Molecular Subtyping Resource: a user-friendly tool for rapid biological discovery from transcriptional data

Ahmaderaghi, B., Amirkhah, R., Jackson, J., Lannagan, T. R. M., Gilroy, K., Malla, S. B., Redmond, K. L., Quinn, G., McDade, S., Maughan, T., Consortium, ACRC., Leedham, S., Campbell, A. S., Sansom, O. J., Lawler, M., & Dunne, P. D. (2022). Molecular Subtyping Resource: a user-friendly tool for rapid biological discovery from transcriptional data. *Disease Models and Mechanisms*, *15*(3). https://doi.org/10.1242/dmm.049257

Published in:
Disease Models and Mechanisms

Document Version:
Publisher's PDF, also known as Version of record

© 2022. Published by The Company of Biologists Ltd | Disease Models & Mechanisms (2022) 15, dmm049257. doi:10.1242/dmm.049257

RESOURCE ARTICLE

Molecular Subtyping Resource: a user-friendly tool for rapid biological discovery from transcriptional data

Baharak Ahmaderaghi[1,*], Raheleh Amirkhah[1], James Jackson[2], Tamsin R. M. Lannagan[3], Kathryn Gilroy[3], Sudhir B. Malla[1], Keara L. Redmond[1], Gerard Quinn[1], Simon S. McDade[1], ACRCelerate Consortium[4], Tim Maughan[5], Simon Leedham[6], Andrew S. D. Campbell[3], Owen J. Sansom[3,7], Mark Lawler[1,‡] and Philip D. Dunne[1,‡,§]

ABSTRACT

Generation of transcriptional data has dramatically increased in the past decade, driving the development of analytical algorithms that enable interrogation of the biology underpinning the profiled samples. However, these resources require users to have expertise in data wrangling and analytics, reducing opportunities for biological discovery by 'wet-lab' users with a limited programming skillset. Although commercial solutions exist, costs for software access can be prohibitive for academic research groups. To address these challenges, we have developed an open source and user-friendly data analysis platform for on-the-fly bioinformatic interrogation of transcriptional data derived from human or mouse tissue, called Molecular Subtyping Resource (MouSR). This internet-accessible analytical tool, https://mousr.qub.ac.uk/, enables users to easily interrogate their data using an intuitive 'point-and-click' interface, which includes a suite of molecular characterisation options including quality control, differential gene expression, gene set enrichment and microenvironmental cell population analyses from RNA sequencing. The MouSR online tool provides a unique freely available option for users to perform rapid transcriptomic analyses and comprehensive interrogation of the signalling underpinning transcriptional datasets, which alleviates a major bottleneck for biological discovery.

This article has an associated First Person interview with the first author of the paper.

KEY WORDS: Bioinformatics, Data analytics, RNA-seq

INTRODUCTION

In the years since the first whole genome was sequenced, the costs associated with the generation of molecular 'big data' have decreased rapidly, to a point at which the data handling, rather than data generation, is the limiting factor in large biological discovery programmes. Furthermore, large repositories [such as The Cancer Genome Atlas (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga) and Gene Expression Omnibus (Edgar et al., 2002)], now provide free access to publicly available molecular data. Large international molecular subtyping projects have markedly improved our biological understanding of cancer (Sohn et al., 2017), but in doing so they have created a critical bottleneck in terms of data reduction, analysis and interpretation, resulting in an urgent need for solutions that enable rapid biological interrogation of large datasets (Cerami et al., 2012).

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

RESULTS

MouSR interface

The MouSR (https://mousr.qub.ac.uk/) platform is implemented as an open-source application that enables both computational and non-computational users to rapidly go from an existing data matrix, containing integers [whole numbers, i.e. RNA-sequencing (RNA-seq) read counts] or decimals (i.e. estimated read counts), to biologically meaningful results in a user-friendly way. At each step of the process, users have the option to modify outputs, through a series of on-the-fly customisable graphics that can all be downloaded at high resolution for future use. The standard pipeline includes initial data quality control assessments, followed by differential analyses, both single-sample and group-wise gene set enrichment analyses and microenvironment population counters, producing publication-ready data (Fig. 1). The system uses a species-agnostic 'blind embedding' format, where users can upload data derived from any patient sample or *in vitro*/*in vivo* model as

[1]The Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast BT9 7AE, UK. [2]Information Services, Queen's University Belfast, Belfast BT7 1NN, UK. [3]Cancer Research UK Beatson Institute, Glasgow G61 1BD, UK. [4]https://www.beatson.gla.ac.uk/ACRCelerate/teams.html. [5]Oxford Institute of Radiation Oncology, University of Oxford, Oxford OX3 7DQ, UK. [6]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [7]Institute of Cancer Sciences, University of Glasgow, Glasgow OX3 7DQ, UK.
*Present address: School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, 16 Malone Road, Belfast, BT9 5AF, UK.
‡These authors contributed equally to this work

§Author for correspondence (p.dunne@qub.ac.uk)

P.D.D., 0000-0001-9160-283X

Handling Editor: Elaine R. Mardis
Received 12 August 2021; Accepted 24 January 2022

