bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 **The Molecular Subtyping Resource (MouSR): a user-friendly tool for rapid biological**
2 **discovery from human or mouse transcriptional data**

3

4 **Short title:** MouSR: Molecular Subtyping Resource Application

5

6 Baharak Ahmaderaghi[1], Raheleh Amirkhah[1], James Jackson[2], Tamsin RM Lannagan[3], Kathryn
7 Gilroy[3], Sudhir B Malla[1], Keara L Redmond[1], ACRCelerate Consortium, Tim Maughan[4], Simon
8 Leedham[4], Andrew S Campbell[3], Owen J Sansom[3,5], Mark Lawler[1,$], Philip D Dunne[1,*,$]

9 [1]The Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, UK
10 [2] Information Services, Queen's University Belfast, UK
11 [3] Cancer Research UK Beatson Institute, Glasgow, UK
12 [4] University of Oxford, UK
13 [5] Institute of Cancer Sciences, University of Glasgow, UK.
14

15 ***Correspondence:** Philip Dunne, The Patrick G Johnston Centre for Cancer Research, Queen's
16 University Belfast, BT9 7AE, UK. Email: p.dunne@qub.ac.uk. Telephone +4428 90972792.
17 [$] These authors contributed equally as co-senior.

18

19 **Funding:** This research was supported by an International Accelerator Award, ACRCelerate,
20 jointly funded by Cancer Research UK (A26825 and A28223), FC AECC (GEACC18004TAB)
21 and AIRC (22795), a CRUK early detection project grant (A29834) and a Health Data Research
22 UK Substantive Site grant. This manuscript reflects the views of the authors and may not reflect
23 the opinions or views of the funders.

24

25 **Declaration of interests:** ML has received honoraria from Pfizer, EMF Serono and Roche for
26 presentations unrelated to this work; ML has received an unrestricted educational grant from Pfizer
27 for research unrelated to this work. OS has received funding from Novartis, Astra Zeneca, Cancer
28 Research Technology and Redex for research unrelated to this work.

29

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

30 **ABSTRACT**

31 Generation of transcriptional data has dramatically increased in the last decade, driving the
32 development of analytical algorithms that enable interrogation of the biology underpinning the
33 profiled samples. However, these resources require users to have expertise in data wrangling and
34 analytics, reducing opportunities for biological discovery by "wet-lab" users with a limited
35 programming skillset. Although commercial solutions exist, costs for software access can be
36 prohibitive for academic research groups.

37 To address these challenges, we have developed an open source and user-friendly data analysis
38 platform for on-the-fly bioinformatic interrogation of transcriptional data derived from human or
39 mouse tissue, called "MouSR". This internet-accessible analytical tool, https://mousr.qub.ac.uk/,
40 enables users to easily interrogate their data using an intuitive "point and click" interface, which
41 includes a suite of molecular characterisation options including QC, differential gene expression,
42 gene set enrichment and microenvironmental cell population analyses from RNA-Seq. Users are
43 provided with adjustable options for analysis parameters to generate results that can be saved as
44 publication-quality images. To highlight its ability to perform high quality data analysis, we utilise
45 the MouSR tool to interrogate our recently published tumour dataset, derived from genetically
46 engineered mouse models and matched organoids, where we rapidly reproduced the key
47 transcriptional findings.

48 The MouSR online tool provides a unique freely-available option for users to perform rapid
49 transcriptomic analyses and comprehensive interrogation of the signalling underpinning
50 transcriptional datasets, which alleviates a major bottleneck for biological discovery.

51

52

53 ***Index Terms***—*Bioinformatics, Data Analytics Platform, RNA-seq data, Statistical graphics,*
54 *Visualization, Shiny-App, cancer, transcriptomics*

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

55 **INTRODUCTION**

56 In the years since the first whole genome was sequenced, the costs associated with the generation
57 of molecular "big data" have decreased rapidly, to a point where the data handling, rather than data
58 generation, is the limiting factor in large biological discovery programmes. Furthermore, large
59 repositories (such as the TCGA (NCI and National Human Genome Research Institute, 2006) and
60 Gene Expression Omnibus (Edgar et al., 2002)), now provide free access to publicly-available
61 molecular data. Large international molecular subtyping projects have markedly improved our
62 biological understanding of cancer (Sohn et al., 2017), but in doing so they have created a critical
63 bottleneck in terms of data reduction, analysis and interpretation, resulting in an urgent need for
64 solutions that enable rapid biological interrogation of large datasets (Cerami et al., 2012).

65 Given the relative paucity of translational bioinformaticians within many research groups (Gao et
66 al., 2013), there is a need for wet-lab researchers to have access to user-friendly analytic platforms
67 that provide rapid and statistically controlled algorithms to perform common transcriptional
68 analysis tasks, alongside an array of tools for visualising and interrogating the resulting data. For
69 these tools to be widely adopted, they will need to provide the non-computational user with
70 intuitive "point and click" options for transcriptional analyses, rather than programming-based
71 options. To address this need, we have developed the molecular subtyping resource; MouSR;
72 https://mousr.qub.ac.uk/, which enables individual non-computational end-users to pursue lines of
73 investigation on transcriptional data within their domain of interest/area of expertise without the
74 need for expertise in scripting. The MouSR platform enables interrogation of existing publicly-
75 available or in-house transcriptional data and analytics from either human or mouse models, within
76 a standardised molecular stratification environment.

77

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

78 **Methods:**

79 **Access and Requirements:** The MouSR App was built using R (v3.2) and is running on the shiny
80 server (v1.5.16) hosted on the Queen's University Belfast virtual server CentOS 7, 64-bit, Intel
81 Xeon Gold 6130 CPU @2.10 GHZ, 16 Core. The service was given extra security and protection
82 by being placed behind a proxy service, which meant the server itself is never directly exposed to
83 the internet. This configuration may be of benefit, where possible, to other potential users who
84 wish to install their own MouSR version. The system is accessible via all web browsers tested, and
85 on both Linux and Windows systems, via https://mousr.qub.ac.uk/. However, Chrome, Edge and
86 Firefox are the recommended browsers for the best App experience.

87 **File formats:** *Importing Text Files (CSV or TXT)*: This App accepts two commonly used text file
88 formats. For ".csv" comma separated values (Comma delimited), data files are converted to text
89 files, with each line corresponding to a row and all the fields in each line separated by commas,
90 whereas in ".txt" (Tab delimited) all the fields in each line are typically separated by tabs. The
91 uploaded files size is limited to 30 MB.

92 *Transposing data tool for import*: The Transcriptional Data Matrix file to be uploaded in the App
93 has a defined format to follow, however for those users where their file is not in the same
94 orientation as the suggested format in the App, a link to the online transposing tool has been
95 embedded in the introduction section of the App and can be accessed via this link (Data Design
96 Group, 2013).

97 **Tools embedded within the MouSR application:** Principal Component Analysis (PCA): The
98 PCA analysis is performed by the prcomp function in the R stats package (Jolliffe and Cadima,
99 2016). PCA is defined by a transformation of a high dimensional vector space into a low

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

100 dimensional space. It uses linear combinations of the original data to define a new set of variables
101 that are referred to as principal components.

102 Multi-Dimensional Scaling Analysis (MDS): We used the cmdscale function in the R stats package
103 to perform MDS analysis (Young, 2013). Unlike the PCA method that minimises dimensions while
104 preserving covariance of the data, MDS minimises dimensions and preserves distance between
105 data points. However, both methods can provide similar results, if the covariance in data and
106 Euclidean distance measure between data points in high dimension is equal. MDS uses the
107 similarity matrix as input, which has an advantage over PCA as it can be applied directly to
108 pairwise-compared banding patterns. The '% Variance' describes how much of the total variance
109 is explained by each of the components with respect to the whole (the sum). '% Variance' values
110 are shown on the axis labels.

