British Journal of Cancer www.nature.com/bjc

ARTICLE OPEN

Check for updates

Genetics and Genomics

MmCMS: mouse models' consensus molecular subtypes of colorectal cancer

Raheleh Amirkhah[1], Kathryn Gilroy[2], Sudhir B. Malla[1], Tamsin R. M. Lannagan[2], Ryan M. Byrne[1], Natalie C. Fisher[1], Shania M. Corry[1], Noha-Ehssan Mohamed[2], Hojjat Naderi-Meshkin[3], Megan L. Mills[2], Andrew D. Campbell[2], Rachel A. Ridgway[2], Baharak Ahmaderaghi[4], Richard Murray[1], Antoni Berenguer Llergo[5], Rebeca Sanz-Pamplona[6], Alberto Villanueva[7], Eduard Batlle[5,8,9], Ramon Salazar[10], Mark Lawler[1], Owen J. Sansom[2,11] and Philip D. Dunne[1,2]✉

BACKGROUND: Colorectal cancer (CRC) primary tumours are molecularly classified into four consensus molecular subtypes (CMS1–4). Genetically engineered mouse models aim to faithfully mimic the complexity of human cancers and, when appropriately aligned, represent ideal pre-clinical systems to test new drug treatments. Despite its importance, dual-species classification has been limited by the lack of a reliable approach. Here we utilise, develop and test a set of options for human-to-mouse CMS classifications of CRC tissue.

METHODS: Using transcriptional data from established collections of CRC tumours, including human (TCGA cohort; $n = 577$) and mouse ($n = 57$ across $n = 8$ genotypes) tumours with combinations of random forest and nearest template prediction algorithms, alongside gene ontology collections, we comprehensively assess the performance of a suite of new dual-species classifiers.

RESULTS: We developed three approaches: MmCMS-A; a gene-level classifier, MmCMS-B; an ontology-level approach and MmCMS-C; a combined pathway system encompassing multiple biological and histological signalling cascades. Although all options could identify tumours associated with stromal-rich CMS4-like biology, MmCMS-A was unable to accurately classify the biology underpinning epithelial-like subtypes (CMS2/3) in mouse tumours.

CONCLUSIONS: When applying human-based transcriptional classifiers to mouse tumour data, a pathway-level classifier, rather than an individual gene-level system, is optimal. Our R package enables researchers to select suitable mouse models of human CRC subtype for their experimental testing.

British Journal of Cancer (2023) 128:1333–1343; https://doi.org/10.1038/s41416-023-02157-6

INTRODUCTION

Colorectal cancer (CRC) primary tumours can be molecularly classified into four consensus molecular subtypes (CMS1–4) [1]. According to this classification, CMS1 (14% of patients) is enriched for tumours with microsatellite instability (MSI) and immune activation. CMS2 (37% of patients) epithelial-rich tumours represent the canonical subtype and are associated with activation of the WNT/MYC pathways and chromosome instability. CMS3 (13% of patients) tumours display signalling indicative of increased metabolic activity and *KRAS* mutations. Finally, CMS4 (23% of patients) tumours display stromal-rich and mesenchymal features, alongside activation of TGF-β and VEGFR pathways [1]. While CMS classification provides valuable prognostic information, its ability to identify subtype-specific responses to therapies remains an area of active research, with several reverse-translation studies using human pre-clinical models, such as cell lines, organoids and patient-derived xenografts (PDX) [2, 3]. While CMS classification in these models is possible, the reliance of CMS classification on gene expression signals from tumour microenvironment (TME) compartments can undermine attempts to identify the mesenchymal subtype of CRC (CMS4) in cell lines, patient-derived organoids and PDXs [4, 5]. To address this, Eide et al. developed a CMS classifier specifically designed for human pre-clinical models, named CMScaller, which used a filtered set of

[1]The Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, UK. [2]Cancer Research UK Beatson Institute, Glasgow, UK. [3]Wellcome-Wolfson Institute for Experimental Medicine, Queen's University Belfast, Belfast, UK. [4]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, UK. [5]Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology, Barcelona, Spain. [6]Unit of Biomarkers and Susceptibility, Oncology Data Analytics Program (ODAP), Catalan Institute of Oncology (ICO), Oncobell Program, Bellvitge Biomedical Research Institute (IDIBELL) and CIBERESP, L'Hospitalet de Llobregat, Barcelona, Spain. [7]Chemoresistance and Predictive Factors Group, Program Against Cancer Therapeutic Resistance (ProCURE), Catalan Institute of Oncology (ICO), Oncobell Program, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet del Llobregat, Barcelona, Spain. [8]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Barcelona, Spain. [9]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [10]Department of Medical Oncology, Catalan Institute of Oncology (ICO), Oncobell Program, Bellvitge Biomedical Research Institute (IDIBELL), CIBERONC and Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain. [11]School of Cancer Sciences, University of Glasgow, Glasgow, UK. ✉email: p.dunne@qub.ac.uk

Received: 3 June 2022 Revised: 10 January 2023 Accepted: 12 January 2023
Published online: 30 January 2023

cancer cell-intrinsic, subtype-enriched gene expression markers, giving a surrogate measurement of alignment with CMS subtypes in in vitro and in vivo models [6].

Although translation of human CMS subtypes to human-based pre-clinical models has been addressed, there remains a need to develop and test a classifier that can be used with mouse-based tumour data from genetically engineered mouse models (GEMMs). GEMMs, alongside the armament of human pre-clinical models, represent the most appropriate models to mimic the complexity of human CRC biology. GEMMs in particular provide an ideal system to improve pre-clinical drug testing within a native immunocompetent host [7, 8]. Identifying murine models that recapitulate each CRC subtype feature can therefore de-risk clinical translation of therapeutics, while also providing an excellent opportunity to improve our understanding of the nuanced and complex interactions between cancer epithelial cells and their microenvironment. Currently, there is no reliable and standardised approach for CMS classification using data from mouse tissues. In the absence of such a system, users have relied on converting the human CMS template to mouse orthologues, followed by sample classification using the nearest template prediction (NTP) method (as with the CMScaller), or conversely converting mouse genes to human orthologues and applying the random forest (RF) method used in the original CMSclassifier algorithm [1, 6]. Both approaches rely on overlapping nomenclature for individual genes; as mouse genes with different names to the ones in the human classifier template will be ignored/removed during CMS assignment, or vice versa. In addition, both systems are also fully reliant on the assumption that genes within the classifier will perform the same biological function in both mouse and human tumours and ignore interspecies variability. Recent studies have shown that pathway-based classifications are more robust as they are composed of tens to hundreds of co-ordinately expressed genes, and therefore are protected to some degree from the loss of individual genes or variations in functions, both of which are known to undermine gene-level classifiers [9, 10]. As such, pathway-based approaches consider the collective impact of genes on pathway-level activity rather than being influenced by a single differentially expressed gene. Furthermore, broad biological knowledge-based approaches have previously been shown to be less influenced by non-biological factors [11, 12].

To improve on the current state-of-the-art approach of classifying GEMM tumours, we developed three options for CMS classification in mouse tissue. The first, hereafter named as MmCMS-A, uses mouse orthologues of the human CMS gene template from CMScaller [6]; thus, it has a sole emphasis on individual genes. Given the benefits of pathway-level approaches for classification, over gene level, we proposed two further options (MmCMS-B and MmCMS-C) that use biological knowledge-based information from either gene ontology (GO) (MmCMS-B) or a compendium of signatures from biological signalling collections and microenvironment populations (MmCMS-C). Most importantly, to ensure the field can utilise these mouse CMS classification approaches, we developed an R package, namely MmCMS, which provides a publicly available tool to classify samples according to all three options, enabling users to assess the alignment of GEMM tumours to human CMS subtype.

METHODS

Human CRC cohort

The processed TCGA COREAD RNA-Seq dataset ($n = 577$) was downloaded directly from the Guinney et al., *CMSclassifier* study via Synapse (ID: syn2023932), where it has been described previously [1]. Gene symbols and Entrez IDs were matched using *org.Hs.eg.db* R package (v3.8.2) thereafter CMS classification was performed via RF method using *CMSclassifier* R package (version 1.0.0).

