

1 **C3: An R package for cross-species compendium-based cell-type**
2 **identification**

3

4 Md Humayun Kabir^{1,2,3}, Djordje Djordjevic^{2,4}, Michael D. O'Connor^{1,5}, Joshua W. K. Ho^{2,4,*}

5

6 *¹School of Medicine, Western Sydney University, Campbelltown, NSW, Australia*

7 *²Victor Chang Cardiac Research Institute, Darlinghurst, NSW, Australia*

8 *³Department of Computer Science and Engineering, University of Rajshahi, Bangladesh*

9 *⁴St. Vincent's Clinical School, University of New South Wales, Sydney, NSW, Australia*

10 *⁵Medical Sciences Research Group, Western Sydney University, Campbelltown, NSW,*
11 *Australia*

12

13 *Corresponding author (Ho JWK):

14 E-mail: j.ho@victorchang.edu.au

15 Address: Victor Chang Cardiac Research Institute, 405 Liverpool Street, Darlinghurst, NSW
16 2010, Australia

17 Phone: (61 2) 9295 8645

18 Fax: (61 2) 9295 8601

19

20

21 **Abstract**

22 Cell type identification from an unknown sample can often be done by comparing its gene
23 expression profile against a gene expression database containing profiles of a large number of
24 cell-types. This type of compendium-based cell-type identification strategy is particularly
25 successful for human and mouse samples because a large volume of data exists for these
26 organisms. However, such rich data repositories often do not exist for most non-model
27 organisms. This makes transcriptome-based sample classification in these species
28 challenging. We propose to overcome this challenge by performing a *cross-species*
29 compendium comparison. The key is to utilise a recently published cross-species gene set
30 analysis (XGSA) framework to correct for biases that may arise due to potentially complex
31 homologous gene mapping between two species. The framework is implemented as an open
32 source R package called C3. We have evaluated the performance of C3 using a variety of
33 public data in NCBI Gene Expression Omnibus. We also compared the functionality and
34 performance of C3 against some similar gene expression profile matching tools. Our
35 evaluation shows that C3 is a simple and effective method for cell type identification. C3 is
36 available at <https://github.com/VCCRI/C3>.

37

38 **KEYWORDS:** bioinformatics; transcriptomics; cell type identification; cross-species; gene
39 set analysis

40

41

42 **Introduction**

43 The key question we seek to address in this article is *how can we identify the cell-type of a*
44 *biological sample given its gene expression profile?* This question commonly arises when
45 investigating a novel cell population resulting from differentiation of pluripotent stem cells or
46 isolation of a cell population in a non-model organism. The most popular bioinformatics
47 approach is a compendium-based identification approach, in which the unknown sample's
48 gene expression profile is used as a query profile against a large gene expression
49 compendium consisting of many cell types. A number of tools have been developed to
50 perform such a task, such as GEMINI [1], ProfileChaser [2], ExpressionBlast [3] and
51 CellMortage [4]. All these tools work in a similar fashion: match the query gene expression
52 profile or a gene set against a database of gene expression profiles to identify its best
53 matches. Importantly, most of these tools implicitly assume there is a one-to-one
54 correspondence between genes in the query sample and the compendium sample, which can
55 be violated when comparing data from different species. Beyond supporting filtering for
56 genes with one-to-one homology mapping across species, none of the current tools
57 effectively handle a cross-species query in a statistically rigorous fashion.

58
59 Therefore, when using currently available tools it is important to always use a database of the
60 same species as the query sample. This is often practically impossible because most publicly
61 available data sets are only available for a small number of species. Let's take as an example
62 one of the largest public gene expression repositories, the NCBI Gene Expression Omnibus
63 (GEO) [5]. As of March 2017, there were more than 57,000 GEO series (GSE) generated by
64 microarrays or RNA-Seq. Collectively, these data are a valuable resource for researchers to
65 discover new biological insights. Nonetheless, most of these GSE data sets were generated
66 from just two species: *Homo sapiens* (human) and *Mus musculus* (mouse). In fact, around
67 two thirds of these GSE data sets are derived from human or mouse samples (Figure 1). The
68 other third come from more than 1,300 species, with only 33 species having over 100 GSE
69 (Figure 1). In other words, while it is possible to curate a useful gene expression compendium
70 for human and mouse, it is practically impossible for other species, especially non-model
71 organisms.

