Id1, Spp1 **and** ***Pak3*** **are biomarkers of** ***Smad4*** **and TGF-β1 dependency in conditional intestinal adenoma, organoids and colorectal cancer.**

Mirvat Surakhy[1], Julia Matheson[1], David Barnes[1,3], Emma J. Carter[1], Jennifer Hughes[1], Claudia Bühnemann[1], Sabina Sanegre[1], Hans Morreau[2], Paul Metz[1], Charlotte J. Imianowski[1] and A. Bassim Hassan[1,¶].

[1]Oxford Molecular Pathology Institute, Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford, OX1 3RE, United Kingdom.

[2]Department of Pathology, Leiden University Medical Centre, Leiden, The Netherlands.

[3]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham B15 2TT

Short title;
Smad4 and TGF-β1 dependent functional gene expression in intestinal adenoma

Key words; Smad4, Apc, intestine, adenoma, TGF-β1, Id1, Spp1, Pak3, gene expression, colorectal cancer, biomarkers

Corresponding author:

¶Prof A B Hassan: bass.hassan@path.ox.ac.uk

Oxford Molecular Pathology Institute, Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford, OX1 3RE, UK

Tel: +44 (0)1865 275044

Fax: +44 (0)1865 275515

Abstract

The evolutionarily conserved TGF-β signalling pathway suppresses cell growth, yet loss of function results in cancer phenotypes. Here, we evaluate loss of the TGF-β pathway signal transduction regulator, Mothers against decapentaplegic homolog 4 (*Smad4*) with respect to intestinal adenoma phenotypes and specific TGF-β dependent gene expression. Conditional *Lgr5-Cre* activation (*Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$*Lgr5*CreER^{T2}) resulted in loss of function of *Apc* and *Smad4*, reduced small intestinal adenoma burden, yet discordant development of large caecal adenoma with nuclear localisation of phospho-Smad2/3. *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$ adenoma organoids resisted TGF-β1 induced cell death and EMT (IC_{50} 534pM) compared to *Apc*$^{\Delta/\Delta}$*Smad4*$^{+/+}$ (IC_{50} 24pM). TGF-β1 (390pM) modified adenoma bulk mRNA gene expression (RNA-Seq), with decreased *Id1* and high *Spp1* in *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$. Single cell RNAseq of caecal adenoma identified *Lgr5*^{low}, *Pak3*^{high} and *Id1*^{low} progenitor populations in *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$, that also correlated with poor prognosis in colorectal cancer (TCGA). *Smad4* loss of function-TGF-β1 dependent *Id1*^{low}, *Spp1*^{high} and *Pak3*^{high} in intestinal epithelial progenitor cells are specific gene-pathway biomarkers for further evaluation in intestinal cancers.

160

Introduction

The normal growth suppressing activity of transforming growth factor β (TGF-β) ligands (including activin and nodal) act through specific signal transduction receptors via Smads (R-Smads; Smad2/3), with BMP ligands (including GDF and anti-Mullerian hormone) acting through different receptors (R-Smads; Smad1/5/8) [1-4]. The R-Smads and co-Smads (Smad4) are also frequently mutated in invasive and metastatic intestinal carcinomas where there is Wnt- β-catenin pathway activation [5]. Unusually for ligand-dependent signalling pathways, mutation of the downstream TGF-β-BMP pathway can result in context-dependent and paradoxical gain of function (GOF) of the intestinal cancer phenotype, i.e. the signalling pathway can exert both initial tumour suppressive effects and that following mutation, can also result in 'oncogenic' GOF activation and epithelial-mesenchymal transition (EMT) [2]. The identification of cancer cells that have undergone GOF of the TGF-β-BMP pathway remains a challenge and requires validated mechanism-based biomarkers of the consequences of pathway loss of function [6]. Moreover, RNA based molecular prognostic classifiers of colorectal cancer have developed in line with the clinically applied breast cancer equivalents, and have identified several clusters (e.g. CRISB) that include a stromal based signature that correlate with poor prognosis in colorectal cancer [7-9]. Gene-pathway specific gene expression biomarkers based on driver gene variants and their dependencies are overall less well validated. We set out to evaluate loss of the TGF-β pathway signal transduction regulator, Mothers against decapentaplegic homolog 4 (*Smad4*), *in vivo* and *in vitro*, with respect to a murine intestinal adenoma phenotype and dependent gene expression. In view of the complexity of the interacting pathways that can alter TGF-β1-*Smad4* induced phenotypic effects, especially in immortalised human cancer cell lines, we sought to better define the contextual impact of *Smad4* loss of function (LOF) to specifically evaluate the *Smad4* and TGF-β1 ligand dependency on gene expression biomarkers and to translate these findings with respect to human intestinal cancer.

The TGF-β1 ligand acts as a dimer to bring together specific heterodimeric cell surface type I and II serine/threonine kinase receptors to activate signals by phosphorylation of R-Smad transcription factors. TGF-β binds to TGFBR1 and TGFBR2 to signal through p-Smad2/3, and can also bind to activin receptor type 1A (ACVR1) to transduce signal through p-Smad1/5 [10,11]. Following receptor activation, p-R-Smads form either heterotrimeric or heterodimeric complexes with Smad4, the co-Smad, as it continuously shuttles between the cytoplasm and nucleus [10,12]. Smad4-R-Smad complexes bind with co-factors and DNA [11].

