268 **Table 2: Organism-specific vs. reference pathways.** Number of SEP ($P \leq 0.1$) detected in both the KEGG organism-
269 specific and KEGG reference pathway sets, and those significant in only one of the sets.

Dataset	Common pathways	Organism-specific only	Reference only
Labbé	19	0	6
Yachida	11	1	19
Stevens	5	0	1
Quirós	46	3	28
Fuhrer (yfgm)	27	0	26
Fuhrer (dcus)	27	0	23

270

271 A final consideration when selecting a pathway database is the version of the database one will
272 use. Not all ORA tools will use the latest version of a certain pathway database available. The
273 vast majority of pathway databases will undergo at least yearly updates, with some such as
274 Reactome providing four major releases per year. To investigate how much impact pathway

275 database updates can have on ORA results, we obtained four years' worth of Reactome pathway
276 sets spanning the period from June 2017 to December 2020. We compared three aspects of the
277 Reactome human pathway sets (R-HSA) between each release: the number of pathways, the
278 number of unique compounds in the database, and the mean pathway size (Fig 4b). As
279 expected, the number of new pathways increased gradually from release to release, alongside
280 the number of unique compounds. From 2017 to 2020, over 200 new pathways were added as
281 well as almost 500 new compounds. Interestingly, the mean pathway size gradually increased
282 from release 61 to release 68, after which it steadily decreased, but altogether remained
283 between 17 and 19 compounds on average throughout the course of 14 releases.

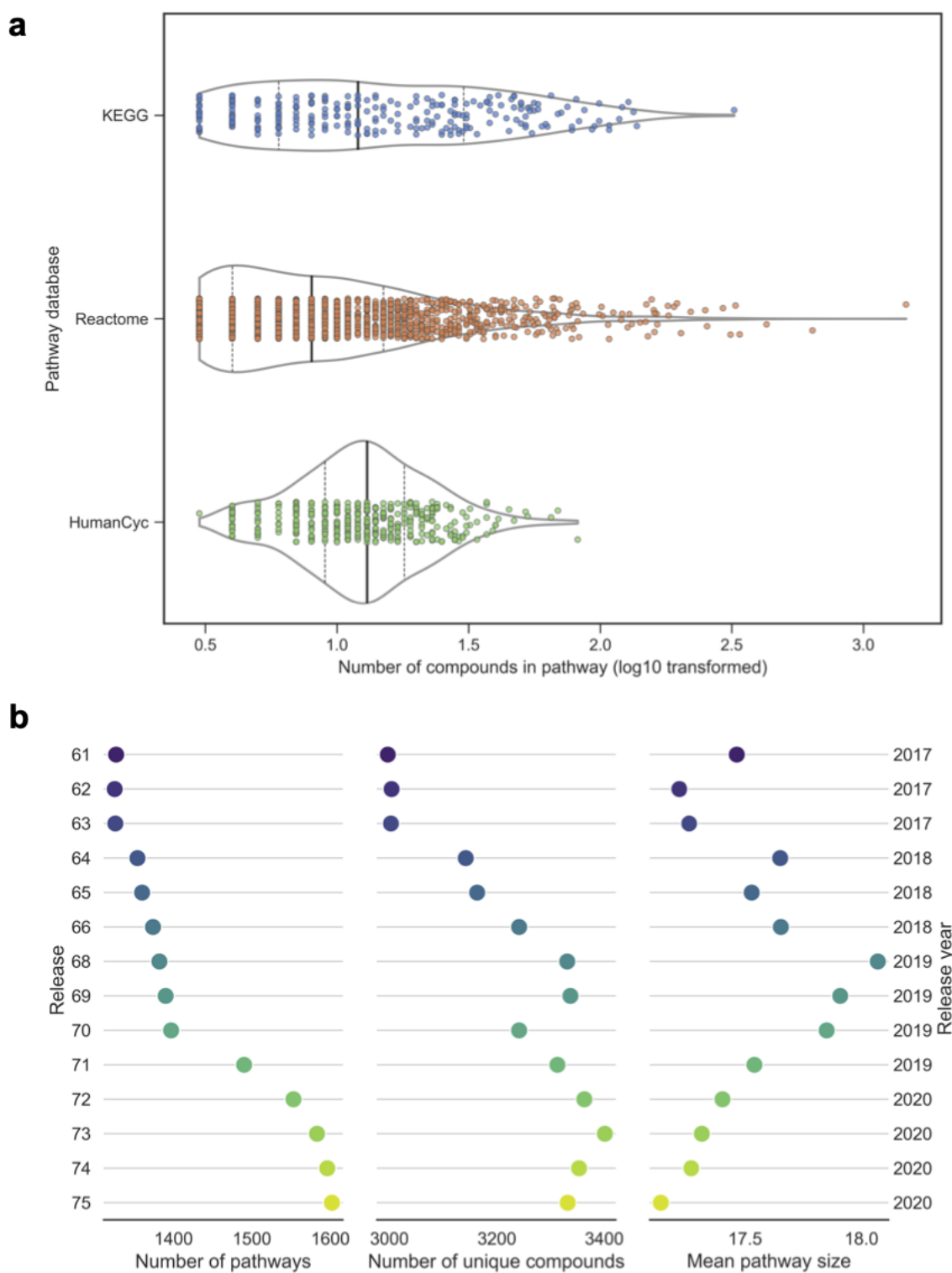


Fig 4: Comparison of pathway databases and database updates. **a** Pathway size distribution of KEGG, Reactome, and HumanCyc databases. Violin plots show the distribution of pathway size (number of compounds, log10 transformed). Bold vertical lines show median, dashed vertical lines show lower and upper quartiles. **b** Comparison of Reactome human pathway set (R-HSA) releases spanning the years 2017 (R61, June 2017) to 2020 (R75, December 2020). Data for release 67 was not available and hence is not shown. Dot colour corresponds to release version, with lighter colours representing newer releases.

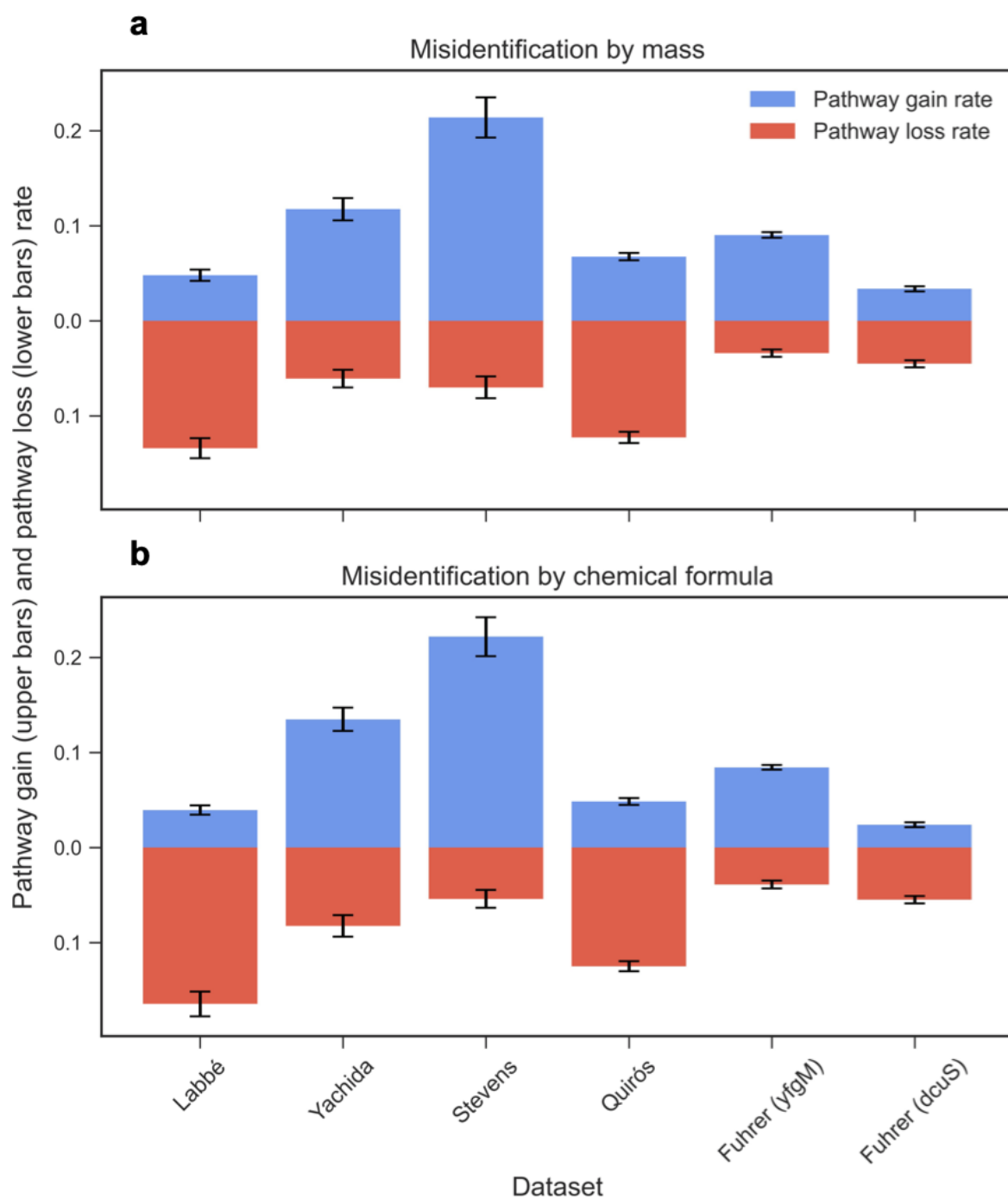
285 **Metabolite misidentification results in both gain and loss of truly significant pathways**