111 Differential gene expression analysis (DESeq2): The differential gene expression analysis is
112 performed based on the negative binomial distribution using the DESeq2 R package (version
113 1.24.0) (Love et al, 2014). DESeq2 is a count-based statistical method that performs an internal
114 normalization where estimated variance-mean is calculated for each gene across all samples.
115 DESeq2 also estimates the gene-wise dispersion and logarithmic fold changes, a dispersion value
116 is estimated for each gene through a model fit procedure, and differential expression is tested,
117 based on a model using the negative binomial generalized linear distribution (Love et al, 2014).
118 We used the DESeq2 package to normalize the data and identify genes, which are differentially
119 expressed between the two main groups selected by the user.

120 Single sample Gene Set Enrichment Analysis classification (ssGSEA): is performed using the
121 GSVA package version 1.32.0. The R package msigdbr version 7.1.1 is also used to retrieve

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

122 mouse/human Hallmark and biological processes (GO_BP) gene sets and applied to the samples
123 (Hänzelmann et al., 2013).

124 Gene Set Enrichment Analysis (GSEA): This method consists of three steps**:** Enrichment Score
125 (ES) is calculated, reflecting the degree to which a set of genes is overrepresented at the top or
126 bottom of the entire ranked list. Second, the statistical significance of the ES is estimated by using
127 an empirical phenotype-based permutation test procedure that preserves the complex correlation
128 structure of the gene expression data. Finally, after an entire database of gene sets is evaluated, the
129 estimated significance level is adjusted to account for multiple hypothesis testing by first
130 calculating Normalised Enrichment Score (NES), based on dividing the actual enrichment score
131 by the mean of enrichment scores against all permutations of the dataset, then calculating the False
132 Discovery Rate (FDR) corresponding to each NES. In this paper, GSEA is performed on log
133 expression ratio using fgsea, an R package which is a fast implementation of pre-ranked GSEA
134 (Sergushichev, 2019).

135 Microenvironment Cell Population counter (MCP): The MCPcounter and Murine MCP (mMCP)
136 counter R packages are used to estimate the quantity of several immune and stromal cell
137 populations from heterogeneous transcriptomic data for human and murine samples, respectively
138 (Petitprez et al., 2020) (Becht et al., 2016).

139 **Packages used:** Shiny (Chang et al., 2018), shinythemes (Chang et al., 2021), shinydashboard
140 (Chang et al., 2018), shinycustomloader (Tanaka and Niichan, 2018), shinycssloaders (Sali et al.,
141 2020), ggplot2 (Wickham et al., 2020), tibble (Müller et al., 2021), DESeq2 (Love et al., 2014),
142 limma (Smyth et.al, 2015), plyr (Wickham, 2020), biomaRt (Durinck et al., 2005), heatmaply
143 (Galili et al., 2021), reshape (Wickham, 2018), plotly (Sievert, 2019), WGCNA (Horvath et al.,

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

144 2021), lattice (Sarkar et al., 2020), pheatmap (Kolde, 2019), RColorBrewer (Neuwirth, 2014),
145 GSVA (Hänzelmann et al., 2013), rlist (Ren, 2016), msigdbr (Dolgalev, 2020), tidyverse
146 (Wickham, 2021), mMCPcounter (Petitprez et al., 2020), MCPcounter (Becht et al., 2016),
147 magrittr (Bache et al., 2020), dplyr (Wickham et al., 2021), ggrepel (Slowikowski et al., 2021),
148 readxl (Wickham et al., 2019), DT (Xie et al., 2021), colourpicker (Attali and Griswold, 2020),
149 fgsea (Sergushichev, 2019), enrichplot (Yu and Hu, 2021).

150 **File outputs and modifiable formats:** The customised options for plots that are common in all
151 App sections are: turning labels on/off, justifying height and width of the plot and having different
152 downloading format (png/svg). Additionally, some analytical panels have extra customised
153 options in their filtering criteria, in order to provide an easier-to-use environment for the users,
154 such as changing the colours, size of labels and points, assigning output filename, adding legend
155 and changing scale.

156

157 **Data availability:** The complete source code for the MouSR App can be launched on any system
158 that has R/RStudio installed and is available to download at https://github.com/Dunne-
159 Group/MouSR (Site will go live on publication acceptance). In addition, we are open to mutually-
160 beneficial collaborations to further develop the MouSR analytic options.