Disease Models & Mechanisms (2022) 15, dmm049257. doi:10.1242/dmm.049257

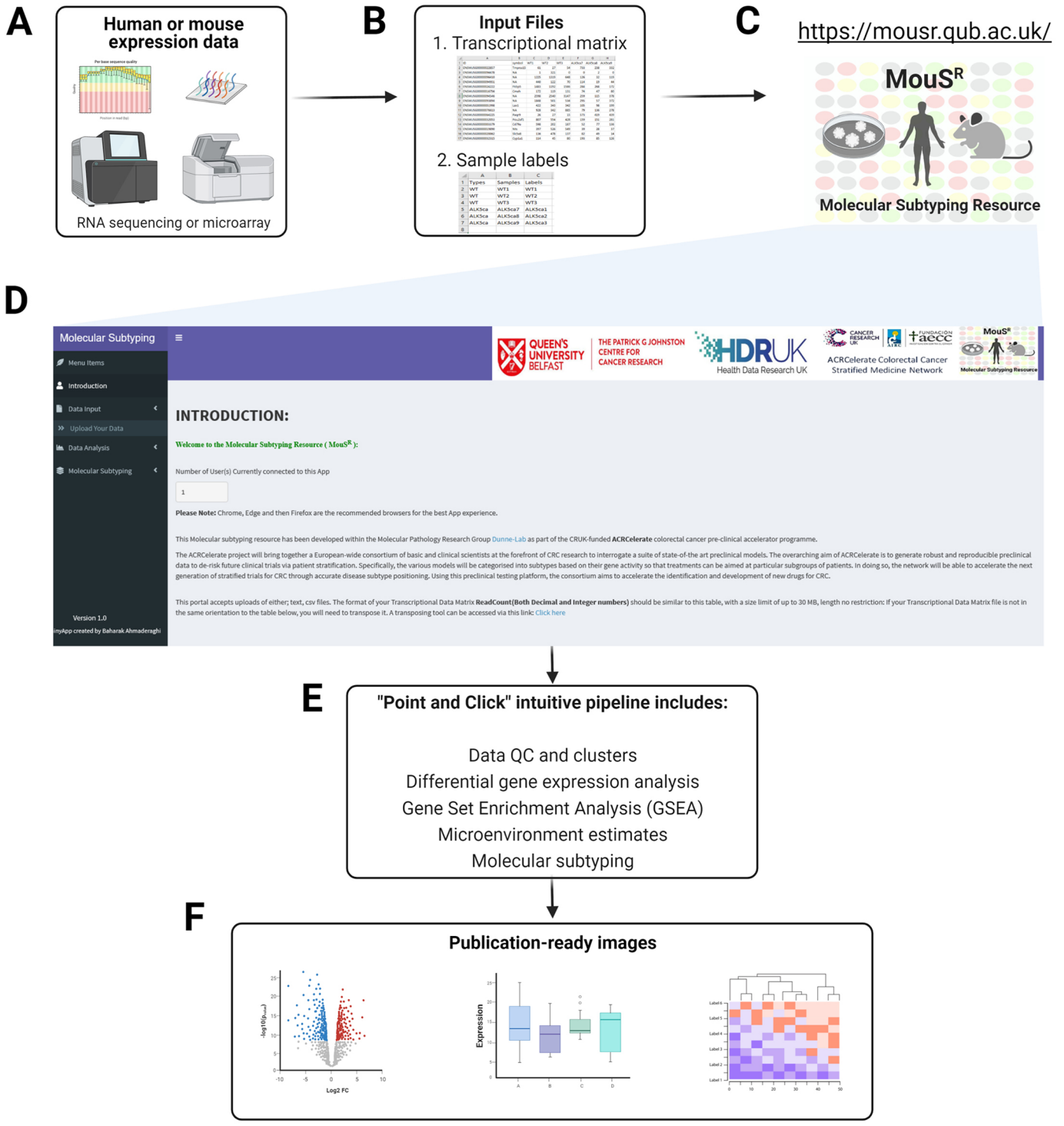

Fig. 1. The MouSR workflow and outputs. (A-F) Utilising transcriptional data derived from human or mouse tissue/cells (A), users are required to have a transcriptional data matrix and sample information as the input for the MouSR pipeline (B), accessible via https://mousr.qub.ac.uk/ (C). From the Introduction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

either ungrouped individual samples or multiple samples within experimental groups. The species-specific selection options available for downstream analyses in MouSR enable users to perform biological discovery/validation on transcriptional data derived from human or mouse origin.

To highlight the functionality of the MouSR system, in this paper, we will focus on colorectal cancer (CRC), using our previously published and comprehensively characterised genetically engineered mouse model tumour and organoid datasets, comprising n=28 samples from matched tumour tissue (n=4

Disease Models & Mechanisms (2022) 15, dmm049257. doi:10.1242/dmm.049257

groups) and genotype-matched organoids (n=3 groups) derived from n=4 different genotypes and available from ArrayExpress study E-MTAB-6363 (Jackstadt et al., 2019).

Accessing the MouSR website presents the user with a General Introduction landing page, providing an overview of the application with instructions and exemplar formats that are required for use. The user interface structure has two main sections, namely (1) Data Input and (2) Data Analysis, which will be described briefly, followed by an analysis of the CRC mouse exemplar dataset (ArrayExpress E-MTAB-6363) from Jackstadt and colleagues (Jackstadt et al., 2019; https://github.com/Dunne-Group/MouSR/tree/main/Data). For convenience, these data are included here as supplementary files (Datasets 1 and 2). In addition, a tutorial video is also included at each point throughout the app to summarise the main features of the MouSR tool.

Data input

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

In terms of data types for input 1, MouSR has been successfully tested using human and mouse data derived from a variety of microarray and RNA-seq platforms and is adaptable enough to accept data that have been processed using a range of pipelines resulting in either integers (whole numbers, i.e. RNA-seq read counts) or decimals (i.e. estimated read counts). However, as DESeq2 requires integer counts as input, for users who select decimal input the differential expression options and group-wise gene set enrichment analysis (GSEA) will not be accessible. Additionally, the MouSR system has been designed to accept the most common file formats, including .csv and .txt data files that utilise various separators including comma, semicolon or tabs. Input 2 includes a summary of basic information that relates to sample labels and groups.

Prior to uploading their data, users must ensure that both files are in the recommended format described on the Introduction page, which is aligned with a standard data matrix output containing gene ID and gene symbol columns followed by sample values and is in the format selected by the user according to their data types (default set as tab delimited/.txt file). To ensure that users with data in an orientation not supported by MouSR can use the tool, we have created a transpose link on the Introduction page, which will adjust the transcriptional matrix using the Transpose CSV Tool (https://www.convertcsv.com/transpose-csv.htm).

Once the files are in the correct format, the user is required to upload their two files into input 1 and input 2 (Fig. 2A), using either a drag/drop from a folder or by navigating to the file location using the browse function. When files are selected, a progress bar will immediately begin to indicate that the file is uploading until the upload is complete.

Given the flexibility and the point-and-click design for downstream analyses, users must have their files in the correct format to proceed. At this stage, by clicking ‘Check Input Files’, users can verify whether their data are correctly loaded or whether modifications are required. If their files have the correct format, then the ‘Continue To Submit’ button will be activated. By clicking on it, the user interface becomes active, triggering the computational analysis on the background server with input detail and input summary being displayed when complete, including information on number of samples, sample names and number of expression values identified. Our exemplar files took <20 s from submission to display of input details, confirming that it consists of 24,751 individual genes across 28 samples across seven experimental groups (Fig. 2B).