286 Next, we investigated some factors which are specific to metabolomics data, such as metabolite
287 misidentification and assay chemical bias. A major bottleneck in untargeted metabolomics is the
288 identification of compounds. In untargeted metabolomics, it is commonplace to putatively
289 identify (“annotate”) metabolites based on their physicochemical properties (e.g. m/z ratio,
290 polarity) and similarity to compounds in spectral databases, and then confirm the identities of
291 compounds of interest using chemical reference standards. Consequently, a large proportion of
292 compounds in untargeted metabolomics assays are expected to have a degree of uncertainty in
293 their identification, ranging from Metabolomics Standards Initiative (MSI) confidence levels 2-4
294 [21].

295 To compare the effects of metabolite misidentification on the number and identity of
296 significant pathways detected using ORA, we introduce two new statistics: the pathway loss rate
297 and the pathway gain rate (see Methods). The former describes how, as the data are degraded,
298 some pathways are “lost” (no longer identified as significant) and others are “gained” (newly
299 identified as significant). These are analogous to false-negative and false-positive rates, but
300 account for the fact that we do not know the truly enriched pathways. For the purposes of this
301 simulation, we make the assumption that all pathways significant at 0% misidentification are
302 the “true” SEPs, and we compare these to the SEPs obtained at varying levels of simulated
303 misidentification. The pathway loss rate refers to the proportion of SEPs present at 0%
304 misidentification that are no longer present at $f\%$ misidentification, and the pathway gain rate
305 refers to the number of SEPs not originally present at 0% misidentification which become
306 significant at $f\%$ misidentification.

307 We simulated the effects of metabolite misidentification on ORA using KEGG pathways
308 by replacing the true metabolites with false ones in two different ways: a) by similar molecular
309 weight (20ppm window), and b) by identical chemical formula (see Methods). For both
310 approaches, we calculated the pathway loss and gain rate for each dataset at 4% simulated
311 misidentification, which although there are few published estimates of misidentification rates in

312 metabolomics studies, endeavours to simulate a representative scenario (Fig 5). All the example
313 datasets had a pathway loss and gain rate greater than zero at 4% simulated misidentification
314 either by molecular weight or formula. Such findings suggest that even at a misidentification
315 rate as low as 4%, it is likely that some pathways are significant simply as an effect of
316 misidentification, and other pathways are not detected as significantly enriched due to the noise
317 in the data caused by the misidentification. Pathway loss and gain rates from 1-5% are shown in
318 Fig S2. Pathway loss and gain rate results were similar for both misidentification by molecular
319 weight and formula, likely owing to the fact that compounds with identical chemical formula
320 share the same molecular weight.



321

322 *Fig 5: Metabolite misidentification. The effect of compound misidentification by a molecular weight (20ppm window)*
323 *and b chemical formula on the mean pathway loss rate (red bars) and mean pathway gain rate (blue bars) averaged*
324 *over 100 random resamplings at 4% misidentification. Error bars represent standard error of the mean.*

325

326

327

328

329 **The polarity of compounds in a metabolomics experiment influences the pathways**

330 **discoverable using ORA**

331 The analytical platform and specific assay used for a metabolomics study can be expected to
332 introduce bias into the pathways which might be detected by ORA. One common characteristic
333 in which assays differ is their ability to detect compounds of different polarity, often depending
334 on the type of chromatography used. Hydrophilic interaction chromatography is typically
335 optimised for the detection of polar compounds, whereas reverse-phase liquid or gas
336 chromatography are usually more advantageous for non-polar compounds. While it is
337 increasingly common for metabolomics experiments to incorporate multiple types of
338 chromatography, many datasets still consist of metabolites measured using just a single type of
339 chromatography. We would expect to observe differences in SEPs based on the polarity of
340 compounds in the dataset. We simulated the effect of using different types of chromatography
341 by splitting the compounds in each dataset into two halves based on the median logP coefficient,
342 to achieve an approximately even number of polar and non-polar compounds on each side. We
343 then performed ORA using KEGG pathways on the polar and non-polar halves of each dataset
344 and compared the results (Fig S3 shows an example using the Labbé dataset). In the Labbé
345 example, only a single KEGG pathway, *Pyrimidine metabolism*, was enriched in both the polar
346 and non-polar halves of the dataset. All remaining significant pathways (9 in total) were only
347 found in either the polar or non-polar half. While this might be expected, it is a clear
348 demonstration that ORA results are highly influenced by the chemistry probed by the assay, and
349 especially the type of chromatography employed.

350

351

352 Discussion

353 As metabolomics continues to grow as a field of study with a multitude of applications within
354 various disciplines, deriving meaningful conclusions from such data becomes increasingly
355 important. ORA is one of the most popular approaches used to draw functional interpretations
356 from metabolomics data. However, to date, there have been no published investigations of the
357 consequences of varying input parameters on ORA results derived from metabolomics data.
358 Understanding the sensitivity of ORA to tuning parameters, especially how it is influenced by
359 metabolomics-specific factors, will play a crucial role in its successful application. In the present
360 study, we sought to investigate the effects of varying inputs on ORA results, which we
361 demonstrated using *in-silico* simulations applied to five untargeted metabolomics datasets.

362 One of the most salient findings was the difference in the number of SEPs detected when
363 using an assay-specific versus a nonspecific background set. The use of a nonspecific
364 background set, such as all compounds present in the KEGG reference or human pathway set,
365 for example, resulted in a drastic increase in the number of SEPs. In many ORA tools, use of a
366 nonspecific background is typically the default option, and one that may lead users to believe
367 that this is the 'correct' procedure. It is crucial however to understand that the consequence of
368 not specifying a background set, which should contain all compounds that are realistically
369 observable, is that an assumption is being made that the compounds in the default background
370 set are all equally likely to be detected in the experiment [22]. Such an assumption is highly
371 unlikely to be true given that most technologies can only detect a small fraction of the
372 metabolome and may lead to false-positive pathways. Additionally, the size of the background
373 set is an important consideration, with larger sets generally yielding higher numbers of SEPs.
374 Mass-spectrometry based approaches can usually detect a larger number of compounds than
375 NMR-based methods, for example, at least for typical 1D NMR methods that are most commonly
376 used for profiling [23]. Users need to consider whether their metabolomics dataset is large
377 enough to provide sufficient statistical power such that ORA results can be considered useful.

378 The list of compounds of interest (often corresponding to metabolites differentially
379 present between conditions in experiments) is an essential input for ORA and we have
380 demonstrated that the way these compounds are selected greatly impacts PA results. It is
381 important to select a threshold that strikes a balance between selecting too few compounds,
382 therefore resulting in low power for the detection of significant pathways, or selecting
383 compounds too liberally and losing power by introducing noise into the analysis. Visualisation
384 of the curve of number of significant pathways vs. the number of compounds of interest (Fig 3)
385 can be a useful tool to determine the stability of the analysis to significance thresholds. Multiple
386 testing correction should always be applied to all metabolite-level statistics before filtering
387 them to produce the list of compounds of interest. We examined two of the most popular
388 multiple testing correction methods: Bonferroni and BH FDR correction. As expected,
389 Bonferroni correction tended to be more stringent, resulting in fewer compounds of interest,
390 although this does not necessarily always correspond to fewer SEPs.

