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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

Authors:

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

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

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

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

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

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

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

Key words: pathway analysis, over-representation analysis, pathway enrichment, model interpretation, bioinformatics

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Author summary

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

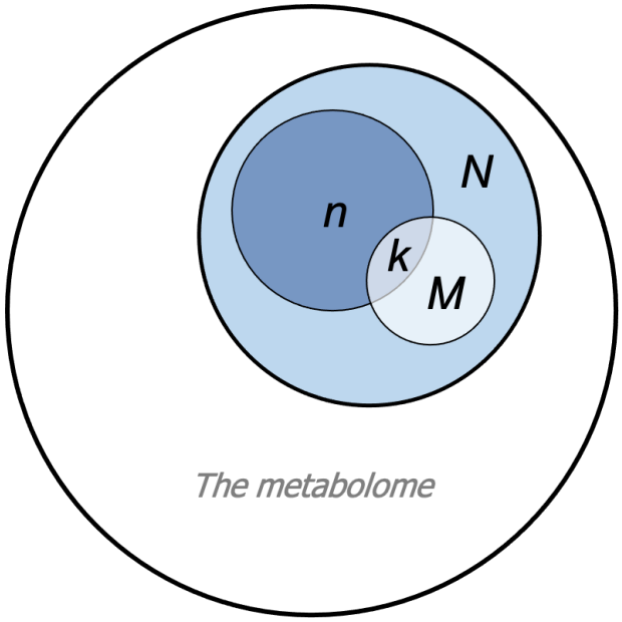

Fig 1: ***Over Representation Analysis (ORA)*** *Venn diagram representing ORA parameters corresponding to Equation 1. N represents compounds forming the background set, which covers part of the full metabolome. M represents compounds in the pathway of interest. n represents compounds of interest (i.e.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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

Results

Nonspecific background sets result in erroneously high levels of enriched pathways

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

184

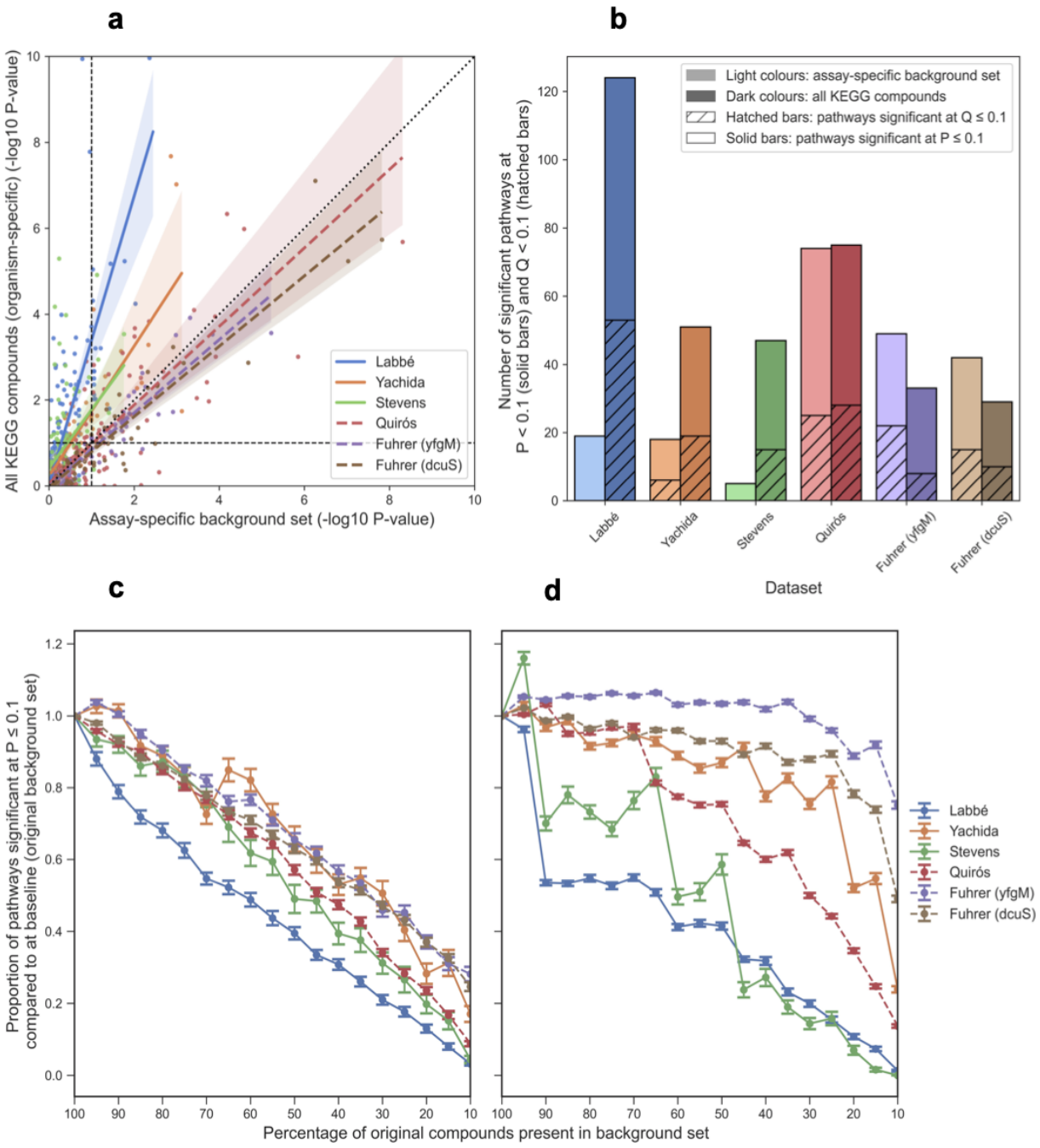

Fig 2: ***Effect of background set. a*** *Scatter plot of -log₁₀ p-values of pathways when using an assay-specific background set consisting of all measurable compounds in each dataset (x-axis) compared to using a non-specific background set containing of all compounds annotated to at least one KEGG 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 ≤0.1 (solid bars) and the number of pathways significant at q < 0.1 (hashed 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

