Generalized Reporter Score-based Enrichment Analysis for Omics Data

Chen Peng[1,2,3], Qiong Chen[1,2,3], Shangjin Tan[4,5], Xiaotao Shen[6], Chao Jiang[1,2,3*]

[1]Zhejiang Provincial Key Laboratory of Cancer Molecular Cell Biology, Life Sciences Institute, Zhejiang University, Hangzhou, Zhejiang 310058, China

[2]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, National Clinical Research Center for Infectious Diseases, First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310009, China

[3]Center for Life Sciences, Shaoxing Institute, Zhejiang University, Shaoxing, Zhejiang 321000, China

[4]BGI Research, Wuhan, Hubei 430074, China

[5]BGI Research, Shenzhen, Guangdong 518083, China

[6]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

[*]Correspondence: jiang_chao@zju.edu.cn

Abstract

Enrichment analysis contextualizes biological features in pathways to facilitate a systematic understanding of high-dimensional data and is widely used in biomedical research. The emerging reporter score-based analysis (RSA) method shows more promising sensitivity, as it relies on *p-values* instead of raw values of features. However, RSA cannot be directly applied to multi-group experimental designs and is often misused due to the lack of a proper tool. Here, we propose the Generalized Reporter Score-based Analysis (GRSA) method for multi-group and longitudinal omics data. A comparison with other popular enrichment analysis methods demonstrated that GRSA had increased sensitivity across multiple benchmark datasets. We applied GRSA to microbiome, transcriptome, and metabolome data and discovered new biological insights in omics studies. Finally, we demonstrated the application of GRSA beyond functional enrichment using a taxonomy database. We implemented GRSA in an R package, ReporterScore, integrating with a powerful visualization module and updatable pathway databases (https://github.com/Asa12138/ReporterScore). We believe the ReporterScore package will be a valuable asset for broad biomedical research fields.

Introduction

Functional enrichment analysis is a popular bioinformatic method that helps understand the biological significance of large omics datasets, such as transcriptomic, metagenomic, and metabolic data. We gain insights into the underlying biological processes and functions by identifying enriched functional categories, such as gene ontology terms or biological pathways, and formulate hypotheses for downstream experimental investigations[1].

Methods for functional enrichment analysis can be roughly divided into three categories based on underlying statistical methods: (i) overrepresentation analysis (ORA), (ii) functional class scoring (FCS), and (iii) pathway topology-based (PT)[2]. Common enrichment analysis methods in omics research are shown in Table 1. The algorithm of

45 reporter score-based analysis (RSA) was originally developed by Patil and Nielsen in 2005
46 to identify metabolites associated with the metabolic network's regulatory hotspots[3]. The
47 RSA has recently gained popularity due to its extended application in functional
48 enrichment analysis in microbiome research[4]. RSA is an FCS method based on parsing the
49 *p-values* of selected statistical analyses without a priori cut-off (threshold-free). The
50 rationale is that the *p-value* can be considered a standardized statistic reflecting the
51 differences between different genes or features, regardless of the mean expression
52 values. The pathways with significantly lower *p-values* than the background *p-value*
53 distribution are enriched[3].
54
55 However, RSA is often misused due to a lack of specific tools and a systematic
56 understanding of the algorithm[5]. In addition, the sign (plus or minus) of the reporter score
57 of each pathway in classic RSA does not represent the increasing or decreasing trend of
58 the pathway expression; rather, reporter scores (including negative value) less than a
59 specified threshold indicate that the corresponding pathway is not significantly enriched.
60 This often leads to misinterpretations of the results.
61
62 Inspired by the classic RSA, we developed the improved Generalized Reporter Score-
63 based Analysis (GRSA) method, implemented in the R package ReporterScore, along with
64 comprehensive visualization methods and pathway databases. GRSA is a threshold-free
65 method that works well with all types of biomedical features, such as genes, chemical
66 compounds, and microbial species. GRSA works in the mixed (classic RSA) and directed
67 modes (enhanced RSA). The directed mode uses signs of the reporter score to distinguish
68 up-regulated or down-regulated pathways, which is more intuitive. Importantly, the GRSA
69 supports multi-group and longitudinal experimental designs, because of the included
70 multi-group-compatible statistical methods (for a full list of supported methods, please
71 see Supplementary Table 1). The ReporterScore package also supports custom
72 hierarchical and relational databases (e.g., table containing the correspondence between
73 pathways and genes), providing extra flexibility for advanced users. In this study, we

74 described the comprehensive utility of GRSA. We benchmarked GRSA against other
75 popular enrichment methods across multiple datasets, and demonstrated the
76 applications of GRSA on diverse omics datasets.
77
78 **Table 1.** Some commonly used enrichment analysis methods.

Category	Method	Tools	Notes	Reference
ORA	Hypergeometric test / Fisher's exact test	DAVID (website), clusterProfiler (R package)	The most common methods used in enrichment analysis.	[6,7]
FCS	Gene set enrichment analysis (GSEA)	GSEA (website), clusterProfiler (R package)	A computational method that determines whether a set of genes shows statistically significant and concordant differences between two biological states.	[8,9]
FCS	Generalized reporter score-based analysis (GRSA/RSA)	**ReporterScore (R package developed in this study)**	Find significant metabolites (first report), pathways, and taxonomy based on the *p-values* for multi-omics data.	[3,5]
PT	Reporter feature analysis	/	Integrates bio-molecular network topology with transcriptome data to identify the key biological features.	[10]
PT	Topology-based pathway enrichment analysis (TPEA)	TPEA (R package)	Integrates topological properties and global upstream/downstream positions of genes in pathways.	[11]

79

Results

Workflow overview

The ReporterScore package has built-in KEGG pathway, module, gene, compound, and GO databases and also allows customized databases, making it compatible with feature abundance tables from diverse omics data. A complete gene abundance table can be used for transcriptomic, scRNA-seq, and related gene-based omics data of a specific species. For metagenomic and metatranscriptomic data, which involve many different species, a KO abundance table can be used, generated using Blast, Diamond, or KEGG official mapper software[12] to align the reads or contigs to the KEGG[13] or the EggNOG database[14]. An annotated compound abundance table can be used for metabolomic data, but the standardization of compound IDs (e.g., convert compound IDs to KEGG IDs) is required.

The workflow of GRSA in the ReporterScore package is shown in Fig. 1, using metagenomic data as an example. The KO abundance table (rows are KOs and columns are samples) and metadata table (rows are samples and columns are experimental design groups) were used as the input for GRSA. Importantly, the input data should not be prefiltered to retain the background information. First, the *p-values* for all KOs were calculated by a proper statistical method. Then, in the classic mode, the *p-values* were directly converted to Z-scores (Fig. 1a, equation [i]). In the directed mode, the *p-values* were divided by 2, converted to Z-scores, and assigned plus or minus signs, denoting up- and down-regulated KOs (Fig. 1a, equation [ii-iv]). Next, the Z-score of the pathway j (Z_{path_j}) was calculated by summing the Z scores of KOs within the pathway j and dividing by the square root of the number of KOs (k_j) in the pathway j (Fig. 1a, equation [v]). The Z_{path_j} is further standardized by the background pathway Z-score distribution, generated by randomly sampling k_j KOs from the total KO pool (Fig. 1a, equation [vi]). The standardized pathway Z-score is defined as the reporter score of a pathway ($ReporterScore_j$). The details of the GRSA algorithm are described in the Methods section.

We designed the ReporterScore package to be user-friendly. In one step, the function `reporter_score` calculates the reporter scores for a feature abundance table with associated metadata. The included assorted visualization methods can be used to explore the entire pathways and features within pathways (Fig. 1b-h). The demo code is included in the Methods section.