Data analysis

The Data Analysis section consists of four main subsections: (1) principal component analysis (PCA) and multi-dimensional scaling (MDS), (2) differential gene expression analysis (DGEA), (3) mouse and/or human GSEA, and (4) mouse and/or human microenvironment cell population counter (mMCP/MCP-counter).

PCA and MDS

PCA and MDS are dimensionality reduction methods that represent an initial step in assessing characteristics of any dataset (Jolliffe and Cadima, 2016; Young, 2013). These options give the user an immediate overview of clustering of samples according to their experimental labels (described further in the Materials and Methods section).

PCA 2D/3D plots

PCA plots enable the user to look at the principal components that describe the largest variability between samples in the dataset, where each data point corresponds to an individual sample. In MouSR, users can also select any pair of the first three principal components for their static PCA plot and choice of customisable options, including the ability to choose a different colour for defined groups, turning labels on/off, justifying height and width of the plot, changing size of the labels/points and having different downloading format (png/svg). In the 3D PCA plot option, MouSR exploits the functionality offered by the Plotly package (https://plotly-r.com) to generate an interactive plot with adjustable features, giving users the option to rotate and zoom the graphic, and isolate certain samples, alongside the ability to obtain sample information by hovering the mouse pointer over each data point. Instructions are displayed on the left-hand side of the plot, under the Plotly mode bar control, and hovering over the 3D PCA graphic itself will also reveal the built-in adjustable options above the sample labels on the top right. Furthermore, the colours of the data points are linked to the earlier 2D PCA colour option. Using our CRC mouse exemplar files (ArrayExpress E-MTAB-6363; https://github.com/Dunne-Group/MouSR/tree/main/Data; Jackstadt et al., 2019), samples related to each experimental group are identifiable using the same colours in both the 2D (Fig. 2C) and 3D (Fig. 2D) options.

MDS 2D plot

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

DGEA

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

Disease Models & Mechanisms (2022) 15, dmm049257. doi:10.1242/dmm.049257

Fig. 2. Data import and exploratory analysis. (A) Two input files are required to begin the analytical pipeline in the app – a gene expression matrix and a metadata that includes sample group labelling. (B) Following data upload, the data summary on samples will be displayed for quick review. (C,D) Exploratory visualisation of data will be provided in a form of 2D principal component analysis (PCA) plot (C) and a 3D PCA plot (D) with labelled sample groups.

Heatmap

MouSR has been developed to ensure that a choice of different filtering options is provided. In the 'Differential categories comparison analysis' panel, the comparison between two main categories is embedded from the labelling information the user uploaded in input 2, with the names of groups appearing as a list in dropdown menus for both Category A and Category B. This design provides the user with options to compare two individual groups, or to perform the comparison on up to ten groups at each time, as users can pool up to five experimental groups in Category A versus up to five experimental groups in Category B. The default heatmap plot is generated based on log2FoldChange [−2,2] and an adjusted *P*-value cutoff of 0.05, using the 'heatmaply' package (version 1.1.1) (https://cran.r-project.org/web/packages/heatmaply/index.html); however, all of these options can be adjusted by the user (Fig. 3A). Clicking the submit button will initiate the MouSR app to run DESeq2 on the data, producing customisable heatmaps, tables, boxplots and volcano plots. Once complete, the heatmap plot details how many genes are either up- or downregulated under these conditions. However, as with most features in MouSR, users have the option to adjust these to their own desired values, followed by clicking on the 'Create Plot' button to trigger the heatmap to be updated in real time. Users have the option to perform the clustering according to sample names or gene names or both, in order to visualise the differentially expressed genes.

This interactive tool allows the inspection of a specific value by hovering the mouse over a cell, as well as zooming into a specific section of the figure by clicking and dragging around the relevant area. The differential gene expression data can also be displayed or downloaded as a table for analysis in other downstream tools, with samples as columns and gene names as rows. To demonstrate the utility of these features, we utilised the exemplar files to reproduce