186

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

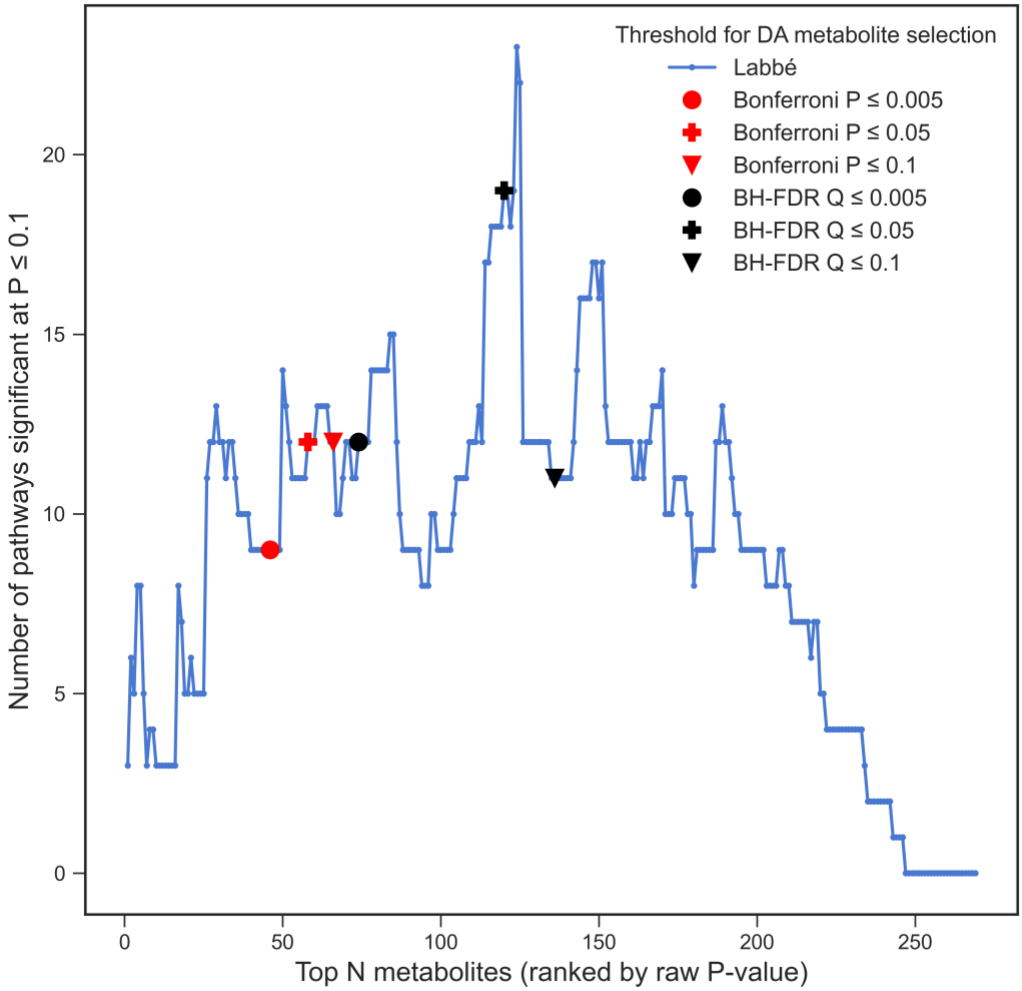

Fig 3: ***Number of DA metabolites****. The effect of the number of DA metabolites in the list of metabolites of interest on the number of significant pathways (p ≤ 0.1) in the Labbé et al. dataset. Results corresponding to Bonferroni thresholds are denoted by red markers while those corresponding to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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Pathway database choice is key

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

low concordance between the results of the three databases in terms of biological function. Similar observations were also made in the other datasets.

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

Table 2: Organism-specific vs. reference pathways. *Number of SEP (P ≤ 0.1) detected in both the KEGG organism-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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