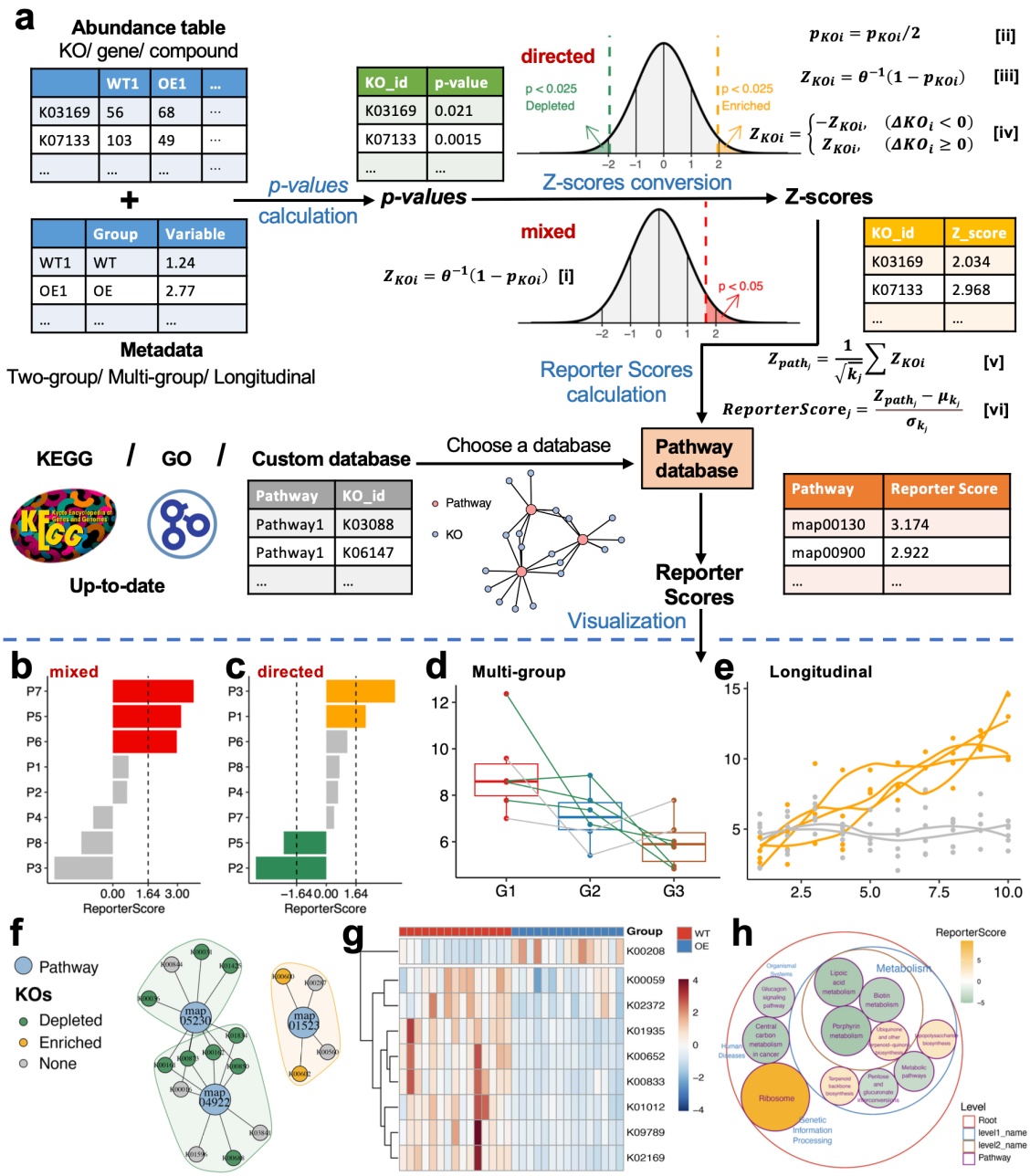


Fig. 1| The overall workflow of GRSA in the ReporterScore package. a GRSA workflow consists of four parts: Calculate the *p-values* for KOs between two or multiple groups by

various statistical methods; convert the *p-values* of KOs to Z-scores by inverse normal distribution and assignment of a plus or minus sign to each Z-score in the directed mode; mapping KOs to annotated pathways and calculating the reporter score for each pathway; visualize the results of GRSA. KO_i represents a certain KO; p_{KO_i} is the *p-value* of KO_i; Z_{KO_i} is the Z-score transformed from p_{KO_i}; ΔKO_i is the abundance difference of between groups. A total of k_j KOs were annotated to the corresponding pathway. μ_{k_j} and σ_{k_j} are the mean and the standard deviation of the background Z-score distribution Z_{path_null}, respectively. The ReporterScore package provides various visualization methods for the GRSA result: **b** The bar chart shows reporter scores of pathways in the mixed mode. The red color indicates significantly enriched pathways, with reporter scores greater than 1.64, corresponding to a *p-value* of 0.05. **c** The bar chart shows reporter scores of pathways in the directed mode. The orange and green colors indicate up-regulated and down-regulated pathways with absolute reporter scores greater than 1.64. **d**-**e** The line charts show the pattern of a selected pathway in the directed mode with a multi-group (**d**) and a longitudinal design (**e**), each line represents the trend of the average abundance of one KO. Line colors indicate whether the KO is significantly enriched (orange), depleted (green), or neither (grey). **f** The network plot shows the KOs present in selected pathways; some KOs can be shared by several pathways. Big dots represent pathways, and small dots represent KOs. The colors of small dots represent the trend of KOs. The colors of the shades encircling pathways denote whether the pathway is overall up-regulated (orange) or down-regulated (green). **g** The heatmap displays the abundance of each KO in a pathway for different samples (columns). **h** The circular packing plot shows the hierarchical relationship of selected pathways; the size of the circle indicates the absolute value of the reporter score, and the color of the circle indicates that the pathway is overall up-regulated (orange) and down-regulated (green).

Applying GRSA to multi-group and longitudinal omics data

An important feature of GRSA is the newly developed directed mode. The key difference between the directed mode and the mixed mode (classic RSA) is that in the directed mode,

145 the reporter score's plus or minus sign indicates the pathway's increasing or decreasing
146 trend (Fig. 1c). However, in the mixed mode, the signs of the reporter score do not
147 indicate the trends of the pathways. We performed GRSA on the public ex_KO_profile
148 dataset (a metagenomic dataset) in two modes (Supplementary Fig. 1a). For pathways
149 enriched in the directed mode, most KOs within the pathway shared the same trend. The
150 pathway with consistently increasing (decreasing) KOs would acquire a significantly bigger
151 (smaller) aggregated Z-score than the background (Supplementary Fig. 1b, blue and red
152 boxes). If KOs within a pathway had opposing trends, the signed Z-scores would cancel
153 each other, leading to insignificant results (Supplementary Fig. 1b, orange box). In
154 comparison, in the mixed mode, the increasing and decreasing trend of the enriched
155 pathway cannot be determined (Supplementary Fig. 1c). Therefore, the directed mode
156 helps find pathways with consistently changing KOs. For simplicity, we used GRSA in the
157 directed mode henceforth.
158
159 Another major advantage of GRSA is the full support of multi-group and longitudinal
160 omics data. The ReporterScore package calculates the *p-value* for each feature between
161 groups using differential abundance analysis or correlation analysis. The Kruskal-Wallis
162 test or ANOVA assesses if the feature abundance varies significantly across multiple
163 groups. The default correlation analysis treats group assignments as ordinal (e.g., groups
164 "G1", "G2", and "G3" will be converted to 1, 2, and 3), so the correlation analysis could
165 evaluate if the feature abundance linearly increases or decreases. Moreover, the
166 ReporterScore package also supports any specified patterns. For example, groups "G1",
167 "G2", and "G3" can be set as 1, 10, and 100 if an exponentially increasing trend is expected.
168 To explore potential patterns within the data, clustering methods, such as C-means
169 clustering, can be used.
170
171 As a general rule, the users must ensure the selected statistical methods are applicable
172 to the datasets and experimental designs. We applied GRSA with different statistical
173 methods on multiple benchmark datasets. For the classic two-group design, the Jaccard

174 similarity exceeded 0.84 for parametric methods and 0.78 for non-parametric methods,
175 but the Jaccard similarity between parametric and non-parametric methods was lower
176 than 0.63 (Supplementary Fig. 2a). The differences mainly stemmed from parametric
177 versus non-parametric methods. For the multi-group data, users can choose the
178 differential abundance analyses if they aim to enrich significantly altered pathways across
179 groups. Correlation analysis is the preferred choice if the goal is to enrich pathways that
180 show consistent increasing or decreasing patterns (Supplementary Fig. 2b,c). Lastly, GRSA
181 also supports other statistical tests, such as 'DESeq2', 'Edger', 'Limma', 'ALDEX', and
182 'ANCOM'[15], to calculate the reporter scores. The demo code is shown in the Methods
183 section.
184

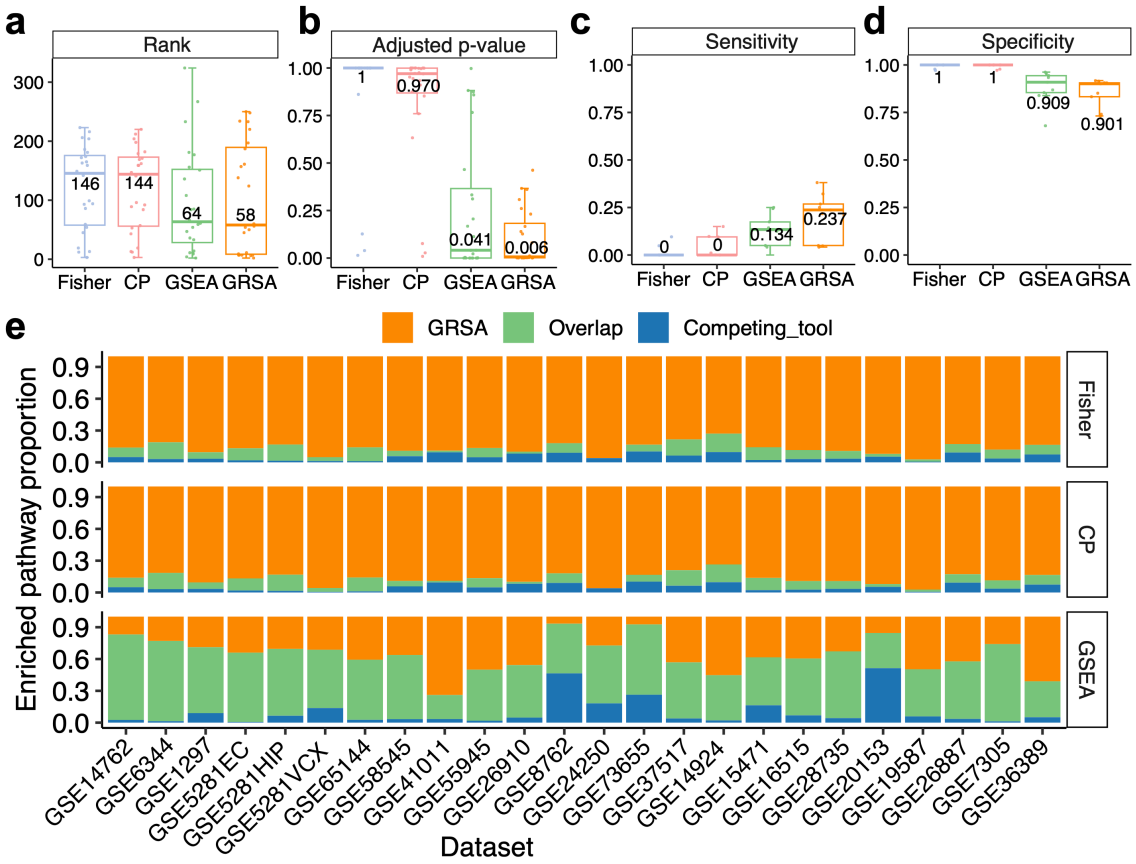
185 GRSA showed higher sensitivity than other commonly used enrichment 186 analysis methods