161 If you use MouSR in your work, please remember to cite this paper.

162

163

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

164 **RESULTS**

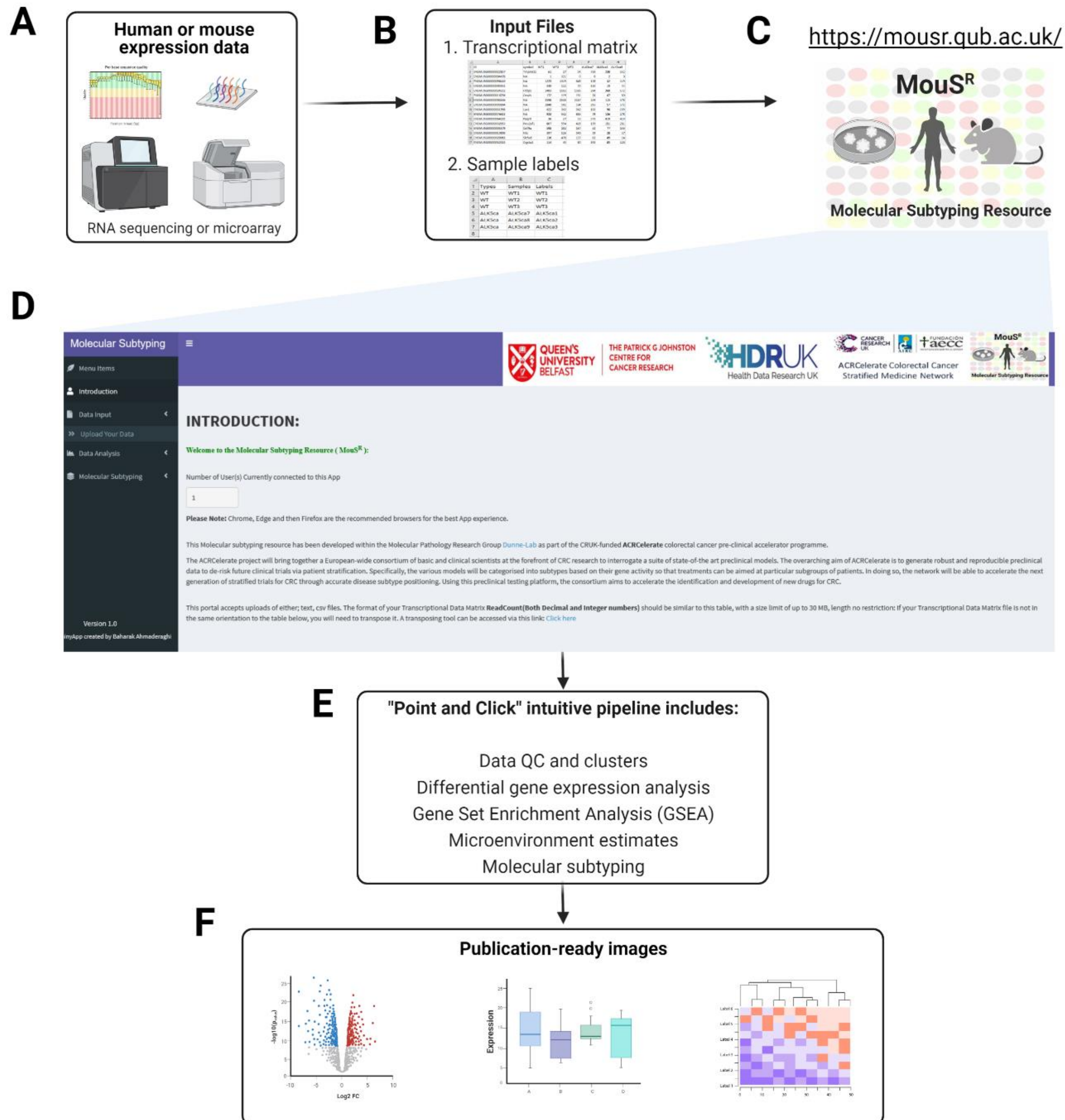
165 **MouSR Interface:**

166 The MouSR (https://mousr.qub.ac.uk/) platform is implemented as an open-source application that
167 enables non-computational users to rapidly go from an existing data matrix (multiple formats) to
168 biologically meaningful results in a user-friendly way. At each step of the process, users have the
169 option to modify outputs, through a series of on-the-fly customisable graphics that can all be
170 downloaded at high resolution for future use. The standard pipeline includes initial data QC
171 assessments, followed by differential analyses, both single sample and group-wise gene set
172 enrichment analyses and microenvironment population counters, producing publication-ready data
173 (Figure 1). The system uses a species-agnostic "blind embedding" format, where users can upload
174 data derived from any patient sample or *in vitro/in vivo* model as either ungrouped individual
175 samples or multiple samples within experimental groups. The species-specific selection options
176 available for downstream analyses in MouSR enable users to perform biological
177 discovery/validation on transcriptional data derived from human or mouse origin.

178 To highlight the functionality of the MouSR system, in this paper we will focus on colorectal
179 cancer (CRC), using our previously published and comprehensively characterised genetically
180 engineered mouse model tumour and organoid datasets, comprising of n=28 samples from matched
181 tumour tissue (n=4 groups) and genotype-matched organoids (n=3 groups) derived from n=4
182 different genotypes (Jackstadt et al., 2019) (Jackstadt, 2019).

183

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



184

185 **Figure 1: The MouSR workflow and outputs**: Utilising transcriptional data derived from either
186 human or mouse tissue/cells (A) users are required to have a transcriptional data matrix and sample
187 information as the input for the MouSR pipeline, accessible via https://mousr.qub.ac.uk/ (B and
188 C). From the introduction page (D), users upload their files and are then presented with a series of
189 point-and-click options for initial data QC, differential analysis, molecular signaling and
190 microenvironment characterisation (E) that can be saved as high-resolution image files for further
191 use (F). As an example of the adaptable nature of the system at each stage, users have options for
192 bespoke formatting, design and labelling of the resulting plots, which can all be downloaded and
193 saved in a publication-ready format. Figure created using BioRender.

194

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

195 Accessing the MouSR website presents the user with a general introduction landing page,
196 providing an overview of the application with instructions and exemplar formats that are required
197 for use. The user interface structure has two main sections, namely 1) Data Input and 2) Data
198 analysis, which will be described briefly, followed by an analysis of the CRC mouse exemplar
199 dataset from Jackstadt and colleagues (Jackstadt et al., 2019) (Jackstadt, 2019) (Ahmaderaghi,
200 2021) and for convenience these data are included here as supplementary files (Supplementary
201 files 1 and 2). In addition, a tutorial video is also includes to summarise the main features of the
202 MouSR tool (Supplementary file 3).

203

204 **Data Input**

205 The Data Input section is designed to have flexibility in terms of acceptable file/data formats, to
206 enable users to upload their own data derived using a variety of transcriptional profiling platforms
207 and normalisation procedures. Users are required to have two separate files; a transcriptional data
208 matrix (input 1) and a sample information file (input 2) that will enable data analysis and
209 generation of results (Figure 1).

210 In terms of data types for input 1, MouSR has successfully been tested using human and mouse
211 data derived from a variety of microarray and RNAseq platforms and is adaptable enough to accept
212 data that has been processed using a range of pipelines resulting in either integers (whole numbers,
213 i.e. RNAseq read counts) or decimals (i.e. estimated read counts). However, as DESeq2 requires
214 non-normalised counts as input, for users who select decimal input the differential expression
215 options will not be accessible. Additionally, the MouSR system has been designed to accept the
216 most common file formats; including csv, txt, and also to accept data files with various separators;

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

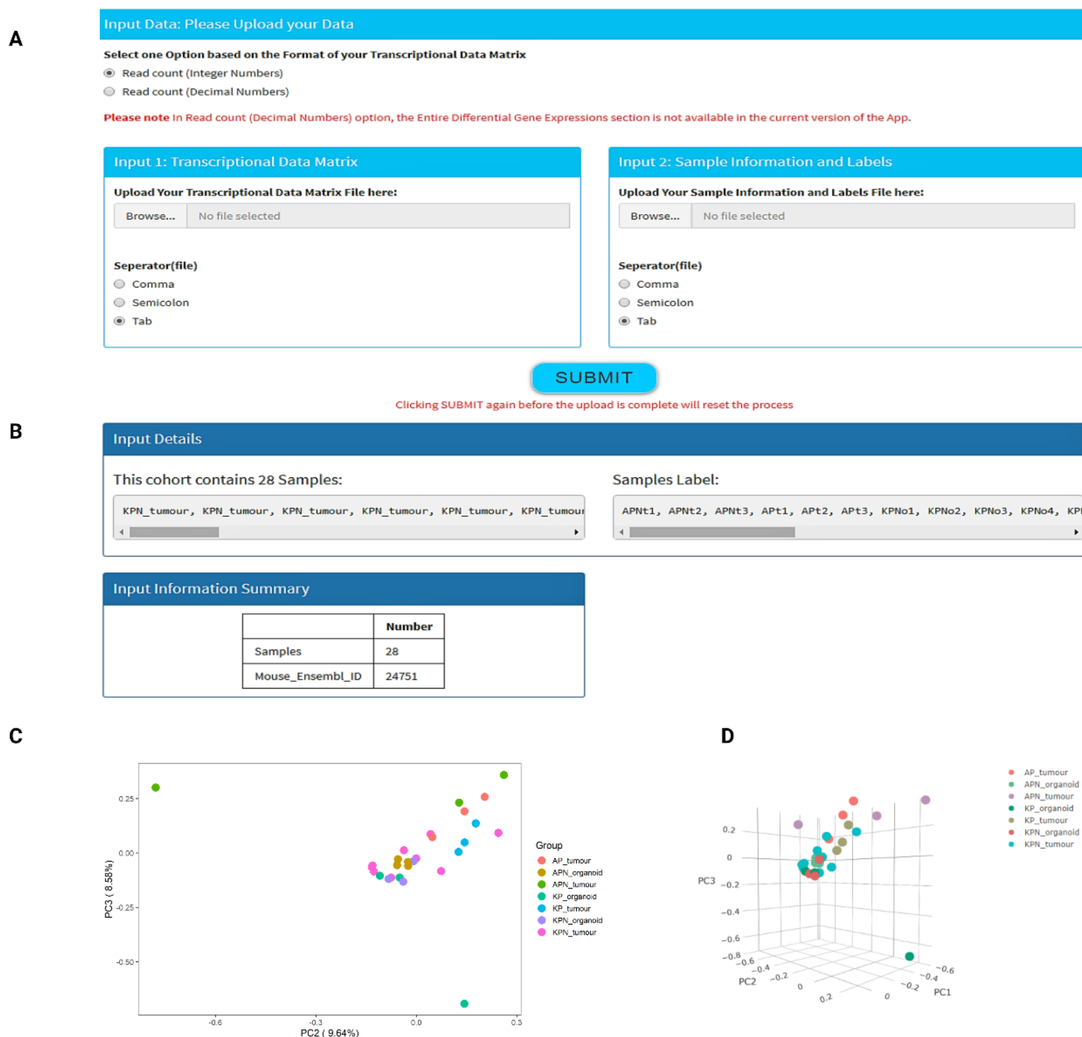
217 including comma, semicolon or tabs. The data required for input 2 is a summary of basic
218 information that relates to sample labels and groups.