391 Unlike other fields (e.g. transcriptomics), the level of uncertainty surrounding
392 compound identities remains a critical issue in metabolomics studies. While it is not possible to
393 find a benchmark level of metabolite misidentification typically found in metabolomics studies,
394 most studies will contain at least a small percentage of misidentified compounds [24]. The level
395 of misidentification will vary depending on the analytical platform used and remains a key
396 bottleneck, more so in MS-based studies, where the number of metabolites detected often
397 exceeds that of NMR-based studies [25]. In this study, we simulated metabolite
398 misidentification by randomly swapping a small percentage of compounds in each of the
399 datasets with compounds of either a similar molecular weight (± 20 ppm) or an identical
400 chemical formula. Even at a low level of misidentification of 4%, we found appreciable pathway
401 loss and gain rates for all datasets. Hence, we suggest that ORA is sensitive to even low levels of
402 metabolite misidentification, resulting in the emergence of false-positive and false-negative
403 SEPs in the results.

404 Another essential input of ORA is the pathway database or list of metabolite sets used.
405 The inherent differences between pathway databases will undoubtedly impact the PA results,
406 regardless of the method used [26]. In the case of ORA, which is based on the hypergeometric
407 formula, pathway size will influence results by rendering smaller pathways more significant and
408 larger pathways less significant [27]. The number of pathways tested using ORA will also
409 directly impact the adjusted significance level if multiple testing correction methods are
410 applied, and the more pathways tested the more statistical power is lost. A related caveat is that
411 the most widely used multiple testing approaches (e.g. Bonferroni, BH FDR) do not account for
412 correlations between pathways and therefore such methods may be too conservative and
413 undermine pathway significance [2].

414 A further important consideration for pathway database evaluation is the type of
415 compound identifiers used in the pathway. KEGG and BioCyc use database-specific identifiers,
416 whereas Reactome uses ChEBI identifiers. It is necessary to convert the identifiers present in a
417 metabolomics dataset to their database-specific equivalent, which often results in loss of
418 information as not all identifiers will necessarily map directly to a database compound or be
419 annotated to a pathway [28]. For example, in the Stevens et al. dataset, over 900 compounds
420 were assigned to Metabolon identifiers, but less than half of these compounds could be mapped
421 to KEGG identifiers. Another characteristic of metabolomics (and in particular lipidomics) is the
422 discrepancy between the chemical precision of identification between the pathway databases
423 and the dataset. For instance, in databases classes of lipids are often gathered into a single
424 element (e.g. “a triglyceride”) while lipidomics allows more in-depth annotation (e.g. “TG
425 16/18/18”). Computational solutions based on chemical ontologies exist to establish a link
426 between dataset elements and pathway database ones [29], but this will also have an impact on
427 the pathway enrichment results since several data elements will map to a single node in the
428 pathway database.

429 The incompleteness of pathway databases, together with the evolution of pathway
430 definitions between releases, are key factors highlighting the necessity of using an up-to-date

431 resource; not doing so can have a detrimental effect on PA results [30]. Furthermore, the
432 magnitude of changes across database releases demonstrated in this work suggests that ORA
433 results are somewhat short-lived and perhaps valid only at a given time, hence they should be
434 periodically revised using an updated database. Frainay et al. examined the coverage of
435 analytes in the human metabolic network and found poor coverage of pathways involving
436 eicosanoids, vitamins, heme, and bile acid metabolism [31]. Finally, although an extensive
437 comparison of pathway databases is beyond the scope of this paper, several excellent studies
438 have examined this in detail to which we refer the interested reader [26,32,33]. A general
439 recommendation is to use multiple pathway databases and derive a consensus signature across
440 these.

441 In this work we have focused on ORA, but many other PA methods exist [1,34]. While
442 functional class scoring and topology-based methods can overcome certain limitations
443 associated with ORA, such as the need to select compounds of interest, or not taking metabolite-
444 level statistics into account, many of our findings are also relevant to these other methods.
445 Pathway database selection, metabolite misidentification rate, and assay chemical bias will
446 impact the majority of metabolomics PA methods. Alongside the present work, further studies
447 examining the input parameters of other PA methods for metabolomics data will be invaluable
448 in establishing a set of best-practice guidelines for their application.

449 This study is limited by the lack of availability of a ground-truth dataset where the
450 identities of enriched pathways have been experimentally confirmed. Such a dataset would have
451 made it possible to investigate a wider variety of performance metrics for ORA. Another
452 limitation is that in the majority of examples, a p-value threshold of $P \leq 0.1$ was used without
453 multiple testing correction to select SEPs. As metabolomics experiments usually identify far
454 fewer compounds than transcriptomic experiments identify genes, ORA based on metabolites
455 appears to have much lower power to identify significant pathways and as such in the example
456 datasets few, if any, pathways remained significant after multiple testing correction was
457 applied.

458 The purpose of the present research was to evaluate the suitability of ORA for
459 metabolomics pathway analysis and assess the effects of varying input data and parameters. We
460 have investigated the three main input parameters: the background set, the list of compounds of
461 interest, and the pathway database, as well as metabolomics-specific considerations such as
462 metabolite misidentification and assay chemical bias. By means of *in-silico* simulations using
463 experimental datasets, all of the aforementioned variables have been shown to introduce
464 varying levels of bias and uncertainty into ORA results, which has significant implications for
465 those using ORA to analyse metabolomics data. In particular, use of an assay-specific
466 background set is often ignored, yet has a critical effect on the output. Overall, this study has
467 been the first detailed investigation into the application of ORA to metabolomics data, with
468 wide-ranging findings that have implications not only to ORA but also a variety of other PA
469 methods in metabolomics.

470 We therefore offer the community a set of recommendations for application, as well as
471 recommended minimal reporting criteria, which may contribute to the future development of
472 best-practice guidelines for the application of ORA to metabolomics data.

473

474 **Suggested recommendations for the application of ORA to metabolomics data:**

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

483

484 **Suggested recommended minimal reporting criteria. Users should report:**

- 485 1. The statistical test/approach used for pathway analysis (e.g. Fisher's exact test)
- 486 2. The tool (and version) used to perform ORA.

- 487 3. The pathway database, the corresponding compound identifier type (e.g. KEGG, ChEBI,
488 BioCyc, etc.), its release number and which organism-specific pathway set was used (if
489 any).
- 490 4. Which compounds form the background set.
- 491 5. The multiple testing correction methods applied for i) selection of DA metabolites and
492 ii) selection of SEP, alongside the adjusted p-value thresholds used.
- 493

494 Materials and methods

495 1. Obtaining the list of metabolites of interest

496 1.1 Summary of experimental datasets used

497 Five publicly available untargeted metabolomics datasets were used in this work (Table 1). We selected a
 498 diverse range of datasets encompassing various organisms, biofluids, and experimental conditions. For
 499 consistency, all datasets used in this work are mass-spectrometry (MS) based. The first dataset is
 500 available at MTBLS135 from the MetaboLights repository and consists of 12 Hi-Myc genotype and 12
 501 wild-type *Mus musculus* plasma samples [35]. The second dataset from Yachida et al. 2019 consists of 149
 502 healthy control and 148 colorectal cancer human stool samples (stages I-IV). The third dataset is
 503 available at MTBLS136 and consists of 667 control samples and 332 estrogen users [37]. The fourth
 504 dataset is from Quirós et al. 2017 from which we compared 8 HeLa cell replicates treated with actinonin
 505 to 8 HeLa cell replicates treated with doxycycline. The final dataset is available from EBI BioStudies (S-
 506 BSST5) and consists of >3,800 single-gene *E. coli* knockouts [39]. We selected two knockout strains to
 507 investigate from this dataset: $\Delta yfgM$ and $\Delta dcuS$. It is important to note that two datasets, Quirós et al.
 508 2017 and Fuhrer et al. 2017, did not use any separation step in their analytical platform, and therefore
 509 there may be a higher degree of uncertainty in the metabolite identifications.

510 **Table 1: Summary of experimental datasets used in this work.** An asterisk (*) besides the MS platform indicates no
 511 chromatography/electrophoresis was used in the assay.