187 Next, we evaluated the performance of GRSA and compared it with other commonly used
188 enrichment analysis methods (Table 1) on several benchmark datasets (Supplementary
189 Tables 2 and 3). PT-based methods may better identify biologically meaningful pathways
190 than non-PT-based methods in certain scenarios[16]. However, PT-based methods require
191 comprehensive topological structures of pathways, limiting their applications in other
192 non-human organisms[17]. Therefore, we focused on the comparison against non-PT
193 enrichment analysis methods.
194

195 Nguyen et al. proposed several approaches to compare enrichment methods[16], we
196 adapted their methodologies and evaluated the performance of GRSA against the other
197 popular enrichment analysis methods using the identical pathway database: Fisher's
198 exact test (`fisher.test`; R base package), an improved form of Fisher's exact test
199 (`enricher`; `clusterProfiler` package), and gene set enrichment analysis (GSEA).
200 We first compared the efficacy of different methods in identifying the target disease
201 pathways using the 24 gene expression datasets associated with known human diseases
202 (Supplementary Table 2). The rationale is that since each dataset is affiliated with a

203 specific disease-linked KEGG pathway (the target pathway), an optimal enrichment
204 analysis method should rank the target pathway among the top of all 342 pathways and
205 enrich the target pathway with a smaller adjusted *p-value*. Notably, GRSA exhibited the
206 highest performance among the four methods (Fig. 2a,b). Generally, threshold-free
207 methods, such as GRSA and GSEA, outperformed the conventional ORA methods.
208
209 We next evaluated the proficiency of different methods in detecting pathways disturbed
210 by gene knockout experiments. In a gene knockout experiment, the knockout gene is a
211 confirmed source of perturbation. Nguyen et al. considered pathways that included the
212 knockout genes as true positives and pathways not including the target genes as true
213 negatives, so the enrichment of such pathways would be considered false positives.
214 Under these assumptions, we could calculate the sensitivity and specificity of a method.
215 GRSA showed the highest median sensitivity among the methods considered, although its
216 specificity is slightly reduced compared to others (Fig. 2c,d). We prioritize the sensitivity
217 of a method because for pathways that included the knockout gene, deleting it should
218 have a sizable impact on the pathway (sensitivity); however, for pathways that do not
219 include the genes in the database, given the potential incompleteness of pathway and
220 gene databases, simply attributing these enriched pathways as false positives (specificity)
221 may not always be appropriate.
222
223 Lastly, we evaluated the ability of different methods to enrich biologically meaningful
224 pathways[11]. We compared the proportions of pathways identified by the GRSA, the
225 competing tool, and both, using the number of all significant pathways as the
226 denominator (Fig. 2e). GRSA consistently identified a larger proportion of pathways than
227 the two ORA methods and largely overlapped with GSEA. In 21 of 24 gene expression
228 datasets associated with human diseases, GRSA identified more enriched pathways than
229 GSEA. For example, in the renal cell carcinoma dataset (GSE6344), pathways related to
230 cytokine−cytokine receptor interaction[18], IL−17 signaling[19], and PI3K−Akt signaling[20] were
231 only enriched by GRSA (Supplementary Fig. 3a). In the endometrial cancer dataset

232 (GSE7305), GRSA identified pathways related to cancer, Toll−like receptor[21], and cortisol
233 synthesis[22], which have been shown to be involved in the pathological characteristics of
234 endometrial cancer (Supplementary Fig. 3b). Therefore, GRSA can potentially identify
235 more pathways biologically relevant to the studied diseases.



236

237 **Fig. 2| Comparisons of GRSA and other commonly used methods of enrichment**
238 **analyses. a-b** The box charts show the ranks (**a**) and adjusted *p-values* (**b**) of target
239 pathways derived by four methods on 24 gene expression datasets. Numbers represent
240 the median values for each method. **c-d** The box charts show the sensitivity (**c**) and
241 specificity (**d**) of four methods on nine wild-type/knockout gene expression datasets.
242 Numbers represent the median values for each method. **e** Proportion of enriched
243 pathways in GRSA and other enrichment analysis methods based on 24 gene expression
244 datasets. Orange pathways were only identified by GRSA, blue pathways were only
245 identified by the competing tool and green pathways were identified by GRSA and the
246 competing tool. Fisher: fisher.test; CP: improved fisher.test used by clusterProfiler; GSEA:

247 gene set enrichment analysis by clusterProfiler; GRSA: generalized reporter score-based
248 analysis.
249

250 Case Study 1: The functional analysis and age-related dynamics of the skin 251 microbiota

252 Next, we showcased the versatile applications of GRSA with different types of omics data.
253 For microbiome data, we collected the KO profile of the IHSMGC (integrated Human Skin
254 Microbial Gene Catalog) dataset published by Wang et al.[23] and re-analyzed the data using
255 the GRSA method. The previous study calculated the pathway abundance by aggregating
256 the abundances of features within a pathway and then performed differential abundance
257 analysis to identify the most significantly differentially abundant functional modules
258 related to cutotypes. We applied the GRSA to find the functional differences between the
259 two cutotypes. The results were largely consistent. As an example, modules related to the
260 biosynthesis of thiamine, phylloquinone, and cobalamin were enriched in the *M-cutotype*,
261 while modules related to tetrahydrofolate, menaquinone, pantothenate, and ubiquinone
262 were enriched in the *C-cutotype* (Fig. 3a). In addition, the *M-cutotype* was enriched with
263 a large number of modules related to the metabolism of sulfur, phenylacetate (aromatic
264 compound), and amino acids, while the *C-cutotype* was enriched with modules related to
265 carbohydrate metabolism (Supplementary Fig. 4). Importantly, GRSA also identified
266 pathways not found in the previous study. The *M-cutotype* was enriched with modules
267 related to nucleotide metabolism, such as the degradation and de novo biosynthesis of
268 purine (Supplementary Fig. 4), indicating that the *M-cutotype* microbiota may have a
269 higher nucleotide turnover rate and stronger proliferation[24].
270

271 The previous study divided samples into five aging groups and found that the prevalence
272 of the *M-cutotype* significantly increased with age. However, they did not perform age-
273 related functional analysis. We re-analyzed the multi-group data using GRSA to explore
274 the functional dynamics related to aging. The larger positive reporter scores indicate that
275 the module had an overall increasing trend with respect to age, such as 'M00866', related

276 to lipid A biosynthesis (Fig. 3b), while modules with negative reporter scores show an
277 overall decreasing trend with respect to age, such as 'M00061', related to D-Glucuronate
278 degradation (Fig. 3c). We next analyzed the chronological trend of the functional modules
279 at the KEGG level B (Fig. 3d), which better reflects the overall metabolic activities of the
280 microbiome. We found that the carbohydrate metabolism activity of the skin microbiota
281 decreased with aging, while the lipid, amino acid, and nucleotide metabolism activity
282 increased with aging. These results suggest that the energy sources of the skin microbiota
283 significantly change with aging.
284
285 The vitamin biosynthesis-related functional modules also showed differences with
286 respect to aging (Fig. 3d). For the glycan metabolism-related functional modules,
287 biosynthesis of KDO2-lipid A and CMP-KDO increased with aging. KDO2-lipid A is an
288 essential lipopolysaccharide (LPS) component in most gram-negative bacteria, which has
289 endotoxin activity and stimulates host immune responses through Toll-like receptor 4
290 (TLR4)[25]. CMP-KDO is an important intermediate in the synthesis of KDO2-lipid A, and
291 CMP-KDO synthesis is the key rate-limiting step for introducing KDO into LPS[26]. These
292 results suggest that the microbiota of aging skins likely accumulated endotoxins, which
293 can stimulate host inflammation. In addition, we found degradation pathways of several
294 sulfated glycosaminoglycans (chondroitin sulfate, dermatan sulfate, and keratan sulfate)
295 decreased in aging skin. Sulfated glycosaminoglycans play a key role in regulating skin
296 physiology, and there is ample evidence that their properties and functions change over
297 time and with extrinsic skin aging[27,28]. Total sulfated glycosaminoglycan abundance was
298 reduced in aging skin[29], which may lead to the decreased degrading ability of the skin
299 microbiota for the sulfated glycosaminoglycans.