Disease Models & Mechanisms (2022) 15, dmm049257. doi:10.1242/dmm.049257

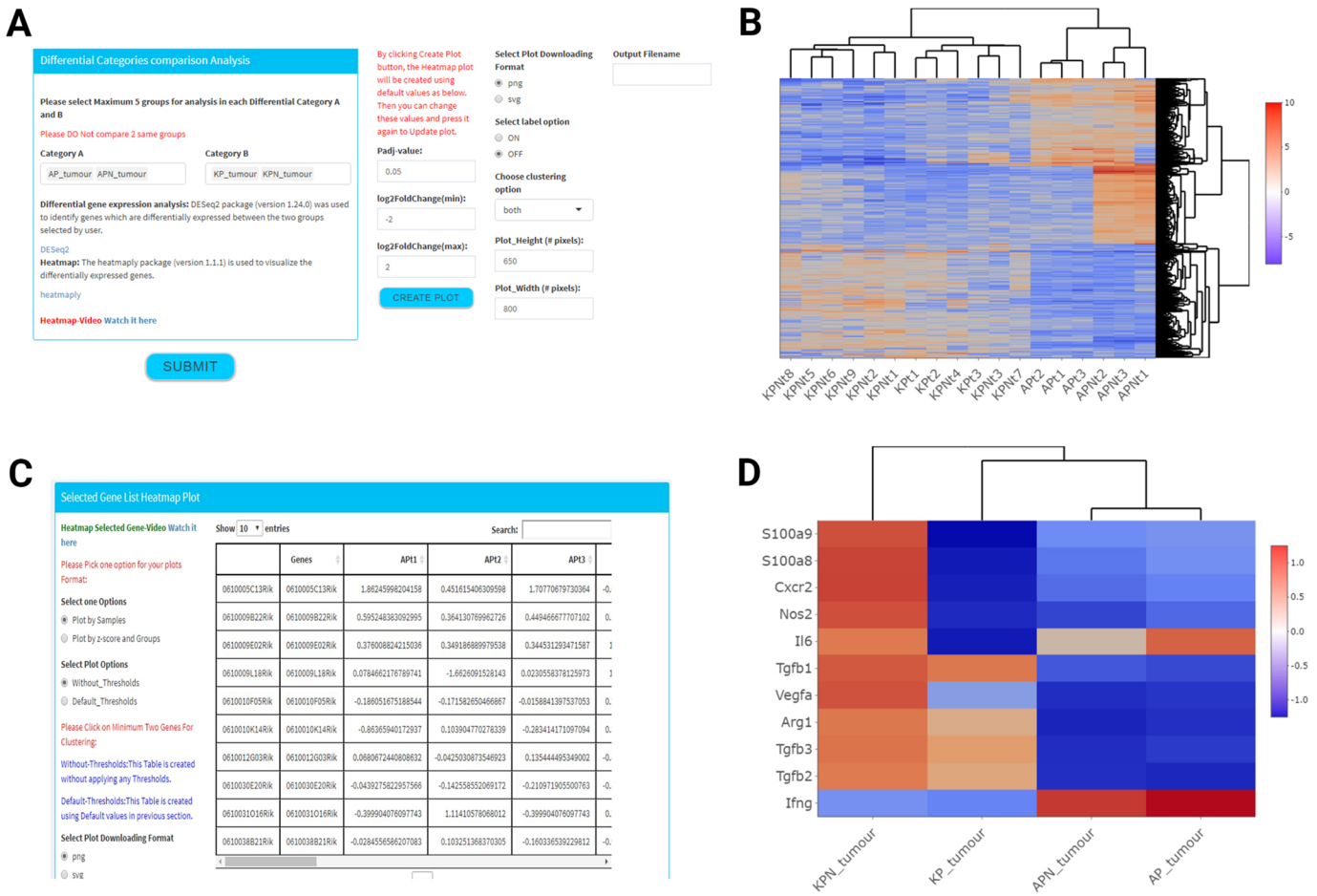

Fig. 3. Differential gene expression analysis and visualisation options. (A) Illustration of the various filtering options in the Heatmap panel for the gene expression in all the samples from the chosen groups. (B) Heatmap depicts comparison of AP, APN versus KP, KPN genotypes across tumours. (C) The selected gene side bar for various filtering options is visible to the left of the table. The table includes the genes and annotations uploaded by the user in the first two columns, followed by columns of expression values under each sample annotation. (D) Heatmap indicates reproducible results, compared to data from Jackstadt et al. (2019) (Fig. 6I) across the selected genetically engineered mouse model tumours.

some of the main findings from the original Jackstadt et al. study (Jackstadt et al., 2019). Using MouSR, we first performed DGEA comparing primary tumour data from four mouse genotypes [villinCre^{ER} *Apc*$^{fl/+}$ *Trp53*$^{fl/fl}$ (AP), villinCre^{ER} *Apc*$^{fl/+}$ *Trp53*$^{fl/fl}$ *Rosa26*$^{N1icd/+}$ (APN) versus villinCre^{ER} *Kras*$^{G12D/+}$ *Trp53*$^{fl/fl}$ (KP), villinCre^{ER} *Kras*$^{G12D/+}$ *Trp53*$^{fl/fl}$ *Rosa26*$^{N1icd/+}$ (KPN)], and plotted the resulting differential genes using the heatmap tool (Fig. 3B).

Heatmap for selected genes

This section provides the ability for users to create heatmaps for specific genes of interest (minimum of two genes) by clicking on the gene name in the table or by using the search bar on the top-right corner of the table. Once selected, the heatmap will be created in real time as more genes are selected/deselected. Furthermore, there are two ways to create a heatmap, based on either individual sample values obtained during the differential analysis, or by creating an experimental group summary *z*-score (scale between 1.25 and −1.25) according to each group analysed (Fig. 3C). The *z*-scores are computed on a gene-by-gene basis by subtracting the mean and then dividing by the s.d. The table is created using two different options; the first is for every gene uploaded in the original matrix, without applying any thresholds (Without_Thresholds), and the second (Default_Thresholds) is based on default values log2FoldChange [−2,2] and an adjusted *P*-value of 0.05 in the previous section or will reflect any modifications to the default thresholds selected by the user during the previous differential step. From the original study, a number of specific markers were found to be differentially expressed between these models [Fig. 6I in Jackstadt et al. (2019)], namely *S100a9*, *S100a8*, *Cxcr2*, *Nos2*, *Il6*, *Tgfb1*, *Vegfa*, *Arg1*, *Tgfb3*, *Tgfb2* and *Ifng*. Assessment of expression levels for these individual genes produced a result in less than 30 s that was consistent with the original study, confirming the utility of the MouSR application (Fig. 3D).

Volcano plot

Using the open-source tool VolcaNoseR (Luijsterburg and Goedhart, 2020) as inspiration, we have incorporated an interactive and customised volcano plot into MouSR using two different options. The first option is ‘using-plotly’, which exploits functions within the Plotly package (https://plotly-r.com) that give the user the option to obtain essential information by hovering the mouse pointer over a dot showing the name of a corresponding gene. The second option, called ‘selectedGenes’, gives the users the option to annotate the volcano plot with up to ten genes names (case sensitive), which generates a new plot in real time. As a default, the volcano plot shows the log2 of the fold change [−5,5] on the *x*-axis and minus log10 of the *P*-value on the *y*-axis, with a significance threshold of 0.01 (Fig. 4A). However, users have the option to adjust the size of data points, alongside options to modify parameters to