<i>Author</i>	<i>Title</i>	<i>Organism</i>	<i>Analytical platform</i>	<i>Sample type</i>	<i>Total number of metabolites mapping to KEGG compounds</i>	<i>Study accession code/data availability</i>
Labbé et al.	High-fat diet fuels prostate cancer progression by rewiring the metabolome and amplifying the MYC program	<i>Mus musculus</i>	UPLC-MS/MS	Tissue	269	MTBLS135
Yachida et al.	Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer	<i>Homo sapiens</i>	CE-TOF MS	Stool	286	Supplementary table S13 of https://doi.org/10.1038/s41591-019-0458-7
Stevens et al.	Serum metabolomic profiles associated with postmenopausal hormone use	<i>Homo sapiens</i>	UPLC-MS/MS	Serum	362	MTBLS136
Quirós et al.	Multi-omics analysis identifies ATF4 as a key regulator of the mitochondrial stress response in mammals	<i>Homo sapiens</i> (HeLa cells)	Flow injection TOF MS*	HeLa cell	1110	Supplementary table S8 of https://doi.org/10.1083/jcb.201702058

Fuhrer et al.	Genomewide landscape of gene-metabolome associations in <i>Escherichia coli</i>	<i>Escherichia coli</i>	Flow injection TOF MS*	E. coli	2468	S-BSST5
---------------	---	-------------------------	------------------------	---------	------	---------

512

513 **1.2 Post-processing of metabolomics datasets**

514 All metabolomics datasets and corresponding metadata used in this study are publicly available
515 from the MetaboLights repository [40], the BioStudies database [41], or in the supplementary
516 information of the original publication (Table 1). Details of metabolomics data pre-processing,
517 as well as sample preparation, data acquisition, and compound identification can be found in the
518 original publication for each dataset. For the purposes of this study, the pre-processed raw
519 metabolite abundance matrices consisting of n samples by m metabolites were downloaded as
520 .csv or .xlsx files and post-processed identically. Missing abundance values were imputed using
521 the minimum value of each metabolite divided by 2. All abundance values in the matrix were
522 then \log_2 transformed and features (metabolites) were auto-scaled by subtracting the mean and
523 dividing by the standard deviation.

524

525 **1.3 Metabolite identifier harmonisation**

526 In order to map compounds to the three pathway databases investigated in this study (KEGG,
527 Reactome, and BioCyc), metabolite identifiers in each dataset were converted to the
528 corresponding identifier type. For the conversion of compound names to KEGG identifiers, the
529 MetaboAnalyst 4.0 [42] ID conversion tool was used
530 (<https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml>). For Reactome,
531 KEGG compounds were mapped to ChEBI identifiers using the Python bioservices package (v
532 1.7.1) [43]. For BioCyc, the web-based metabolite translation service
533 (<https://metacyc.org/metabolite-translation-service.shtml>) was used to convert from KEGG to
534 BioCyc identifiers.

535

536 **1.4 Selection of differentially abundant metabolites**

537 The list of metabolites of interest was determined using a series of two-tailed student's t-tests to
538 determine whether each metabolite in the dataset was significantly associated with the
539 outcome of interest. p-values were adjusted using the Benjamini-Hochberg False discovery rate
540 (BH FDR) procedure [44] to account for multiple testing. Significantly differentially abundant
541 (DA) metabolites were then selected based on a q-value threshold of $q \leq 0.05$. To investigate the
542 effect of the list of input metabolites on the number of significant pathways, we used both BH
543 FDR and Bonferroni methods for p-value adjustment and tested several cut-off thresholds
544 (adjusted $p \leq 0.005$, 0.05, or 0.1) for the selection of DA metabolites using each method.

545

546 2. Performing pathway enrichment

547 2.1 Pathway database details

548 For the purposes of this paper, the pathway sets used contained only compounds (including
549 small molecules, metabolites and drugs). KEGG pathways and their corresponding compounds
550 were downloaded using the KEGG REST API (<https://www.kegg.jp/kegg/rest/keggapi.html>) in
551 October 2020, corresponding to KEGG release 96. Reactome pathways release 75 were
552 downloaded from <https://reactome.org/download-data>. BioCyc pathways v24.5 were exported
553 from <https://biocyc.org/> using the SmartTables function.

554

555 2.2 ORA implementation

556 ORA was implemented using a custom script that utilised the `scipy stats fisher_exact` function
557 (right-tailed) to calculate pathway p-values. Only pathways containing at least 3 compounds
558 were used as input for ORA. p-values were calculated if the parameter k (number of
559 differentially abundant metabolites in the i^{th} pathway) was ≥ 1 .

560

561 3. *In-silico* simulation details

562 3.1 Implementation details

563 All simulations were performed using Python (v 3.8). Simulations with an element of
564 randomisation were repeated 100 times, and results are reported as the mean of 100 random
565 samplings of the simulation, alongside the standard error of the mean.

566

567 **3.2 Simulating metabolite misidentification**

568 Chemical formula and molecular weight information for each metabolite was obtained using the
569 KEGG REST API. For each level of metabolite misidentification, we randomly selected $f\%$ ($f=0, 1,$
570 $\dots X\%$) of compounds that had at least one other compound with a molecular weight within
571 ± 20 ppm (approximately isobaric compound) present in the KEGG pathway set. For each
572 randomly selected compound, one of its isobaric compounds was randomly selected and the
573 identifier of this compound then replaced the original identifier in the dataset, thereby
574 simulating misidentification by mass. Similarly, for misidentification by chemical formula,
575 compounds that had at least one other compound with an identical chemical formula present in
576 the KEGG pathway set were randomly selected, and compound identifiers replaced.
577 Replacement compounds must be present in at least one KEGG pathway but must not already
578 form part of the original background list, to avoid introducing duplicate compounds.

579

580 **3.3 Quantifying changes in results**

581 To illustrate how lists of significant pathways change at varying levels of metabolite
582 misidentification, we define two performance statistics: the pathway loss rate and the pathway
583 gain rate. The pathway loss rate represents the proportion of the original pathways (0%
584 misidentification) significant at $p \leq 0.1$ that are no longer significant at $f\%$ misidentification.
585 The pathway gain rate represents the proportion of pathways that were not significant at 0%
586 misidentification but become significant at $f\%$ misidentification.

587 Let A and B be sets of pathways from ORA such that:

588 $A = \{\text{Pathways significant at } 0\% \text{ metabolite misidentification } (p \leq 0.1)\}$

589 $B_f = \{\text{Pathways significant at } f\% \text{ metabolite misidentification } (p \leq 0.1)\}$

590 The *pathway loss rate* and *pathway gain rate* at $f\%$ metabolite misidentification are then
591 defined as:

$$\text{Pathway loss rate}(A, B_f) = 1 - \frac{|A \cap B_f|}{|A|} \quad (2)$$

592

593

$$\text{Pathway gain rate}(A, B_f) = \frac{|B_f - A|}{|A|} \quad (3)$$

594

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

597

598

599 References

- 600 1. Nguyen TM, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: A
601 comprehensive review and assessment. *Genome Biol.* 2019;20. doi:10.1186/s13059-019-1790-4
- 602 2. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: Current approaches and outstanding
603 challenges. Ouzounis CA, editor. *PLoS Computational Biology*. Public Library of Science; 2012. p.
604 e1002375. doi:10.1371/journal.pcbi.1002375
- 605 3. Karnovsky A, Li S. *Pathway Analysis for Targeted and Untargeted Metabolomics*. Methods in
606 Molecular Biology. Humana Press Inc.; 2020. pp. 387–400. doi:10.1007/978-1-0716-0239-3_19
- 607 4. Marco-Ramell A, Palau-Rodriguez M, Alay A, Tulipani S, Urpi-Sarda M, Sanchez-Pla A, et al.
608 Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics
609 data. *BMC Bioinformatics*. 2018;19: 1. doi:10.1186/s12859-017-2006-0
- 610 5. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: State of the art.
611 *Frontiers in Physiology*. Frontiers Research Foundation; 2015. doi:10.3389/fphys.2015.00383
- 612 6. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic
613 network architecture. *Nat Genet.* 1999;22: 281–285. doi:10.1038/10343
- 614 7. Drăghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. Global functional profiling of gene
615 expression. *Genomics*. 2003;81: 98–104. doi:10.1016/S0888-7543(02)00021-6
- 616 8. Xie C, Jauhari S, Mora A. Popularity and performance of bioinformatics software: the case of gene
617 set analysis. *BMC Bioinformatics*. 2021;22: 191. doi:10.1186/s12859-021-04124-5
- 618 9. Beauclercq S, Nadal-Desbarats L, Hennequet-Antier C, Gabriel I, Tesseraud S, Calenge F, et al.
619 Relationships between digestive efficiency and metabolomic profiles of serum and intestinal
620 contents in chickens. *Sci Rep.* 2018;8: 6678. doi:10.1038/s41598-018-24978-9
- 621 10. Guo YS, Tao JZ. Metabolomics and pathway analyses to characterize metabolic alterations in
622 pregnant dairy cows on D 17 and D 45 after AI. *Sci Rep.* 2018;8: 1–8. doi:10.1038/s41598-018-
623 23983-2
- 624 11. Michonneau D, Latis E, Curis E, Dubouchet L, Ramamoorthy S, Ingram B, et al. Metabolomics
625 analysis of human acute graft-versus-host disease reveals changes in host and microbiota-derived
626 metabolites. *Nat Commun.* 2019;10: 1–15. doi:10.1038/s41467-019-13498-3
- 627 12. McGeachie MJ, Dahlin A, Qiu W, Croteau-Chonka DC, Savage J, Wu AC, et al. The metabolomics of
628 asthma control: A promising link between genetics and disease. *Immun Inflamm Dis.* 2015;3: 224–
629 238. doi:10.1002/iid3.61
- 630 13. Zhang P, Zhang W, Lang Y, Qu Y, Chen J, Cui L. 1H nuclear magnetic resonance-based metabolic
631 profiling of cerebrospinal fluid to identify metabolic features and markers for tuberculosis
632 meningitis. *Infect Genet Evol.* 2019;68: 253–264. doi:10.1016/j.meegid.2019.01.003
- 633 14. Rosato A, Tenori L, Cascante M, De Atauri Carulla PR, Martins dos Santos VAP, Saccenti E. From
634 correlation to causation: analysis of metabolomics data using systems biology approaches.
635 *Metabolomics*. Springer New York LLC; 2018. p. 37. doi:10.1007/s11306-018-1335-y
- 636 15. Kanehisa M, Goto S. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. *Nucleic Acids Research*.
637 Oxford University Press; 2000. pp. 27–30. doi:10.1093/nar/28.1.27
- 638 16. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway
639 knowledgebase. *Nucleic Acids Res.* 2020;48: D498–D503. doi:10.1093/nar/gkz1031
- 640 17. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, et al. The BioCyc collection of
641 microbial genomes and metabolic pathways. *Brief Bioinform.* 2018;20: 1085–1093.
642 doi:10.1093/bib/bbx085
- 643 18. Cottret L, Frainay C, Chazalviel M, Cabanettes F, Gloaguen Y, Camenen E, et al. MetExplore:
644 Collaborative edition and exploration of metabolic networks. *Nucleic Acids Res.* 2018;46: W495–
645 W502. doi:10.1093/nar/gky301
- 646 19. Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in ingenuity pathway