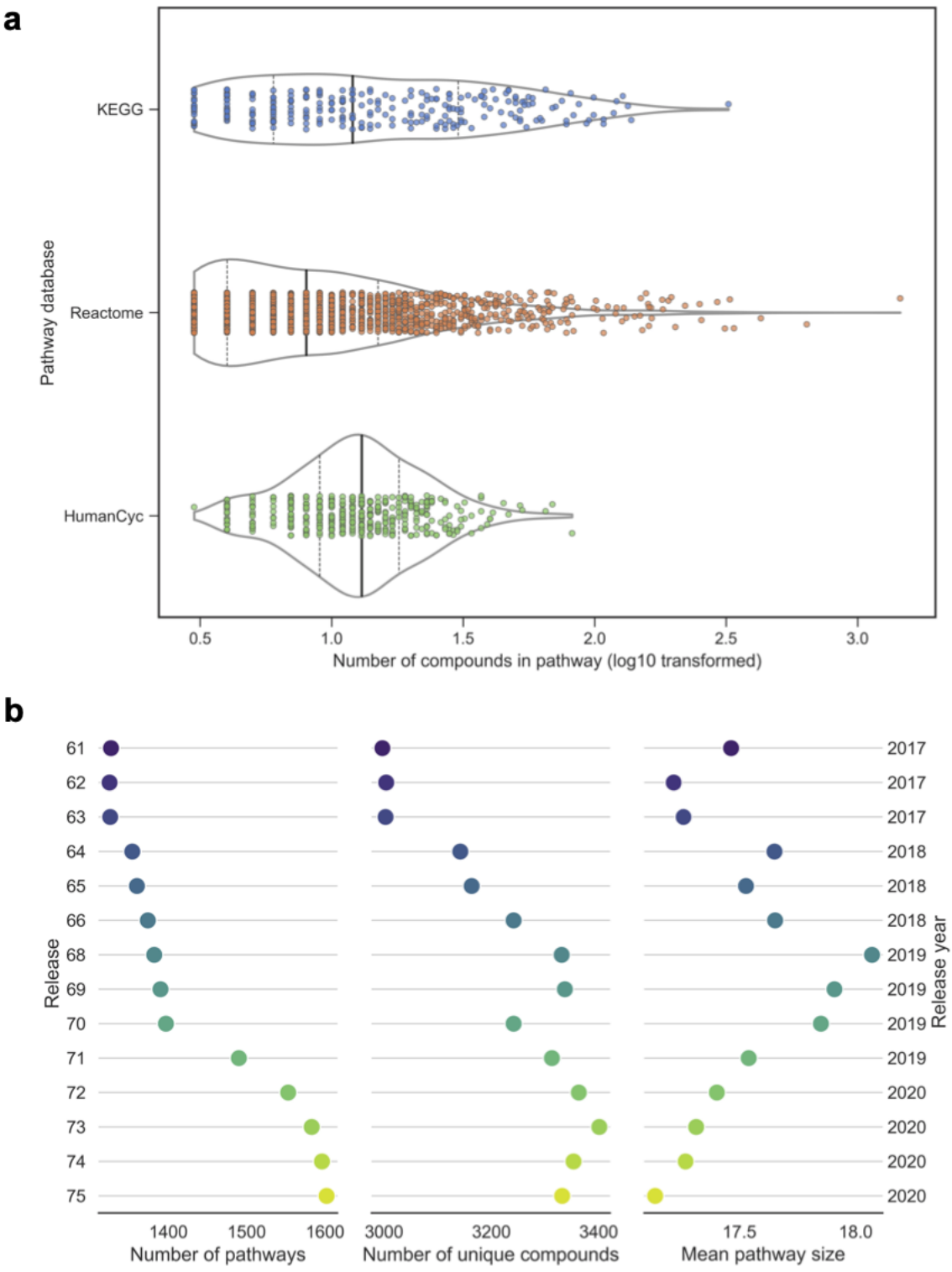

Fig 4: ***Comparison of pathway databases and database updates. a*** *Pathway size distribution of KEGG, Reactome, and HumanCyc databases. Violin plots show the distribution of pathway size (number of compounds, log10 transformed). Bold vertical lines show median, dashed vertical lines show lower and 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.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

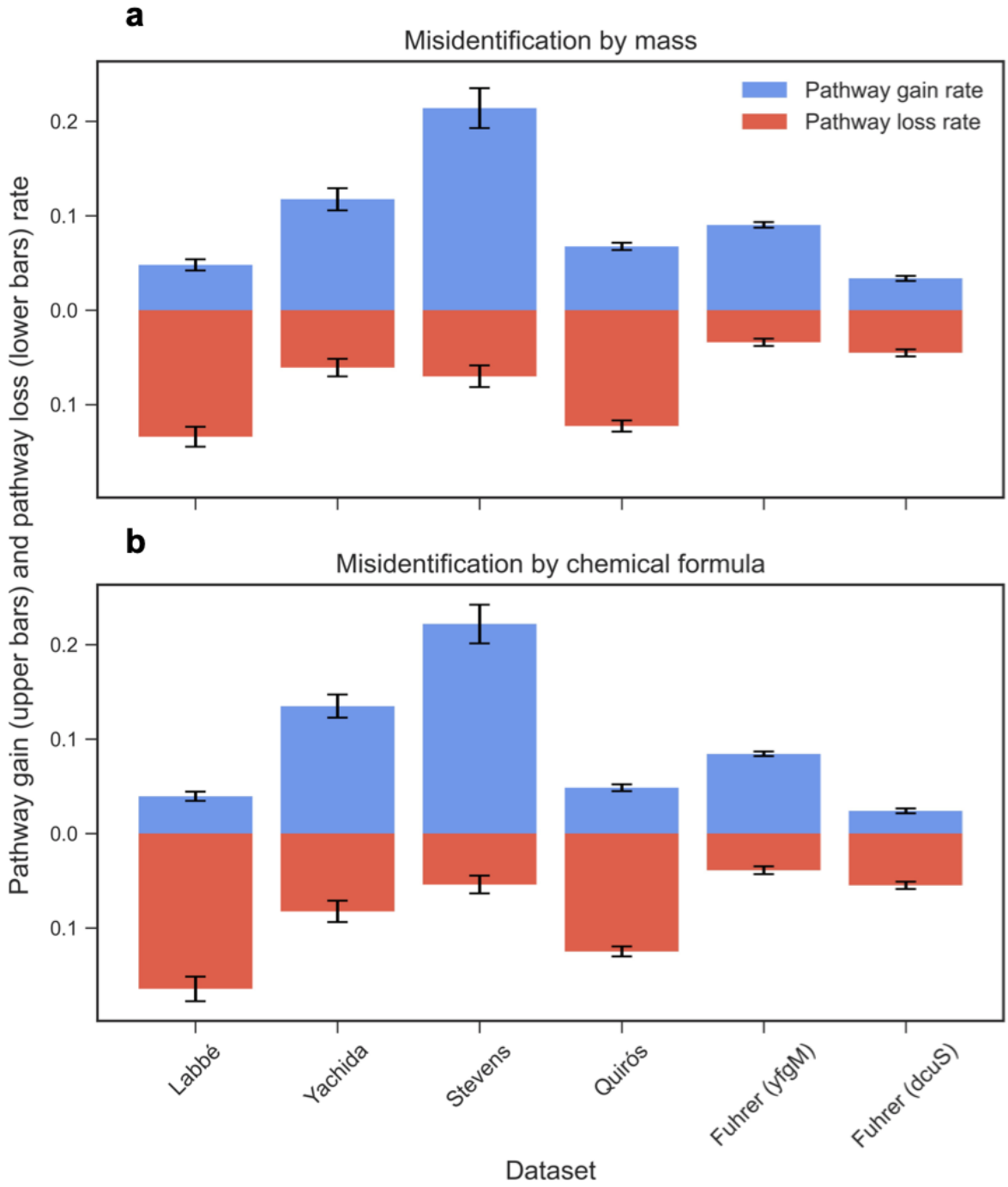

Fig 5: ***Metabolite misidentification****. The effect of compound misidentification by* ***a*** *molecular weight (20ppm window) and* ***b*** *chemical formula on the mean pathway loss rate (red bars) and mean pathway gain rate (blue bars) averaged over 100 random resamplings at 4% misidentification.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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The polarity of compounds in a metabolomics experiment influences the pathways discoverable using ORA