Biological process (BP) subset of GO gene sets was extracted from the Molecular Signature Database (MSigDB) using *msigdbr* R package (v7.0.1). Subsequently, ontology scores were generated for the TCGA dataset using a single sample gene set enrichment analysis (ssGSEA) method from *GSVA* R package (v1.26.0). To determine the CMS-specific GO terms, these ssGSEA scores were averaged for each gene set across samples within each CMS subtype and scaled to *Z* scores where the GO with ssGSEA scaled scores above 0 in a CMS, but below 0 on the others, were selected as the enriched GO term for that CMS. The CMS-specific GO BP gene sets for mouse species were then extracted from the msigdbr R package and used to develop an ontology-based CMS classification for mice.

Mouse models

All animal experiments were performed in accordance with a UK Home Office licence (Project License 70/8646), and were subject to review by the animal welfare and ethical review board of the University of Glasgow. Mice of both sexes were induced with a single injection of 2 mg tamoxifen (Sigma-Aldrich, T5648) by intraperitoneal injection at an age of 6–12 weeks, all experiments were performed on a C57BL/6 background. Mice were sampled at clinical endpoint, which was defined as weight loss and/or hunching and/or cachexia.

Mouse RNA sequencing and analysis

RNA was isolated using either an RNeasy mini kit (Qiagen) or TRIzol reagent (Thermo Fisher Scientific). RNA concentrations were determined using a NanoDrop 200c spectrophotometer (ThermoScientific), and quality was assessed using an Agilent 220 Tapestation using RNA screentape. RNA sequencing was performed using an Illumina TruSeq RNA sample prep kit, then run on an Illumina NextSeq using the High Output 75 cycles kit (2×36 cycles, paired-end reads, single index). Raw sequence quality was assessed using the FastQC algorithm version 0.11.8. Sequences were trimmed to remove adaptor sequences and low-quality base calls, defined as those with a Phred score of <20, using the Trim Galore tool version 0.6.4. The trimmed sequences were aligned to the mouse genome build GRCm38.98 using HISAT2 version 2.1.0, then raw counts per gene were determined using FeatureCounts version 1.6.4. Raw read counts of the small cohort ($n = 18$) which is publicly available at ArrayExpress: E-MTAB-6363 were normalised using *vst* function in *DESeq2* R package (v1.32.0). The models where the batch they were sequenced in was deeply confounded by genotype were removed and data from 51 GEMMs remained. *ComBat_seq* function in *sva* R package (v3.40.0) was used to correct read counts for batch, thereafter, *vst* function in *DESeq2* same as before was used to normalise the data.

Databases

CMS-curated gene sets signatures ($n = 79$) were obtained from Synapse (ID: syn2321865). Cancer hallmarks ($n = 50$) and GO BP gene set (C5 BP) was extracted from MSigDB using '*msigdbr*' R package (v7.4.1).

Ten signatures to estimate the proportion of the eight immune (NK cells, Cytotoxic lymphocyte, T cells, CD8 T cells, B lineage, Monocytic lineage, Neutrophils, Myeloid dendritic cells) and two stromal (Fibroblasts and Endothelial) cell populations in each human sample across CMS subtypes were obtained using the *MCPcounter* R package (v1.2.0); the mouse version of signatures was retrieved from the *mMCPcounter* R package (v0.1.0). Immune-related genes for humans and mice were downloaded from the NanoString panel (https://canopybiosciences.com/product/immunology/).

Statistical analysis

All the statistical analyses were performed in R (v4.1.2) using the *stats* R package, including *cor()* function with method = 'pearson' for Pearson's correlation. The Student *t*-test method embedded in the *geom_signif()* function of *ggsignif* package (v0.6.3) was used to do statistical analysis in violin plots. Boxplots were generated using *ggplot2* (v3.3.5) R package. The *ComplexHeatmap* (v2.8.0) and *circlize* (v0.4.13) packages were used to display heatmaps. We used *glmnet* (v4.1-3) R package to do Least Absolute Shrinkage Selector Operator (LASSO) regression model analysis. The λ or tuning parameter in the LASSO model was selected through the 10-fold cross-validation.

ssGSEA was performed using an R package called *GSVA* (v1.40.1). Alluvial plot to display concordance result was drawn using *riverplot* (v0.10).

The NTP algorithm, with cosine correlation distances, was employed to predict the proximity of each GEM model's expression profile to the four CMS subtypes, using each of the three templates individually (A, B and C),

with an FDR < 0.05 used as a cutoff for statistical significance. To do unsupervised class discovery in the combined mouse cohort ($n = 51$), the gene expression profile converted to GO scores (*msigdbr*, v7.5.1 and *GSVA* R package, v1.44.2) and variables were scaled before applying unsupervised *k*-means clustering. The elbow method was used to identify optimal number of clusters (*factorextra* R package; v1.0.7). *ESTIMATE* R package (v1.0.13) was applied to evaluate the presence of immune and stromal content in each individual mouse tumour sample, following conversion of the mouse gene expression matrix to human orthologs using *biomaRt* package (v2.53.2). Pairwise GSEA analysis was performed using the *fgsea* (v1.10.1) R package on shrunken log2FoldChange of each CMS subtype versus other subtypes for Hallmark gene sets. For human data, differential expression analysis of each CMS group versus others was performed using limma R package (v3.52.2) and GSEA was applied on logFC. Schematics were created using BioRender.

RESULTS

Development and testing of CMS classifier templates for use in mouse tumours

To confirm the concordance between the CMScaller/pre-clinical CMS classifier (NTP method) and the CMSclassifier/original CMS classifier (RF method) in human data, we applied CMScaller on COREAD TCGA RNA-seq transcriptional data ($n = 577$) retrieved from the original CMS article [1]. After removing samples that were unclassified by either RF or CMScaller, we found 91.19% (321/352) concordance between RF and NTP calls (Supplementary Fig. 1a). PCA analysis on the whole transcriptome of these 352 samples demonstrated that samples that gave conflicting calls (indicated as swapped in Supplementary Fig. 1a) between RF and CMScaller were in the boundary of CMS subtypes assigned by the RF method (Supplementary Fig. 1a). To confirm that the discrepancies in classification call were confined to samples with lower CMS probability scores, when we set a more stringent CMS classification probability cutoff (>0.8) for the RF method, the classifications for the two methods increased to 100% concordance, $n = 93$ (Supplementary Excel File 1, Sheet 1), demonstrating that CMScaller provides excellent CMS classification concordance for samples that display the strongest CMS transcriptional traits, as indicated by high subtype RF classification scores.

While these data confirm the suitability of using either the RF CMSclassifier or NTP CMScaller methods for CMS classification of human tumour data, to assess the performance of these methods on mouse tumour model classification, we next assembled transcriptional data from two independent GEMM tumour cohorts (Table 1). Tamoxifen-regulated Cre-*loxP* system was used to generate all models and introduced via an intraperitoneal injection. The small cohort has been previously described by Jackstadt et al. and composed of 18 intestinal primary tumours across 4 genotypes that represent both the serrated (KPN: *Kras*$^{G12D/+}$ *Trp53*$^{fl/fl}$ *Notch1*$^{Tg/+}$; KP: *Kras*$^{G12D/+}$ *Trp53*$^{fl/fl}$) and tubular (APN: *Apc*$^{fl/+}$ *Trp53*$^{fl/fl}$ *Notch1*$^{Tg/+}$; AP: *Apc*$^{fl/+}$ *Trp53*$^{fl/fl}$) tumour histologies [13]. The large independent cohort ($n = 39$) contained a set of independent KP and KPN tumours alongside 4 additional genotypes including *Apc*$^{fl/+}$ (A); *Apc*$^{fl/+}$ *Kras*$^{G12/+}$ (AK); *Braf*$^{V600E/+}$ *Trp53*$^{fl/fl}$ (BP) and *Braf*$^{V600E/+}$ *Trp53*$^{fl/fl}$ Notch1$^{Tg/+}$ (BPN). Median latency age of A, AK AP, APN, KP, KPN, BP and BPN models is 215, 67, 185, 161, 171, 184, 190 and 174 days respectively, developing small intestine tumours primarily, with the exception of seven mice (AK = 5, A = 2) which formed tumours in the colon. For more characterisation of the samples see Supplementary Excel File 2.