219 Prior to uploading their data, users must ensure both files are in the recommended format described
220 on the introduction page, which is aligned with a standard data matrix output containing gene ID
221 and gene symbol columns followed by sample values and is in the format selected by the user
222 according to their data types (default set as tab delimited/.txt file). To ensure that users with data
223 in an orientation not supported by MouSR can use the tool, we have created a transpose link on
224 the introduction page, which will adjust the transcriptional matrix using the Transpose CSV Tool
225 (Data Design Group, 2013).

226 Once the files are in the correct format, the user is required to upload their two files into input 1
227 and input 2 (Figure 2A) using either a drag/drop from a folder, or by navigating to the file location
228 using the browse function. When files are selected, a progress bar will immediately begin to
229 indicate that the file is uploading until the upload is complete. By clicking on the submit button,
230 the user interface becomes active, triggering the computational analysis on the background server,
231 with input detail and input summary being displayed when complete, including information on
232 number of samples, sample names, and number of expression values identified.

233 Given the flexibility and the point-and-click design for downstream analyses, users must have their
234 files in the correct format to proceed. At this stage, users can verify if their data are correctly
235 loaded, or if modifications are required. Our exemplar files took <20 seconds from submission to
236 display of input details, confirming that it consists of 24751 individual genes across 28 samples
237 across 7 experimental groups (Figure 2B).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



238

239 **Figure 2: Data import and exploratory analysis.** A) Two input files are required to begin the
240 analytical pipeline in the App – a gene expression matrix and a meta data that includes sample
241 group labelling. B) Following data upload, the data summary on samples will be displayed for
242 quick review. C) Exploratory visualisation of data will be provided in a form of 2D PCA plot and
243 D) 3D PCA plot with labelled sample groups.

244

245

246

247

248

249

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

250 **Data Analysis**

251 The *Data Analysis* section consists of four main subsections: 1) Principal Component Analysis
252 (PCA) & Multi-Dimensional Scaling (MDS), 2) Differential Gene Expression Analysis (DGEA),
253 3) mouse and/or human Gene Set Enrichment Analysis (GSEA), and 4) mouse and/or human
254 options Microenvironment Cell Population counter (mMCP/MCP-counter).

255

256 ***PCA and MDS*****:** Principal Component Analysis (PCA) and Multi-Dimensional Scaling (MDS) are
257 dimensionality-reduction methods that represent an initial step in assessing characteristics of any
258 dataset (Jolliffe and Cadima, 2016) (Young, 2013). These options give the user an immediate
259 overview of clustering of samples according to their experimental labels (described further in the
260 Methods section).

261 *PCA 2D/3D Plots*: PCA plots enable the user to look at the principal components that describe the
262 largest variability between samples in the dataset, where each data point corresponds to an
263 individual sample. In MouSR users can also select any pair of the first three principal
264 components for their static PCA plot and avail of customisable options, including the ability to
265 choose a different colour for defined groups, turning labels on/off, justifying height and width
266 of the plot, changing size of the labels/points, and having different downloading format
267 (png/svg). In the 3D PCA plot option, MouSR exploits the functionality offered by the Plotly
268 package (Sievert, 2019) to generate an interactive plot with adjustable features, giving users the
269 option to rotate and zoom the graphic, isolate certain samples, alongside the ability to obtain
270 sample information by hovering the mouse pointer over each data point. Instructions are displayed
271 on the left-hand side of the plot, under the Plotly mode bar control, and hovering over the 3D PCA

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

272 graphic itself will also reveal the built-in adjustable options above the sample labels on the top
273 right. Furthermore, the colours of the data points are linked to the earlier 2D PCA colour option.
274 Using our CRC mouse exemplar files (Jackstadt et al., 2019) (Jackstadt, 2019) (Ahmaderaghi,
275 2021), samples related to each experimental group are identifiable using the same colours in both
276 the 2D (Figure 2C) and 3D options (Figure 2D).

277

278 *MDS 2D Plot:* The MDS plot has been added as a further option to project high dimensional data
279 down to two dimensions, while preserving relative distances between observations (described in
280 the Methods section). Again, the colour of the plot is linked to the 2D PCA.

281

282 **Differential Gene Expression Analysis:**

283 A primary objective of many gene expression experiments is to detect and analyse transcripts that
284 display differential expression levels across different samples or experimental conditions. In
285 MouSR, such analyses are made easy via a series of intuitive customisable options that enable
286 selection of bespoke groups, thresholds, and filtering criteria.

287 *Heatmap:* MouSR, has been developed to ensure that a choice of different filtering options are
288 provided. In the group comparison analysis panel, the comparison between two main groups is
289 embedded from the labelling information the user uploaded in input 2, with the names of groups
290 appearing as a list in dropdown menus for both group A and group B. This design provides the
291 user with options to compare two individual groups, or to perform the comparison on up to 10
292 groups at each time, as users can pool up to 5 experimental groups in Group A versus up to 5

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

293 in Group B. The default Heatmap plot is generated based on log2FoldChange [-2,2] and an
294 adjusted p-value cutoff of 0.05, using the "heatmaply" package (version 1.1.1) (Galili et al.,
295 2021) , however all of these options can be adjusted by the user (Figure 3A). Clicking the submit
296 button will initiate the MouSR app to run DESeq2 on the data, producing customisable heatmaps,
297 tables, boxplots, and volcano plots. Once complete, the heatmap plot details how many genes
298 are either up- or down-regulated under these conditions. However, as with most features in
299 MouSR, users have the option to adjust these to their own desired values, followed by clicking
300 on "Update Plot" button to trigger the heatmap to be updated in real time. Users have the option
301 to perform the clustering according to sample names or gene names or both, in order to visualize
302 the differentially expressed genes. This interactive tool allows the inspection of a specific value
303 by hovering the mouse over a cell, as well as zooming into a specific section of the figure by
304 clicking and dragging around the relevant area. The differential gene expression data can also be
305 displayed or downloaded as a table for analysis in other downstream tools, with samples as
306 columns and gene names as rows. To demonstrate the utility of these features, we utilise the
307 exemplar files to reproduce some of the main findings from the original Jackstadt et al. study
308 (Jackstadt et al., 2019). Using MouSR, we first perform differential gene expression analysis
309 comparing primary tumour data from four mouse genotypes (AP, APN vs KP, KPN), and plot
310 the resulting differential genes using the heatmap tool (Fig 3B).

311 *Heatmap for selected Genes:* This section provides the ability for users to create heatmaps for
312 specific genes of interest (minimum of two genes) by clicking on the gene name in the table or by
313 using the search bar on the top right corner of the table. Once selected, the heatmap will be created
314 in real time as more genes are selected/deselected. Furthermore, there are two ways to create a
315 heatmap, based on either individual sample values obtained during the differential analysis, or by