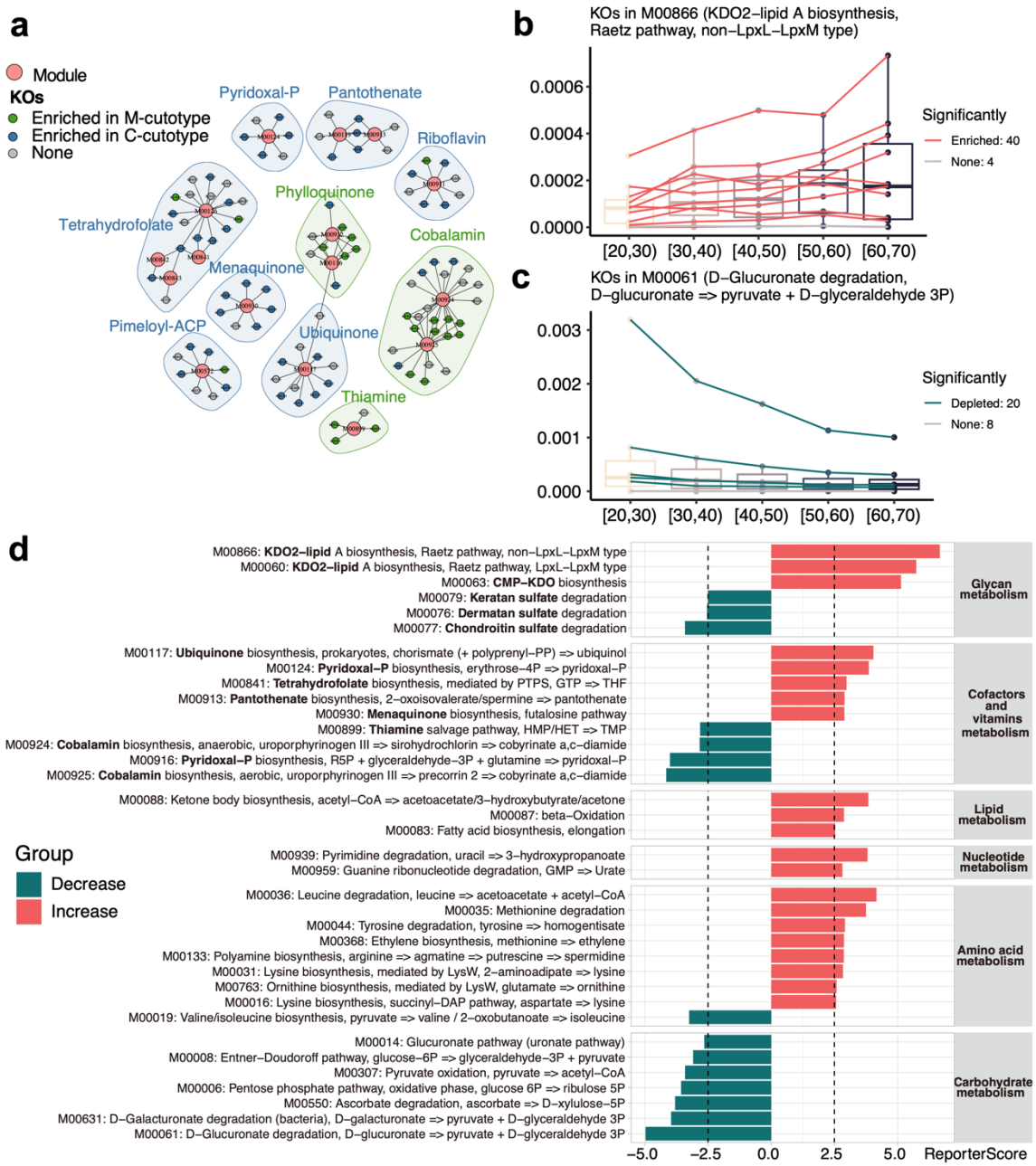


Fig. 3| Application of GRSA to the skin microbiome of the IHSMGC dataset. a The network of KO-Module enriched in the *M-cutotype* (green) and *C-cutotype* (blue). Only modules related to vitamin biosynthesis were shown. Big dots represent modules; small dots represent KOs. The colors of small dots represent cutotypes. Shades indicate modules involved in the biosynthesis of the same vitamin. The colors of shades denote modules enriched in the *M-cutotype* (green) or enriched in the *C-cutotype* (blue). **b**-**c** The Box charts of modules 'M00866' (**b**) and 'M00061' (**c**) across ages. The colors of the lines

308 represent the trend of KO's relative abundance in the module. 'M00866' had the biggest
309 positive reporter score (increasing), while 'M00061' had the biggest absolute value of
310 negative score (decreasing). **d** The Bar charts show significantly enriched modules over
311 aging; the reporter score threshold of 2.5 corresponds to a confidence of about 0.995,
312 and these modules are grouped based on the KEGG level B. Colors denote up-regulated
313 (red) or down-regulated (green) modules with aging.
314

315 Case Study 2: The functional transcriptional dynamics during
316 cardiomyocyte differentiation

317 We applied GRSA to the transcriptomic dataset published by Liu et al. in 2017[30]. The study
318 used the WGCNA method to analyze the temporal transcriptomic changes during the
319 differentiation of cardiomyocytes from 2 hiPSC lines and 2 hESC lines at 4 timepoints
320 (pluripotent stem cells at day 0, mesoderm at day 2, cardiac mesoderm at day 4, and
321 differentiated cardiomyocytes at day 30). Significant changes were observed in the four
322 stages of differentiation among all cell lines. For example, genes in module 1 were highly
323 expressed only in differentiated cardiomyocytes (stage CM), and their enriched Gene
324 Ontology (GO) terms of Biological Process (BP) were related to heart functions, such as
325 regulation of cardiac contraction and muscular system processes. However, WGCNA did
326 not assume the patterns to be linear so that genes can be only highly expressed at day 2
327 during mesoderm development, for example.
328

329 In addition to linearly increasing or decreasing patterns, GRSA allows users to specify any
330 expected patterns for enrichment analysis. To start, we used the fuzzy C-means clustering
331 method to identify the main gene expression patterns (Fig. 4a) and then used these
332 patterns for GRSA to obtain significantly enriched pathways in each pattern (using the
333 RSA_by_cm function in the ReporterScore package). For example, 'Heart process
334 (GO:0003015)' was a significantly enriched GO term for Cluster 6, which was highly
335 expressed only in stage CM (day 30). We identified many genes consistent with the
336 expression pattern of Cluster 6 (Fig. 4b).

337 GRSA results for all clusters are shown in Fig. 4c. Cluster 2 was highly expressed only at
338 day 0, and its enriched GO terms were mainly related to the mitotic cell cycle, which was
339 expected for stem cell self-renewal processes. Cluster 5 had the highest expression level
340 on day 2 and was mainly enriched in various transcription and translation processes.
341 Many known transcription factors, such as EOMES, MIXL1, and WNT3A, were assigned to
342 this cluster and play important roles in mesoderm development. Cluster 4 was highly
343 expressed at day 0 and day 2 and showed a gradually decreasing trend, its function
344 overlapped with Clusters 2 and 5. Cluster 1, highly expressed at day 4, was related to
345 mesoderm formation, such as morphogenesis and organ development. Clusters 3 and 6
346 were primarily up-regulated in differentiated cardiomyocytes (CM stage), and they were
347 related to heart functions, such as regulation of heart contraction and muscle system
348 processes, similar to module 1 in the previous study. Interestingly, the biological
349 processes of hiPSCs/hESCs at day 2 (Cluster 5) induced various RNA-related metabolisms,
350 which were not found in the previous study, indicating that complex transcriptional
351 regulations are involved for further mesoderm formation. Therefore, using the identified
352 expression patterns across groups, we successfully identified pathways and modules
353 essential to different stages of the cardiomyocyte differentiation processes.