72
73 We propose to alleviate this lack of species-specific compendia by performing a *cross-*
74 *species* cell identification, where a query profile is matched against a database of samples

75 which come from different organisms. A key challenge to implementing such a cross-species
76 analysis scheme is that many pairs of species, especially those that are evolutionary distant,
77 can have complex “many-to-many” homologous gene relationships. Failure to properly
78 account for the homology gene mapping can lead to statistical biases [6].

79

80 In this article, we present a new open source R package – C3 – that implements this cross-
81 species compendium-based cell type identification approach using a recently developed
82 cross-species gene set analysis method called XGSA [6]. XGSA has been shown to reduce
83 the false positive bias while still maintain good statistical power for gene sets affected by
84 highly complex homology structures. Using C3, we can harness the large collection of human
85 and mouse public data as a resource to identify unknown cell types for a wide variety of
86 species. We demonstrate the effectiveness of C3 using a large collection of GEO data. We
87 also compare its performance with other similar tools.

88

89 **Methods**

90 C3: a new R package for cross-species cell-type identification

91 C3 is an open source R package for identifying an unknown cell-type from its gene
92 expression profile based on a large compendium of gene expression data that can be derived
93 from different species. A key aspect of this approach is that it is most useful when the
94 compendium represents many different tissue or cell types, preferably from a well-studied
95 organism such as human or mouse. Examples of public data sources that can be used to form
96 this kind of compendium include ENCODE [7, 8] and GTEx [9]. The full description of the
97 method implemented in C3 is described in detail in the rest of this section, but an overview of
98 the framework can be found in Figure 2. Briefly, C3 first identifies genes considered to be
99 specifically-expressed genes in the query and the compendium profiles, by removing genes
100 ubiquitously expressed across these expression profiles. Next, C3 performs XGSA between
101 the query gene set and each of the compendium gene sets to account for “many-to-many”
102 gene relationships, and thereby determine which compendium gene sets are statistically
103 enriched in the query gene set. A p -value is reported for each compendium sample. The cell-
104 types of the most highly ranked compendium gene sets (according to p -value) are then used
105 to predict the cell-type of the query profile. C3 is available at <https://github.com/VCCRI/C3>.

106

107 The human and mouse gene expression compendia

108 For both mouse and human, we constructed a large compendium of tissue-specific genes
109 using RNA data from the ENCODE project. ENCODE gene expression data, summarised as
110 FPKMs, were obtained for human (hg19; 144 tissues or cell lines) [7] and for mouse (mm9;
111 94 tissues or cell types) [8]. Most tissues or cell types in the ENCODE data set are
112 represented by more than one replicate. We combined replicates of the same tissue or cell
113 type by calculating the mean expression value for each gene. If a compendium is constructed
114 from multiple data sources, we only consider genes that are common among all data sets.

115

116 Identification of specifically expressed genes in the query and compendium data

117 Using the compendium data, for each sample in the compendium we identified sets of highly-
118 expressed genes that are specific to each sample using two parameters: n – the number of
119 highly expressed genes to consider for marker gene status; t – the proportion of samples a
120 marker gene can appear in before it is discarded as non-unique/non-specific. Using these two
121 parameters we could identify then remove genes that are consistently highly expressed
122 (within the top n highly expressed genes in each sample) in more than $t \times 100\%$ of samples.
123 The goal of this step is to remove ubiquitously expressed genes such as housekeeping genes.
124 The remaining gene sets should be enriched for cell-type specific genes. To identify the
125 highly-expressed specific genes within the query data set, first we identified the top n highly
126 expressed genes. We then removed the ubiquitously expressed genes identified by the
127 compendium from the top n expressed genes. When the query sample species is different
128 from the species used to create the compendium, we use XGSA to identify the homologs of
129 the set of ubiquitously expressed genes for the query cell species. We then remove this set of
130 gene homologs from the query cell top expressed genes.

131

132 XGSA

133 To provide the required input for XGSA, all genes names are first converted to ENSEMBL
134 gene IDs. XGSA then applies a simple statistical method that computes a conservative p -
135 value based on Fisher's Exact test. This approach takes into account the homology gene
136 mapping structure between two cross-species gene sets [6]. If the two compared gene sets are
137 from the same gene sets, the resulting p -value is identical to that of a standard gene set test
138 based on a Fisher's Exact test. The package then performs Benjamini-Hochberg multiple
139 testing corrections on the raw p -values, and reports and visualises the $-\log_{10}$ of the corrected
140 p -values.