The RF method in the CMSclassifier package was designed for human samples and uses 273 genes to assign CMS subtypes. To enable the use of this method with mouse data, we converted the entire mouse gene matrix to human orthologues using biomaRt [14]. During the conversion of the mouse matrix, 16 genes of the 273 genes used to predict CMS calls in humans were mismatched in both cohorts (Supplementary Table S1). Applying the RF method to our $n = 18$ and $n = 39$ mouse model matrices produced 56% unknown samples in both datasets (Supplementary Fig. 1b, c). Of note, to test the functionality of CMScaller in the same mouse cohorts, we next converted the human CMScaller template genes ($n = 529$; CMS1 = 126, CMS2 = 82, CMS3 = 84, CMS4 = 237) to mouse orthologues ($n = 533$; CMS1 = 128, CMS2 = 80, CMS3 = 90, CMS4 = 235), which as anticipated resulted in a small number of dropouts ($n = 26$ missing genes, Supplementary Table S2) due to lack of recognised orthologues, though overall the number of genes in mouse CMS template increased due to the existence of multiple mapping mouse genes for the individual human genes (Fig. 1a and Supplementary Excel File 3). Using this CMScaller method in our mouse data, we found fewer unknown samples, 17% and 36%, respectively (Supplementary Fig. 1b, c) and therefore selected this NTP-based approach as our initial dual-species classifier, termed MmCMS-A.

Identification of CMS-related GO BP terms in human TCGA data (MmCMS-B)

To complement the gene-level approach in MmCMS-A, we again utilised the RF classifications used in the original CMS classifier development within the human TCGA COREAD data ($n = 577$) to identify the GO BP that are most significantly associated with individual human CMS classes, using ssGSEA to derive the enrichment score across all samples. For each gene set (GO BP), the mean enrichment score was calculated for each human CMS subtype, with scaled *Z*-score >0 in one CMS subtype but <0 in the other three CMS subtypes being selected as distinct features for each particular CMS subtype. This identified $n = 172$, $n = 64$,

Table 1. Summary of mouse models used in this study.

	Model name	Genotype of mouse model
Small cohort ($n = 18$)	AP	villinCre^{ER} Apc$^{fl/+}$ Trp53$^{fl/fl}$ ($n = 3$)
	APN	villinCre^{ER} Apc$^{fl/+}$ Trp53$^{fl/fl}$ Notch1$^{Tg/+}$ ($n = 3$)
	KP	villinCre^{ER} Kras$^{G12D/+}$ Trp53$^{fl/fl}$ ($n = 3$)
	KPN	villinCre^{ER} Kras$^{G12D/+}$ Trp53$^{fl/fl}$ Notch1$^{Tg/+}$ ($n = 9$)
Large cohort ($n = 39$)	A	villinCre^{ER} Apc$^{fl/+}$ ($n = 6$)
	AK	villinCre^{ER} Apc$^{fl/+}$ Kras$^{G12D/+}$ ($n = 6$)
	BP	villinCre^{ER} Braf$^{V600E/+}$ Trp53$^{fl/fl}$ ($n = 4$)
	BPN	villinCre^{ER} Braf$^{V600E/+}$ Trp53$^{fl/fl}$ Notch1$^{Tg/+}$ ($n = 7$)
	KP	villinCre^{ER} Kras$^{G12D/+}$ Trp53$^{fl/fl}$ ($n = 6$)
	KPN	villinCre^{ER} Kras$^{G12D/+}$ Trp53$^{fl/fl}$ Notch1$^{Tg/+}$ ($n = 10$)

In this study, we investigated the presence of CMS subtypes in two panels of 18 and 39 GEMMs with 4 and 6 different genotypes, respectively.

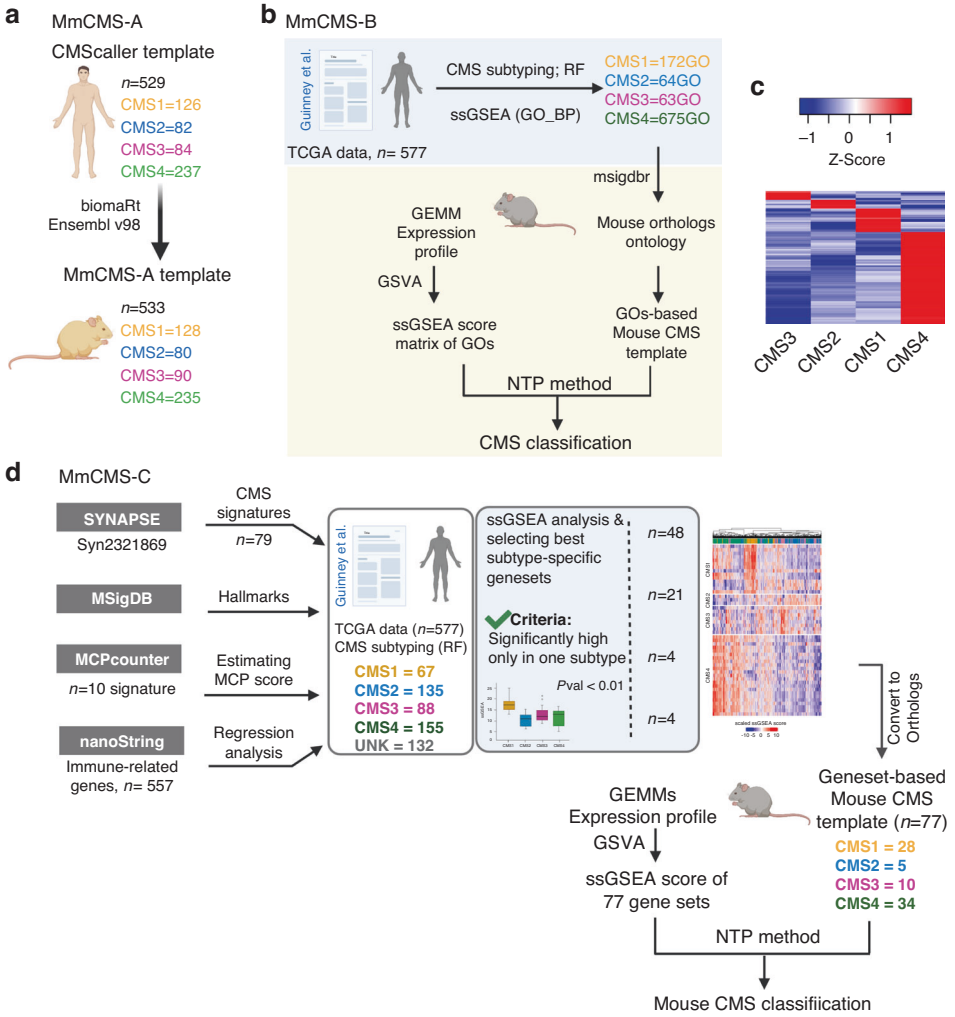

Fig. 1 Three different approaches for mouse CMS subtyping. a The schematic shows the approach of converting the human CMS template to mouse orthologues (MmCMS-A option). **b** Schematic of developing a gene ontology-based classifier to call CMS subtypes in mouse tissues. **c** Heatmap of ssGSEA for selected GO BP terms based on *z*-score > 0 in a CMS subtype and *z*-score < 0 for other subtypes in human data (MmCMS-B option). **d** Schematic of developing MmCMS-C classifier based on four biologically informed signature collections (Fig. 2a heatmap shown as an exemplar).

$n = 63$ and $n = 675$ specific GO terms associated with CMS1, CMS2, CMS3 and CMS4, respectively (Fig. 1b, blue background; Fig. 1c). To test if these CMS class-specific GO terms represent surrogate markers for human samples called with high probability using RF, we again used CMS classifications from the stringent RF threshold (probability cutoff > 0.8) as before and compared them to CMS classifications using this new NTP ontology-based method, where we observe 95% concordance with the RF-based calls (Supplementary Excel File 1, Sheet 2). In line with the generation of MmCMS-A, the mouse-equivalent GO terms of these human gene ontologies were identified using the 'msigdb' R package and used as the MmCMS-B template for CMS classification in mouse data using NTP method (Fig. 1b, yellow background).