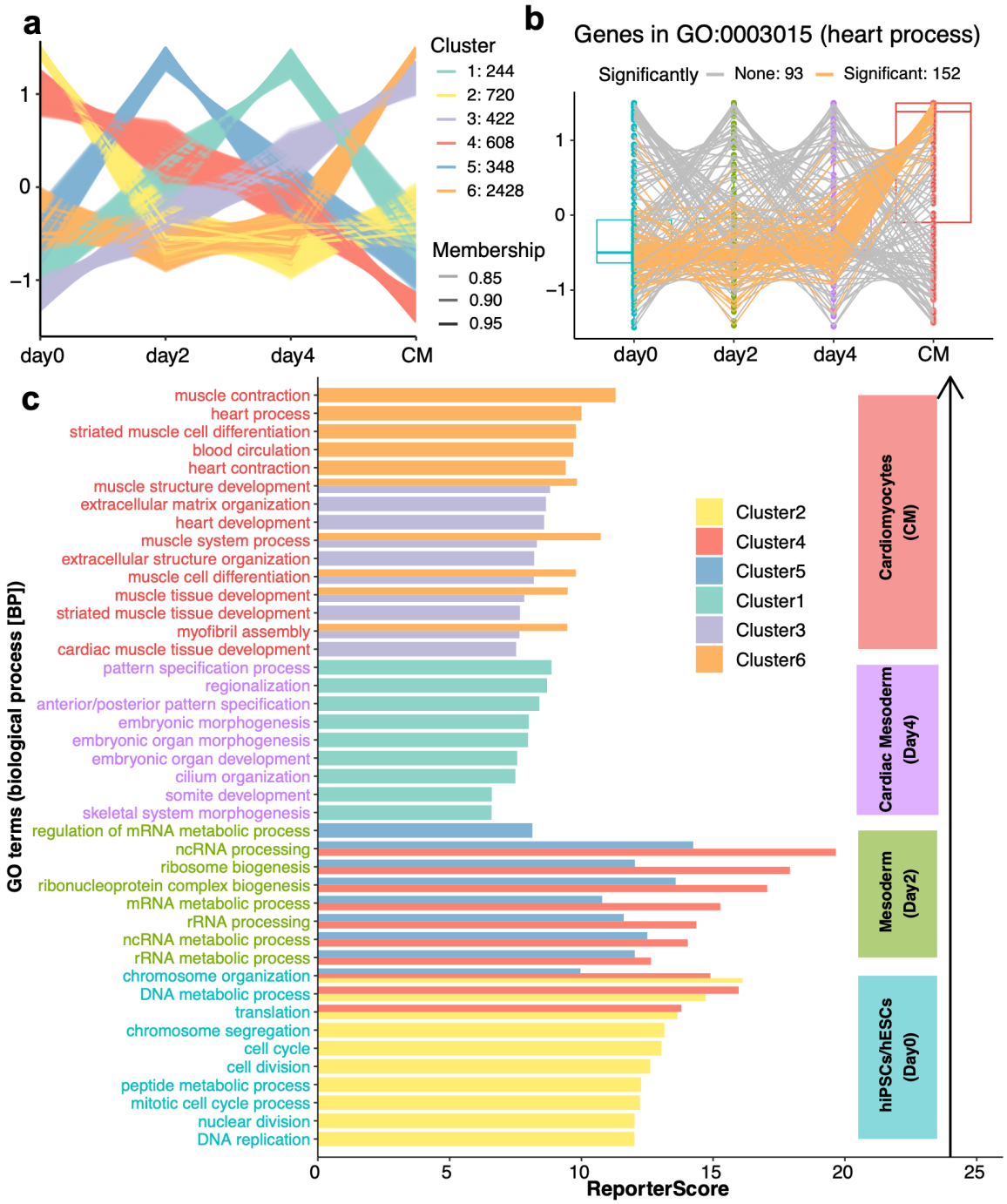


Fig. 4| Application of GRSA to the transcriptomic dataset on the cardiomyocyte differentiation processes. a C-means clustering result of gene abundance profiles across four differentiation stages. The genes with membership scores greater than 0.8 were displayed. The alpha (transparency) of each line was related to the value of its membership score, and the y-axis represent the standardized abundance. **b** The box chart

360 shows the abundance of genes in 'GO:0003015' (heart process) across four time points;
361 the colors of the lines represent the correlative significance of each gene with Cluster 6
362 within the GO term. 'GO:0003015' is a representative term of Cluster 6. **c** The bar chart
363 shows significantly enriched GO terms for each clustering pattern corresponding to the
364 differentiation stages. The colors of the bars represent the cluster information, and the
365 representative GO terms with high reporter scores in each cluster were shown. The text
366 labels on the left were colored according to the stages with the highest expression. In
367 general, Cluster 2 corresponds to day 0, Clusters 4 and 5 correspond to day 2, Cluster 1
368 corresponds to day 4, and Clusters 3 and 6 correspond to CM. Note that only pathways
369 with significant positive scores are shown. The negative score of specified patterns would
370 indicate anti-correlative patterns, which should have already been identified by c-means
371 analysis, such as cluster 3 vs. 4.
372

373 Case Study 3: The systematic maternal metabolomic changes correlated 374 with gestational age

375 We next applied GRSA to metabolomic data from a Danish pregnancy cohort in which
376 female participants had blood drawn weekly from pregnancy to the postpartum period
377 for untargeted metabolomics analysis[31]. Using gestational age as the study variable, they
378 modeled a metabolic clock and found that several marker metabolites increased linearly
379 with gestational age.
380

381 We found several important pathways upregulated with gestational age: steroid hormone
382 biosynthesis, cortisol synthesis and secretion, and Oocyte meiosis (Fig. 5a). Multiple
383 steroid hormones were up-regulated with increasing gestational age (Fig. 5b), including
384 progesterone that interacts with the hypothalamic-pituitary-adrenal axis (HPA axis)[32] and
385 estriol-16-glucuronide produced by the placenta[33]. At the same time, two androgen-
386 related steroid hormones were down-regulated: dehydroepiandrosterone sulfate and
387 androsterone 3-glucuronide, as the concentration of androgens plays important
388 physiological functions during pregnancy[34]. We also found that pathways related to the

389 metabolism of aromatic amino acids were down-regulated with increasing gestational
390 age (Fig. 5a), which has been reported[35].
391
392 Importantly, we identified several up-regulated pathways related to human diseases not
393 mentioned in the previous study. Cushing syndrome happens when the body has too
394 much of the hormone cortisol for a long time, which could be induced by a healthy
395 pregnancy[36]. The up-regulation of pathways related to breast cancer was also noticeable,
396 as Pregnancy-Associated Breast Cancer (PABC) accounts for 7% of all breast cancer in
397 young women[37]. More potential discoveries can be made if the metabolite-pathway
398 database is improved.
399

400 Case Study 4: The application of GRSA beyond functional enrichment 401 analysis

402 The algorithm of GRSA suggests that any features organized in a hierarchical relationship
403 can be used as an enrichment database. For example, we can perform taxonomic
404 enrichment analysis using the phylogenetic relationships of microbes, such as genus-
405 species relationships. To demonstrate, with the `custom_modulelist` function, we used
406 the species abundance table from the IHSMGC dataset and looked for genera enriched in
407 the two cutotypes. We found that *Psychrobacter*, *Paracoccus*, *Chryseobacterium*,
408 *Elizabethkingia*, *Deinococcus*, and *Microbacterium* were enriched in the *M-cutotype*,
409 while *Acidipropionibacterium*, *Staphylococcus*, *Corynebacterium*, and *Cutibacterium* were
410 enriched in the *C-cutotype* (Fig. 5c), some of which were highly consistent with the
411 differential species modules found by co-occurrence network in the previous study.
412 However, we additionally found some genera, such as *Brevundimonas* and *Rhodobacter*,
413 enriched in the *M-cutotype*, while *Pahexavirus* (phages of *Propionibacterium* and
414 *Cutibacterium*) was enriched in the *C-cutotype* (Fig. 5c), probably due to the higher
415 sensitivity of GRSA.
416

Two species, *Moraxella osloensis* and *Cutibacterium acnes*, were used to define the cutotype in the previous study. Interestingly, while the *Cutibacterium* genus was a good biomarker between cutotypes, the *Moraxella* genus was not, as the included species did not share the same trend (Fig. 5d). Therefore, in addition to functional enrichment analysis, GRSA can be extended to any hierarchical relational data structure.