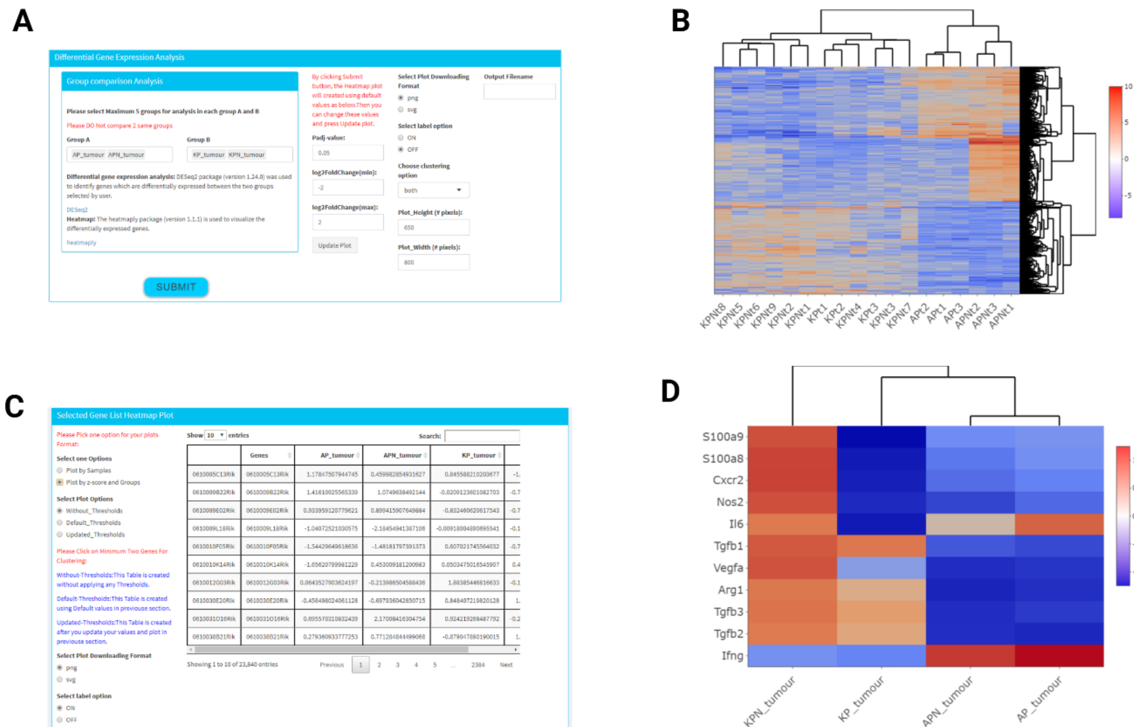
bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

316 creating an experimental group summary Z-score (scale between 1.25 and -1.25) according to each
317 group analysed (Figure 3C). The table is created using three different options, the first one is for
318 every gene uploaded in the original matrix, without applying any thresholds
319 (Without_Thresholds), the second option (Default_Thresholds) is based on default values
320 log2FoldChange [-2,2] and Padj-value: 0.05 in the previous section. The third option
321 (Updated_Thresholds) will reflect any modifications to the default thresholds selected by the
322 user during the previous differential step. From the original study, a number of specific markers
323 were found to be differentially expressed between these models (Figure 6I in Jackstadt et al.
324 (Jackstadt et al., 2019)), namely S100a9, S100a8, Cxcr2, Nos2, Il6, Tgfb1, Vegfa, Arg1, Tgfb3,
325 Tgfb2 and Ifng. Assessment of gene expression levels for these individual genes produced a result
326 in less than 30 seconds that was consistent with the original study, confirming the utility of the
327 MouSR application (Figure 3D).

328

329

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



330

331 **Figure 3: Differential gene expression analysis and visualisation options.** A) Illustration of the
332 various filtering options in the Heatmap panel for the gene expression in all the samples from the
333 chosen groups. B) Heatmap depicts comparison between AP, APN vs KP, KPN across tumors. C)
334 The selected gene side bar for various filtering options is visible to the left of the table. D) Heatmap
335 indicates a reproducible result, compared to data from (Jackstadt et al., 2019) (Figure 6I) across
336 the selected GEMM tumors.

337

338 *Volcano plot:* Using the open-source tool VolcaNoseR (Luijsterburg and Goedhart, 2020) as
339 inspiration, we have incorporated an interactive and customized volcano plot into MouSR using
340 two different options. The first option is “using-plotly”, which exploits functions within the Plotly
341 package (Sievert, 2019) that give the user the option to obtain essential information by hovering
342 the mouse pointer over a dot showing the name of a corresponding gene. The second option, called
343 “selectedGenes”, gives the users the option to annotate the volcano plot with up to 10 genes names
344 (case-sensitive), which generates a new plot in real-time. As a default, the volcano plot shows the

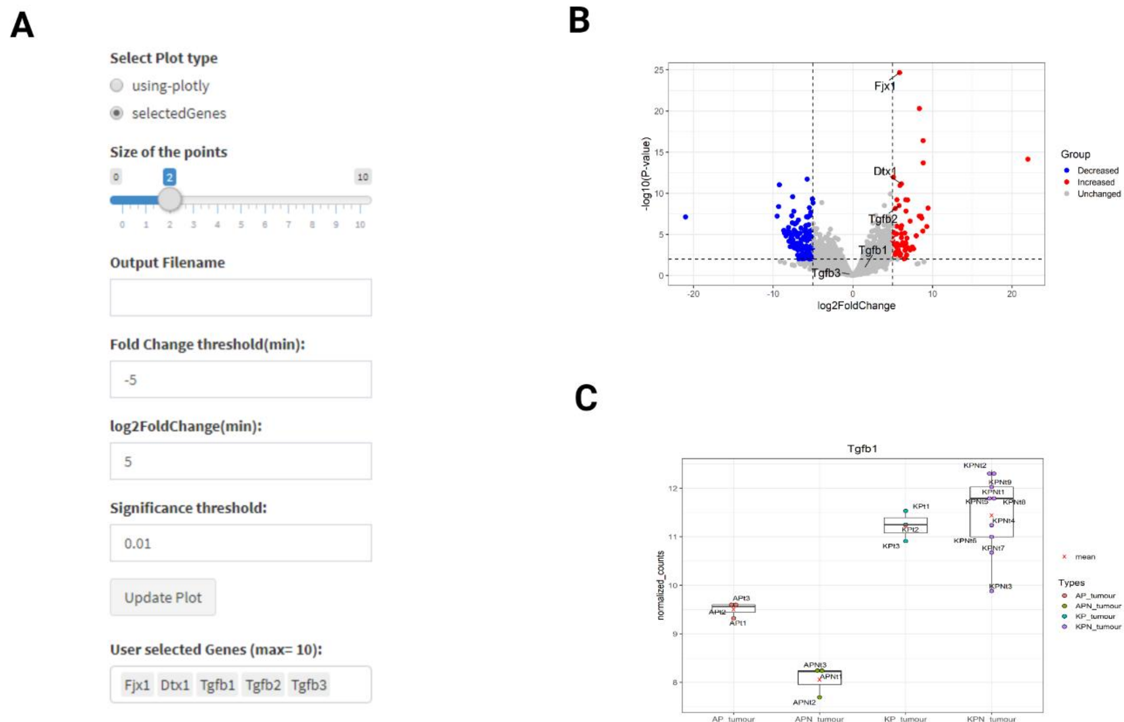
bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

345 log2 of the fold change [-5,5] on the x-axis and minus log10 of the p-value on the y-axis, with a
346 Significance threshold of 0.01 (Figure 4A). However, users have the option to adjust the size of
347 data points, alongside options to modify parameters to their own desired values, followed by
348 clicking on "Update Plot" button to trigger the volcano plot to be updated in real time. Using
349 our exemplar dataset, we examined the transcriptome of KPN versus KP organoids, where we
350 produce a volcano plot that demonstrates increased expression in Fjx1, Dtx1 and Tgfb2 (Figure
351 4B), similar to the published data (Figure 6A in Jackstadt et al. (Jackstadt et al., 2019)).

352 *Gene Expression Levels:* In this subsection, users can again select a specific gene name, via the
353 table or the search option, to produce a boxplot of normalized count for expression values in
354 each experimental group. In the original study, there was a focus on elevated Tgfb gene expression
355 in the KP and KPN models compared to the AP and APN models. Again, using the intuitive
356 MouSR system, we utilise the normalised counts for Tgfb1 expression plotted to reproduce this
357 main finding (Figure 4C).

358

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



359

360 **Figure 4: Gene Expression Levels and Volcano plot options.** A) Volcano plot filtering options, 361 by hovering over the plot using Plotly, the information related to each gene can be accessed 362 immediately. B) Volcano plot displaying differentially expressed genes with highlighted key genes 363 in text between KPN and KP organoids (reproducible result compared to data from (Jackstadt et 364 al., 2019) Figure.6A). C) Boxplot displays normalised counts for Tgfb1 Expression compared 365 between groups.