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

Suggested recommendations for the application of ORA to metabolomics data:

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

Suggested recommended minimal reporting criteria. Users should report:

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3. 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).
4. Which compounds form the background set.
5. The multiple testing correction methods applied for i) selection of DA metabolites and ii) selection of SEP, alongside the adjusted p-value thresholds used.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Materials and methods

1. Obtaining the list of metabolites of interest

1.1 Summary of experimental datasets used

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

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

Author	***Title***	***Organism***	***Analytical platform***	***Sample type***	***Total number of metabolites mapping to KEGG compounds***	***Study accession code/data availability***
Labbé et al.	High-fat diet fuels prostate cancer progression by rewiring the metabolome and amplifying the MYC program	*Mus musculus*	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	*Homo sapiens*	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	*Homo sapiens*	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	*Homo sapiens* (HeLa cells)	Flow injection TOF MS*	HeLa cell	1110	Supplementary table S8 of https://doi.org/10.1083/jcb.201702058

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

1.2 Post-processing of metabolomics datasets

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

1.3 Metabolite identifier harmonisation

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

1.4 Selection of differentially abundant metabolites

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

2. Performing pathway enrichment

2.1 Pathway database details

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

2.2 ORA implementation

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

3. In-silico simulation details

3.1 Implementation details

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

3.2 Simulating metabolite misidentification

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

3.3 Quantifying changes in results

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

590 The *pathway loss rate* and *pathway gain rate* at f % metabolite misidentification are then

591 defined as:

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

592

593

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Nguyen TM, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: A comprehensive review and assessment. Genome Biol. 2019;20. doi:10.1186/s13059-019-1790-4

2. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: Current approaches and outstanding challenges. Ouzounis CA, editor. PLoS Computational Biology. Public Library of Science; 2012. p. e1002375. doi:10.1371/journal.pcbi.1002375

3. Karnovsky A, Li S. Pathway Analysis for Targeted and Untargeted Metabolomics. Methods in Molecular Biology. Humana Press Inc.; 2020. pp. 387–400. doi:10.1007/978-1-0716-0239-3_19

4. Marco-Ramell A, Palau-Rodriguez M, Alay A, Tulipani S, Urpi-Sarda M, Sanchez-Pla A, et al. Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. BMC Bioinformatics. 2018;19: 1. doi:10.1186/s12859-017-2006-0

5. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: State of the art. Frontiers in Physiology. Frontiers Research Foundation; 2015. doi:10.3389/fphys.2015.00383

6. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999;22: 281–285. doi:10.1038/10343

7. Drăghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. Global functional profiling of gene expression. Genomics. 2003;81: 98–104. doi:10.1016/S0888-7543(02)00021-6

8. Xie C, Jauhari S, Mora A. Popularity and performance of bioinformatics software: the case of gene set analysis. BMC Bioinformatics. 2021;22: 191. doi:10.1186/s12859-021-04124-5

9. Beauclercq S, Nadal-Desbarats L, Hennequet-Antier C, Gabriel I, Tesseraud S, Calenge F, et al. Relationships between digestive efficiency and metabolomic profiles of serum and intestinal contents in chickens. Sci Rep. 2018;8: 6678. doi:10.1038/s41598-018-24978-9

10. Guo YS, Tao JZ. Metabolomics and pathway analyses to characterize metabolic alterations in pregnant dairy cows on D 17 and D 45 after AI. Sci Rep. 2018;8: 1–8. doi:10.1038/s41598-018-23983-2

11. Michonneau D, Latis E, Curis E, Dubouchet L, Ramamoorthy S, Ingram B, et al. Metabolomics analysis of human acute graft-versus-host disease reveals changes in host and microbiota-derived metabolites. Nat Commun. 2019;10: 1–15. doi:10.1038/s41467-019-13498-3

12. McGeachie MJ, Dahlin A, Qiu W, Croteau-Chonka DC, Savage J, Wu AC, et al. The metabolomics of asthma control: A promising link between genetics and disease. Immun Inflamm Dis. 2015;3: 224–238. doi:10.1002/iid3.61

13. Zhang P, Zhang W, Lang Y, Qu Y, Chen J, Cui L. 1H nuclear magnetic resonance-based metabolic profiling of cerebrospinal fluid to identify metabolic features and markers for tuberculosis meningitis. Infect Genet Evol. 2019;68: 253–264. doi:10.1016/j.meegid.2019.01.003

14. Rosato A, Tenori L, Cascante M, De Atauri Carulla PR, Martins dos Santos VAP, Saccenti E. From correlation to causation: analysis of metabolomics data using systems biology approaches. Metabolomics. Springer New York LLC; 2018. p. 37. doi:10.1007/s11306-018-1335-y

15. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. Oxford University Press; 2000. pp. 27–30. doi:10.1093/nar/28.1.27

16. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020;48: D498–D503. doi:10.1093/nar/gkz1031

17. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, et al. The BioCyc collection of microbial genomes and metabolic pathways. Brief Bioinform. 2018;20: 1085–1093. doi:10.1093/bib/bbx085

18. Cottret L, Frainay C, Chazalviel M, Cabanettes F, Gloaguen Y, Camenen E, et al. MetExplore: Collaborative edition and exploration of metabolic networks. Nucleic Acids Res. 2018;46: W495–W502. doi:10.1093/nar/gky301

19. Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in ingenuity pathway

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

analysis. Bioinformatics. 2014;30: 523–530. doi:10.1093/bioinformatics/btt703

20. Domingo-Fernández D, Hoyt CT, Bobis-Álvarez C, Marín-Llaó J, Hofmann-Apitius M. ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. npj Syst Biol Appl. 2019;5: 1–8. doi:10.1038/s41540-018-0078-8

21. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics. 2007;3: 211–221. doi:10.1007/s11306-007-0082-2

22. Cavill R, Jennen D, Kleinjans J, Briedé JJ. Transcriptomic and metabolomic data integration. Brief Bioinform. 2016;17: 891–901. doi:10.1093/bib/bbv090

23. Emwas AHM. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. Methods Mol Biol. 2015;1277: 161–193. doi:10.1007/978-1-4939-2377-9_13