Development of mouse CMS template (MmCMS-C) based on combining gene sets/pathways that best characterise each human CMS subtype in a supervised approach

While MmCMS-B is focussed solely on GO BP signatures, for MmCMS-C we generated a classifier based on four biologically informed signature collections (Fig. 1d). First, we compiled the $n = 79$ gene sets used to characterise biological signalling in the original CMS study from the Synapse database (https://doi.org/10.7303/syn2623706). As with MmCMS-B, we refined these 79 signatures into only those with individual CMS class-specific expression (*t*-test; *p* value < 0.01) and signatures only kept if one subtype was significantly higher when compared to each of the other subtypes in turn, resulting in 48 of the 79 gene sets being used (Fig. 2a and Supplementary Fig. 2a). Next, using the 50 MSigDB hallmark gene sets, we identified 21 with significant expression (*t*-test; *p* value < 0.01) across CMS groups (Fig. 2b and Supplementary Fig. 2b). In the third step, we used the microenvironment cell population (MCP)-counter signatures, and in line with previous studies, we found cytotoxic lymphocyte and NK cells are significantly enriched in CMS1, whereas fibroblast and endothelial cells are enriched in CMS4, thus 4 signatures from MCPcounter [15] were included (Supplementary Fig. 2c; *t*-test; *p* value < 0.01). Finally, given the importance of inflammatory lineages in development and classification, we assessed immune-related genes ($n = 557$; from a NanoString panel) for their associations with each CMS subtype, filtered first using the LASSO regression model (Fig. 2c). Based on coefficient >0, overall 44 immune-related genes (CMS1 = 14, CMS2 = 8, CMS3 = 9, CMS4 = 13) were found as the best predictors of individual CMS classes. As with Options B, these co-ordinately expressed immune genes for each CMS subtype were then grouped for ssGSEA, and enrichment scores were assessed across subtypes which were

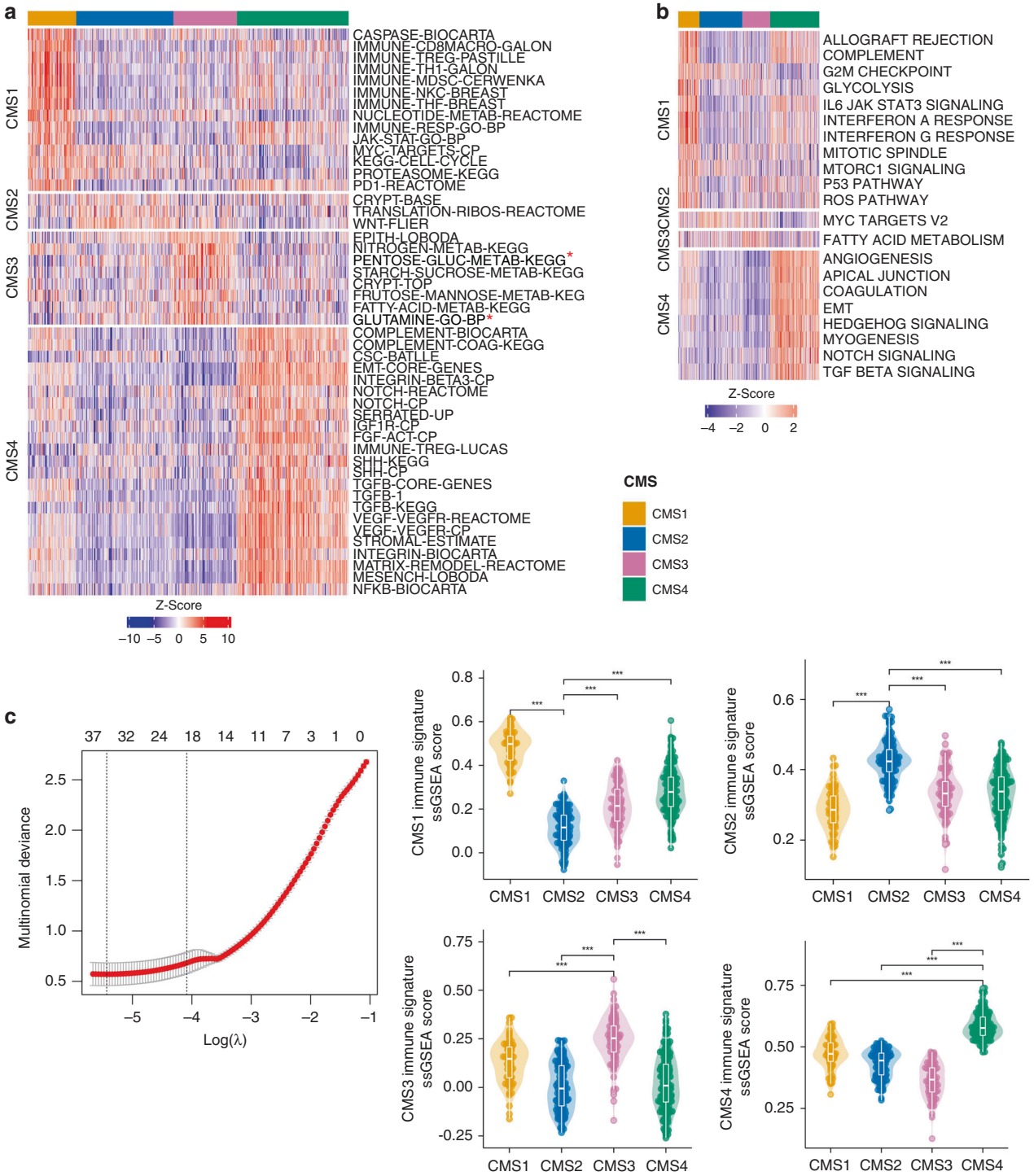

Fig. 2 Identification of pathways and biology that are best characteristic of each human CMS subtype for Option C. a ssGSEA scores heatmap of 48 CMS-related signatures (CMS1 = 14, CMS2 = 3, CMS3 = 8, CMS4 = 23), from original the CMS article [1], that are significantly different (pairwise *t*-test; see Supplementary Fig. 2b) across CMS subtypes in human dataset. Scores are converted to *Z* scores. *PENTOSE_GLUC_METAB_KEGG and GLUTAMINE_GO_BP are defined with different names in the CMS curated gene sets signatures ($n = 79$) from Synapse (ID: syn2321865) but the genes are the same. **b** ssGSEA scores heatmap of 21 selected hallmark pathways that are significantly different (pairwise *t*-test; see Supplementary Fig. 2b) across CMS subtypes in the human dataset. Scores are converted to *Z* scores. **c** Selection of the λ parameter in the LASSO model by 10-fold cross-validation based on minimum criteria. Red dots show the average deviance values for each model with a given λ. The vertical black lines define the optimal values of λ, where the model provides the best fit to the data. A λ value of 0.01682636 (lambda.1se), was chosen. Violin plot and pairwise *t*-test used to display enrichment of selected immune gene set across CMS subtypes (CMS1 $n = 14$, CMS2 = 8, CMS3 = 9, CMS4 = 13). ***$p \le 0.001$.

significantly enriched (*t*-test; *p* value <0.01) (Fig. 2c). Overall, this four-step MmCMS-C approach identified 77 CMS class-specific gene sets (CMS1 = 28; CMS2 = 5; CMS3 = 10; CMS4 = 34). When tested in the same way as MmCMS-A and B, using the NTP method on TCGA data, MmCMS-C was found to have 98% concordance with the RF-based high probability calls, threshold = 0.8 (Supplementary Excel File 1, Sheet 3).

To enable mouse classification, the biomaRt [14] and msigdb [16] R packages were used to obtain the mouse version of 48 gene sets and 21 hallmark pathways, respectively, with mouse MCP signatures retrieved from the mouse-specific mMCP-counter package [17]. Individual orthologues of immune-related genes were obtained from the mouse NanoString panel (Supplementary Table S3), with 39 mouse genes aligned to the 44 human immune genes identified using regression analysis, and then grouped into signature scores as before.

MmCMS classification of mouse tumours in previously characterised cohort

To assess the performance of our three options for classifying mouse tumours, two different cohorts of GEMMs as described above were used (Table 1 and Fig. 3a). As there is no CMS 'ground truth' or reference for mouse tumour data, we utilised tumours from $n = 18$ mouse models across four genotypes (KPN, KP, APN, AP; Table 1), which we have previously shown to correlate with signalling associated with stromal CMS4 tumours (KPN and KP) or epithelial-rich CMS2/3 tumours (AP and APN). PCA on the dataset revealed distinct groups according to genotype (Supplementary Fig. 3a). The NTP-based algorithm was employed to predict CMS classification of GEMM tumours, using each of the three templates individually (MmCMS-A, B and C), with an FDR < 0.05 used as a cutoff for significant calls. Within the small cohort, both MmCMS-A and MmCMS-C returned three unknown calls; however, MmCMS-B classified all mouse tumours (Fig. 3b). Comparing these unknown classification rates to those returned by the existing RF method used for human tumour classification, which returned 10 unknown classification calls, re-emphasised the value of using a pathway/GO-based approaches (MmCMS-B, -C) compared to the individual gene-based methods (MmCMS-A and RF). Overall, these CMS grouping using our new R-based MmCMS classifier were all in line with previously published subtype associations for these models (Fig. 3b, c). There was broad consensus across all three options for samples classified as CMS4, indicating how distinct this subtype is compared to the others; however, samples classified as CMS3 using MmCMS-C were classified as either CMS2 or unclassified using MmCMS-B and MmCMS-A (Fig. 3c). Characterisation of these GEMM tumours using ssGSEA shows that, as with human tumours, all samples assigned as CMS4 display high levels of enrichment for TGFβ signalling, EMT, angiogenesis, Notch and Hedgehog signalling. In line with human CMS biology, samples classified as CMS3 using MmCMS-C display high expression of metabolic pathways, such as bile acid, xenobiotic, fatty acid, heme metabolism and glycolysis. Samples classified as CMS2 have high expression of MYC and E2F targets which are well-identified signalling molecules in the CMS2 subtype. One sample was consistently classified as CMS1, which displayed high expression of interferon-gamma response and interferon-alpha response (Fig. 3c and Supplementary Fig. 3b).