Chromatin regions alter quickly (within 10 minutes) following TGF-β stimulation, with Smad complexes binding to *cis*-regulatory elements with AP-1 footprints [13,14].

Smad4 and R-Smad proteins consist of two highly conserved globular domains, named Mad homology domains, both at the N-terminus (MH1) and C-terminus (MH2) separated by a less well-conserved serine and proline-rich linker region critical for protein-protein interactions and stability [3,11,14]. The MH1 domain is involved in DNA binding through a conserved identical hairpin structure present in all members except Smad2 [10,15,16]. The majority of SMAD4 cancer somatic mutations are missense, although nonsense, splice site, frameshift, and in-frame insertion/deletion mutations exist. For example, there are hot spot mutations within the conserved R-SMAD binding surface in the MH2 domain, but missense mutations have also been reported in the MH1 domain along the DNA-binding interface [5].

The specialised intestinal epithelium is comprised of crypt-villus functional cell units that provide an ideal experimental system for the investigation of specific components of the TGF-β signalling pathway [17]. Differentiated cells in the crypt and villous comprise of four epithelial lineages: enterocytes, goblet, entero-endocrine and Paneth cells [17]. Leucine-rich repeat containing G-protein-coupled receptor (*Lgr5*) positive crypt base columnar cells (CBC) are stem cells that interact with Paneth cells to maintain the stem cell niche through *Wnt* activation and β-catenin stability [18,19]. *Lgr5* positive CBC cells divide, renew and create transit amplifying cells that undergo 4 to 5 cycles of cell division before differentiating as they ascend the villus. By using a *Lgr5* driven Cre recombinase, genes with associated loxP sites can be disrupted in CBCs once the Cre is activated. BMP signalling is active at the crypt-villous border towards the villous tip and is controlled by a ligand gradient to promote differentiation towards secretory and enteroendocrine cells [20-22]. BMP concentrations are regulated in the intestinal crypts by several different BMP antagonists, such as Gremlin, Noggin and Chordin-like1 [4]. TGF-β type I/II receptors and ligands are present and expressed in the muscularis, lamina propria and the differentiated compartment of both the small intestine and colon [23]. TGF-β is also frequently present in the intestinal tumour microenvironment and is released either by cancer cells, tumour stromal cell or immune cells [24]. Intestinal adenoma form following stabilisation of β-catenin/ Tcf4 when *Apc* LOF no longer regulates the degradation of β-catenin. Subsequent mutational events are selected for during the progression of adenoma to invasive carcinoma in human, e.g. *RAS*/*MAPK* GOF, *PTEN* and *TP53* LOF, and later acquired LOF mutation of the TGF-β-BMP ligand activated receptor pathway, such a mutation of receptors and Smad2/3/4 [25]. We report the reverse genetic experiment *Apc*$^{fl/fl}$ model using *Lgr5CreER*^{T2} that also incorporates loss of function of *Smad4* (*Smad4*$^{fl/fl}$ alleles). These

data suggested a context dependency to the *Smad4* LOF phenotype leading to further investigation of the TGF-β and *Smad4* dependent effects on bulk gene expression in cultured adenoma organoids and single cell sequencing of adenoma *in situ* that identifies functional gene-pathway specific expression biomarkers relevant for human intestinal cancers.

Results

Combined Apc and Smad4 conditional loss of function results in discordant adenoma phenotypes

To investigate the genetic interactions between *Apc* and *Smad4* in the intestine, we combined genotypes with a *Cre* driven by the promoter *Lgr5* and a ROSA26-loxSTOP reporter allele (*Lgr5-EGFP-IRES-CreER*^{T2} or *Lgr5CreER*^{T2})[26] (Fig. 1a). Heterozygote alleles were intercrossed (*Apc*$^{fl/+}$*Smad4*$^{fl/+}$*Lgr5CreER*^{T2}) on a C57Bl/6J background (backcrossed to C57Bl/6J for >10 generations) to generate homozygotes and heterozygote littermates that differ in Smad4 conditional alleles [27,28]. Conditional recombination followed tamoxifen injection (4OHT) in adult animals (day 0), and the mice were monitored daily (Fig. 1a). The floxed allele of *Smad4* (*Smad4*$^{\Delta/\Delta}$) resulted in loss of the MH1 DNA binding domain and the production of a truncated 43kDa protein which was not capable of translocating to the nucleus (see below, Supplementary Fig. 1). The germline homozygote loxP alleles are denoted *Smad4*$^{fl/fl}$ and *Apc*$^{fl/fl}$, and where somatic tissues have been directly tested for the presence of the resulting Cre-induced recombined alleles using PCR, we denoted the alleles as *Smad4*$^{\Delta/\Delta}$ and *Apc*$^{\Delta/\Delta}$. Recombination rates in intestinal epithelial cells were lower overall with the *Lgr5* transgene (5-6%), but highly specific for *Lgr5* CBC stem cells. Rapid adenoma development (within 35 days) occurred as expected and animals were culled if they exceeded predefined humane endpoints. Comparison of overall survival (Kaplan-Meier) between *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$*Lgr5CreER*^{T2} and *Apc*$^{fl/fl}$*Smad4*$^{fl/+}$*Lgr5CreER*^{T2} with *Apc*$^{fl/fl}$*Smad4*$^{+/+}$*Lgr5CreER*^{T2} controls showed a modest but significant improvement in survival (p=0.011, Fig. 1a). These observations suggest that introducing *Smad4* LOF in intestinal adenoma results in a tumour suppressor phenotype, yet analysis of intestinal adenoma distribution showed a marked discordance in growth phenotypes in *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$*Lgr5CreER*^{T2}. A reduction in small intestinal adenoma occurred with growth promotion of adenoma in the caecum (Fig. 1b). As the distribution of adenoma on a C57BL/6 background is predominantly in the small intestine rather than the colon, the small intestinal adenoma usually result in anaemia, small intestinal obstruction and decreased survival [29,30]. The significantly reduced small intestinal adenoma burden in *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$*Lgr5CreER*^{T2} compared to controls may have normally improved