Disease Models & Mechanisms (2022) 15, dmm049257. doi:10.1242/dmm.049257

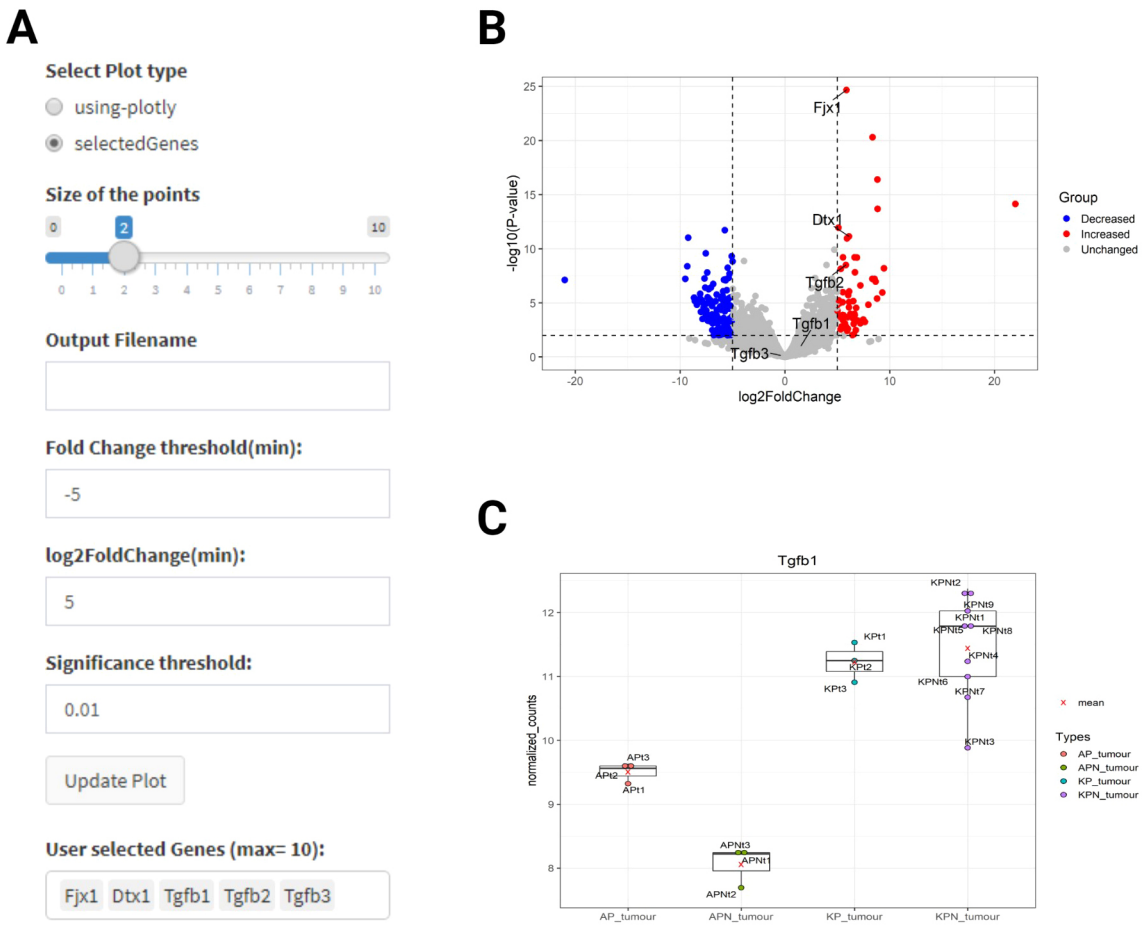

Fig. 4. Gene expression levels and volcano plot options. (A) Volcano plot filtering options, by hovering over the plot using Plotly, the information related to each gene can be accessed immediately. (B) Volcano plot displaying differentially expressed genes with highlighted key genes in text between KPN and KP organoids [reproducible results compared to data from Jackstadt et al. (2019) (Fig. 6A)]. (C) Boxplot displays normalised counts for *Tgfb1* expression compared between groups.

their own desired values, followed by clicking on the 'Update Plot' button to trigger the volcano plot to be updated in real time. Using our exemplar dataset, we examined the transcriptome of KPN versus KP organoids, to produce a volcano plot that demonstrates increased expression in *Fjx1*, *Dtx1* and *Tgfb2* (Fig. 4B), similar to the published data [Fig. 6A in Jackstadt et al. (2019)].

Gene expression levels

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

Mouse/human-specific GSEA classification

Since its introduction, GSEA (Mootha et al., 2003; Subramanian, 2005) has become an essential part of the genomic analysis compendium of tools, owing to its ability to measure and compare similarities or differences in experimentally validated biological signatures in transcriptional datasets. In MouSR, we provide options for both the original pairwise GSEA method (Category A versus B) and the modified single sample classification (ssGSEA), using the fgsea and GSVA packages (Hänzelmann et al., 2013), respectively, for both the Hallmark and Gene Ontology collections. Furthermore, in order to facilitate simultaneous classification between human- and mouse-derived data, we have extended our framework to provide an option for the users to choose between human or mouse analytical packages, based on their transcriptional data. Users also have the option of uploading their own bespoke list of genes or pathways of interest as an .rdata file. For users with a gene list from a spreadsheet, we have also created a side link that will convert a .txt file to .rdata, making this more user friendly for non-computational users.

GSEA plot

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

Mouse/human-specific microenvironment cell population counter (mMCP/MCP-counter)

The MCP algorithm gives an estimate of predefined immune and stromal cell populations from heterogeneous transcriptomic data

Disease Models & Mechanisms (2022) 15, dmm049257. doi:10.1242/dmm.049257

(Becht et al., 2016). MouSR includes dual species templates to ensure that users can assess either mouse or human data. For human, these populations include eight immune populations [$CD3^+$ T cells, $CD8^+$ T cells, cytotoxic lymphocytes, natural killer (NK) cells, B lymphocytes, cells originating from monocytes (monocytic lineage), myeloid dendritic cells and neutrophils] and two stromal populations (endothelial cells and fibroblasts) (Becht et al., 2016). For mouse, these populations include 12 immune cell types (T cells, $CD8^+$ T cells, NK cells, B-derived cells, memory B cells, monocytes/macrophages, monocytes, granulocytes, mast cells, eosinophils, neutrophils and basophils) and four stromal populations (vessels, lymphatics, endothelial cells and fibroblasts) (Petitprez et al., 2020).

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

DISCUSSION

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

Data analytical pipelines require specific skill sets, such as data informatics and specific programming, which are not currently in the armamentarium of traditional 'wet-lab' scientists. An increased focus on biomarker development and target-drug discovery for personalised medicine requires results generated by gene expression profiling to be interrogated using high-performance computing and potentially with advanced artificial intelligence or machine-learning algorithms, again requiring the use of complex bioinformatics tools. For the non-computational biologist, MouSR, with its intuitive structure and user-friendly navigation, enables rapid point-and-click publication-ready analysis of highly complex information, where significant volumes of data can be analysed using multiple methodologies on a single app. MouSR provides a unique opportunity for non-specialist users to analyse their data using customised easy-to-use bioinformatic tools, while also having dual functionality embedded within the app to investigate disease-specific models and algorithms that offer deeper insights to facilitate simultaneous classification between human- and mouse-derived data. The user can choose to deploy all the features within our intuitive transcriptional analysis pipeline for comprehensive work-up, or in other instances the user might decide to utilise only a selection of the available options within MouSR for their bespoke analysis requirements.