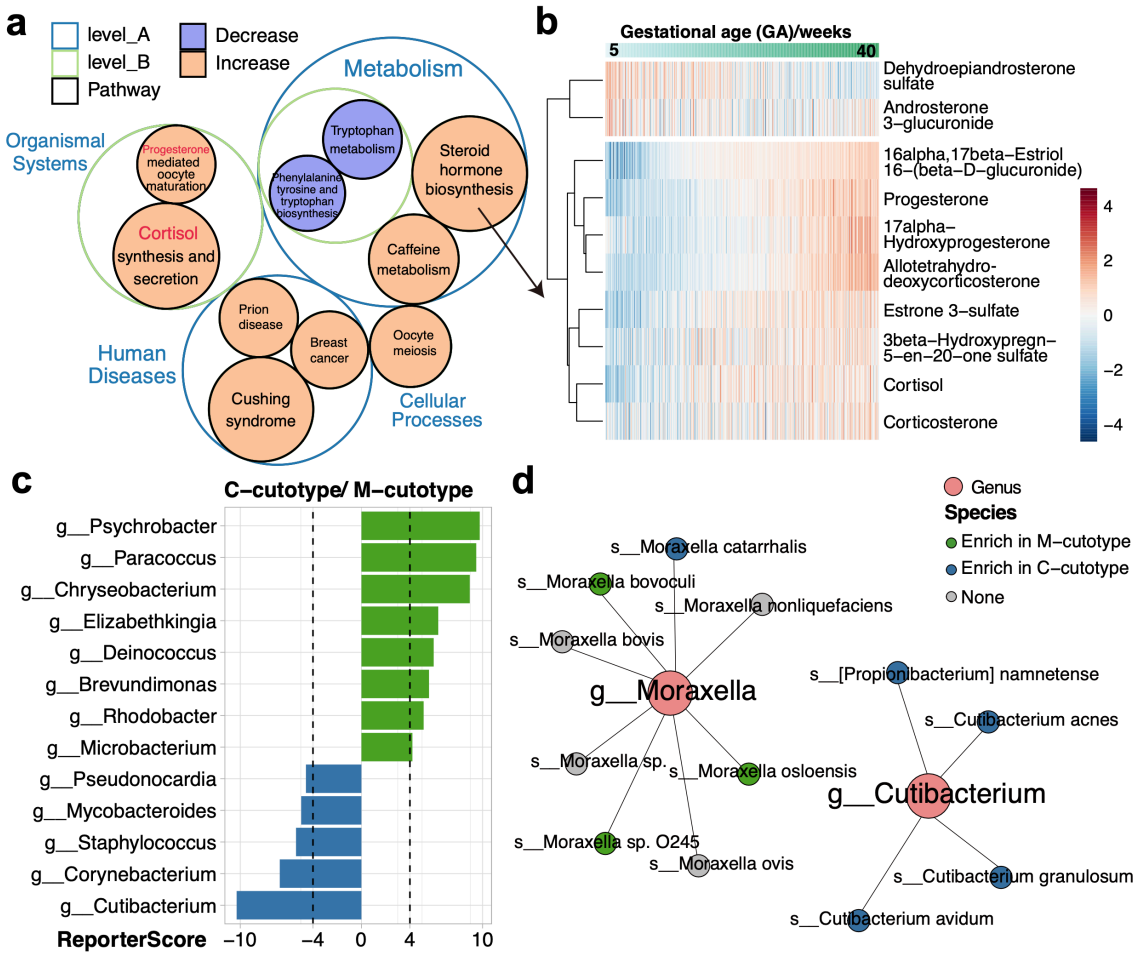


Fig. 5| Application of GRSA in the metabolic data of the Danish pregnancy cohort and the taxonomic enrichment analysis of the IHSMGC dataset. a The circle packing chart shows the hierarchical relationships of significantly enriched pathways identified by GRSA in the metabolomic study. The size of the circle indicates the absolute value of the reporter score, and the color of the circle indicates the sign of the reporter score. The positive reporter score indicates the pathway was increased (orange), while the negative reporter score indicates the pathway was decreased (purple). **b** The heatmap shows the abundance of metabolites in the pathway 'steroid hormone biosynthesis'. The columns

431 are samples ordered by the increasing gestational age. **c** The bar chart shows significantly
432 enriched genera in the *C-cutotype* and *M-cutotype*. **d** The network plot shows the species
433 in *g_Moraxella* and *g_Cutibacterium* enriched in the *M-cutotype* (green) or *C-cutotype*
434 (blue).
435

436 Discussion

437 We developed the ReporterScore package and demonstrated broad applications of the
438 GRSA enrichment analyses. We improved the classic RSA method for easier interpretation
439 of the plus and minus signs of the reporter scores. More importantly, we expanded the
440 scope of GRSA from two-group designs to multi-group and longitudinal designs. We
441 demonstrated these new features with metagenomic, transcriptomic, and metabolomic
442 data (Fig. 3-5). Lastly, we showed that the GRSA is not limited to functional enrichment
443 analysis. Notably, all figures were generated using the visualization module in the
444 ReporterScore package.
445
446 GRSA considers all KOs involved in the pathway compared to hypergeometric tests that
447 only consider a pre-defined list (e.g., KO/gene with *p-value* < 0.05). Thus, GRSA is more
448 sensitive and can comprehensively assess the feature abundance differences in the
449 pathway (Fig. 2). Compared with GSEA, which considers the magnitude of gene changes
450 for ranking, GRSA uses *p-values* for ranking and permutation. We considered sensitivity a
451 crucial metric for enrichment methods because an effective method should maximize the
452 identification of the expected pathways (Fig. 2c). In contrast, we cannot precisely define
453 the true negative results due to the incompleteness of databases. Thus, the specificity of
454 a method cannot be accurately determined (Fig. 2d). Moreover, the primary role of
455 functional enrichment analysis is to guide for interpreting the omics data. Additional
456 experimental verification is essential to substantiate the analysis results.
457
458 Multi-omic studies are increasingly prevalent and so are the demands for functional
459 enrichment analysis of all types of omics data[38,39,40,41]. The GRSA applies to all types of

460 omics datasets as long as a relevant relational database is available. As demonstrated by
461 case studies, we confirmed previous key findings and acquired new biological insights. For
462 example, applying GRSA on the IHSMGC dataset suggested different functional profiles
463 between aging and young skin microbiota. Biosynthesis of KDO2-lipid A and CMP-KDO
464 increased while degradation pathways of several sulfated glycosaminoglycans decreased
465 in older skin microbiota, which may be linked to changes in the skin's physiological
466 properties. Further studies are needed to investigate the underlying mechanisms and
467 their implications for skin health. Applying GRSA on the transcriptomic data of
468 cardiomyocyte differentiations revealed that hiPSCs/hESCs at day 2 specialized in various
469 RNA-related metabolisms, suggesting the involvement of complex transcriptional
470 regulation in further mesoderm formation. Finally, applying GRSA to the metabolomic
471 data from the Danish pregnancy cohort showed that several pathways related to human
472 diseases were up-regulated with gestational age, including Cushing syndrome and PABC.
473
474 The GRSA offers the option for user-specified patterns for enrichment analysis, allowing
475 for rapid testing of educated hypotheses in complex multi-group studies. This is
476 demonstrated in our analysis of the transcriptomic data during cardiomyocyte
477 differentiations. The GRSA offers applications beyond functional enrichment analysis.
478 Applying GRSA in the taxonomic enrichment analysis of the IHSMGC dataset identified
479 key genera that significantly differed between the two cutotypes. The results were highly
480 consistent with the microbial co-occurrence network analysis in the previous study, but
481 performing GRSA was much easier and faster than the network analysis.
482
483 In summary, we believe the GRSA and the ReporterScore package can greatly facilitate
484 the functional enrichment analyses of diverse omics data, with higher sensitivity,
485 compatibility with multi-group and longitudinal designs, and flexibility with customized
486 databases for creative applications beyond functional enrichment analyses.

487 ## References

488 1. Zhao, K. & Rhee, S. Y. Interpreting omics data with pathway enrichment analysis.
489 *Trends in Genetics* **39**, 308–319 (2023).
490 2. Mubeen, S., Tom Kodamullil, A., Hofmann-Apitius, M. & Domingo-Fernández, D.
491 On the influence of several factors on pathway enrichment analysis. *Briefings in*
492 *Bioinformatics* **23**, bbac143 (2022).
493 3. Patil, K. R. & Nielsen, J. Uncovering transcriptional regulation of metabolism by
494 using metabolic network topology. *Proceedings of the National Academy of Sciences of*
495 *the United States of America* **102**, 2685–2689 (2005).
496 4. Zhu, F. *et al.* Metagenome-wide association of gut microbiome features for
497 schizophrenia. *Nature Communications* **11**, 1612 (2020).
498 5. Liu, L., Zhu, R. & Wu, D. Misuse of reporter score in microbial enrichment analysis.
499 *iMeta* **2**, e95 (2023).
500 6. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics
501 data. *Innovation (Cambridge (Mass.))* **2**, 100141 (2021).
502 7. Naylor, D. *et al.* Deconstructing the Soil Microbiome into Reduced-Complexity
503 Functional Modules. *mBio* **11**, 10.1128/mbio.01349–20 (2020).
504 8. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data
505 using g:Profiler, GSEA, cytoscape and EnrichmentMap. *Nature Protocols* **14**, 482–517
506 (2019).
507 9. Zhu, C.-M. *et al.* Alternate succession of aggregate-forming cyanobacterial genera
508 correlated with their attached bacteria by co-pathways. *Science of The Total Environment*
509 **688**, 867–879 (2019).
510 10. Oliveira, A. P., Patil, K. R. & Nielsen, J. Architecture of transcriptional regulatory
511 circuits is knitted over the topology of bio-molecular interaction networks. *BMC Systems*
512 *Biology* **2**, 17 (2008).
513 11. Yang, Q. *et al.* Pathway enrichment analysis approach based on topological
514 structure and updated annotation of pathway. *Briefings in Bioinformatics* **20**, 168–177
515 (2019).

516 12. Kanehisa, M. & Sato, Y. KEGG Mapper for inferring cellular functions from protein
517 sequences. *Protein Science* **29**, 28–35 (2020).
518 13. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a
519 reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457–
520 462 (2016).
521 14. Huerta-Cepas, J. *et al.* eggNOG 5.0: A hierarchical, functionally and
522 phylogenetically annotated orthology resource based on 5090 organisms and 2502
523 viruses. *Nucleic Acids Research* **47**, D309–D314 (2019).
524 15. Nearing, J. T. *et al.* Microbiome differential abundance methods produce different
525 results across 38 datasets. *Nature Communications* **13**, 342 (2022).
526 16. Nguyen, T.-M., Shafi, A., Nguyen, T. & Draghici, S. Identifying significantly
527 impacted pathways: A comprehensive review and assessment. *Genome Biology* **20**, 203
528 (2019).
529 17. Ma, J., Shojaie, A. & Michailidis, G. A comparative study of topology-based
530 pathway enrichment analysis methods. *BMC Bioinformatics* **20**, 546 (2019).
531 18. Ocvirk, S. & O'Keefe, S. J. D. Dietary fat, bile acid metabolism and colorectal cancer.
532 *Seminars in Cancer Biology* **73**, 347–355 (2021).
533 19. Jarocki, M. *et al.* Interleukin 17 and Its Involvement in Renal Cell Carcinoma.
534 *Journal of Clinical Medicine* **11**, 4973 (2022).
535 20. Banumathy, G. & Cairns, P. Signaling pathways in renal cell carcinoma. *Cancer*
536 *Biology & Therapy* **10**, 658–664 (2010).
537 21. Arabian, S., Boostan, A. & Darzi, S. The role of toll-like receptors (TLRs) and their
538 therapeutic applications in endometrial cancer. *Clinical & Translational Oncology: Official*
539 *Publication of the Federation of Spanish Oncology Societies and of the National Cancer*
540 *Institute of Mexico* **25**, 859–865 (2023).
541 22. Larsson, S. C., Lee, W.-H., Kar, S., Burgess, S. & Allara, E. Assessing the role of
542 cortisol in cancer: A wide-ranged Mendelian randomisation study. *British Journal of*
543 *Cancer* **125**, 1025–1029 (2021).