24. Creek DJ, Dunn WB, Fiehn O, Griffin JL, Hall RD, Lei Z, et al. Metabolite identification: are you sure? And how do your peers gauge your confidence? Metabolomics. 2014;10: 350–353. doi:10.1007/s11306-014-0656-8

25. Dunn WB, Erban A, Weber RJM, Creek DJ, Brown M, Breitling R, et al. Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics. Springer; 2013. pp. 44–66. doi:10.1007/s11306-012-0434-4

26. Stobbe MD, Houten SM, Jansen GA, van Kampen AHC, Moerland PD. Critical assessment of human metabolic pathway databases: A stepping stone for future integration. BMC Syst Biol. 2011;5: 165. doi:10.1186/1752-0509-5-165

27. Karp PD, Midford PE, Caspi R, Khodursky A. Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. BMC Genomics 2021 221. 2021;22: 1–11. doi:10.1186/s12864-021-07502-8

28. Pham N, van Heck RGA, van Dam JCJ, Schaap PJ, Saccenti E, Suarez-Diez M. Consistency, inconsistency, and ambiguity of metabolite names in biochemical databases used for genome-scale metabolic modelling. Metabolites. 2019;9: 28. doi:10.3390/metabo9020028

29. Poupin N, Vinson F, Moreau A, Batut A, Chazalviel M, Colsch B, et al. Improving lipid mapping in Genome Scale Metabolic Networks using ontologies. Metabolomics. 2020;16: 44. doi:10.1007/s11306-020-01663-5

30. Wadi L, Meyer M, Weiser J, Stein LD, Reimand J. Impact of outdated gene annotations on pathway enrichment analysis. Nature Methods. Nature Publishing Group; 2016. pp. 705–706. doi:10.1038/nmeth.3963

31. Frainay C, Schymanski EL, Neumann S, Merlet B, Salek RM, Jourdan F, et al. Mind the gap: Mapping mass spectral databases in genome-scale metabolic networks reveals poorly covered areas. Metabolites. 2018;8. doi:10.3390/metabo8030051

32. Labena AA, Gao YZ, Dong C, Hua H li, Guo FB. Metabolic pathway databases and model repositories. Quantitative Biology. Higher Education Press; 2018. pp. 30–39. doi:10.1007/s40484-017-0108-3

33. Mubeen S, Hoyt CT, Gemünd A, Hofmann-Apitius M, Fröhlich H, Domingo-Fernández D. The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. Front Genet. 2019;10: 1203. doi:10.3389/fgene.2019.01203

34. Fang X, Liu Y, Ren Z, Du Y, Huang Q, Garmire LX. Lilikoi V2.0: a deep learning–enabled, personalized pathway-based R package for diagnosis and prognosis predictions using metabolomics data. Gigascience. 2021;10: 1–11. doi:10.1093/gigascience/giaa162

35. Labbé DP, Zadra G, Yang M, Reyes JM, Lin CY, Cacciatore S, et al. High-fat diet fuels prostate cancer progression by rewiring the metabolome and amplifying the MYC program. Nat Commun. 2019;10: 1–14. doi:10.1038/s41467-019-12298-z

36. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al. Metagenomic and

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nature Medicine. Nature Publishing Group; 2019. pp. 968–976. doi:10.1038/s41591-019-0458-7

37. Stevens VL, Wang Y, Carter BD, Gaudet MM, Gapstur SM. Serum metabolomic profiles associated with postmenopausal hormone use. Metabolomics. 2018;14: 97. doi:10.1007/s11306-018-1393-1

38. Quirós PM, Prado MA, Zamboni N, D'Amico D, Williams RW, Finley D, et al. Multi-omics analysis identifies ATF4 as a key regulator of the mitochondrial stress response in mammals. J Cell Biol. 2017;216: 2027–2045. doi:10.1083/jcb.201702058

39. Fuhrer T, Zampieri M, Sévin DC, Sauer U, Zamboni N. Genomewide landscape of gene–metabolome associations in Escherichia coli. Mol Syst Biol. 2017;13: 907. doi:10.15252/msb.20167150

40. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, et al. MetaboLights: A resource evolving in response to the needs of its scientific community. Nucleic Acids Res. 2020;48: D440–D444. doi:10.1093/nar/gkz1019

41. Sarkans U, Gostev M, Athar A, Behrangi E, Melnichuk O, Ali A, et al. The BioStudies database-one stop shop for all data supporting a life sciences study. Nucleic Acids Res. 2018;46: D1266–D1270. doi:10.1093/nar/gkx965

42. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. Nucleic Acids Res. 2018;46: W486–W494. doi:10.1093/nar/gky310

43. Cokelaer T, Pultz D, Harder LM, Serra-Musach J, Saez-Rodriguez J. BioServices: a common Python package to access biological Web Services programmatically. Bioinformatics. 2013;29: 3241–3242. doi:10.1093/bioinformatics/btt547

44. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995;57: 289–300. Available: http://www.jstor.org/stable/2346101

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Author contributions:

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

Acknowledgements:

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

Competing interests statement:

All authors declare they have no conflict of interest.

Funding:

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

Data Availability

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

Code Availability

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

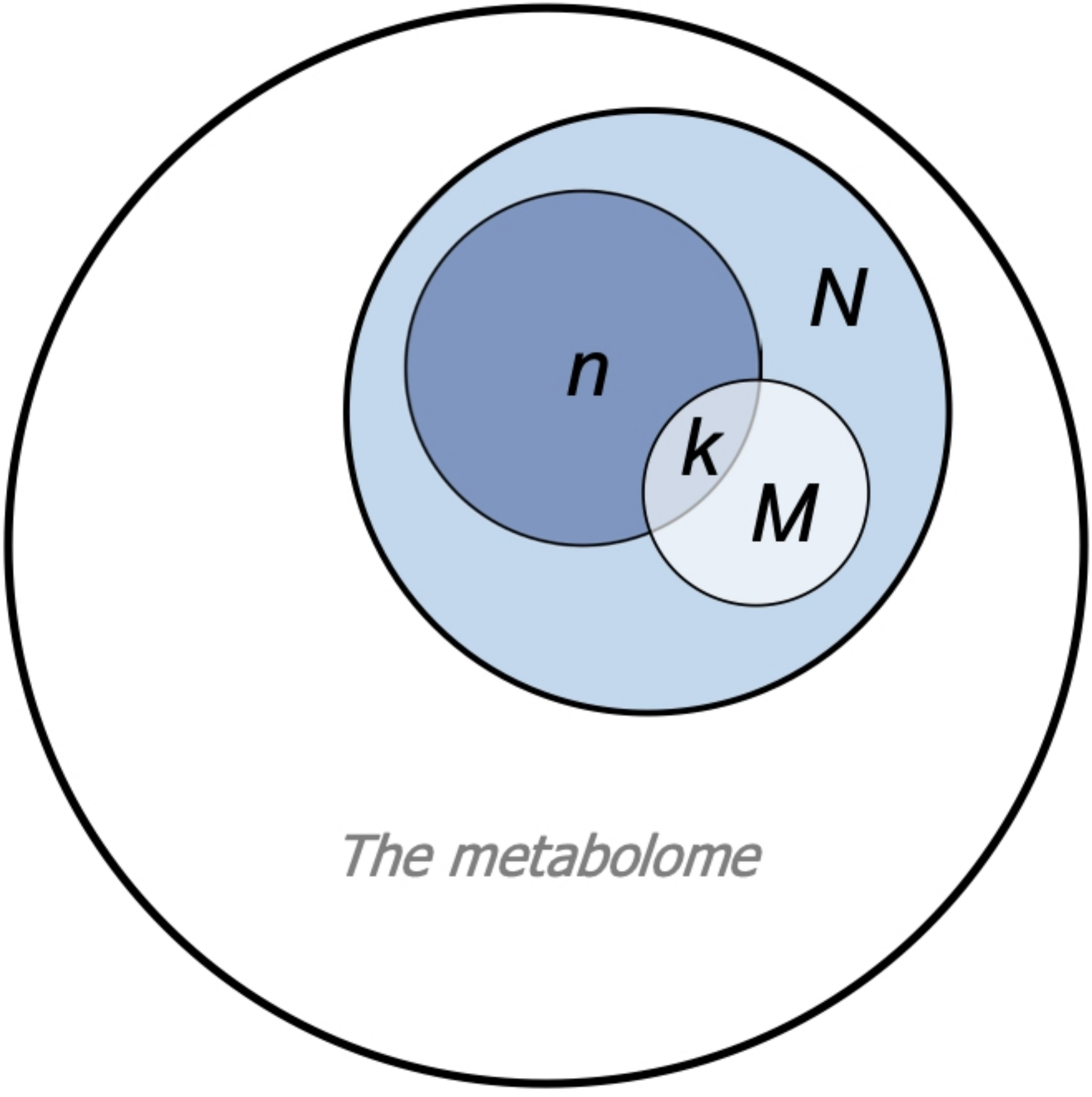

Fig 1