141

142 Comparison with ExpressionBlast

143 For the comparison with ExpressionBlast, we used brain, kidney and liver sample data sets
144 from the *R. norvegicus* species [10]. We identified the specific highly expressed genes for
145 each of the sample tissue types using our C3 package by setting parameter values as $n = 1000$
146 and $t = 0.10$. Among these specific highly expressed genes, we have selected the top 100
147 expressed genes based on their expression values. We used this set of highly expressed tissue
148 specific genes with \log_2 expression values as the input to the ExpressionBlast web tool. In
149 this way we have tested each of the three tissue types against both the human and mouse
150 organism using ExpressionBlast.

151

152

153 **Results**

154 Evaluation of C3

155 To evaluate the performance of C3, we collected gene expression profiles from four GEO
156 data series (GSE43013 [10], GSE74754 [11], GSE78770 [12], and GSE53393 [13]), which
157 collectively contain data from 13 different species (*B. taurus*, *C. familiaris*, *C. porcellus*, *E.*
158 *caballus*, *E. europaeus*, *F. catus*, *M. musculus*, *O. cuniculus*, *R. norvegicus*, *S. scrofa*, *D.*
159 *rerio*, *T. truncates*, and *M. mulatta*) across five different tissue types (brain, kidney, liver,
160 blood, and skeletal muscle). We tested whether C3 could correctly identify the cell type of
161 the samples when compared against a human compendium or a mouse compendium
162 constructed from ENCODE data [7, 8]. For comprehensiveness, we tested two combinations
163 of parameters in C3 (n and t). The summary result is shown in Figure 3 and the detailed
164 results are shown in the Supplementary materials [see Supplementary Tables 1-2]. Overall,
165 baring a few exceptions which will be discussed below, C3 was able to consistently identify
166 the correct or the most closely related cell type across all species (Figure 3).

167

168 GSE43013 [10] contains a gene expression data set from three different tissue types (brain,
169 kidney and liver) in 33 mammalian species, among which 10 have homology mapping
170 information available via ENSEMBL. C3 could correctly identify the cell types in all the
171 brain and liver samples across all 10 species. For the kidney data, C3 correctly identified the
172 cell type when compared against the mouse compendium across 10 species, but was much
173 less effective when compared against the human compendium. Interestingly, this comparison
174 against the human compendium resulted in most of the kidney gene sets being identified as

175 liver samples ahead of the human kidney samples. As both of these tissues are highly
176 vascularised, it may be that gene expression profiles from blood and blood vessel cells within
177 the kidney samples confound the analysis against the human compendium.

178

179 We also tested three more GSE datasets that contained data from 3 additional species; *D.*
180 *rerio* (GSE74754; brain) [11], *T. truncates* (GSE78770; blood) [12], and *M. mulatta*
181 (GSE53393; skeletal muscle) [13]. Through these analysis C3 correctly identified the cell
182 types of *D. rerio* brain and *T. truncates* blood. The *M. mulatta* skeletal muscle samples were
183 correctly identified by C3 when they compared to the mouse compendium but were not as
184 effectively identified using the human compendium (top hit was heart/tongue sample) (Figure
185 3). As with the kidney, skeletal muscle is also highly vascularised – and this could be the
186 cause of the misidentification of the *M. mulatta* skeletal muscle sample using the human
187 compendium.

188

189 Overall, a total of 160 C3 analyses were performed (80 against the mouse compendium and
190 80 against the human compendium) using two combinations of n/t parameters (i.e., 500/0.05
191 and 1000/0.1). Notably, all the cell type identity predictions made by C3 using the mouse
192 compendium were correct for at least one of the parameter combinations (i.e., typically at
193 least 1000/0.1 if not also 500/0.05). For comparison against the human compendium: correct
194 predictions were made for 67.5% of the queries, and for a further 25% of the queries the
195 correct prediction was ranked second or third by C3 (i.e., the correct prediction was in the top
196 3 positions 92.5% of the time using the human compendium). Only 1 out of the 80
197 predictions made by C3 using the human compendium (0.625%; *F. cattus*, *kidney*) did not
198 include the correct identification in the top 5 predictions. Notably, only two cell types were
199 not predicted correctly (i.e., as the top prediction): kidney and skeletal muscle. These tissues
200 are both highly vascularised, and this may be a confounding factor when comparing against
201 human samples. However, as shown in Figure 3, all the kidney and skeletal muscle datasets
202 were correctly identified when compared against the mouse compendium.