366

367

368 Mouse/Human-specific Gene Set Enrichment Analysis classification (GSEA):

369 Since its introduction, GSEA (Mootha et al., 2003) (Subramanian, 2005) has become an essential 370 part of the genomic analysis compendium of tools, due to its ability to measure and compare 371 similarities or differences in experimentally validated biological signatures in transcriptional 372 datasets. In MouSR, we provide options for both the original pairwise GSEA method (group A v

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

373 B) and the modified single sample classification (ssGSEA), using the fgsea and GSVA packages
374 (Hänzelmann et al., 2013) respectively, for both the Hallmark and Gene Ontology collections.
375 Furthermore, in order to facilitate simultaneous classification between human- and mouse-derived
376 data, we have extended our framework to provide an option for the users to choose between human
377 or mouse analytical packages, based on their transcriptional data. Users also have the option of
378 uploading their own bespoke list of genes or pathways of interest as an .rdata file. For users with
379 a gene list from a spreadsheet, we have also created a side link that will convert a .txt file to .rdata,
380 making this more user-friendly for non-computational users.

381 *GSEA plot:* In the GSEA plot section, users have the option to compare their two groups (selected
382 during the differential analysis) with any specific gene sets within the Hallmark or Gene Ontology
383 collections, which produces an enrichment plot and an indication of the number of leading-edge
384 genes. For the Gene Ontology option, as the collection comprises over 7000 gene sets, only the
385 first 50 pathways based on Enrichment Score (ES) will be available. The GSEA algorithm ranks
386 genes based their expression, focussing on enrichment differences between samples belonging to
387 two classes, labelled A or B.

388

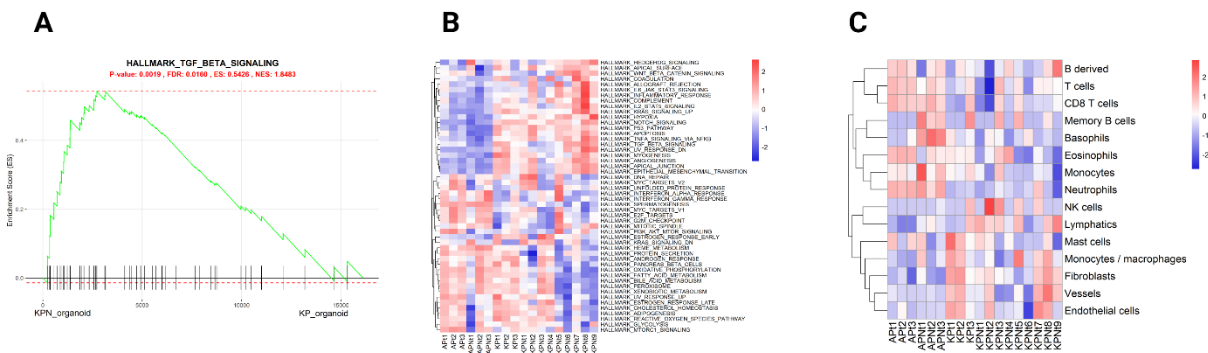
389 **Mouse/Human-specific Microenvironment Cell Population counter (mMCP/MCP-counter)**

390 The MCP algorithm gives an estimate of predefined immune and stromal cell populations from
391 heterogeneous transcriptomic data (Becht et al., 2016). MouSR includes dual species templates to
392 ensure users can assess either mouse or human data. For human, these populations include 8
393 immune populations (CD3+ T cells, CD8+ T cells, cytotoxic lymphocytes, NK cells, B
394 lymphocytes, cells originating from monocytes (monocytic lineage), myeloid dendritic cells,

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

395 neutrophils), and 2 stromal populations (endothelial cells and fibroblasts) (Becht et al., 2016). For
396 mouse, these populations include 12 immune cell types (T cells, CD8$^+$ T cells, NK cells, B-derived
397 cells, memory B cells, monocytes/macrophages, monocytes, granulocytes, mast cells, eosinophils,
398 neutrophils, and basophils) and 4 stromal populations (vessels, lymphatics, endothelial cells, and
399 fibroblasts) (Petitprez et al., 2020).

400 Using the exemplar data, we perform GSEA using the Hallmarks collection on KP and KPN
401 tumour samples, where, in line with the original publication, we observe an enrichment for
402 TGF_BETA_SIGNALLING in KPN compared to KP tumour (Figure 5A). In addition to the pair-
403 wise method, MouSR also enables users to perform single sample assessment using ssGSEA
404 (Figure 5B) and MCP (Figure 5C) to assess enrichment in individual samples regardless of the
405 experimental group. Given the adaptability of the MouSR tool, we will continue to add new options
406 for data analyses, therefore features in a testing phase will be indicated as such (i.e. beta version).



407
408 **Figure 5: Gene set enrichment analysis and MCP analysis.** A) Enrichment plot for TGF-beta
409 signalling Hallmark gene set for KPN vs KP organoid groups with p-value, FDR value, enrichment
410 score (ES) and normalised enrichment scores (NES). The X-axis is all the genes in the data
411 experiment pre-ranked by the metric, where each black bar is the gene in this gene set (pathway)
412 and the Y-axis details the level of enrichment via an Enrichment Score (ES). B) Single sample
413 gene set enrichment analysis for individual samples displayed in a heatmap. C) mMCP analysis
414 with infiltrating cell population estimates visualised in a heatmap.

415

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

416 **Discussion**

417 Every day significant amounts of molecular data are created from biological samples at an ever-
418 reducing cost, shifting the challenge from data acquisition to data analysis and interpretation that
419 can inform deeper understanding of biological processes. Such understanding is essential in order
420 to improve our understanding of disease and identify mechanistic signaling that can aid in
421 diagnosis, prediction of disease outcomes or in the development of new therapeutic strategies
422 (Gambardell, 2020). An example of how important interrogation of molecular data can be is
423 reflected in the worldwide response to the COVID-19 pandemic, where rapid interpretable data
424 underpinned a meaningful mitigation response to the pandemic's impact on health and society (Lai
425 et al., 2020).

426 Data analytical pipelines require specific skill sets, such as data informatics and specific
427 programming, which are not currently in the armamentarium of traditional "wet-lab" scientists.
428 An increased focus on biomarker development and target-drug discovery for personalized
429 medicine requires results generated by gene expression profiling to be interrogated using high-
430 performance computing and potentially with advanced Artificial Intelligence (AI) / Machine
431 Learning (ML) algorithms, again requiring the use of complex bioinformatics tools. For the non-
432 computational biologist, MouSR, with its intuitive structure and user-friendly navigation, enables
433 rapid point-and-click publication-ready analysis of highly complex information, where significant
434 volumes of data can be analysed using multiple methodologies on a single app. MouSR provides
435 a unique opportunity for non-specialist users to analyse their data using customised easy-to-use
436 bioinformatic tools, while also having dual functionality embedded within the App to investigate
437 disease-specific models and algorithms that offer deeper insights to facilitate simultaneous
438 classification between human- and mouse-derived data. The user can choose to deploy all the

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

439 features within our intuitive transcriptional analysis pipeline for comprehensive work-up, or in
440 other instances the user might decide to utilise only a selection of the available options within
441 MouSR for their bespoke analysis requirements.

442 The application is internet accessible, but by making our source code freely available, MouSR
443 provides an open-source option for individual users or institutes to install their own instance on
444 local computers/servers. Furthermore, given the remarkable growth in the R programming
445 language community, the MouSR tool provides an adaptable template for further development that
446 will is not limited by recurring software fees. As a clear demonstration of the utility of MouSR,
447 utilising our previously published data, we rapidly reproduced a number of the main molecular
448 findings from the original study in a matter of minutes.