Validation of MmCMS in extended GEMM tumour cohort

Following assessment in this initial cohort of histologically distinct tumours, with tubular/epithelial-rich (AP, APN) and serrated/stroma-rich (KP, KPN) genotypes, we next tested each of the individual classifier options in an independent and more heterogeneous cohort of 39 mouse tumours across 6 genotypes. The NTP method, the same as above, was used to predict CMS classification that measures the proximity of each sample with the template by calculating distance d and assigning the CMS label that it is closest to in the template (Supplementary Fig. 3b).

As can be seen in Fig. 3d, MmCMS-A, B and C returned $n = 14$, $n = 2$ and $n = 6$ unknown samples respectively. MmCMS-A returned unknown calls for all $Apc^{fl/+}$ samples; however, when using MmCMS-B and C, all samples with $Apc^{fl/+}$ genotype were assigned as CMS2, with biological characterisation using GSEA indicating that these samples have enriched signalling hallmarks related to proliferation including G2M checkpoint, E2F targets and MYC targets (Supplementary Fig. 3c). The only $Apc^{fl/+}$ genotype sample (A2) that remained unclassified by MmCMS-C appeared as an outlier when assessed by PCA, as it did not cluster with other $Apc^{fl/+}$ samples (Supplementary Fig. 3d). GSEA reveals that samples classified as CMS2 by MmCMS-A display features inconsistent with human CMS2 tumours, and are more aligned with human CMS3 classification, including high expression of metabolic pathway and low expression of proliferation-related hallmarks, indicating limited ability of the gene-level approach to identifying CMS2 tumours (Fig. 3d and Supplementary Fig. 3c). Furthermore, genes associated with immune response, colon epithelial differentiation, goblet cell-like, and stroma, which represent CMS1, CMS2, CMS3, and CMS4, respectively, were obtained from [18], converted to mouse orthologues using biomart and examined in the GEMMs to determine if they support the CMS calls assigned by the classifiers (Fig. 3e and Supplementary Fig. 3e). Referring to MmCMS-C calls particularly, this analysis reveals a strong association between CMS4 samples and stroma signature. The CMS2 samples in the larger cohort ($n = 39$) are repressed for immune response and stroma signatures but have high enrichment for colon epithelial differentiation as expected as well as goblet cell-like signatures (Fig. 3e). Although all CMS3 samples have universal enrichment for goblet cell-like signatures, some samples with BP, BPN, AK genotype also display elevated immune response and colon epithelial differentiation signatures. Moreover, the result demonstrates high enrichment of only immune response signature for CMS1 samples in the small cohort as expected; however, in the larger cohort there is also some level of expression for colon epithelial differentiation and goblet cell-like signatures; although these inconsistencies may be explained due to limited samples classified as CMS1 using any method (Fig. 3e and Supplementary Fig. 3e).

Interestingly, these results also revealed intra-genotype variation in CMS classifications, particularly within the non-*Apc* models, indicating that mice with the same genotype at induction can develop tumours with heterogeneity in terms of CMS biology. (Fig. 3b–d and Supplementary Table S4). Our group and others have previously described the importance of the TME in CMS classification, particularly with CMS1/4 [5, 19]. Therefore, we performed an assessment of stromal/immune percentages using ESTIMATE [20] R package for the large mouse cohort ($n = 39$), compared to MmCMS-C results, which confirmed a strong association between these histological features and CMS classification (Fig. 3f). These findings indicate that while intra-genotype subtype heterogeneity exists in mouse tumours, similar to our previous reports in human tumour data, this can be explained to some extent by the histology of the established tumour.

MmCMS-C provides an optimal classifier for CMS2-like mouse tissues

To test how well our GEMM classifications align with the biological characteristics associated with human CMS subtypes, we next measured the biological traits of immune-related, metabolic, proliferation and stromal signalling associated with CMS calls in human TCGA data and compared them directly to the CMS classification calls according to each of our three MmCMS options in both independent GEMM cohorts. Mean ssGSEA scores were calculated across samples of each human CMS subtype, using the same TCGA samples used in Fig. 1, alongside mean ssGSEA scores