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

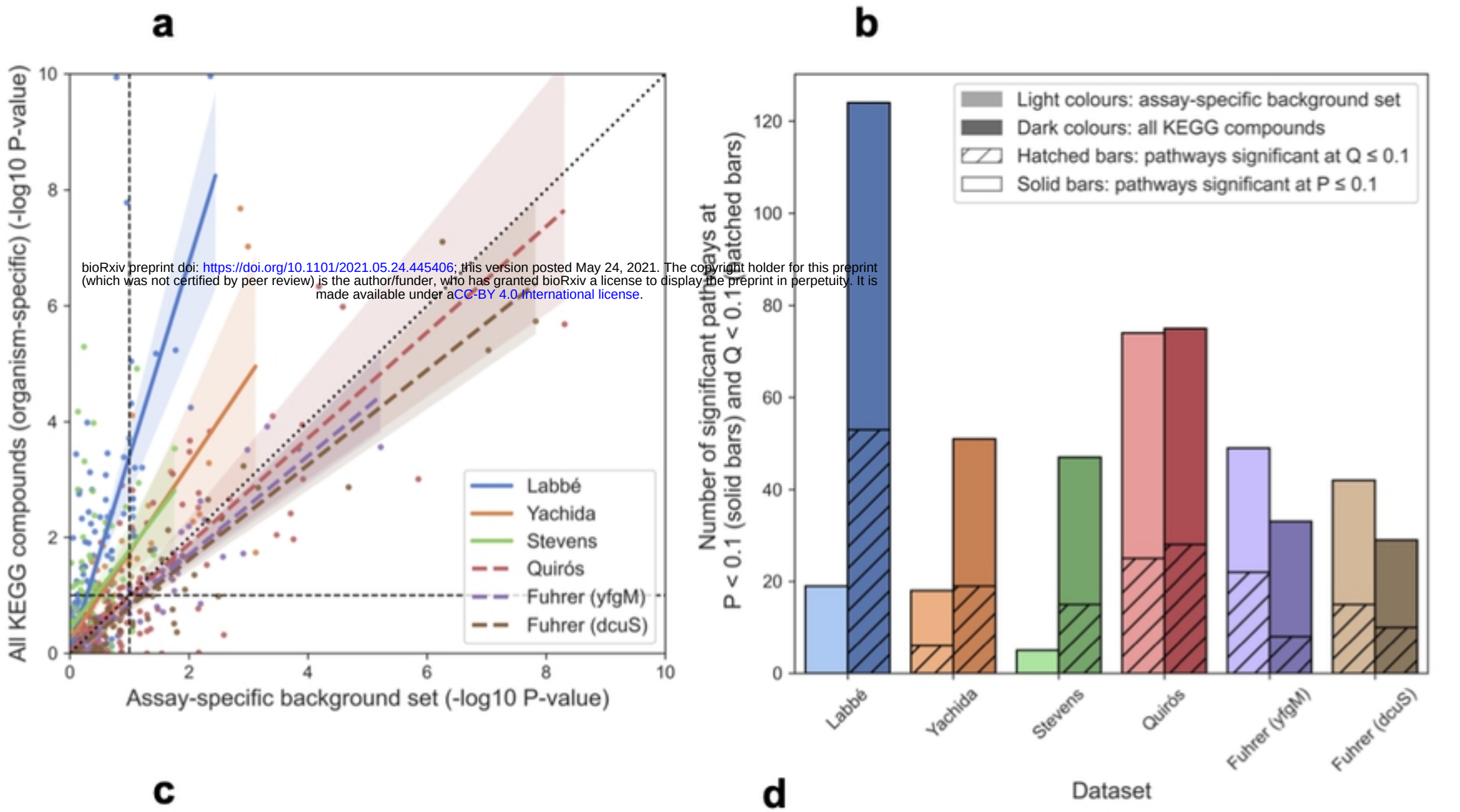

Fig 2

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

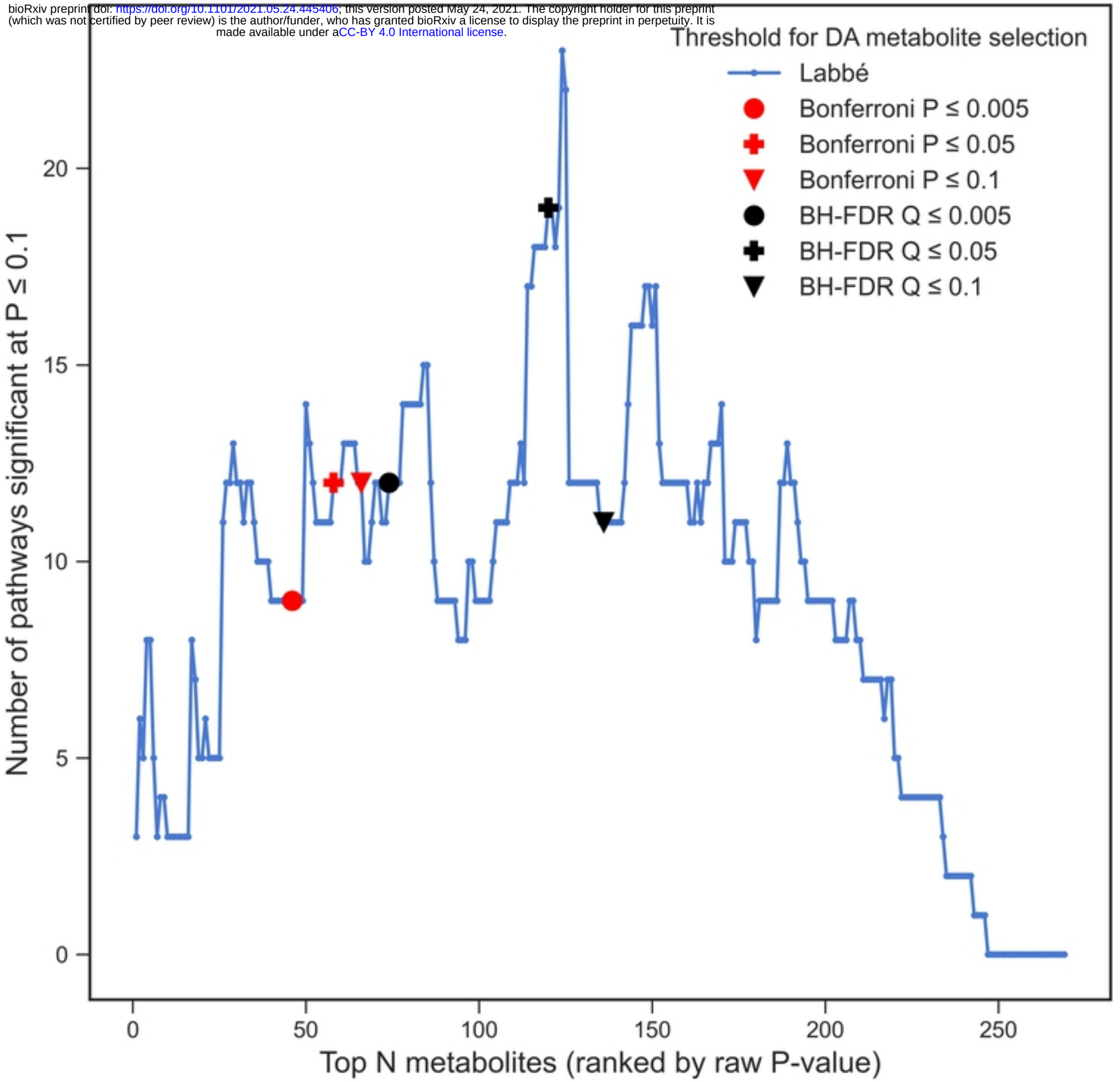

Fig 3

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