449 During its development, decisions were made to broaden the range of analyses that MouSR could
450 offer, which in turn leads to a number of limitations which we acknowledge. Multiple analysis
451 methods in MouSR have been created using different libraries under different version of R; finding
452 the best R version that can suit them all and at the same time can accommodate the shiny server
453 version and CentOS server can be challenging. For instance, there is a recently published library
454 called "Interactive Complex Heatmap" (Zuguang Gu and Hübschmann, 2021) , which provides an
455 easy-to-use tool for constructing highly customizable heatmaps, especially for analysing DESeq2
456 results. However, based on the version of R and shiny that we are using, we preferred to deploy
457 the "heatmaply" package. It is worth noting that although this single purpose tool can provide
458 highly customizable heatmaps, it does not have the same breath of capabilities or versatility in
459 comparison with MouSR. Additionally, MouSR (as a free, open-source tool), can provide more
460 analysis methods than other existing free analytical apps currently available. For instance, DEApp
461 (Li and rade, 2017) is only focused on differential expression analysis of count-based NGS data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

462 In addition, TCC-GUI (Wei Su et al., 2019) uses differential expression pipelines with robust
463 normalization and simulation data generation under various conditions, however it does not
464 include gene set enrichment analysis and MCP/mMCP counter analysis. The START tool (Nelson
465 et al., 2017), while having a number of specific functionalities, again does not include gene
466 enrichment analysis and MCP/mMCP counter analysis. Finally, the GENAVi application (Reyes
467 et al., 2019) can provide certain analyses, however does not include MCP/mMCP counter analysis,
468 multi-group comparison or dual functionality for both human and mouse derived data when
469 compared to MouSR. Despite these limitations, the MouSR architecture design provides a
470 structure that offers, in the future, the possibility to implement new types of bespoke analysis
471 pipelines and graphical outputs with precise functionalities within the open-source R programming
472 language, facilitating access to thousands of statistical packages which are continually released
473 and updated globally.

474

475 In summary, MouSR is a freely-available tool that provides a user-friendly graphical interface for
476 biological characterisation and interrogation of transcriptional datasets. Approaches such as ours
477 help remove a bottleneck in biological discovery for users with limited programme skills, enabling
478 them to perform statistically-controlled bioinformatics analyses to make valid biologically-
479 informed conclusions more precisely.

480

481

482

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

483 **Acknowledgments**

484 This project has been developed as a part of the CRUK-funded ACRCelerate colorectal cancer
485 pre-clinical accelerator program. The ACRCelerate project brings together a European-wide
486 consortium of basic and clinical scientists at the forefront of CRC research to interrogate a suite of
487 state-of-the art preclinical models. The overarching aim of ACRCelerate is to generate robust and
488 reproducible preclinical data to de-risk future clinical trials via patient stratification. We
489 acknowledge 'VolcaNoseR' for their Volcano plot and 'DEApp' for their first page template
490 inspiration.

491 We thank the collaborative network across Belfast, Oxford and Glasgow for feedback and
492 discussions during development. We also thank the IT department at Queen's University and
493 Crispin Miller at the Beatson institute for advice and support.

494

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

495 **References**

496

497 Ahmaderaghi, B. (2021). *MouSR.* [Online]
498 Available at: https://github.com/Dunne-Group/MouSR/Data

499 Attali, D. and Griswold, D. (2020). *colourpicker.* [Online]
500 Available at: https://cran.r-project.org/web/packages/colourpicker/index.html

501 Bache,M.S., Wickham,H.,Henry,L. and RStudio. (2020). *magrittr.* [Online]
502 Available at: https://cran.r-project.org/web/packages/magrittr/index.html

503 Becht, E., Giraldo, N., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P.,
504 Sautès-Fridman, C., Fridman, WH. et al. (2016). Estimating the population abundance of tissue-
505 infiltrating immune and stromal cell populations using gene expression. *Genome Biology,* pp. 17, 218.

506 Cerami,E., Gao,J., Dogrusoz,U.,Gross,EB., Sumer,OS., Aksoy,BA., Jacobsen,A., Byrne,JC. , Heuer, ML.,
507 Larsson,E. et al. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring
508 Multidimensional Cancer Genomics Data. *Cancer Discovery 2,* p. 40.

509 Chang,W. , Cheng,J., Allaire,JJ. , Sievert,C., Schloerke,B. , Xie,Y., Allen,J., McPherson,J. , Dipert,A. et
510 al. (2018). *Shiny: Web Application Framework for R.* [Online]
511 Available at: https://cran.r-project.org/web/packages/shiny/index.html

512 Chang,W., Ribeiro,BB., RStudio, Almasaeed Studio and Adobe Systems Incorporated.(2018).
513 *shinydashboard.* [Online]
514 Available at: https://cran.r-project.org/web/packages/shinydashboard/index.html

515 Chang,W., RStudio, Park,T., Dziedzic,L., Willis,N., Google Corporation , McInerney,M., Adobe Systems
516 Incorporated, Canonical Ltd.(2021). *shinythemes.* [Online]
517 Available at: https://cran.r-project.org/web/packages/shinythemes/index.html

518 Data Design Group. (2013). *Transpose CSV Tool(convertcsv).* [Online]
519 Available at: https://www.convertcsv.com/transpose-csv.htm

520 Dolgalev, I.(2020). *msigdbr.* [Online]
521 Available at: https://cran.r-project.org/web/packages/msigdbr/index.html

522 Durinck,S., Huber,W., Davis,S., Pepin,F., Buffalo,VS. and Smith,M .(2005). *biomaRt.* [Online]
523 Available at: https://bioconductor.org/packages/release/bioc/html/biomaRt.html

524 Edgar,R. , Domrachev,M. and Lash,AE.(2002). Gene Expression Omnibus: NCBI gene expression and
525 hybridization array data repository. *Nucleic Acids Research,* pp. 207-10.

526 Galili,T., O'Callaghan,A., Sidi,J. and Benjamini,Y.(2021). *heatmaply.* [Online]
527 Available at: https://cran.r-project.org/web/packages/heatmaply/index.html

528 Gambardell, V.(2020). Personalized Medicine: Recent Progress in Cancer Therapy. *Cancers (Basel),*
529 Volume 12(4): 1009.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

530 Gao,J. , Aksoy,BA., Dogrusoz,U., Dresdner,G., Gross,B., Sumer,SO., Sun,Y., Jacobsen,A., Sinha,R.,
531 Larsson,EK. et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the
532 cBioPortal. *science Signaling,* pp. 6, pl1.

533 Hänzelmann, S., Castelo,R. and Guinney,J. (2013). GSVA: gene set variation analysis for microarray and
534 RNA-Seq data. *BMC Bioinformatics.*

535 Langfelder,P., Horvath,S., Cai,C., Dong,J., Miller,J., Song,L., Yip,A. and Zhang,B. (2021). *WGCNA.*
536 [Online]
537 Available at: https://cran.r-project.org/web/packages/WGCNA/index.html

538 Jolliffe,TI. and Cadima, J. (2016). Principal component analysis: a review and recent developments.
539 *Philosophical transactions of the royalsociety.*

540 Kolde, R. (2019). *pheatmap.* [Online]
541 Available at: https://cran.r-project.org/web/packages/pheatmap/index.html

542 Lai,AG., Pasea,L., Banerjee,A. Hall,G., Denaxas,S., Chang,WH.,Katsoulis,M., Williams,B.,Pillay,D.,
543 Noursadeghi,M. et al. (2020). Estimated impact of the COVID-19 pandemic on cancer services and
544 excess 1-year mortality in people with cancer and multimorbidity: near real-time data on cancer care,
545 cancer deaths and a population-based cohort study. *BMJ Open.*

546 Leonelli, S. (2019). The challenges of big data biology. *Elife,* p. 8. 10.7554/eLife.47381.

547 Li,Y., and Andrade,J.(2017). DEApp: an interactive web interface for differential expression analysis of
548 next generation sequence data. *Source Code for Biology and Medicine,Springer Nature journal.*

549 Love,M., Ahlmann-Eltze,C., Forbes,K., Anders,S. and Huber,W.(2014). *DESeq2.* [Online]
550 Available at: https://bioconductor.org/packages/release/bioc/html/DESeq2.html

551 Love, MI., Huber,W. and Anders,S. (2014). Moderated estimation of fold change and dispersion for
552 RNA-seq data with DESeq2. *Genome Biology,* p. 15(12):550.