- 647 analysis. *Bioinformatics*. 2014;30: 523–530. doi:10.1093/bioinformatics/btt703
- 648 20. Domingo-Fernández D, Hoyt CT, Bobis-Álvarez C, Marín-Llaó J, Hofmann-Apitius M. ComPath: an
649 ecosystem for exploring, analyzing, and curating mappings across pathway databases. *npj Syst*
650 *Biol Appl*. 2019;5: 1–8. doi:10.1038/s41540-018-0078-8
- 651 21. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum
652 reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG)
653 Metabolomics Standards Initiative (MSI). *Metabolomics*. 2007;3: 211–221. doi:10.1007/s11306-
654 007-0082-2
- 655 22. Cavill R, Jennen D, Kleinjans J, Briedé JJ. Transcriptomic and metabolomic data integration. *Brief*
656 *Bioinform*. 2016;17: 891–901. doi:10.1093/bib/bbv090
- 657 23. Emwas AHM. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with
658 particular focus on metabolomics research. *Methods Mol Biol*. 2015;1277: 161–193.
659 doi:10.1007/978-1-4939-2377-9_13
- 660 24. Creek DJ, Dunn WB, Fiehn O, Griffin JL, Hall RD, Lei Z, et al. Metabolite identification: are you sure?
661 And how do your peers gauge your confidence? *Metabolomics*. 2014;10: 350–353.
662 doi:10.1007/s11306-014-0656-8
- 663 25. Dunn WB, Erban A, Weber RJM, Creek DJ, Brown M, Breitling R, et al. Mass appeal: Metabolite
664 identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*. Springer;
665 2013. pp. 44–66. doi:10.1007/s11306-012-0434-4
- 666 26. Stobbe MD, Houten SM, Jansen GA, van Kampen AHC, Moerland PD. Critical assessment of human
667 metabolic pathway databases: A stepping stone for future integration. *BMC Syst Biol*. 2011;5: 165.
668 doi:10.1186/1752-0509-5-165
- 669 27. Karp PD, Midford PE, Caspi R, Khodursky A. Pathway size matters: the influence of pathway
670 granularity on over-representation (enrichment analysis) statistics. *BMC Genomics* 2021 221.
671 2021;22: 1–11. doi:10.1186/s12864-021-07502-8
- 672 28. Pham N, van Heck RGA, van Dam JCJ, Schaap PJ, Saccenti E, Suarez-Diez M. Consistency,
673 inconsistency, and ambiguity of metabolite names in biochemical databases used for genome-
674 scale metabolic modelling. *Metabolites*. 2019;9: 28. doi:10.3390/metabo9020028
- 675 29. Poupin N, Vinson F, Moreau A, Batut A, Chazalviel M, Colsch B, et al. Improving lipid mapping in
676 Genome Scale Metabolic Networks using ontologies. *Metabolomics*. 2020;16: 44.
677 doi:10.1007/s11306-020-01663-5
- 678 30. Wadi L, Meyer M, Weiser J, Stein LD, Reimand J. Impact of outdated gene annotations on pathway
679 enrichment analysis. *Nature Methods*. Nature Publishing Group; 2016. pp. 705–706.
680 doi:10.1038/nmeth.3963
- 681 31. Frainay C, Schymanski EL, Neumann S, Merlet B, Salek RM, Jourdan F, et al. Mind the gap: Mapping
682 mass spectral databases in genome-scale metabolic networks reveals poorly covered areas.
683 *Metabolites*. 2018;8. doi:10.3390/metabo8030051
- 684 32. Labena AA, Gao YZ, Dong C, Hua H li, Guo FB. Metabolic pathway databases and model
685 repositories. *Quantitative Biology*. Higher Education Press; 2018. pp. 30–39. doi:10.1007/s40484-
686 017-0108-3
- 687 33. Mubeen S, Hoyt CT, Gemünd A, Hofmann-Apitius M, Fröhlich H, Domingo-Fernández D. The
688 Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling.
689 *Front Genet*. 2019;10: 1203. doi:10.3389/fgene.2019.01203
- 690 34. Fang X, Liu Y, Ren Z, Du Y, Huang Q, Garmire LX. Lilikoi V2.0: a deep learning-enabled,
691 personalized pathway-based R package for diagnosis and prognosis predictions using
692 metabolomics data. *Gigascience*. 2021;10: 1–11. doi:10.1093/gigascience/giaa162
- 693 35. Labbé DP, Zadra G, Yang M, Reyes JM, Lin CY, Cacciatore S, et al. High-fat diet fuels prostate cancer
694 progression by rewiring the metabolome and amplifying the MYC program. *Nat Commun*.
695 2019;10: 1–14. doi:10.1038/s41467-019-12298-z
- 696 36. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al. Metagenomic and

- 697 metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal
698 cancer. *Nature Medicine*. Nature Publishing Group; 2019. pp. 968–976. doi:10.1038/s41591-019-
699 0458-7
- 700 37. Stevens VL, Wang Y, Carter BD, Gaudet MM, Gapstur SM. Serum metabolomic profiles associated
701 with postmenopausal hormone use. *Metabolomics*. 2018;14: 97. doi:10.1007/s11306-018-1393-1
- 702 38. Quirós PM, Prado MA, Zamboni N, D’Amico D, Williams RW, Finley D, et al. Multi-omics analysis
703 identifies ATF4 as a key regulator of the mitochondrial stress response in mammals. *J Cell Biol*.
704 2017;216: 2027–2045. doi:10.1083/jcb.201702058
- 705 39. Fuhrer T, Zampieri M, Sévin DC, Sauer U, Zamboni N. Genomewide landscape of gene–metabolome
706 associations in *Escherichia coli*. *Mol Syst Biol*. 2017;13: 907. doi:10.15252/msb.20167150
- 707 40. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, et al. MetaboLights: A
708 resource evolving in response to the needs of its scientific community. *Nucleic Acids Res*. 2020;48:
709 D440–D444. doi:10.1093/nar/gkz1019
- 710 41. Sarkans U, Gostev M, Athar A, Behrangi E, Melnichuk O, Ali A, et al. The BioStudies database—one
711 stop shop for all data supporting a life sciences study. *Nucleic Acids Res*. 2018;46: D1266–D1270.
712 doi:10.1093/nar/gkx965
- 713 42. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: Towards more
714 transparent and integrative metabolomics analysis. *Nucleic Acids Res*. 2018;46: W486–W494.
715 doi:10.1093/nar/gky310
- 716 43. Cokelaer T, Pultz D, Harder LM, Serra-Musach J, Saez-Rodriguez J. BioServices: a common Python
717 package to access biological Web Services programmatically. *Bioinformatics*. 2013;29: 3241–
718 3242. doi:10.1093/bioinformatics/btt547
- 719 44. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach
720 to Multiple Testing. *J R Stat Soc Ser B*. 1995;57: 289–300. Available:
721 <http://www.jstor.org/stable/2346101>
- 722

723 **Author contributions:**

724 TE, JGB, CW, FJ, and CF conceived and designed the study. CW performed the analysis. FV extracted KEGG
725 pathway data. CW and TE wrote the manuscript with input from CF, NP, PRM, JC, RPJL, FJ, and JGB. All
726 authors contributed to the interpretation of results and approved the final manuscript.