203

204 Comparison with other similar software programs

205 A comparison of the features of C3 and other similar methods is illustrated in Table 1. The
206 four similar methods discussed are primarily web-based with only GEMINI offering a Python
207 command-line version. GEMINI lacks the ability to perform cross-species cell type
208 identification. It uses level 3 gene expression datasets from The Cancer Genome Atlas

209 (TCGA) project [14]. CellMontage can compare only the expression data from similar
210 microarray platforms. As a result, neither of these methods could be included in our
211 comparative analysis. ProfileChaser supports cross-species analyses using NCBI
212 HomoloGene for only 6 species, and uses only the set of genes that have one-to-one human
213 homology mapping. However, ProfileChaser searches only the curated GEO DataSets (GDS)
214 (support only 1,815 GDS) for similar biological conditions based on differential gene
215 expression from reduced set of gene expression features. We were unable to meaningful
216 include this tool in our comparative analysis.

217

218 The only C3 alternative we are aware of that can compare a transcriptomic profile to a
219 compendium of data across species in order to identify an unknown cell type is
220 ExpressionBlast [3]. ExpressionBlast is a web-based tool that takes a maximum of one
221 hundred differentially expressed genes with their expression values, and compares it to
222 microarray data from 8 different species on GEO. For cross-species comparisons,
223 ExpressionBlast uses homologous gene groups from InParanoid and handles multiple
224 homologs using the closest expression value of the input gene. In contrast, C3 is an open
225 source R package that takes gene expression profiles as input. C3 leverages XGSA to
226 perform cross-species analysis between any of species in the growing list of species in
227 Ensembl Compara (currently 93 species).

228

229 To compare the performance of ExpressionBlast with C3, we analysed the brain, kidney and
230 liver sample data from *R. norvegicus* (GSE43013) [10] using both methods, as the rat is one
231 of the eight species supported by ExpressionBlast. For C3, we tested against the human and
232 mouse compendiums with parameter values $n=1000$ and $t=0.10$. For ExpressionBlast, we
233 inputted the 100 highly expressed tissue specific genes with their $\log_2(FPKM+1)$ expression
234 values. The summary results for C3 and ExpressionBlast are shown in Table 2, and the
235 detailed results are presented in Supplementary Table 2 (for C3) and Supplementary Figure 1
236 (for ExpressionBlast). From the comparative test results, it is clear that C3 can identify cell
237 type at least as accurately as ExpressionBlast. Nonetheless, C3 is has markedly greater
238 flexibility than ExpressionBlast in that it can handle the whole query gene expression profile,
239 it can be applied to data from a wide range of organisms, and its R package enables it to be
240 easily incorporated into any analytical pipeline.

241

242 **Discussion**

243 This work highlights the utility of cross-species analysis in cell-type identification using a
244 gene expression compendium-based approach. This is particularly important when
245 considering that the majority (two thirds) of transcriptomic data in the GEO database is from
246 human and mouse, with the remaining third of data shared between over 1,000 organisms
247 (Figure 1), most of which have very scant genomic resources. Our aim with C3 was to
248 leverage the many published data sets from the well characterised human and mouse
249 organisms to identify an unknown cell type from a potentially poorly characterised organism.

250

251 Recently we have used this approach to identify that a novel PAX7+ cell population in lizard
252 *Anole carolinensis* is highly similar to muscle satellite cells in human and mouse [15]. As
253 another real-life application, we have recently used the C3 approach to demonstrate that a
254 ROR1+ cell population derived from human pluripotent stem cells is similar to lens epithelial
255 cells in both human and mouse [16]. Both examples highlight the power of C3 in determining
256 or confirming the identity of a cell type using a compendium of gene expression profiles from
257 different species.