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

During its development, decisions were made to broaden the range of analyses that MouSR could offer, which in turn leads to a number of limitations that we acknowledge. The MouSR tool only controls data from the point of input into the application, leaving the user responsible for the source of the data and the pre-processing steps performed. We have included the DESeq2 tool as an initial step in our pipeline, to ensure that it aligns with the guidelines for running GSEA; however, similar to all biological findings, the validity of any downstream analysis remains entirely dependent on the quality of the input data. In addition, multiple analysis methods in MouSR have been created using different libraries under a different version of R; finding the best R version that can suit them

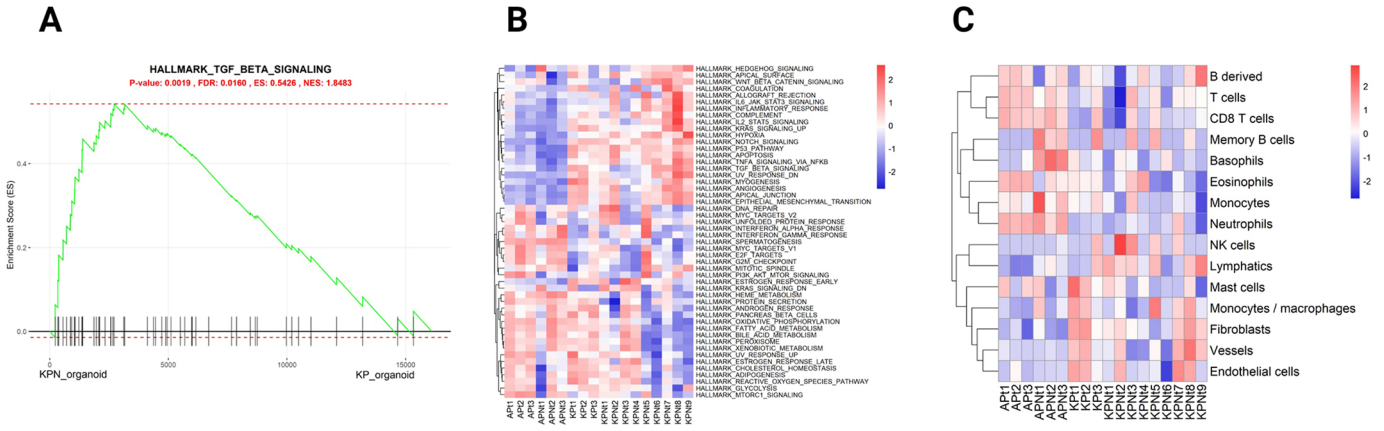

Fig. 5. Gene set enrichment analysis (GSEA) and MCP analysis. (A) Enrichment plot for TGF_BETA_SIGNALLING Hallmark gene set for KPN versus KP organoid groups, with *P*-value, FDR value, enrichment score (ES) and normalised enrichment score (NES). The *x*-axis is all the genes in the data experiment pre-ranked by the metric, where each black bar is the gene in this gene set (pathway); the *y*-axis details the level of enrichment via an ES. (B) Single-sample GSEA for individual samples displayed in a heatmap. (C) Murine microenvironment cell population (mMCP) analysis with infiltrating cell population estimates visualised in a heatmap.

all and at the same time accommodate the shiny server version and CentOS server can be challenging. For instance, there is a recently published library called 'Interactive Complex Heatmap' (Zuguang Gu and Hübschmann, 2022), which provides an easy-to-use tool for constructing highly customisable heatmaps, especially for analysing DESeq2 results. However, based on the version of R and shiny that we are using, we preferred to deploy the 'heatmaply' package. It is worth noting that although this single-purpose tool can provide highly customisable heatmaps, it does not have the same breadth of capabilities or versatility in comparison with MouSR.

Furthermore, MouSR (as a free, open-source tool) can provide more analysis methods than other existing free analytical apps currently available. For instance, DEApp (Li and Andrade, 2017) is largely focused on differential expression analysis of count-based next-generation sequencing data. In addition, TCC-GUI (Wei Su et al., 2019) uses differential expression pipelines with robust normalisation and simulation data generation under various conditions; however, it does not include GSEA and MCP/mMCP-counter analysis. The START tool (Nelson et al., 2017), while having a number of specific functionalities, again does not include GSEA and MCP/mMCP-counter analysis. Finally, the GENAVi application (Reyes et al., 2019) can provide certain analyses; however, it does not include MCP/mMCP-counter analysis, multi-group comparison or dual functionality for both human- and mouse-derived data when compared to MouSR. The MouSR architecture design provides a structure that offers, in the future, the possibility to implement new types of bespoke analysis pipelines and graphical outputs with precise functionalities within the open-source R programming language, facilitating access to thousands of statistical packages that are continually released and updated globally.

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

MATERIALS AND METHODS

Access and requirements

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

File formats

There are numerous pipelines that exist for processing microarray and RNA-seq data, with no 'standard' method being capable of universally transcending all experimental conditions. A wealth of literature exists on pipelines for pre-processing of microarray and RNA-seq data (Olson, 2006) (Conesa et al., 2016), and information on how to access, search and download these data types from the NCBI gene expression omnibus is detailed extensively at https://www.ncbi.nlm.nih.gov/geo/.

Importing text files (.csv or .txt)

The app accepts two commonly used text file formats. For '.csv' comma-separated values (Comma delimited), each line corresponds to a row and all the fields in each line are separated by commas, whereas in '.txt' tab-separated values (Tab delimited), all the fields in each line are typically separated by tabs. The uploaded files size is limited to 30 MB; however, this limitation is only on the online version, and not if users install their own local instance, owing to our server bandwidth limitation. Please note, our exemplar file with 24,751 individual genes across 28 samples is only 2.80 MB. The exemplar file (ArrayExpress E-MTAB-6363; https://github.com/Dunne-Group/MouSR/tree/main/Data; Jackstadt et al., 2019) and accepted format is described in more detail in the 'Data input' subsection of the Results.

Transposing data tool for import

The transcriptional data matrix file to be uploaded in the app has a defined format to follow; however, for users for whom their file is not in the same orientation as the suggested format in the app, a link to the online transposing tool has been embedded in the Introduction section of the app and can be accessed via https://www.convertcsv.com/transpose-csv.htm.