727

728 **Acknowledgements:**

729 The authors gratefully acknowledge the help of the Reactome support team based at the Ontario Institute
730 for Cancer Research, for providing previous release files of their database.

731

732 **Competing interests statement:**

733 All authors declare they have no conflict of interest.

734

735 **Funding:**

736 This research was funded in whole, or in part, by the Wellcome Trust [222837/Z/21/Z]. For the purpose
737 of open access, the author has applied a CC BY public copyright licence to any Author Accepted
738 Manuscript version arising from this submission. CW is supported by a Wellcome Trust PhD Studentship
739 [222837/Z/21/Z]. RPJL receives support from the UK Medical Research Council (MR/R008922/1). JC is
740 supported by a state-funded PhD contract (MESRI (Minister of Higher Education, Research and
741 Innovation)). FJ is supported by the French Ministry of Research and National Research Agency as part of
742 the French MetaboHUB, the national metabolomics and fluxomics infrastructure (Grant ANR-INBS-0010),
743 and MetClassNet project (ANR-19-CE45-0021 and DFG: 431572533). TE gratefully acknowledges partial
744 support from BBSRC grant BB/T007974/1, NIH grant R01 HL133932-01 and the NIHR Imperial
745 Biomedical Research Centre (BRC).

746

747 **Data Availability**

748 The metabolomics and metadata reported in this paper are available via their respective MetaboLights or
749 BioStudies identifiers, or in the supplementary information of the relevant paper, detailed in Table 1.

750

751 **Code Availability**

752 The software developed in this study is available via a Jupyter notebook interface to enable reproduction
753 of the simulations. The notebook, usage guidelines, dependencies, and processed metabolomics data are
754 available via <https://github.com/cwieder/metabolomics-ORA>.