258

259 C3 can only correctly identify the cell type of an unknown transcriptomic profile if a similar
260 cell type is represented in the compendium. With this in mind, the quality, variety and size of
261 the compendium is paramount and future work should investigate larger compendiums such
262 as based on ARCHS4 [17], as well as domain specific compendiums such as for identifying
263 cancer subtypes.

264

265 **Conclusion**

266 Overall, we demonstrated that C3 can prioritise identification of the correct corresponding
267 cell type as the most significant hit. We believe C3 should facilitate rapid cell type
268 identification for less characterised species, or for poorly characterised cell types obtained
269 from stem cell differentiation strategies.

270

271

272

273 **Authors' contributions**

274 J.W.K.H. initiated the project; M.H.K. designed the method, implemented the package,
275 performed evaluation and wrote the manuscript; D.D. contributed to method design and
276 software testing; M.D.O'C and J.W.K.H. supervised the whole project and revised the
277 manuscript. All authors read and approved the final manuscript.

278

279 **Competing interests**

280 The authors declare no competing financial interests.

281

282 **Acknowledgements**

283 M.H.K. is supported by a UWS Postgraduate Research Award (International). J.W.K.H is
284 supported by a Career Development Fellowship by the National Health and Medical Research
285 Council (1105271) and a Future Leader Fellowship by the National Heart Foundation of
286 Australia (100848).

287

288

289 **References**

- 290 [1] DeFreitas T, Saddiki H, Flaherty P. GEMINI: a computationally-efficient search engine
291 for large gene expression datasets. *BMC Bioinformatics* 2016;17:102.
- 292 [2] Engreitz JM, Chen R, Morgan AA, Dudley JT, Mallewar R, Butte AJ. ProfileChaser:
293 searching microarray repositories based on genome-wide patterns of differential expression.
294 *Bioinformatics* 2011;27:3317-8.
- 295 [3] Zinman GE, Naiman S, Kanfi Y, Cohen H, Bar-Joseph Z. ExpressionBlast: mining large,
296 unstructured expression databases. *Nat Methods* 2013;10:925-6.
- 297 [4] Fujibuchi W, Kiseleva L, Taniguchi T, Harada H, Horton P. CellMontage: similar
298 expression profile search server. *Bioinformatics* 2007;23:3103-4.
- 299 [5] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI
300 GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* 2013;41:D991-5.
- 301 [6] Djordjevic D, Kusumi K, Ho JW. XGSA: A statistical method for cross-species gene set
302 analysis. *Bioinformatics* 2016;32:i620-i8.
- 303 [7] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the
304 human genome. *Nature* 2012;489:57-74.
- 305 [8] The Mouse ENCODE Consortium. An encyclopedia of mouse DNA elements (Mouse
306 ENCODE). *Genome Biol* 2012;13:418.
- 307 [9] The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*
308 2013;45:580-5.
- 309 [10] Fushan AA, Turanov AA, Lee SG, Kim EB, Lobanov AV, Yim SH, et al. Gene
310 expression defines natural changes in mammalian lifespan. *Aging Cell* 2015;14:352-65.
- 311 [11] Mayrhofer M, Gourain V, Reischl M, Affaticati P, Jenett A, Joly JS, et al. A novel brain
312 tumour model in zebrafish reveals the role of YAP activation in MAPK- and PI3K-induced
313 malignant growth. *Dis Model Mech* 2017;10:15-28.
- 314 [12] Morey JS, Neely MG, Lunardi D, Anderson PE, Schwacke LH, Campbell M, et al.
315 RNA-Seq analysis of seasonal and individual variation in blood transcriptomes of healthy
316 managed bottlenose dolphins. *BMC Genomics* 2016;17:720.
- 317 [13] Chapalamadugu KC, Vandevort CA, Settles ML, Robison BD, Murdoch GK. Maternal
318 bisphenol a exposure impacts the fetal heart transcriptome. *PLoS One* 2014;9:e89096.
- 319 [14] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast
320 tumours. *Nature* 2012;490:61-70.

321 [15] Palade J, Djordjevic D, Hutchins ED, George RM, Cornelius JA, Rawls A, et al.
322 Identification of satellite cells from anole lizard skeletal muscle and demonstration of
323 expanded musculoskeletal potential. *Dev Biol* 2018;433:344-56.