Tools embedded within the MouSR application

PCA

The PCA analysis is performed by the prcomp function in the R stats package (Jolliffe and Cadima, 2016). PCA is defined by a transformation of a high-dimensional vector space into a low-dimensional space. It uses linear combinations of the original data to define a new set of variables that are referred to as principal components.

MDS

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

DGEA (DESeq2)

DGEA is performed based on the negative binomial distribution using the DESeq2 R package (version 1.24.0) (https://bioconductor.org/packages/release/bioc/html/DESeq2.html; Love et al., 2014). DESeq2 is a count-based statistical method that performs an internal normalisation where estimated variance-mean is calculated for each gene across all samples. DESeq2 also estimates the gene-wise dispersion and logarithmic fold changes; a dispersion value is estimated for each gene through a model fit procedure, and differential expression is tested, based on a model using the negative binomial generalised linear distribution (https://bioconductor.org/packages/release/bioc/html/DESeq2.html; Love et al., 2014). We used the DESeq2 package to normalise the data and identify genes that are differentially expressed between the two main groups selected by the user.

ssGSEA

ssGSEA was performed using the GSVA package version 1.32.0. The R package msigdbr version 7.1.1 was also used to retrieve mouse/human Hallmark and biological processes (GO_BP) gene sets and applied to the samples (Hänzelmann et al., 2013).

GSEA

This method consists of three steps (Subramanian, 2005). First, ES is calculated, reflecting the degree to which a set of genes is over-represented at the top or bottom of the entire ranked list. Second, the statistical significance of the ES is estimated by using an empirical phenotype-based permutation test procedure that preserves the complex correlation structure of the gene expression data. Finally, after an entire database of gene sets is evaluated, the

estimated significance level is adjusted to account for multiple hypothesis testing by first calculating the normalised ES (NES), based on dividing the actual ES by the mean of ESs against all permutations of the dataset, then calculating the false discovery rate (FDR) corresponding to each NES. In this study, GSEA was performed on log expression ratio using fgsea, an R package that is a fast implementation of pre-ranked GSEA (see below).

MCP

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

Packages used

Shiny (https://cran.r-project.org/web/packages/shiny/index.html; https://cran.r-project.org/web/packages/shinydashboard/index.html), shinythemes (https://cran.r-project.org/web/packages/shinythemes/index.html), shinydashboard (https://cran.r-project.org/web/packages/shiny/index.html; https://cran.r-project.org/web/packages/shinydashboard/index.html), shinycustomloader (https://cran.r-project.org/web/packages/shinycustomloader/index.html), shinycssloaders (https://cran.r-project.org/web/packages/shinycssloaders/index.html), shinyalert (https://cran.r-project.org/web/packages/shinyalert/index.html), ggplot2 (https://cran.r-project.org/web/packages/ggplot2/index.html), tibble (https://cran.r-project.org/web/packages/tibble/index.html), DESeq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html; Love et al., 2014), limma (https://bioconductor.org/packages/release/bioc/html/limma.html), plyr (https://cran.r-project.org/web/packages/plyr/index.html), biomaRt (https://bioconductor.org/packages/release/bioc/html/biomaRt.html), heatmaply (https://cran.r-project.org/web/packages/heatmaply/index.html), reshape (https://cran.r-project.org/web/packages/reshape/index.html), plotly (https://plotly-r.com), WGCNA (https://cran.r-project.org/web/packages/WGCNA/index.html), lattice (https://cran.r-project.org/web/packages/lattice/index.html), pheatmap (https://cran.r-project.org/web/packages/pheatmap/index.html), RColorBrewer (https://cran.r-project.org/web/packages/RColorBrewer/index.html), GSVA (Hänzelmann et al., 2013), rlist (https://cran.r-project.org/web/packages/rlist/index.html), msigdbr (https://cran.r-project.org/web/packages/msigdbr/index.html), tidyverse (https://cran.r-project.org/web/packages/tidyverse/index.html), mMCPcounter (Petitprez et al., 2020), MCPcounter (Becht et al., 2016), magrittr (https://cran.r-project.org/web/packages/magrittr/index.html), dplyr (https://cran.r-project.org/web/packages/dplyr/index.html), ggrepel (https://cran.r-project.org/web/packages/ggrepel/index.html), readxl (https://cran.r-project.org/web/packages/readxl/index.html), DT (https://cran.r-project.org/web/packages/DT/index.html), colourpicker (https://cran.r-project.org/web/packages/colourpicker/index.html), fgsea (http://bioconductor.org/packages/release/bioc/html/fgsea.html) and enrichplot (https://bioconductor.org/packages/release/bioc/html/enrichplot.html) packages were used in this study.

File outputs and modifiable formats

The customised options for plots that are common in all app sections are turning labels on/off, justifying height and width of the plot, and having different downloading format (png/svg). Additionally, some analytical panels have extra-customised options in their filtering criteria, in order to provide an easier-to-use environment for the users, such as changing the colours, or sizes of labels and points, assigning output filenames, adding legends and changing scales.

Acknowledgements

This project has been developed as a part of the Cancer Research UK-funded ACRCelerate CRC pre-clinical accelerator program. The ACRCelerate project brings together a European-wide consortium of basic and clinical scientists at the forefront of CRC research to interrogate a suite of state-of-the art preclinical models. The overarching aim of ACRCelerate is to generate robust and reproducible preclinical data to de-risk future clinical trials via patient stratification. We acknowledge 'VolcaNoseR' for their Volcano plot and 'DEApp' for their first-page template inspiration. We thank the collaborative network across Belfast, Oxford and Glasgow for feedback and discussions during development. We also thank the IT department at Queen's University Belfast and Crispin Miller at the Beatson Institute Glasgow for advice and support.

Competing interests

M.L. has received honoraria from Pfizer, EMF Serono and Roche for presentations unrelated to this work, and received an unrestricted educational grant from Pfizer for research unrelated to this work. O.J.S. has received funding from Novartis, Astra Zeneca, Cancer Research Technology and Redex for research unrelated to this work.