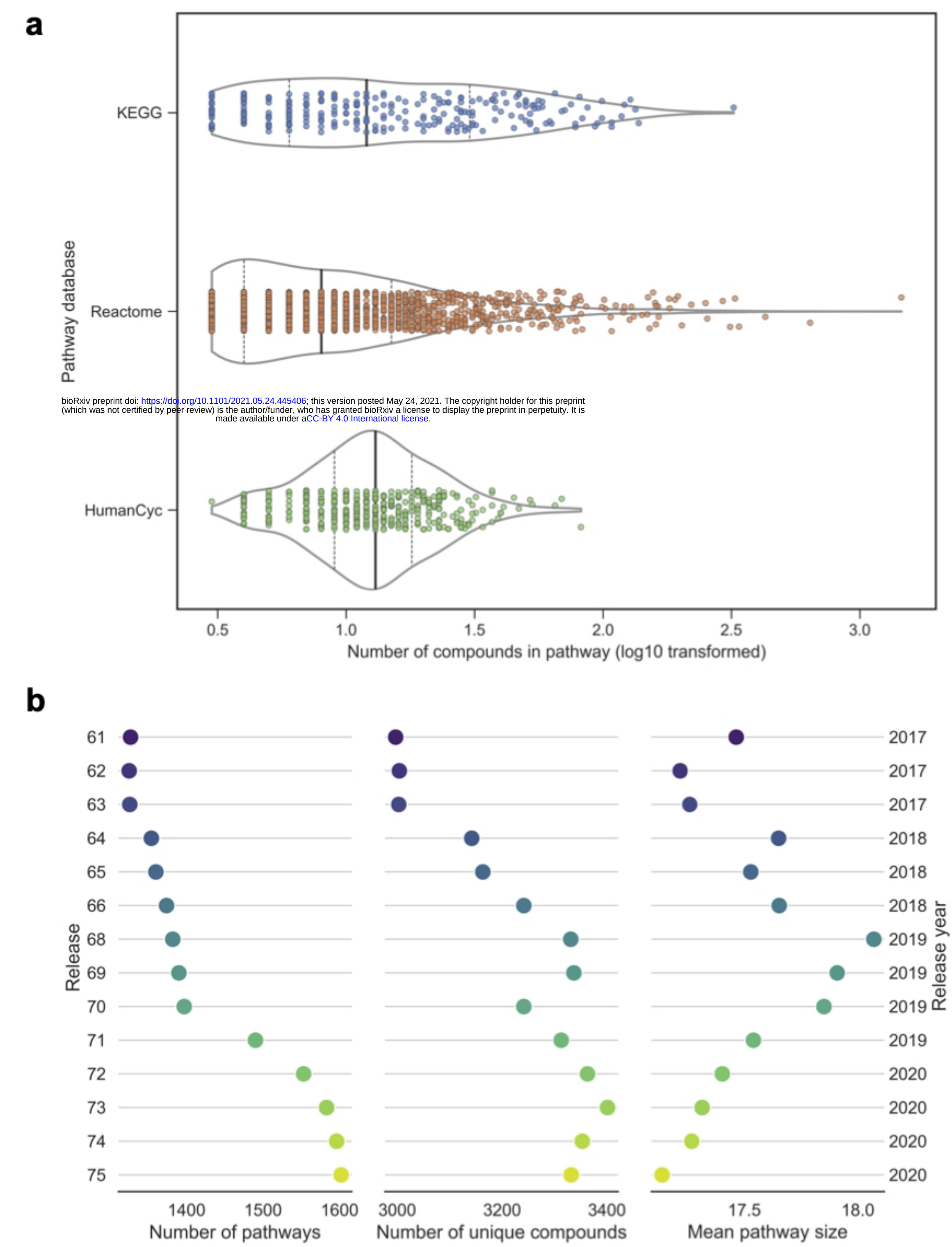

Fig 4

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.24.445406; this version posted May 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

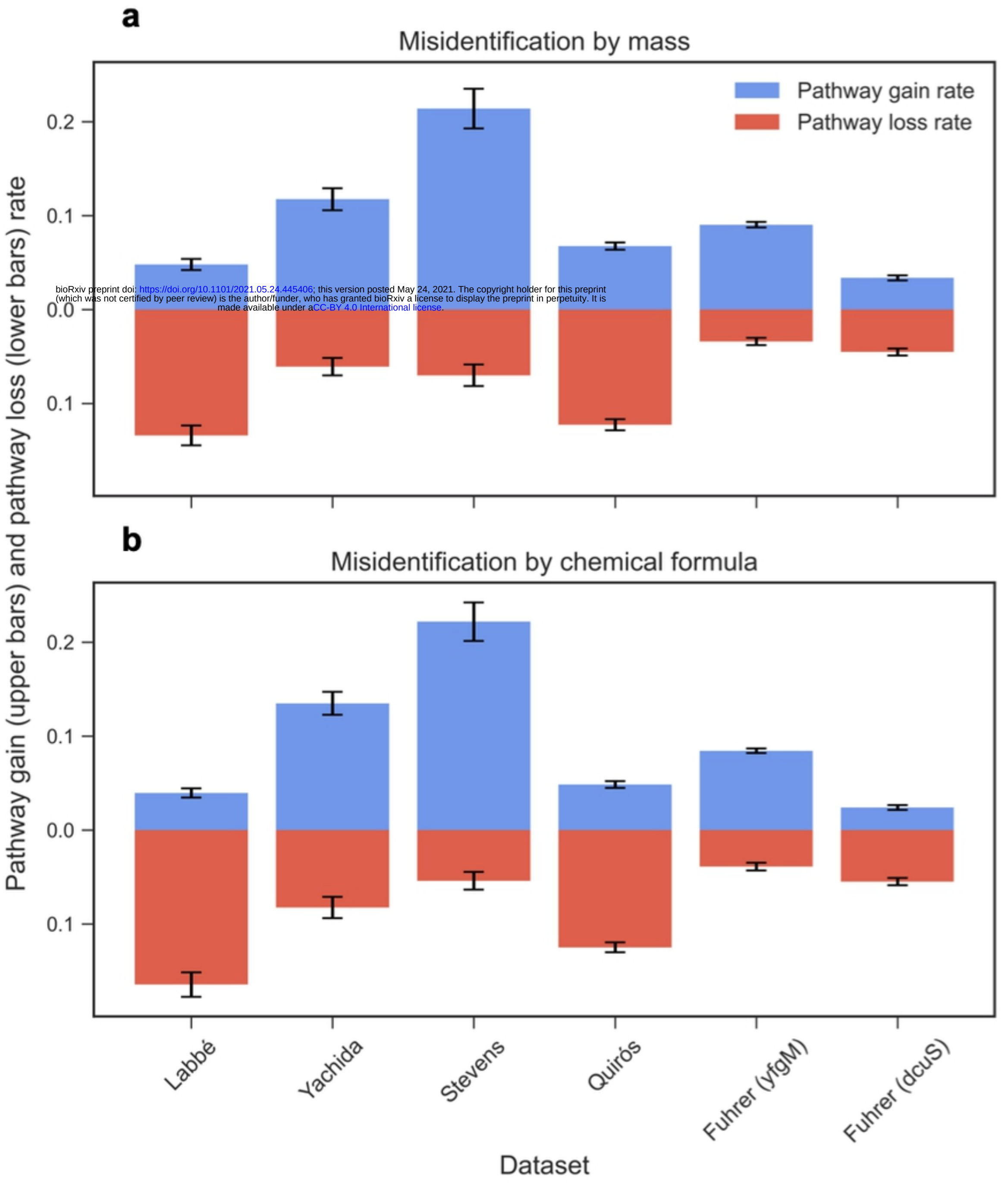

Fig 5