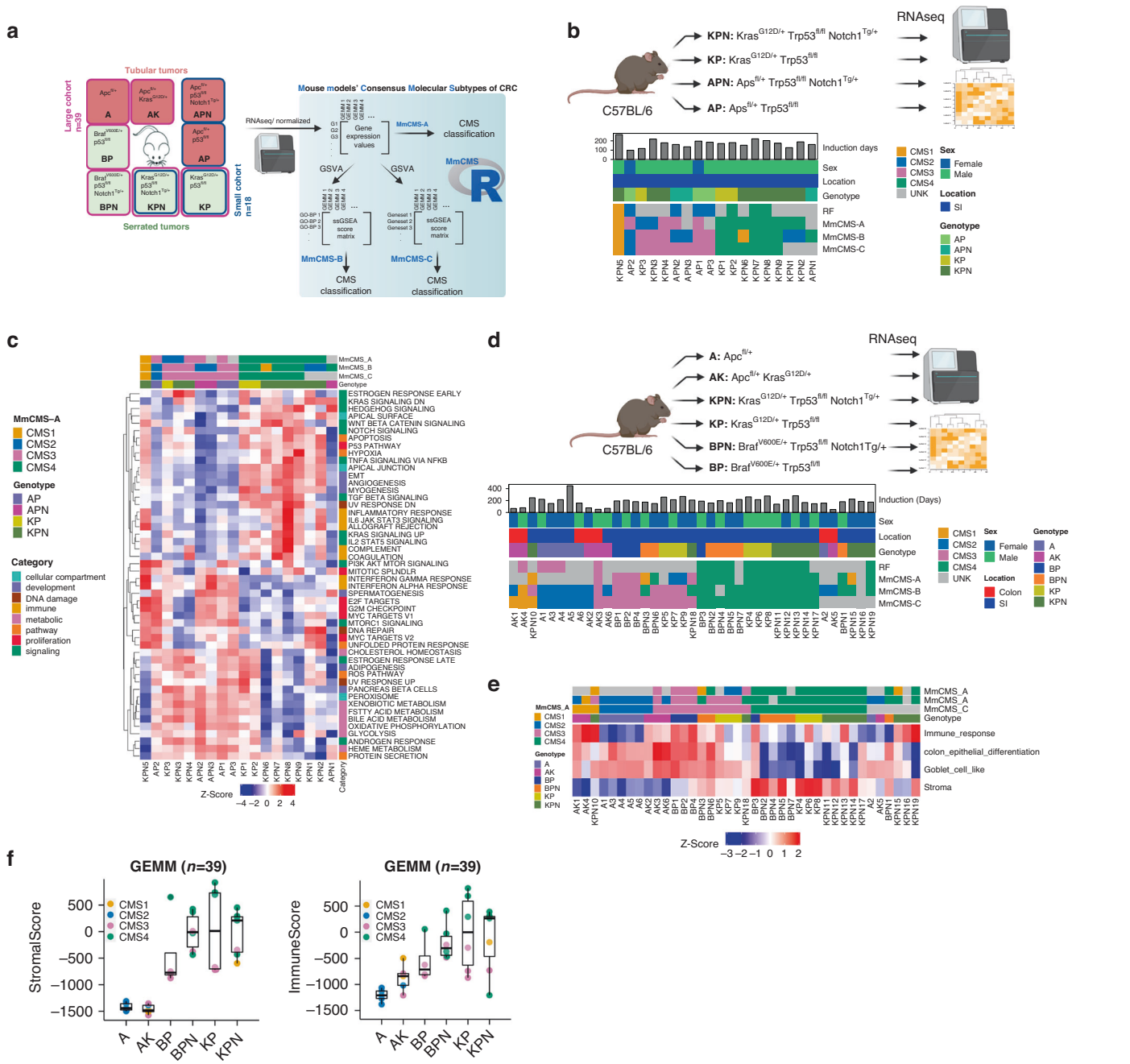

Fig. 3 Molecular characterisation of GEM models. a Left part of schematic shows all the GEMMs used in this study. The border colour indicates the genotypes that are included in each cohort, blue for small cohort and pink for large cohort. The squares with both border colours show the presence of that particular genotypes in both cohorts, but the samples are different. The squares with peach background are representative of the tubular tumour models and with green background show the serrated tumour models. Right part of schematic shows the input file for each three options. MmCMS-A option use normalised expression values to call CMS for mouse tissues, but MmCMS-B and -C work on ssgsea score matrix. All the analysis process to convert gene expression values to ssgsea score matrix will be automatically done in the R package. Users just need to provide the package with normalised expression values as input file while genes are row names and samples are in columns. **b** Comparison of CMS classification results using our three options (MmCMS-A, MmCMS-B and MmCMS-C) in small GEMM cohort ($n = 18$). Grey colour indicates unclassified samples. The CMS calls are aligned by genotype, location, sex and duration of tamoxifen induction. **c** Heatmap of Hallmark ssGSEA score across samples from the small cohort of GEMMs (scores are z-score scaled). **d** CMS classification of 39 GEM models using our three options. Grey colour indicates unclassified samples. The CMS calls are aligned by genotype, tumour location, sex and duration of tamoxifen induction. **e** ssGSEA scores heatmap of immune response, colon epithelial differentiation, goblet cell-like, and stroma-related gene sets across GEMMs that are aligned by genotype and CMS classification result from the three options in the large cohort ($n = 39$). **f** Boxplot shows stromal score (left) and immune score (right) across GEMM genotypes annotated with CMS colours. Horizontal line represents median values, boxes indicate the inter-quartile range and bars denote the maximum and minimum values.

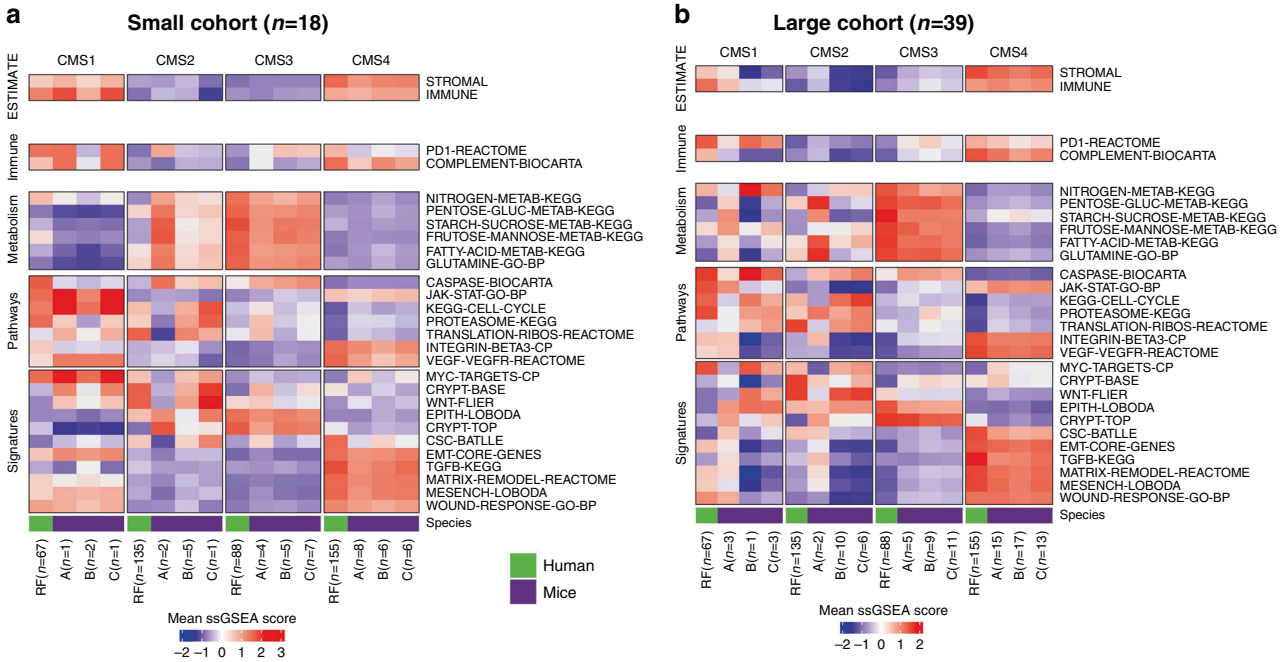

Fig. 4 Pathway-based classification is more reliable, particularly in calling CMS2-like mouse tissues. a Comparison of mean ssGSEA scores of hallmarks (especially those associated with immune, proliferation, metabolism and stromal infiltration signatures) in mouse CMS calls ($n = 18$ GEMM cohort) using the three classification options and human CMS calls. **b** Comparison of mean ssGSEA scores of hallmarks in mouse CMS calls ($n = 39$ GEMM cohort) using the three classification options and human CMS calls.

for MmCMS-A, B and C predictions in the $n = 18$ and $n = 39$ GEMM cohorts (Fig. 4a, b). Using the human RF calls as the ground truth, followed by cross-comparison and correlation analysis of samples assigned as CMS2 by all three mouse options, we find a strong correlation with MmCMS-B ($r = 0.79$, $p = 0.0000005$) and MmCMS-C ($r = 0.81$, $p = 0.0000001$) and no correlation with MmCMS-A in both cohorts (Fig. 4 and Supplementary Table S5). In addition, we found limited associations for biological traits in human CMS1 with the CMS1 calls for any of our mouse classifier options, again however this may be due to the small numbers of CMS1 classifications in mouse tumours. In samples classified as CMS3 and CMS4, all 3 MmCMS options show a significant positive correlation with related human CMS subtypes, although again MmCMS-C classification calls display higher association to human traits compared to MmCMS-A (Supplementary Table S5).

MmCMS-B and C (biological knowledge-based approaches) are less influenced by non-biological factors

To assess how much non-biological factors, such as normalisation methods would affect the CMS classification result of three options, we generated a larger collection of GEM models by combining both cohorts used in this study. APN and AP models were excluded, as the batch they were sequenced in was deeply confounded by genotype, resulting in a collection of transcriptomic data from $n = 51$ tumour samples, including 6 genotypes: A, AK, BP, BPN, KP and KPN. After batch correction using ComBat_seq, two different methods of normalisation, namely vst and quantile, were applied and thereafter CMS classification was performed using the 3 options. The results show 100% concordance between both methods for CMS calls assigned by MmCMS-B and MmCMS-C; however, in line with limitations of gene-level classifiers, concordance with the gene-level MmCMS-A classifier was reduced to 92% (Supplementary Fig. 4a, b). This suggests broad biological knowledge-based approaches based on overall gene ranking across biological pathways, rather than individual genes, are more robust and less likely to be influenced by non-biological factors [12].

Individual gene-level classifiers are not reliable for dual-species classification

Our study suggests that individual gene-level classifiers, both the RF and MmCMS-A approaches, derived from human tumours perform poorly when applied to data derived from mouse CRC tumours (Fig. 3). Therefore, we next assessed if the reverse was also true; if individual genes associated with MmCMS-C classified mouse tumours could be used in a reverse-translational way to distinguish CMS within human tumours. Using the 20 most discriminatory genes (Fold change > 4; p value < 0.05) within mouse tumours assigned using MmCMS-C (Supplementary Fig. 5a), we calculated the median value of these 'meta-signatures' within each CMS subtype in the human TCGA cohort. In line with our human-to-mouse findings, although gene-level classification can be used to distinguish CMS4 tumours from all other subtypes, individual genes associated with mouse CMS1-3 displayed inconsistent subtype associations when applied in human (Supplementary Fig. 5b), further reinforcing the importance of using pathway-level data for dual-species tumour classification.