Author contributions

Conceptualization: B.A., M.L., P.D.D.; Methodology: B.A.; Software: B.A., R.A., G.Q., S.M.; Formal analysis: B.A., R.A.; Resources: J.J., T.R.M.L., K.G., S.B.M., K.L.R., T.M., S.L., A.S.C., O.J.S., P.D.D.; Writing - original draft: B.A., M.L., P.D.D.; Writing - review & editing: B.A., R.A., J.J., T.R.M.L., K.G., S.B.M., K.L.R., G.Q., S.M., T.M., S.L., A.S.C., O.J.S., M.L., P.D.D.; Supervision: M.L., P.D.D.; Project administration: P.D.D.; Funding acquisition: O.J.S., M.L., P.D.D.

Funding

This research was supported by an International Accelerator Award, ACRCelerate, jointly funded by Cancer Research UK (A26825 and A28223), Fundación Científica Asociación Española Contra el Cáncer (GEACC18004TAB) and Associazione Italiana per la Ricerca sul Cancro (22795), a Cancer Research UK early detection project grant (A29834) and a Health Data Research UK Substantive Site grant. The manuscript reflects the views of the authors and may not reflect the opinions or views of the funders.

Data availability

The complete source code for the MouSR app can be launched on any system that has R/RStudio installed and is available to download at https://github.com/Dunne-Group/MouSR. In addition, we are open to mutually beneficial collaborations to further develop the MouSR analytic options. Transcriptional data used in this manuscript were downloaded from ArrayExpress under study ID E-MTAB-6363. The four genotypes included APC-deficient models (APN or AP) and KRAS models (KPN or KP).

References

Becht, E., Giraldo, N., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W. H. et al. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218. doi:10.1186/s13059-016-1070-5

Cerami, E., Gao, J., Dogrusoz, U., Gross, E. B., Sumer, O. S., Aksoy, B. A., Jacobsen, A., Byrne, J. C., Heuer, M. L., Larsson, E. et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2**, 401-404. doi:10.1158/2159-8290.CD-12-0095

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X. et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**. doi:10.1186/s13059-016-0881-8

Edgar, R., Domrachev, M. and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207-210. doi:10.1093/nar/30.1.207

Gambardell, V. (2020). Personalized medicine: recent progress in cancer therapy. *Cancers* **12**, 1009. doi:10.3390/cancers12041009

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E. K. et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1. doi:10.1126/scisignal.2004088

Hänzelmann, S., Castelo, R. and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**. doi:10.1186/1471-2105-14-7

Jackstadt, R., van Hooff, S. R., Leach, J. D., Cortes-Lavaud, X., Lohuis, J. O., Ridgway, R. A., Wouters, V. M., Roper, J., Kendall, T. J., Sansom, O. J. et al. (2019). Epithelial NOTCH signaling rewires the tumor microenvironment of colorectal cancer to drive poor-prognosis subtypes and metastasis. *Cancer Cell* **36**, 319-336.e7. doi:10.1016/j.ccell.2019.08.003

Jolliffe, T. I. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society* **374**. doi:10.1098/rsta.2015.0202

Lai, A. G., Pasea, L., Banerjee, A., Hall, G., Denaxas, S., Chang, W. H., Katsoulis, M., Williams, B., Pillay, D., Noursadeghi, M. et al. (2020). Estimated impact of the COVID-19 pandemic on cancer services and excess 1-year mortality in people with cancer and multimorbidity: near real-time data on cancer care, cancer deaths and a population-based cohort study. *BMJ Open* **10**, doi:10.1136/bmjopen-2020-043828

Li, Y. and Andrade, J. (2017). DEApp: an interactive web interface for differential expression analysis of next generation sequence data. *Source Code for Biology and Medicine* **12**, 2. doi:10.1186/s13029-017-0063-4
Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. doi:10.1186/s13059-014-0550-8
Luijsterburg, M. and Goedhart, J. (2020). VolcaNoseR – a web app for creating, exploring, labeling and sharing volcano plots. *Sci. Rep.* **10**, 20560. doi:10.1038/s41598-020-76603-3
Mootha, V. K., Lindgren, C. M., Eriksson, K. L., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E. et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267-273. doi:10.1038/ng1180
Nelson, W. J., Sklenar, J., Barnes, A. P. and Minnier, J. (2017). The START App: a web-based RNAseq analysis and visualization resource. *Bioinformatics* **33**, 447-449. doi:10.1093/bioinformatics/btw624
Olson, N. E. (2006). The microarray data analysis process: from raw data to biological significance. *NeuroRX3* **3**, 373-383. doi:10.1016/j.nurx.2006.05.005
Petitprez, F., Levy, S., Sun, C. M., Meylan, M., Linhard, C., Becht, E., Elarouci, N., Tavel, D., Roumenina, L. T., Ayadi, M. et al. (2020). The murine Microenvironment Cell Population counter method to estimate abundance of tissue-infiltrating immune and stromal cell populations in murine samples using gene expression. *Genome Medicine* **12**, doi:10.1186/s13073-020-00783-w
Reyes, A. L. P., Silva, T. C., Coetzee, S. G., Plummer, J. T., Davis, B. D., Chen, S., Hazelett, D. J., Lawrenson, K., Berman, B. P., Simon, A. et al. (2019). GENAVi: a shiny web application for gene expression normalization, analysis and visualization. *BMC Genomics* **20**, doi:10.1186/s12864-019-6073-7.
Sohn, B. H., Hwang, J. E., Jang, H. J., Lee, J. S., Oh, S. H., Shim, J. J., Lee, K. W., Kim, E. H., Yim, S. Y., Lee, S. H. et al. (2017). Clinical significance of four molecular subtypes of gastric cancer identified by the cancer genome Atlas project. *Clin Cancer Research* **23**, 4441-4449. doi:10.1158/1078-0432.CCR-16-2211.
Subramanian, A. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545-15550. doi:10.1073/pnas.0506580102
Wei Su, W., Sun, J., Shimizu, K. and Kadota, K. (2019). TCC-GUI: a Shiny-based application for differential expression analysis of RNA-Seq count data. *BMC Research Notes* **12**, doi:10.1186/s13104-019-4179-2
Young, F. W. (2013). *Multidimensional Scaling: History, Theory, and Applications*. Psychology Press.
Zuguang Gu, Z. and Hübschmann, D. (2022). Make interactive complex heatmaps in R. *Bioinformatics* **38**, 1460-1462. doi:10.1093/bioinformatics/btab806