324 [16] Murphy P, Kabir MH, Srivastava T, Mason ME, Dewi CU, Lim S, et al. Light-focusing
325 human micro-lenses generated from pluripotent stem cells model lens development and drug-
326 induced cataract in vitro. *Development* 2018;145.

327 [17] Alexander Lachmann DT, Alexandra B. Keenan, Kathleen M. Jagodnik, Hyojin J. Lee,
328 Lily Wang, Moshe C. Silverstein, and Avi Ma'ayan (2017), 'Massive Mining of Publicly
329 Available RNA-seq Data from Human and Mouse', bioRxiv.

330

331

332 **Figure legends**

333

334 **Figure 1 Summary of GSE based on species in NCBI GEO**

335 The pie chart shows the total number of GSE for *H. sapiens* (blue), *M. musculus* (pink) and
336 all other species (orange). The bar plot shows the top 60 species according to the number of
337 GSE in NCBI GEO.

338

339 **Figure 2 Overall workflow diagram of C3**

340

341 **Figure 3 Evaluation of C3**

342 Gene expression profiles of tissues from 13 different organisms were selected from four GEO
343 data sets. These profiles were used to evaluate whether C3 could correctly identify its cell
344 type of the sample when compared against a human ENCODE compendium (Human) or a
345 mouse ENCODE compendium (Mouse). *n*: top number of highly expressed genes; *t*: cut-off
346 threshold value; 1 = Statistically significant and in top position; 2 = Statistically significant
347 but in top 2-3rd position; 3 = Statistically significant but in top 4-5th position; 4 = Not
348 statistically significant but in top position; 5 = Not statistically significant but in top 2-5th
349 position; 6 = Not statistically significant and not in 2-5th position

350

351

352 **Table legends**

353 **Table 1 Comparison of software features of C3 and other similar methods**

	C3	ExpressionBlast	ProfileChaser	GEMINI	CellMontage
Cross-species method	Ensembl BioMart portal, complete homology structure using XGSA	Inparanoid, handles multiple orthologues using closest value of input gene	One-to-one human homolog	Not supported	Not mentioned
How many species	As many as ENSEMBL mapping	8	6	-	-
Input	Gene expression matrix	Max 100 differentially expressed genes with expression values	Gene expression matrix	Gene expression matrix	Gene expression matrix with raw expression values
User interface	R command line	Web	Web	Web and Python command-line	Web
Availability	Open source	Free	Free	Free	Free
Application	General	General	Specific to GDS	Level 3 gene expression from TCGA project	Specific to similar microarray platforms
Dependency	Previously made compendium	Differentially expressed genes	Reduced set of gene expression features	Reduced dimension of expression profile	UniGene names for gene ids

354

355

356 **Table 2 Comparison of cross-species cell type identification using C3 and**
 357 **ExpressionBlast**

	Identified cell type by C3	Identified cell type by ExpressionBlast
<i>R. norvegicus</i> brain with Human compendium	brain	other than brain (no brain sample among top 5)
<i>R. norvegicus</i> brain with Mouse compendium	brain	brain
<i>R. norvegicus</i> kidney with Human compendium	liver at top position and then kidney	liver (no kidney sample among top 5)
<i>R. norvegicus</i> kidney with Mouse compendium	kidney	kidney
<i>R. norvegicus</i> liver with Human compendium	liver	liver
<i>R. norvegicus</i> liver with Mouse compendium	liver	liver

358

359

360 **Supplementary material**

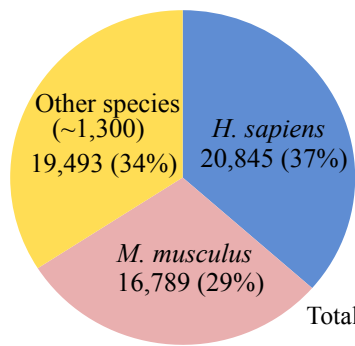
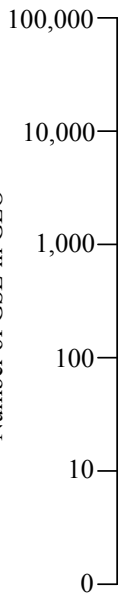
361 **Supplementary Table 1** Detail test result with different species' different cells/tissues with
362 $n=500, t=0.05$

363 **Supplementary Table 2** Detail test result with different species' different cells/tissues with
364 $n=1000, t=0.10$

365 **Supplementary Figure 1** Test result screenshot of *R.norvegicus* sample datasets using
366 ExpressionBlast: (a) and (b) show results for brain dataset with *H. sapiens* and *M. musculus*
367 respectively; (c) and (d) show results for kidney dataset with *H. sapiens* and *M. musculus*
368 respectively; (e) and (f) show results for liver dataset with *H. sapiens* and *M. musculus*
369 respectively.