Unsupervised clustering within mouse tumour tissue identifies clusters that align with CMS

Following cross-comparison with human data and identification of MmCMS-C as the most optimal classifier, pairwise GSEA was performed for each CMS subtype versus other subtypes in both human and mouse data to identify conserved hallmark signalling in both species (Fig. 5a). In CMS2 group, we found positive enrichment of MYC targets, E2F targets and G2M checkpoint (padj < 0.05) which are best characteristic signatures for this subtype in both species. Due to the small number of CMS1 classifications in mouse tumours, no significant hallmark biology could be found in CMS1. Furthermore, although our cohort of $n = 51$ GEMM tumour samples is small in comparison to the dataset size likely required for robust subtype discovery, it does represent one of the largest collections of GEMM tumour data. Therefore, with the caveats of sample size in mind, we next tested if de novo unsupervised clustering of our mouse tumour transcriptional data would

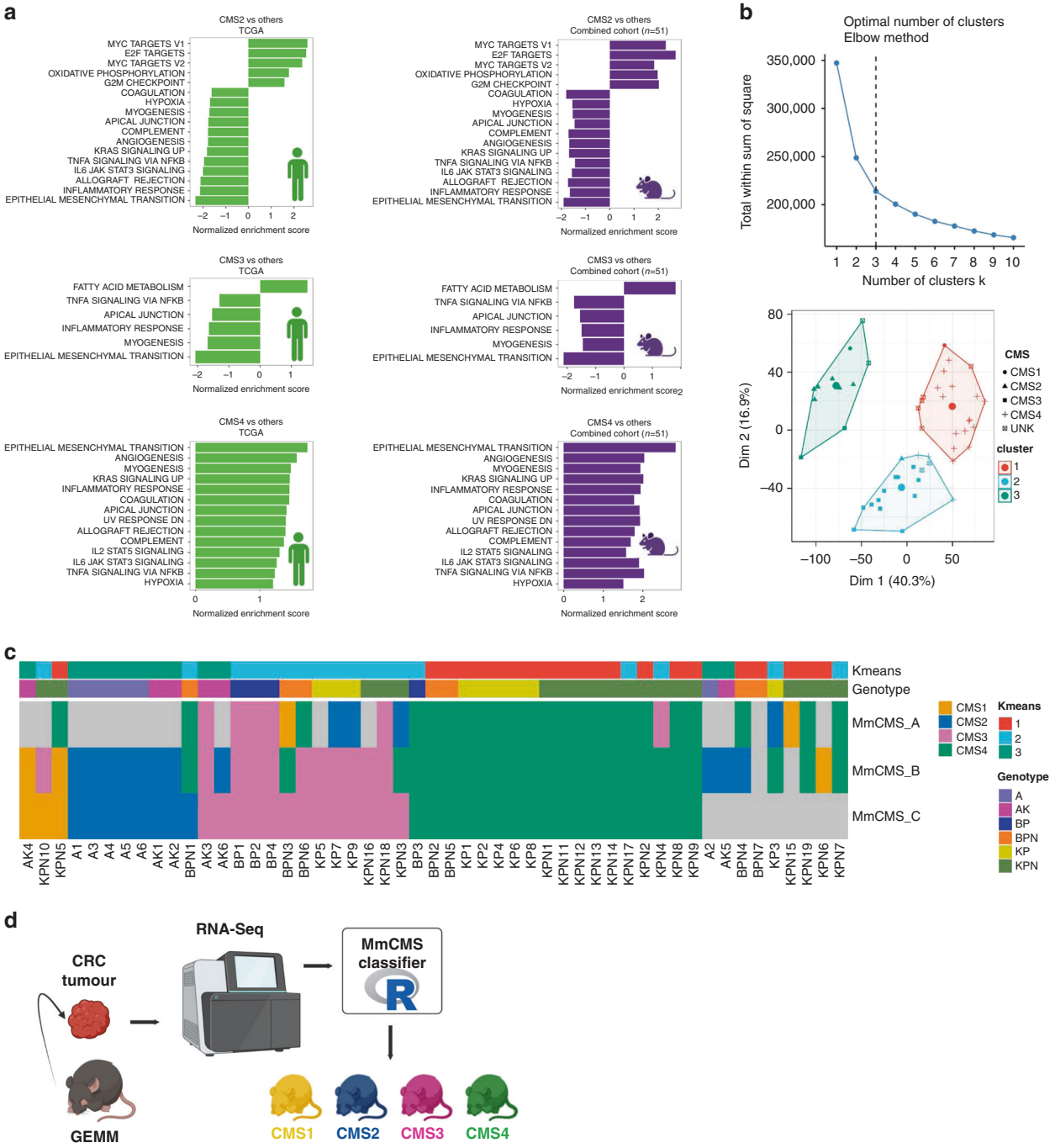

Fig. 5 Unsupervised clustering of our mouse tumour transcriptional data identifies tumour clusters that relate to CMS biology. a Barplots shows the conserved hallmark biologies across both species that are significantly enriched (*padj* < 0.05) in each CMS subtype versus other subtypes. **b** Elbow method on $k = 1{:}10$ indicated $k = 3$ (shown in vertical black dashed line) as optimal number of cluster (left). K-means clusters of $n = 51$ samples are visualised using principal component 1 (written as Dim1) and principal component 2 (written as Dim2) (right). **c** Heatmap depicts the predicted CMS class using each individual MmCMS classifiers with K-means clusters and genotype annotation on the top. **d** Schematic describing the development of the R-based MmCMS classifier.

identify tumour clusters that relate to CMS biology, or indicate the presence of unique mouse-related tumour biology. To this end, we performed class discovery using K-means algorithm along with elbow method, which identified the presence of three clusters as the optimal number of classes (Fig. 5b). Assessment of these three cluster in comparison to the MmCMS calls in our cohort revealed a striking alignment with CMS2, CMS3 and CMS4 subtypes (Fig. 5c). Although our cohort did not contain a large number of CMS1 mouse tumours, these tumours were equally split across all three clusters identified (Fig. 5c). Finally, we utilised the K-means algorithm to identify a fourth cluster in an attempt to segregate CMS1 tumours, which revealed that when set to $k = 4$ clusters, the CMS1 tumours remained split across the same three initial clusters, while CMS3 tumours were further subdivided into two clusters (Supplementary Fig. 6). Taken together, despite the limitations of small sample size, these analyse indicate that there

is general alignment between human CMS tumour clustering with that of the transcriptional traits underpinning current mouse models of CRC.

Overall, this study presents a new approach to CMS classification of mouse tumour tissue, alongside the development of a publically available R-based classification tool (Fig. 5d) that will improve the reproducibility of disease positioning and enable standardisation in pre-clinical molecular subtyping studies.

DISCUSSION

GEMMs represent a valuable tool to test novel treatments that may benefit specific subtypes of tumours, making it essential to ensure the chosen models accurately recapitulate biological signalling and phenotypes underpinning human subtypes [21]. Therefore, accurate and robust classification of mouse CRCs according to human subtypes is a critical step to improve disease positioning of models and translation of findings from the pre-clinical setting. Integrity and robustness in positioning models with human cancer subtypes is critical in the era of stratified medicine, where therapeutic approaches are designed for the biology underpinning specific tumour subtypes. In order to successfully translate pre-clinical efficacies into clinical benefit, testing of therapeutics must be performed in models that are representative of specific patient subtypes. Despite its importance, dual-species classification has been limited by the lack of a reliable and standardised approach, limiting researchers' ability to ensure faithful alignment between human tumours and pre-clinical models. Therefore, to address this, we developed a series of dual-species CMS classification models, named MmCMS, and an accompanying R package, which allows users to rapidly perform CMS classification of mouse tissue using three different options (A–C) of increasing complexity, from gene-level to biological pathways. To ensure that these new classifier options benefit the field, we developed a publicly available R package for MmCMS, which can be downloaded from https://github.com/MolecularPathologyLab/MmCMS. Although we have focussed on CMS in this study, data presented here provide an ideal template for the development and testing of other dual-species classifications, for subtypes such as CRC intrinsic subtypes [5], Braf mutant subtypes [22] and many others.