544 23. Li, Z. *et al.* Characterization of the human skin resistome and identification of two
545 microbiota cutotypes. *Microbiome* **9**, 47 (2021).
546 24. Diehl, F. F. *et al.* Nucleotide imbalance decouples cell growth from cell
547 proliferation. *Nature Cell Biology* **24**, 1252–1264 (2022).
548 25. Wang, X., Quinn, P. J. & Yan, A. Kdo2-lipid A: Structural diversity and impact on
549 immunopharmacology. *Biological Reviews* **90**, 408–427 (2015).
550 26. Heyes, D. J. *et al.* Structure-based Mechanism of CMP-2-keto-3-deoxymanno-
551 octulonic Acid Synthetase. *Journal of Biological Chemistry* **284**, 35514–35523 (2009).
552 27. Wang, S. T., Neo, B. H. & Betts, R. J. Glycosaminoglycans: Sweet as Sugar Targets
553 for Topical Skin Anti-Aging. *Clinical, Cosmetic and Investigational Dermatology* **14**, 1227–
554 1246 (2021).
555 28. Shin, J.-W. *et al.* Molecular Mechanisms of Dermal Aging and Antiaging
556 Approaches. *International Journal of Molecular Sciences* **20**, 2126 (2019).
557 29. Oh, J.-H., Kim, Y. K., Jung, J.-Y., Shin, J.-E. & Chung, J. H. Changes in
558 glycosaminoglycans and related proteoglycans in intrinsically aged human skin in vivo.
559 *Experimental Dermatology* **20**, 454–456 (2011).
560 30. Liu, Q. *et al.* Genome-Wide Temporal Profiling of Transcriptome and Open
561 Chromatin of Early Cardiomyocyte Differentiation Derived From hiPSCs and hESCs.
562 *Circulation Research* **121**, 376–391 (2017).
563 31. Liang, L. *et al.* Metabolic dynamics and prediction of gestational age and time to
564 delivery in pregnant women. *Cell* **181**, 1680–1692.e15 (2020).
565 32. Chrousos, G. P., Torpy, D. J. & Gold, P. W. Interactions between the Hypothalamic-
566 Pituitary-Adrenal Axis and the Female Reproductive System: Clinical Implications. *Annals*
567 *of Internal Medicine* **129**, 229–240 (1998).
568 33. Levitz, M., Kadner, S. & Young, B. K. Intermediary metabolism of estriol in
569 pregnancy. *Journal of Steroid Biochemistry* **20**, 971–974 (1984).
570 34. Makieva, S., Saunders, P. T. K. & Norman, J. E. Androgens in pregnancy: Roles in
571 parturition. *Human Reproduction Update* **20**, 542–559 (2014).

572 35. Cox, B. D. & Calame, D. P. Changes in Plasma Amino Acid Levels during the Human
573 Menstrual Cycle and in Early Pregnancy. A Preliminary Report. *Hormone and Metabolic*
574 *Research* **10**, 428–433 (1978).
575 36. Hamblin, R., Coulden, A., Fountas, A. & Karavitaki, N. The diagnosis and
576 management of Cushing's syndrome in pregnancy. *Journal of Neuroendocrinology* **34**,
577 e13118 (2022).
578 37. As, G. *et al.* Association with pregnancy increases the risk of local recurrence but
579 does not impact overall survival in breast cancer: A case-control study of 87 cases. *Breast*
580 *(Edinburgh, Scotland)* **30**, (2016).
581 38. Jiang, C. *et al.* Dynamic Human Environmental Exposome Revealed by Longitudinal
582 Personal Monitoring. *Cell* **175**, 277–291.e31 (2018).
583 39. Li, F. *et al.* Multi-omic Profiling Reveals that Intra-abdominal-Hypertension-
584 Induced Intestinal Damage Can Be Prevented by Microbiome and Metabolic Modulations
585 with 5-Hydroxyindoleacetic Acid as a Diagnostic Marker. *mSystems* **7**, e01204–21 (2022).
586 40. Wei, X. *et al.* Charting the landscape of the environmental exposome. *iMeta* **1**, e50
587 (2022).
588 41. Gao, P. *et al.* Precision environmental health monitoring by longitudinal exposome
589 and multi-omics profiling. *Genome Research* **32**, 1199–1214 (2022).
590

Methods

Algorithm

The algorithm of GRSA is described as follows, using metagenomic data as an example.

(1) Calculating the *p-values*

A statistical method (the full list of supported statistical methods is in Supplementary Table 1) was used to obtain the *p-values* of the features (i.e., p_{KO_i}, KO_i represents a certain KO; Fig. 1a). We used KO to represent different features in the formulas for simplicity.

(2) Converting the *p-values* into Z-scores

For the classic mixed RSA, we used an inverse normal cumulative distribution function (θ^{-1}) to convert the *p-values* of KOs into Z-scores. Thus, assuming uniformly distributed *p-values* under the random data assumption ranges, the resulting Z-scores will follow a standard normal distribution (Fig. 1a), the formula is:

$$Z_{KO_i} = \theta^{-1}\left(1 - p_{KO_i}\right) \qquad (1)$$

For the new directed RSA, we first divided the *p-values* by 2, transforming the range of *p-values* from $(0,1]$ to $(0,0.5]$:

$$p_{KO_i} = p_{KO_i}/2 \qquad (2)$$

Secondly, we used an inverse normal cumulative distribution function to convert the *p-values* of KOs into Z-scores (equation (1)). When the *p-value* is 0.5, the converted Z-score equals 0. Since the above *p-values* are no greater than 0.5, all converted Z-scores will be greater than 0 (Fig. 1a).

We then determined if a KO is up-regulated or down-regulated and calculated the ΔKO_i. In a differential abundance analysis (two-group design):

$$\Delta KO_i = \overline{KO_{i_{g1}}} - \overline{KO_{i_{g2}}} \qquad (3)$$

$\overline{KO_{i_{g1}}}$ is the average abundance of KO_i in group1, and $\overline{KO_{i_{g2}}}$ is the average abundance of KO_i in group2.

In a correlation analysis (two-group, multi-group, and longitudinal design):

$$\Delta KO_i = \rho_{KO_i} \qquad (4)$$

ρ_{KO_i} is the correlation coefficient between KO_i and the numeric variable.

Finally, assign a plus or minus sign to each Z-score:

$$Z_{KO_i} = \begin{cases} -Z_{KO_i}, & (\Delta KO_i < 0) \\ +Z_{KO_i}, & (\Delta KO_i \geq 0) \end{cases} \qquad (5)$$

Therefore, a KO_i with a Z_{KO_i} greater than 0 is up-regulated, a KO_i with a Z_{KO_i} less than 0 is down-regulated.

(3) Scoring the pathway

We next used the Z-scores of KOs to score the pathway. First, choose a pathway database as the reference. It is of particular interest to note any hierarchy relational table (e.g., KEGG, taxonomy database) can be used as a reference as long as the relationship between the upstream and downstream features (e.g., pathways and KOs) can be represented by a bipartite network (Fig. 1a). For each pathway in the selected database, calculate the Z-score of pathway j (Z_{path_j}) as follows:

$$Z_{path_j} = \frac{1}{\sqrt{k_j}} \sum_{i=1}^{k_j} Z_{KO_i} \qquad (6)$$

where Z_{KO_i} is the Z-score of KO_i within the $path_j$, and k_j denotes the total number of KOs in the $path_j$;

Next, we corrected Z_{path_j} using the randomly sampled $Z_{path_null_j}$ from the background distribution of the Z-scores of all KOs ($Z_{KO_{all}} = \{Z_{KO_1}, Z_{KO_2}, \ldots, Z_{KO_j}\}$) to evaluate the significance of enrichment. Specifically, for a given pathway $path_j$ including k_j KOs, we randomly sampled the same number (j) of KOs from the background $Z_{KO_{all}}$ without replacement, and calculated the $Z_{path_null_j}$, for N times (N = 10,000 in this study[10]). We then standardized Z_{path_j} by subtracting the mean (μ_{k_j}) and dividing by the standard deviation (σ_{k_j}) of the $Z_{path_null_j}$ distribution. The standardized Z_{path_j} is the $ReporterScore_j$. The *p-value* of $ReporterScore_j$ is also estimated by the above permutation. The formulas for the reporter scores and associated *p-values* are:

$$ReporterScore_j = \frac{Z_{path_j} - \mu_{k_j}}{\sigma_{k_j}} \qquad (7)$$

$$p_value_j = \frac{\sum_{n=1}^{N} I\left(Z_{path_j}, Z_{path_null_{jn}}\right) + 1}{N + 1} \qquad (8)$$

where $Z_{path_null_j}$ have the same k to Z_{path_j}, μ_{k_j} is the mean of the randomly generated N $Z_{path_null_j}$, σ_{k_j} is the standard deviation of the randomly generated N $Z_{path_null_j}$.