755

756

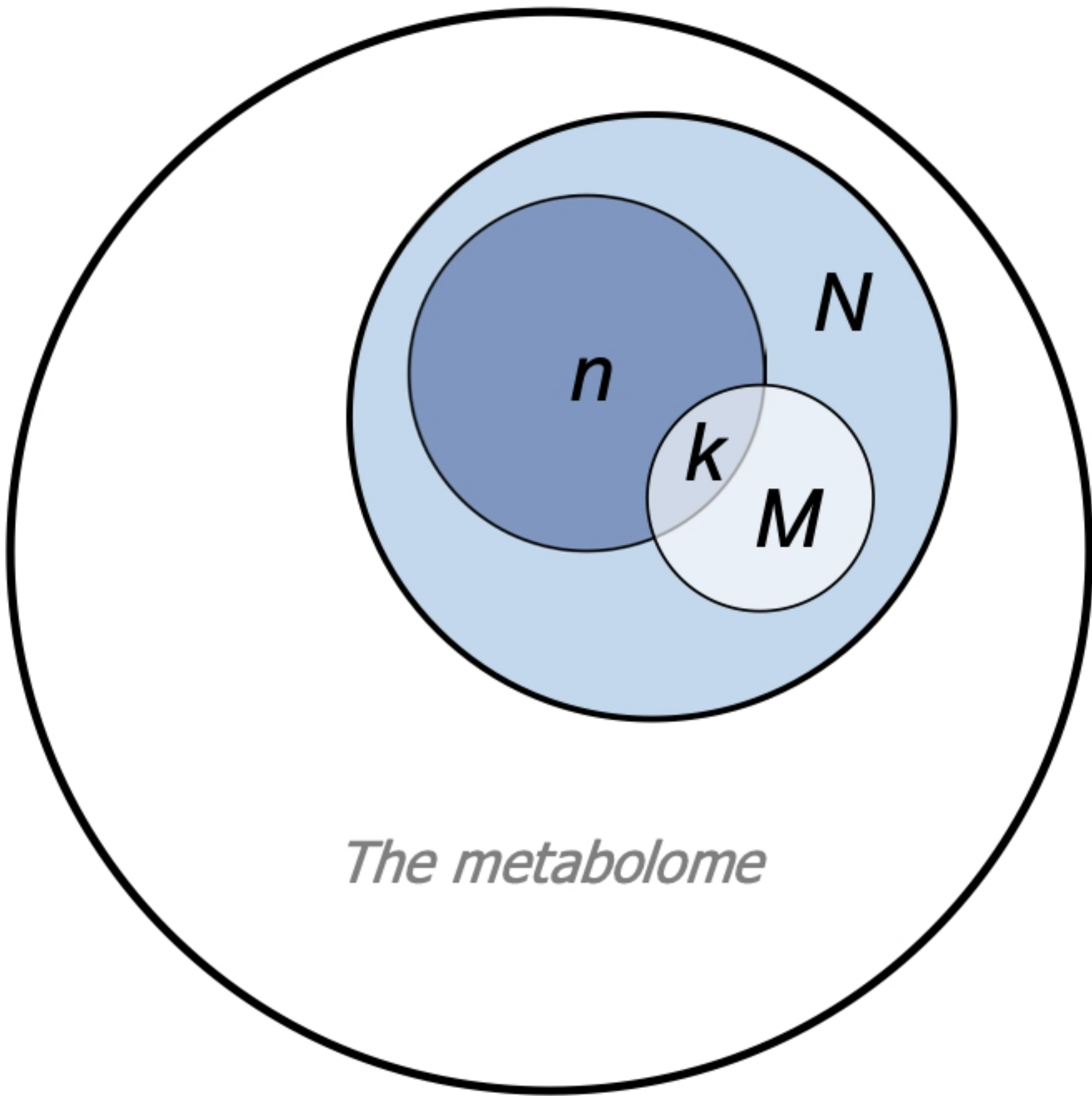


Fig 1

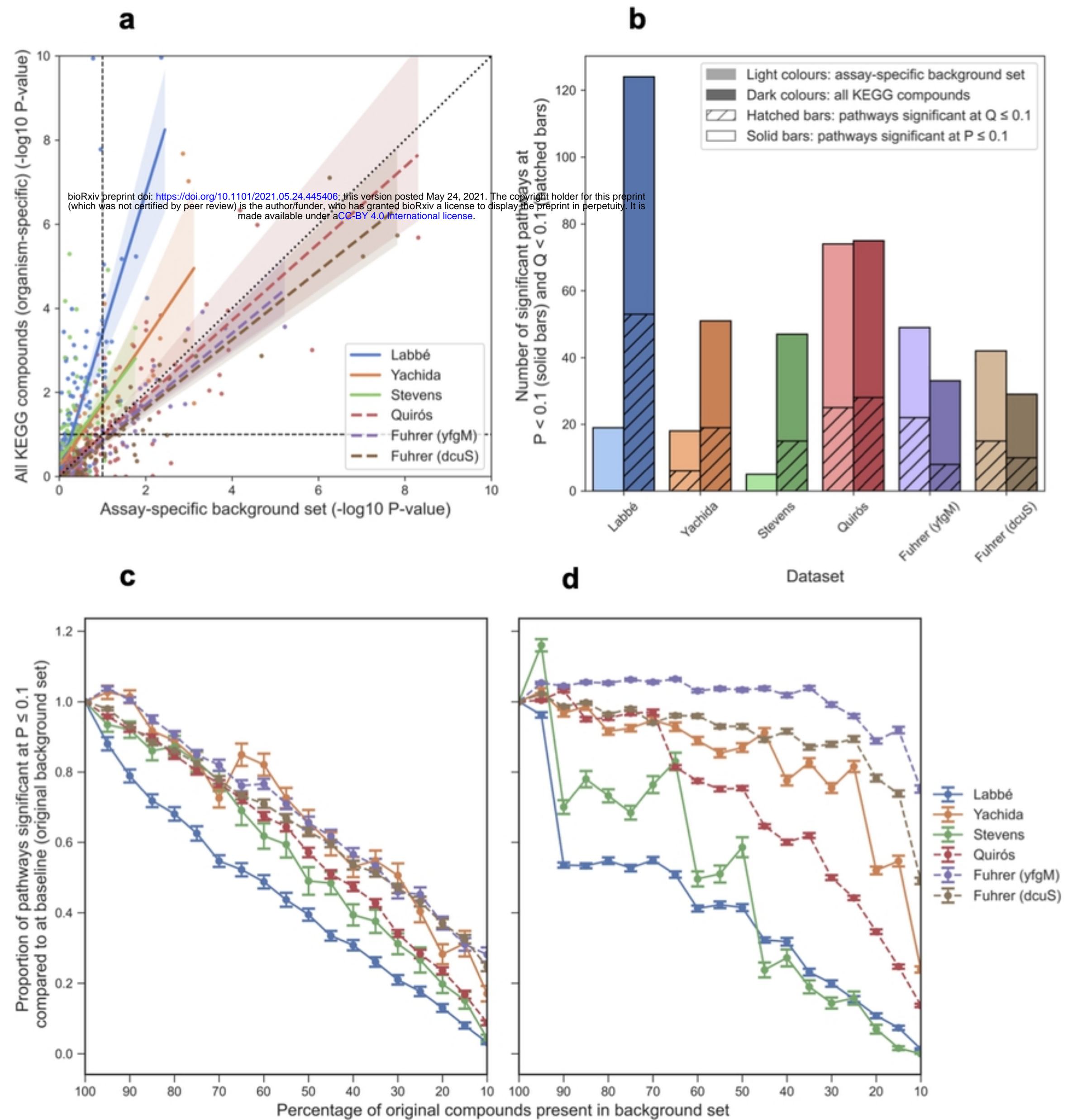


Fig 2

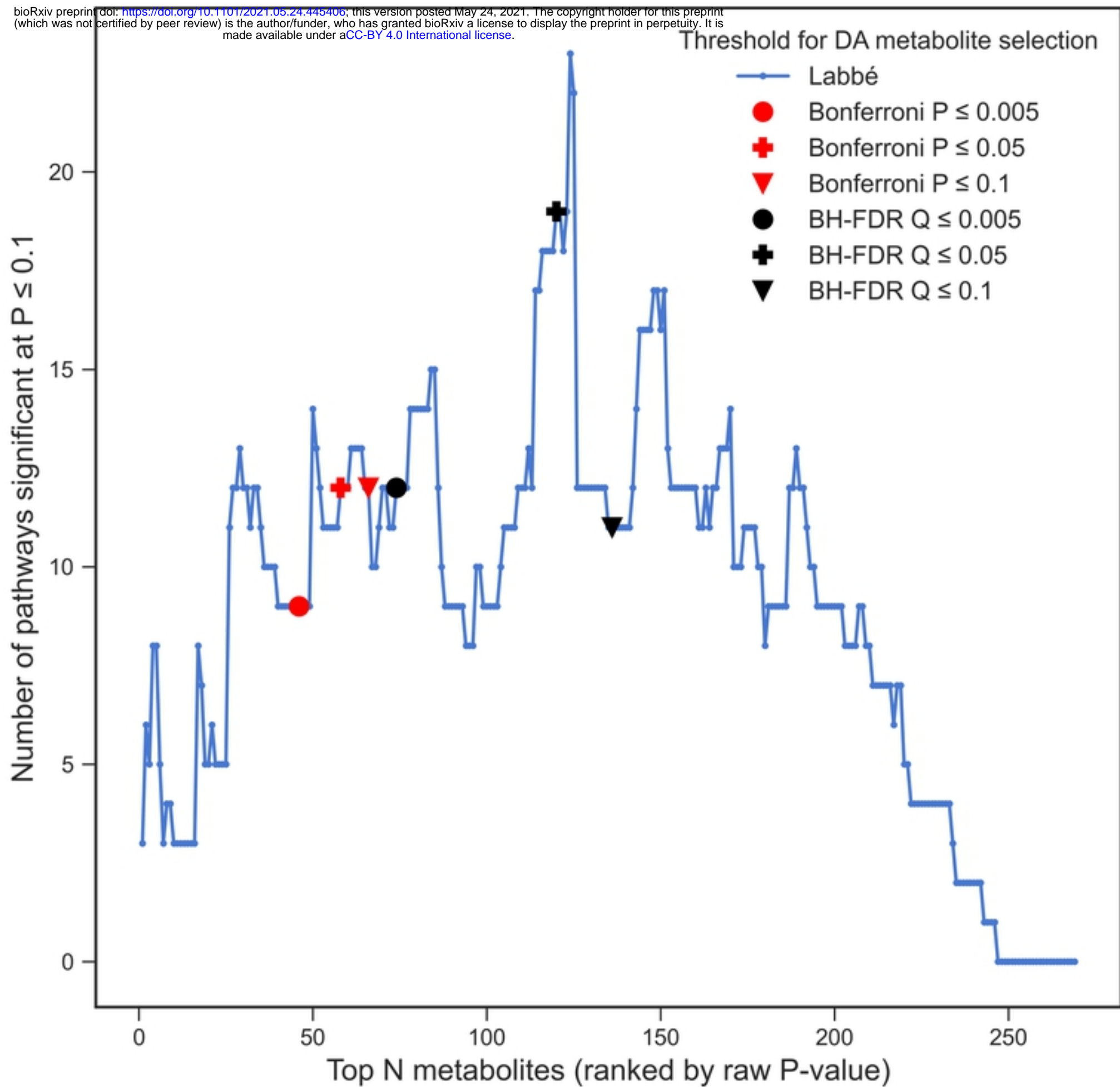