Our gene-level classifier, MmCMS-A, converts the human CMS template, embedded in CMScaller R package, to mouse orthologs and then use the NTP algorithm to carry out mouse CMS classification. The CMScaller package has been developed to enable exploration of the CMS subtypes in human pre-clinical models, particularly in cell lines, organoids and PDX tumours, to overcome the limitation of CMSclassifier's strong dependence on gene expression derived from the TME [1, 6]. As this approach is based on individual genes, any genes lost during the process of obtaining mouse orthologues [23, 24] can affect classification performance, resulting in a higher number of inaccurate or unknown calls, compared to biological knowledge-based approaches. In addition to biological differences between mice and humans, the representation and coverage of individual genes required for robust CMS classification may not be equivalent across different transcriptome profiling platforms [11], which again can lead to poor classifier performance.

Recent studies have shown that classifiers based on biological pathways, rather than individual genes, have the potential to provide a more robust classification, as by using hundreds of co-ordinately expressed genes they become far less sensitive to bias that is associated with missing individual genes [9, 10]. This is based on the understanding that ontology/pathway-level approaches for transcriptional analyses have the advantage of identifying biologically meaningful information associated with a particular subgroup, rather than individual genes which can be confounded by issues such as intratumoural heterogeneity or technical variations associated with molecular profiling [9, 11, 12]. In our MmCMS R package, MmCMS-B and C were developed to overcome the limitations of individual gene-based approaches and are based on ssGSEA scores from broad biological knowledge-based approaches, less influenced by non-biological factors such as normalisation methods. Correlation analysis between each CMS-related pathway mean scores in human samples with each MmCMS individual classifiers shows that MmCMS-B and -C are more similar to human CMS classification, using the original RF classifier, and have higher discriminatory power and classification rates, particularly for CMS2 and CMS3. Our results suggest the presence of intra-genotype CMS subtype heterogeneity, indicating that the same mutations driver events can result in variable downstream transcriptional signalling, emphasising that faithful mouse model alignment with human tumour signalling should not be based on mutation alone.

Coupled with advances in our understanding of the biology underpinning tumour development and progression, the versatility and accessibility of transcriptional signatures has seen them become a fundamental tool in the alignment of clinical phenotypes and biological signalling across human tumours and pre-clinical models. As therapeutics are being tested in a variety of mouse-based in vivo models, it is now even more important to ensure faithful alignment between models and human tumours and that the models we use represent the same biology during forward and reverse-translation studies. Our study provides an important standardised approach for researchers to enable more reproducible and comparable classification of CRC mouse models, aligned to the biology underpinning human CRC subtypes. The identification of mouse tumours that truly mimic each human CRC subtype is essential for the proper interpretation of results, and their translation into effective human clinical trials.

DATA AVAILABILITY

The dataset of 18 GEM models is available via ArrayExpress: E-MTAB-6363. The dataset of 39 GEM models is available via Gene Expression Omnibus: GSE218776. To ensure that these new classifier options benefit the field, we developed a publicly available R package for MmCMS, which can be downloaded from https://github.com/MolecularPathologyLab/MmCMS.

REFERENCES

1. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;21:1350–6.
2. Linnekamp JF, Van Hooff SR, Prasetyanti PR, Kandimalla R, Buikhuisen JY, Fessler E, et al. Consensus molecular subtypes of colorectal cancer are recapitulated in in vitro and in vivo models. Cell Death Differ. 2018;25:616–33.
3. Sawayama H, Miyamoto Y, Ogawa K, Yoshida N, Baba H. Investigation of colorectal cancer in accordance with consensus molecular subtype classification. Ann Gastroenterol Surg. 2020;4:528–39.
4. Dunne PD, Alderdice M, O'Reilly PG, Roddy AC, McCorry AMB, Richman S, et al. Cancer-cell intrinsic gene expression signatures overcome intratumoural heterogeneity bias in colorectal cancer patient classification. Nat Commun. 2017;8:15657.
5. Isella C, Brundu F, Bellomo SE, Galimi F, Zanella E, Porporato R, et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. Nat Commun. 2017;8:15107.
6. Eide PW, Bruun J, Lothe RA, Sveen A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. Sci Rep. 2017;7:16618.
7. Roper J, Hung KE. Priceless GEMMs: genetically engineered mouse models for colorectal cancer drug development. Trends Pharm Sci. 2012;33:449–55.
8. Belmont PJ, Budinska E, Jiang P, Sinnamon MJ, Coffee E, Roper J, et al. Cross-species analysis of genetically engineered mouse models of MAPK-driven colorectal cancer identifies hallmarks of the human disease. Dis Model Mech. 2014;7:613–23.
9. Kim S, Kon M, DeLisi C. Pathway-based classification of cancer subtypes. Biol Direct. 2012;7:21.
10. Pfefferle AD, Herschkowitz JI, Usary J, Harrell JC, Spike BT, Adams JR, et al. Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. Genome Biol. 2013;14:R125.

11. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733–9.
12. Gao F, Wang W, Tan M, Zhu L, Zhang Y, Fessler E, et al. DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis. 2019;8:44.
13. Jackstadt R, van Hooff SR, Leach JD, Cortes-Lavaud X, Lohuis JO, Ridgway RA, et al. Epithelial NOTCH signaling rewires the tumor microenvironment of colorectal cancer to drive poor-prognosis subtypes and metastasis. Cancer Cell. 2019;36:319–36.
14. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21:3439–40.
15. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17:218.
16. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1:417–25.
17. Petitprez F, Levy S, Sun C-M, Meylan M, Linhard C, Becht E, et al. 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 Med. 2020;12:86.
18. Varga J, Nicolas A, Petrocelli V, Pesic M, Mahmoud A, Michels BE, et al. AKT-dependent NOTCH3 activation drives tumor progression in a model of mesenchymal colorectal cancer. J Exp Med. 2020;217:e20191515.
19. Dunne PD, McArt DG, Bradley CA, O'Reilly PG, Barrett HL, Cummins R, et al. Challenging the cancer molecular stratification dogma: intratumoral heterogeneity undermines consensus molecular subtypes and potential diagnostic value in colorectal cancer. Clin Cancer Res. 2016;22:4095–104.
20. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
21. Cancer models for reverse and forward translation. Nat Cancer. 2022;3:135.
22. Barras D, Missiaglia E, Wirapati P, Sieber OM, Jorissen RN, Love C, et al. BRAF V600E mutant colorectal cancer subtypes based on gene expression. Clin Cancer Res. 2017;23:104–15.
23. Fang G, Bhardwaj N, Robilotto R, Gerstein MB. Getting started in gene orthology and functional analysis. PLoS Comput Biol. 2010;6:e1000703.
24. Mazza R, Strozzi F, Caprera A, Ajmone-Marsan P, Williams JL. The other side of comparative genomics: genes with no orthologs between the cow and other mammalian species. BMC Genomics. 2009;10:604.

ACKNOWLEDGEMENTS

We thank members of the CRUK-funded International Accelerator (ACRCelerate) consortium and the MRC-funded National Mouse Genetics Network Cancer Cluster for their valuable contributions to this manuscript.

AUTHOR CONTRIBUTIONS

Conceptualisation: RA, PDD. Methodology: RA, KG, SBM, TRML, PDD. Formal analysis: RA, KG, SBM, TRML, RMB, NCF, SMC, N-EM, HN-M, MLM, ADC, RAR, BA, RM, ABL, RS-P, AV, PDD. Interpretation: RA, KG, SBM, TRML, RMB, NCF, SMC, N-EM, HN-M, MLM, ADC, RAR, RM, EB, RS, ML, OJS, PDD. Data curation: RA, KG, SBM, TRML, RMB, SMC, N-EM, MLM, ADC, RAR, ML, OJS, PDD. Visualisation: RA, SBM, PDD. Writing original draft: RA, PDD. Review and editing: all authors. Supervision: OJS, PDD.

FUNDING

This research was supported by a CRUK early detection project grant (A29834), an International Accelerator Award, ACRCelerate, jointly funded by Cancer Research UK (A26825 and A28223), FC AECC (GEACC18004TAB) and AIRC (22795), the CRUK Glasgow Centre (A25142), a CRUK core funding (A21139, A17196 and A31287) and the Cancer Cluster programme within the MRC National Mouse Genetics Network (MC_PC_21042).

COMPETING INTERESTS

OJS has received grants from Astra Zeneca, Boehringer Ingelheim, Cancer Research Horizons and Novartis, none of which are directly relevant to this study. The other authors declare no competing interests.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

All animal experiments were performed in accordance with a UK Home Office licence (Project License 70/8646), and were subject to review by the animal welfare and ethical review board of the University of Glasgow.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41416-023-02157-6.

Correspondence and requests for materials should be addressed to Philip D. Dunne.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.