For the classic mixed RSA:

$$I(a,b) = \begin{cases} 1 & (a \geq b) \\ 0 & (a < b) \end{cases} \qquad (9)$$

For the new directed RSA:

$$I(a,b) = \begin{cases} 1 & \left(ReporterScore_j \geq 0 \ \& \ a \geq b\right) \\ 1 & \left(ReporterScore_j < 0 \ \& \ a \leq b\right) \\ 0 & (Otherwise) \end{cases} \qquad (10)$$

Demo codes

An example code tailored for one step GRSA of a KO abundance table is as follows.

```
library(ReporterScore)
# Load the KO abundance table
KO_abundance=read.table("ko_abundance.tsv", header = TRUE, sep = "\t")
# Get the sample metadata
metadata=read.table("sample_metadata.tsv", header = TRUE, sep = "\t")
# Run RSA analysis
reporter_score_res = reporter_score(KO_abundance, "Group", metadata,
                                    mode="directed", type = "pathway")
# Visualization
plot_report(reporter_score_res, rs_threshold = c(-3,3))
plot_report_circle_packing(reporter_score_res, rs_threshold = c(-3,3))
plot_KOs_in_pathway(reporter_score_res, map_id = "map00780")
plot_KOs_heatmap(reporter_score_res, map_id = "map00780")
plot_KOs_network(reporter_score_res, map_id = c("map05230","map04922"))
```

An example code for GRSA using other statistical tests is as follows.

```
#1. Use a specific statistical test method to get the p-values
ko_pvalue=your_method(KO_abundance)
#2. Convert the p-values of KOs to the Z-scores (select mode: mixed, di
rected)
ko_stat=pvalue2zs(ko_pvalue, mode=choice_of_mode)
#3. Calculate the reporter score of each pathway.
reporter_s=get_reporter_score(ko_stat)
```

Benchmark datasets

Benchmark datasets included one metagenomic KO profile (ex_KO_profile downloaded from https://github.com/wangpeng407/ReporterScore), 24 gene expression profiles of multiple human tissue types with disease, and 9 gene expression profiles of wild-type/knockout mice from the GEO database (https://www.ncbi.nlm.nih.gov/geo/).

We used these benchmark datasets with two-group or multi-group experimental designs to investigate the performance of GRSA, including similarities and differences between the two modes, statistical methods, and comparing GRSA with other commonly used enrichment analysis methods. Details of the datasets can be found in Supplementary Table 2 and Supplementary Table 3.

Case study datasets

Three case studies were re-analyzed using ReporterScore to demonstrate the versatile applications of GRSA, including microbiome, transcriptome, and metabolome.

Skin microbiome data were generated by Wang et al. (2021)[23]. Using the shotgun method, they sequenced 822 skin samples and constructed the complete Human Skin Microbiome Gene Catalog (iHSMGC). A full KO profile based on the KEGG database was provided on the website (https://ftp.cngb.org/pub/SciRAID/Microbiome/humanSkin_10.9M/AbundanceProfile/IHSMGC.KO.normalization.ProfileTable.gz). Metadata with details about including body

site, sex, age and cutotype were obtained via https://static-content.springer.com/esm/art%3A10.1186%2Fs40168-020-00995-7/MediaObjects/40168_2020_995_MOESM2_ESM.xlsx.

Transcriptomic data were extracted from the study by Liu et al. (2017)[30]. They investigated time-course transcriptomic profiling of cardiomyocyte differentiation derived from human hESCs and hiPSCs. The gene expression matrix is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85331.

Metabolomic data were generated by Liang et al. (2020)[31]. They analyzed the untargeted mass-spectrum data of 784 samples from 30 pregnant women and built a metabolic clock with five metabolites that time gestational age. The 264 identified level-1 and level-2 metabolites with HMDB IDs and their log2(intensity) can be found at https://ars.els-cdn.com/content/image/1-s2.0-S009286742030564X-mmc2.xlsx.

Statistical analyses

All statistical analyses were done on the R 4.2.2 platform. The developed ReporterScore package (https://github.com/Asa12138/ReporterScore) was used for GRSA and visualization. The venn diagram and venn network diagram were drawn by the pcutils package (https://github.com/Asa12138/pcutils/).

To compare the performance of different statistical methods in the two-group experimental design, we defined a Jaccard similarity index:

$$Similarity = \frac{|method(i) \cap method(j)|}{|method(i) \cup method(j)|} \qquad (11)$$

where method(i) (method(j)) is the number of significant pathways based on benchmark data sets enriched by different methods.

732 We used fuzzy c-means (FCM) clustering to explore the performance of different
733 statistical test methods in multi-group experimental design (Extended Data Fig. b,c) and
734 gene expression patterns in transcriptome (Fig. 4a). FCM is an unsupervised machine-
735 learning technique that partitions a population into groups or clusters[42]. Three methods
736 (Elbow, Silhouette, and Gap statistic) were used to determine the optimal number of
737 clusters. In FCM, the membership score is the probability that a feature belongs to a
738 cluster. Each feature is assigned to a cluster based on its highest membership score.
739
740 We compared GRSA against other most commonly used enrichment analysis methods.
741 Fisher's exact test is one of the most common functional enrichment analyses, which
742 relies on an arbitrary cutoff of fold change and significance. GSEA is a classic functional
743 class scoring method and analyzes all features based on their differential expression rank
744 without prior feature filtering. GSEA calculates an Enrichment Score (ES) by moving
745 through the ranked features list, increasing the ES if a feature is in the pathway, and
746 decreasing the ES if not. These running sum values are weighted to amplify enrichment in
747 the top- and bottom-ranking features. The ES is normalized to pathway size, yielding a
748 Normalized Enrichment Score (NES). Positive and negative NES indicate enrichment at the
749 top and bottom of the feature list, respectively. Lastly, a permutation-based *p-value* is
750 computed, and multi-test correction is applied, yielding a False Discovery Rate (FDR) or Q
751 value from 0 (significant) to 1 (not significant)[8]. However, the GSEA cannot be directly
752 applied to multi-group or longitudinal datasets.
753
754 In the comparison of GRSA with other commonly used enrichment methods, we first
755 calculated the *p-values* of features by t-test and performed adjustment of the *p-values*
756 using the Benjamini & Hochberg (BH) method to control for False Discovery Rate (FDR).
757 In order to perform the overrepresentation analysis, we selected the genes that have *p-*
758 *values* less than 0.05. And the top 400 genes with the highest unsigned log-fold changes
759 were considered as differentially expressed genes (DEGs)[43]. `fisher.test` (Fisher) was
760 performed by the R base package. Improved `fisher.test` (CP) was performed by

761 enricher function in the clusterProfiler package with the DEGs list. Gene set
762 enrichment analysis (GSEA) was performed by the GSEA function in the clusterProfiler
763 package, and the t-test statistic was used as the metric for ranking[44] of GSEA. GRSA was
764 performed with the BH-adjusted *p-values* of features by t-test. For convenience, the
765 ReporterScore package provides an interface to the above-mentioned enrichment
766 methods: KO_fisher for fisher.test, KO_enrich modified from clusterProfiler
767 based on fisher.test, and KO_gsea modified from GSEA in clusterProfiler. These
768 enrichment methods also support custom databases and are compatible with the format
769 of the input data for the reporter_score function in GRSA, making it easy to make cross-
770 comparisons.
771
772 Each human gene expression dataset in Supplementary Table 2 was associated with a
773 disease with defined mechanism in a KEGG pathway (termed the target pathway). Ideally,
774 a good enrichment analysis method would rank the target pathway at or near the top of
775 the list with a small adjusted *p-value*[16]. We compared the performance of four enrichment
776 analysis methods on these datasets.
777
778 For wild-type/knockout mice gene expression datasets (Supplementary Table 3), we
779 considered the pathways containing the knockout gene as true positives and the
780 pathways without the knockout gene as true negatives[16]. An adjusted *p-value* threshold
781 of 0.05 was used to determine if a pathway is significantly affected. Based on the
782 definitions of true positives (TP), true negatives (TN), false positives (FP), and false
783 negatives (FN), we can calculate the sensitivity and specificity as follows:

784 $$Sensitivity = \frac{TP}{TP + FN} \qquad (12)$$

785 $$Specificity = \frac{TN}{TN + FP} \qquad (13)$$

786

Data availability

The source data for all figures are available from GitHub (https://github.com/Asa12138/GRSA_figures).

Code availability

Code is available as an open-source R package 'ReporterScore', which can be downloaded from GitHub (https://github.com/Asa12138/ReporterScore). The main analysis scripts (Rmarkdown format) and source data are available from GitHub (https://github.com/Asa12138/GRSA_figures).

References

42. Bezdek, J. C., Ehrlich, R. & Full, W. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences* **10**, 191–203 (1984).

43. Shi, L. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* **24**, 1151–1161 (2006).

44. Zyla, J., Marczyk, M., Weiner, J. & Polanska, J. Ranking metrics in gene set enrichment analysis: Do they matter? *BMC Bioinformatics* **18**, 1–12 (2017).

Acknowledgments

We are grateful to our colleagues at the core facility of the Life Sciences Institute, especially the NECHO high-performance computing cluster. This research was partly supported by the grant from NSFC (82173645).

Author contributions

C.P. and C.J. conceived the study. C.P. developed the R package. C.J. and C.P. collected all datasets. C.P. completed the main benchmarking and case study analyses. Q.C., S.T., and

X.S. contributed to the analyses. C.P. and C.J. drafted and revised the manuscript with input from other authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Fig. 1-4

Supplementary Table 1-3