Fig 3

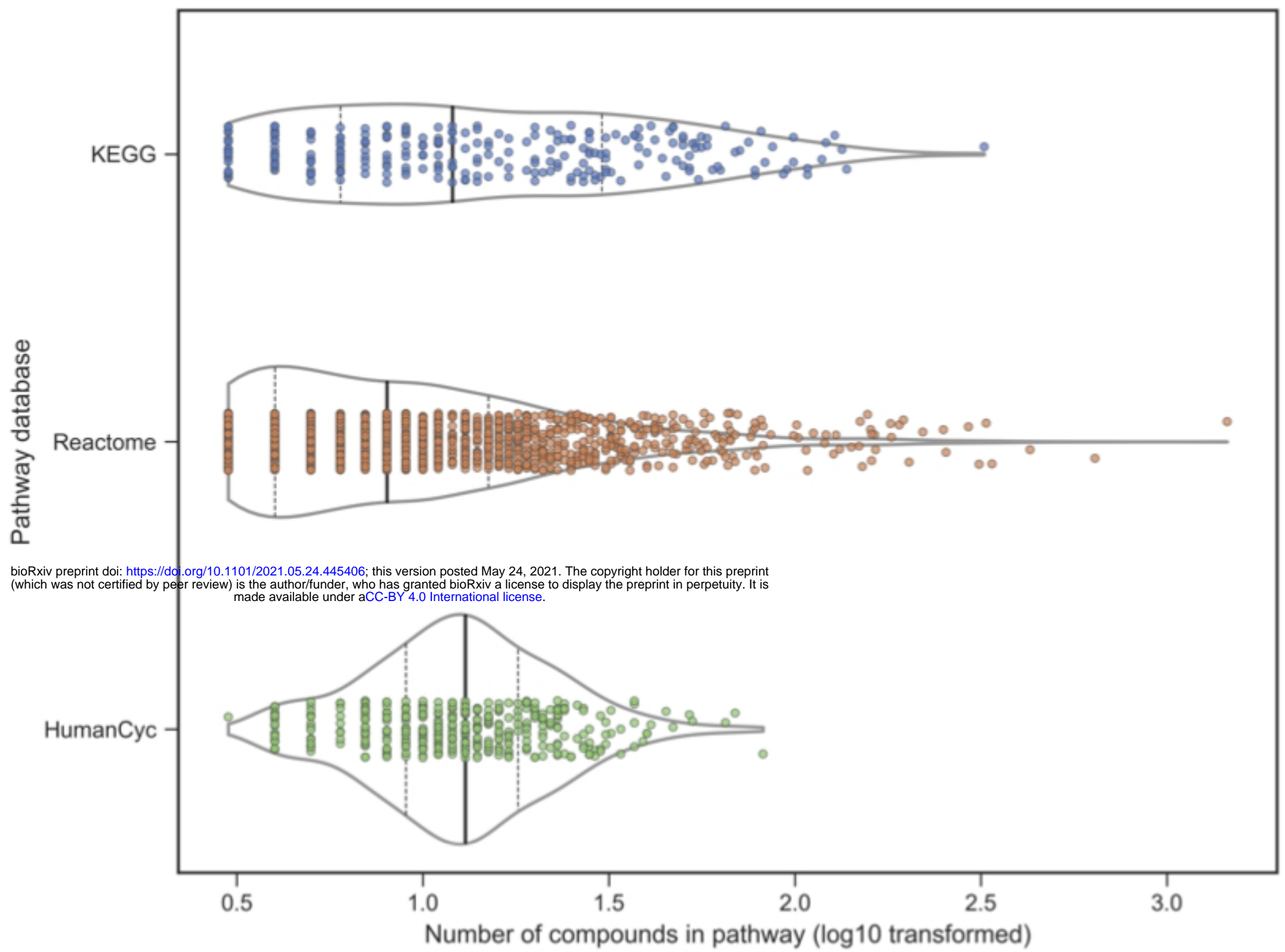
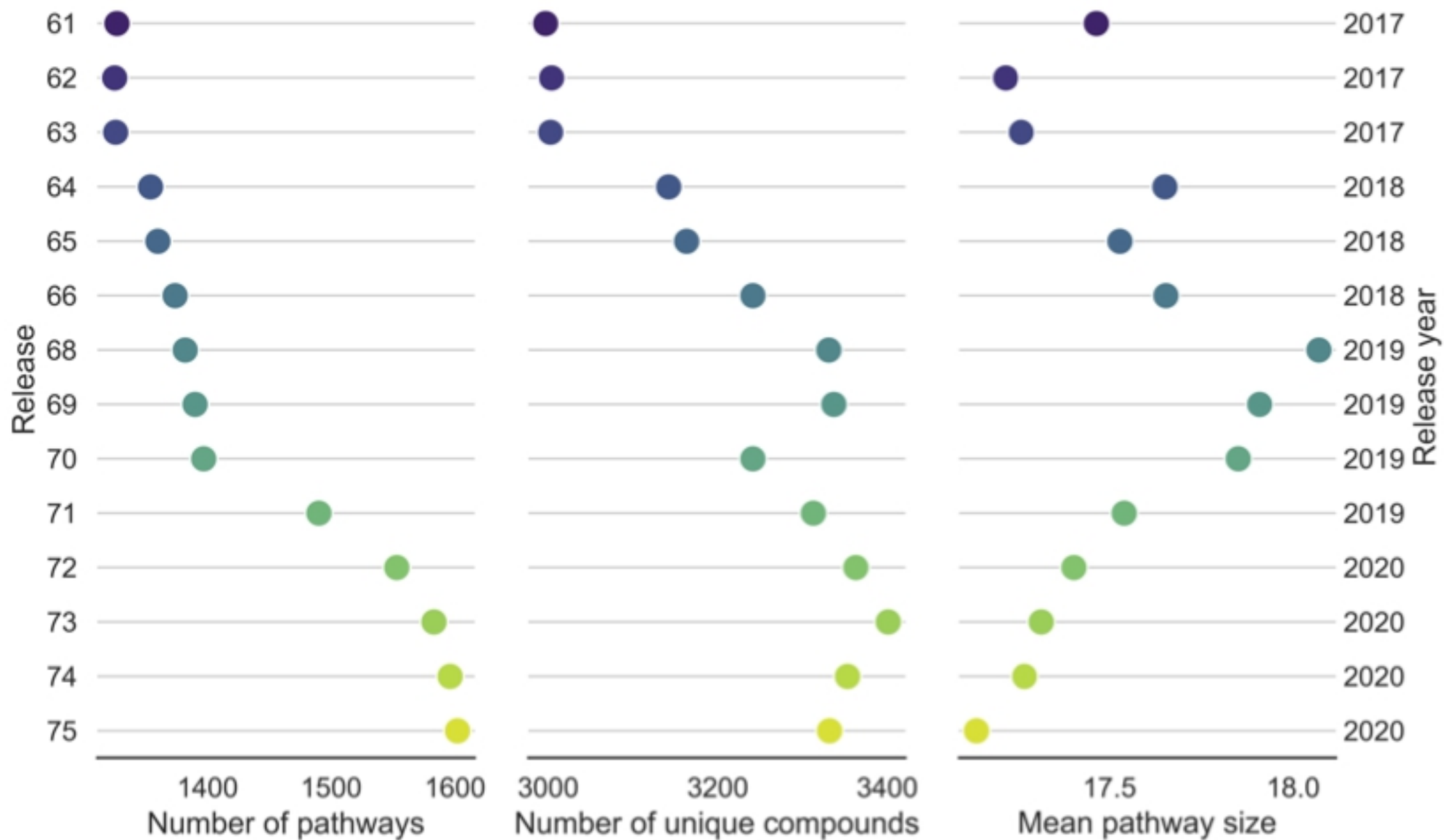
a**b**

Fig 4

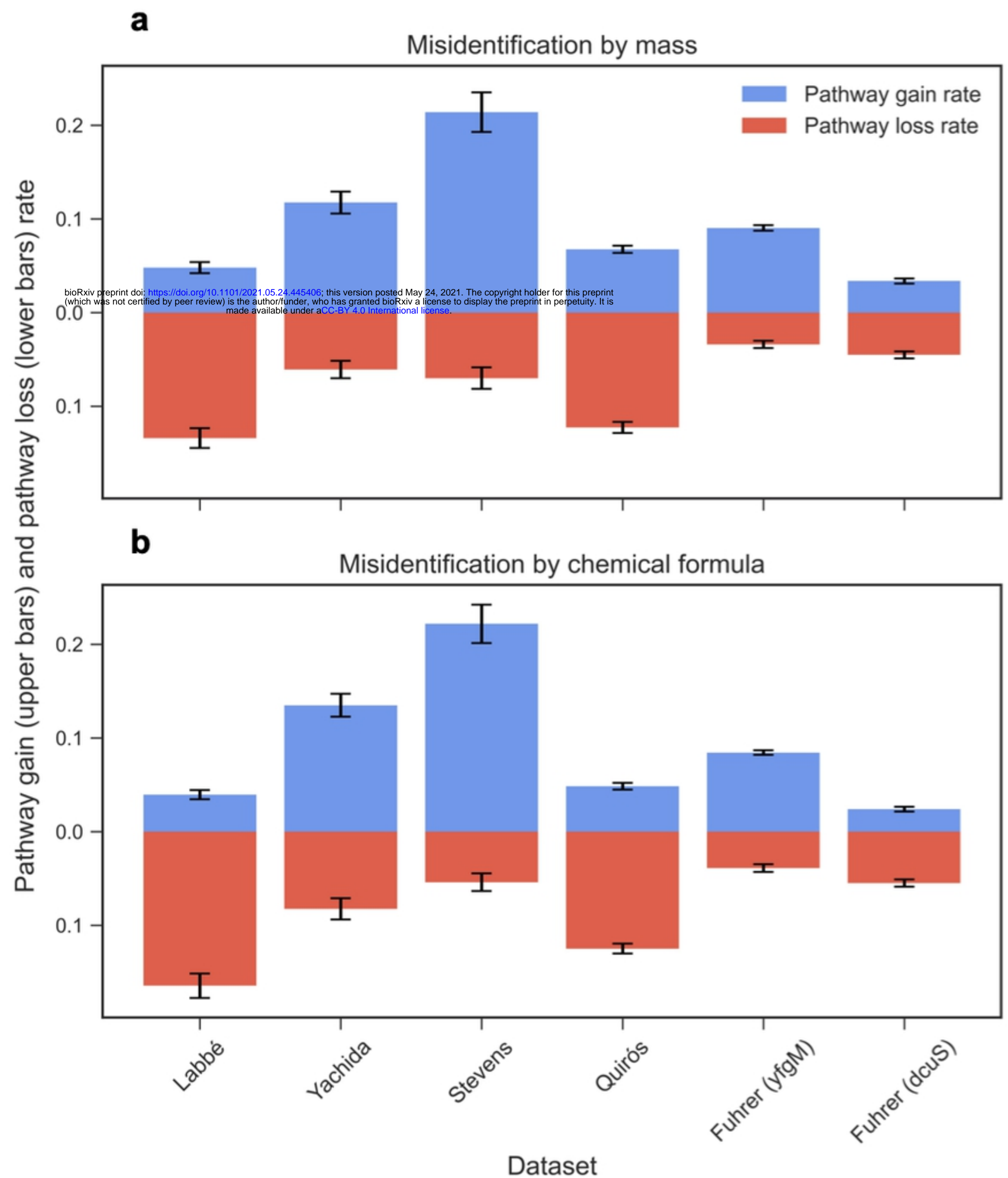


Fig 5