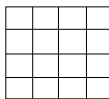
370

Number of GSE in GEO

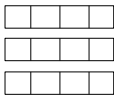


Total = 57,127 GSE

Compendium

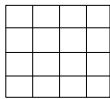


Gene expression data



Marker gene sets

Query data

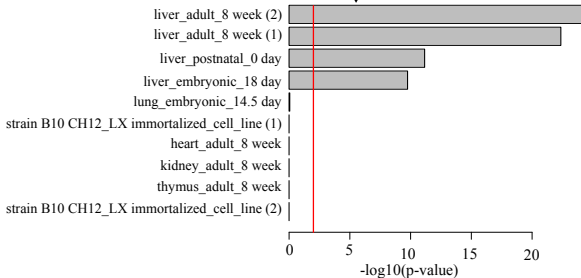


Gene expression matrix

Compendium of cell type specific gene expression profile (e.g., human and mouse ENCODE)

Cross-species gene set analysis (XGSA)

display result



		$n=500, t=0.05$		$n=1000, t=0.10$			
		Human	Mouse	Human	Mouse		
Data set 1 (GSE43013)		Sample name		Human	Mouse		
		<i>B. taurus</i> brain		1	1	1	1
		<i>C. familiaris</i> brain		1	1	1	1
		<i>C. porcellus</i> brain		1	1	1	1
		<i>E. caballus</i> brain		1	1	1	1
		<i>E. europaeus</i> brain		1	1	1	1
		<i>F. catus</i> brain		1	1	1	1
		<i>M. musculus</i> brain		1	1	1	1
		<i>O. cuniculus</i> brain		1	1	1	1
		<i>R. norvegicus</i> brain		1	1	1	1
		<i>S. scrofa</i> brain		1	1	1	1
		<i>B. taurus</i> kidney		2	1	2	1
		<i>C. familiaris</i> kidney		3	1	2	1
		<i>C. porcellus</i> kidney		3	1	2	1
		<i>E. caballus</i> kidney		2	1	2	1
		<i>E. europaeus</i> kidney		3	1	2	1
		<i>F. catus</i> kidney		6	1	2	1
		<i>M. musculus</i> kidney		2	1	2	1
		<i>O. cuniculus</i> kidney		2	1	2	1
		<i>R. norvegicus</i> kidney		2	1	2	1
		<i>S. scrofa</i> kidney		5	1	3	1
		<i>B. taurus</i> liver		1	1	1	1
		<i>C. familiaris</i> liver		1	1	1	1
		<i>C. porcellus</i> liver		1	1	1	1
		<i>E. caballus</i> liver		1	1	1	1
		<i>E. europaeus</i> liver		1	1	1	1
		<i>F. catus</i> liver		1	1	1	1
		<i>M. musculus</i> liver		1	1	1	1
		<i>O. cuniculus</i> liver		1	1	1	1
<i>R. norvegicus</i> liver		1	1	1	1		
<i>S. scrofa</i> liver		1	1	1	1		
Data set 2 (GSE74754)	<i>D. rerio</i> brain (control)	1	1	1	1		
	<i>D. rerio</i> brain (tumour)	1	1	1	1		
Data set 3 (GSE78770)	<i>T. truncatus</i> blood (hua)	1	1	1	1		
	<i>T. truncatus</i> blood (kai)	1	1	1	1		
	<i>T. truncatus</i> blood (keo)	1	1	1	1		
	<i>T. truncatus</i> blood (pele)	1	1	1	1		
Data set 4 (GSE53393)	<i>M. mulatta</i> skeletal muscle (early BPA)	1	1	2	1		
	<i>M. mulatta</i> skeletal muscle (early control)	1	1	2	1		
	<i>M. mulatta</i> skeletal muscle (late BPA)	2	1	2	1		
	<i>M. mulatta</i> skeletal muscle (late control)	2	1	2	1		