553 Luijsterburg,M. and Goedhart,J. (2020). VolcaNoseR – a web app for creating, exploring, labeling and
554 sharing volcano plots. *Scientific Reports,* Volume 10(1):20560, p. 25.

555 Mootha ,VK., Lindgren,CM., Eriksson,KL., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P.,
556 Carlsson,E., Ridderstråle,M., Laurila,E. et al (2003). PGC-1alpha-responsive genes involved in oxidative
557 phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics,* Volume 34, p. 267–
558 273.

559 Müller,K. ,Wickham,H. , Francois,R., Bryan,J and RStudio. (2021). *tibble.* [Online]
560 Available at: https://cran.r-project.org/web/packages/tibble/index.html

561 NCI and National Human Genome Research Institute (2006). *The Cancer Genome Atlas Program.*
562 [Online] Available at: https://www.cancer.gov/about-nci/organization/ccg/research/structural-
563 genomics/tcga

564 Nelson,W.J, Sklenar,J., Barnes,A.P. and Minnier,J. (2017). The START App: a web-based RNAseq
565 analysis and visualization resource. *Bioinformatics Journal-Oxford Academic,* 33(3), p. 447–449.

566 Neuwirth, E.(2014). *RColorBrewer.* [Online]
567 Available at: https://cran.r-project.org/web/packages/RColorBrewer/index.html

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

568 Petitprez, F., Levy, S., Sun,CM., Meylan,M., Linhard,C., Becht, E., Elarouci, N., Tavel, D., Roumenina,
569 LT., Ayadi, M. et al. (2020). The murine Microenvironment Cell Population counter method to estimate
570 abundance of tissue-infiltrating immune and stromal cell populations in murine samples using gene
571 expression. *Genome Medicine.*

572 Jackstadt,R., van Hooff,SR., Leach,JD. , Cortes-Lavaud,X., Lohuis,JO., Ridgway,RA., Wouters,VM.,
573 Roper,J. , Kendall,TJ, Sansom,OJ. et al.(2019). Epithelial NOTCH Signaling Rewires the Tumor
574 Microenvironment of Colorectal Cancer to Drive Poor-Prognosis Subtypes and Metastasis. *Cancer Cell,*
575 Volume 36, p. 319–336.

576 Jackstadt, R. (2019). *ArrayExpress- E-MTAB-6363 - RNA-seq of intestinal cancer GEMMs.* [Online]
577 Available at: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6363/

578 Ren, K. (2016). *rlis.* [Online]
579 Available at: https://cran.r-project.org/web/packages/rlist/index.html

580 Reyes, ALP., Silva, TC., Coetzee, SG., Plummer, JT., Davis, BD., Chen,S., Hazelett, DJ., Lawrenson, K.,
581 Berman, BP., Simon A. Gayther, SA. et al. (2019). GENAVi: a shiny web application for gene expression
582 normalization, analysis and visualization. *BMC Genomics.*

583 Sali,A., Hass,L. and Attali,D. (2020). *shinycssloaders.* [Online]
584 Available at: https://cran.r-project.org/web/packages/shinycssloaders/index.html

585 Sarkar,D., Andrews,F., Wright,K. , Klepeis,N. , Larsson,J. and Murrell,P. (2020). *lattice.* [Online]
586 Available at: https://cran.r-project.org/web/packages/lattice/index.html

587 Sergushichev, A.(2019). An algorithm for fast preranked gene set enrichment analysis using cumulative
588 statistic calculation.. *Preprint at http://biorxiv.org/content/early/2016/06/20/060012.*

589 Sievert, C.(2019). *plotly for R.* [Online]
590 Available at: https://plotly-r.com

591 Slowikowski,K., Schep,A.,Hughes,S., Kien Dang,T., Lukauskas,S., Irisson ,Jo.,Kamvar,ZN.,
592 Ryan,T.,Christophe,D., Hiroaki,Y. et al. (2021). *ggrepel.* [Online]
593 Available at: https://cran.r-project.org/web/packages/ggrepel/index.html

594 Smyth,G., Hu,Y. Ritchie,M., Silver,J., Wettenhall,J.,McCarthy,D., Wu,D., Wei Shi,W., Phipson,B.,
595 Lun,A. et al. (2015). *limma.* [Online] Available at:
596 https://bioconductor.org/packages/release/bioc/html/limma.html

597 Sohn, BH., Hwang,JE.,Jang,HJ., Lee,JS., Oh,SH., Shim,JJ., Lee.KW.,Kim,EH., Yim,SY., Lee,SH.et
598 al.(2017). Clinical Significance of Four Molecular Subtypes of Gastric Cancer Identified by The Cancer
599 Genome Atlas Project. *Clin Cancer Research.*

600 Subramanian, A.(2005). Gene set enrichment analysis: a knowledge-based approach for interpreting
601 genome-wide expression profiles. *Proceedings of the National Academy of Sciences,* Volume 102, p.
602 15545–15550.

603 Tanaka,E. and Niichan.(2018). *shinycustomloader.* [Online]
604 Available at: https://cran.r-project.org/web/packages/shinycustomloader/index.html

605 Wei Su,W.,Sun,J.,Shimizu,K. and Kadota,K.(2019). TCC-GUI: a Shiny-based application for differential
606 expression analysis of RNA-Seq count data. *BMC Research Notes.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.12.456127; this version posted August 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

607 Wickham,H., Chang,W., Henry,L.,Pedersen,TL.,Takahashi,K.,
608 Wilke,C.,Woo,K.,Yutani,H.,Dunnington,D. and RStudio. (2020). *ggplot2.* [Online]
609 Available at: https://cran.r-project.org/web/packages/ggplot2/index.html

610 Wickham,H., Bryan,J., Kalicinski,M., Valery,K., Leitienne,C., Colbert,B, Hoerl,D., Miller,E. and
611 RStudio. (2019). *readxl.* [Online]
612 Available at: https://cran.r-project.org/web/packages/readxl/index.html

613 Wickham,R., François,R., Henry.L., Müller,K.and RStudio (2021). *dplyr.* [Online]
614 Available at: https://cran.r-project.org/web/packages/dplyr/index.html

615 Wickham,H.R.(2018). *reshape.* [Online]
616 Available at:https://cran.r-project.org/web/packages/reshape/index.html

617 Wickham,H.R.(2020) *plyr.* [Online]
618 Available at: https://cran.r-project.org/web/packages/plyr/index.html

619 Wickham,H.R.(2021). *tidyverse.* [Online]
620 Available at: https://cran.r-project.org/web/packages/tidyverse/index.html

621 Xie,Y., Cheng,J., Tan,X. , Allaire,JJ., Girlich,M., Ellis,GF., Rauh,J. Reavis,B ., Gersen,L., Szopka,B. et
622 al. (2021). *DT.* [Online]
623 Available at: https://cran.r-project.org/web/packages/DT/index.html

624 Young,F.W. (2013). *Multidimensional Scaling: History, theory, and applications.* Psychology Press.

625 Yu,G. and Hu, E. (2021). *enrichplot.* [Online]
626 Available at: https://bioconductor.org/packages/release/bioc/html/enrichplot.html

627 Zuguang Gu,Z. and Hübschmann,D.(2021). Make Interactive Complex Heatmaps in R. *Preprint at*
628 *https://www.biorxiv.org/content/10.1101/2021.03.08.434289v3.*

629