survival (Fig. 1a, b). In contrast, examination of the caecum and large intestine revealed markedly enlarged coalesced adenoma where numerical analysis was impossible, hence our reporting of increased caecal weight in *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$*Lgr5CreER*^{T2} (Fig 1b). The adenoma burden in heterozygote *Apc*$^{fl/fl}$*Smad4*$^{+/fl}$*Lgr5CreER*^{T2} compared to *Apc*$^{fl/fl}$*Smad4*$^{+/+}$*Lgr5CreER*^{T2} controls was statistically non-significant and not supportive of either *Smad4* haplo-insufficiency or a dominant negative effect of *Smad4*$^{fl/+}$. Histological grading of small intestinal adenoma showed low-grade dysplasia in all genotypes, whereas more aggressive carcinoma *in situ* was observed in *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$*Lgr5CreER*^{T2} in the caecum (Fig. 1c, author HM pathological review). Examination of H&E stained small intestinal adenoma tissue sections showed that *Apc*$^{fl/fl}$*Smad4*$^{+/+}$ mice also had sessile adenomatosis and *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$ had mainly tubular adenomas (HM). No tissue invasion and metastasis were observed in any mouse and crypt abscesses were detected in both genotypes (Fig. 1c). The efficiency of Cre-mediated *Smad4* allele deletion was confirmed by immunohistochemistry (IHC, Fig. 1c). Smad4 protein was detected in *Apc*$^{\Delta/\Delta}$*Smad4*$^{+/+}$ adenoma and adjacent tissue as expected. Adenoma from *Apc*$^{\Delta/\Delta}$ *Smad4*$^{\Delta/\Delta}$ did not label with Smad4 protein, confirming tissue-specific gene disruption in both the small and large intestine (Fig.1c). The presence of active TGF-β (BMP) signalling supported by p-Smad2 and p-Smad1/5/8 nuclear immunolabeling, respectively, was confirmed by immunoblots (Fig. 1d). Both p-Smad2 and p-Smad 1/5/8 were detected in adenoma from *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$ compared to normal tissue and littermate control adenoma, consistent with either TGF-β alone or with BMP ligand ($p<0.001$, Fig. 1d).

Combined Apc and Smad4 disruption modifies growth of adenoma organoids

We next isolated *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$ (*Smad4*$^{\Delta/\Delta}$) and *Apc*$^{\Delta/\Delta}$*Smad4*$^{+/+}$ (*Smad4*$^{+/+}$) caecal adenoma and generated *in vitro* organoid cultures in Matrigel using standard conditions [26,31]. In culture, *Smad4*$^{\Delta/\Delta}$ adenomas appeared to be smaller compared to *Smad4*$^{+/+}$ adenomas, as measured by both area and the alamarBlue (AB) growth assay (day 2-7, Supplementary Fig. 2). Noggin, a BMP pathway inhibitor, normally prevents enterocyte differentiation and is required for crypt culture maintenance [25,26,31]. Noggin had no effect on adenoma growth (Supplementary Fig. 2). SB-505124, a TGF-βR1 small molecule inhibitor also had no significant inhibitory effect on Smad4$^{+/+}$ and *Smad4*$^{\Delta/\Delta}$ adenoma growth (1 μM) (Supplementary Fig.2). These data suggest that the adenoma culture conditions appeared to be independent of endogenous TGF-β.

To exclude whether there may have been a genotype dependent difference in the floxing efficiency between *Apc* and *Smad4* that may have accounted for the two-fold

differences in growth in culture, we generated adenoma with an alternative conditional Cre (*Villin-CreER^{T2}*). Here, cultured small intestinal organoids from healthy animals bred with non-floxed alleles, *Apc$^{fl/fl}$Smad4$^{+/+}$Villin-CreER^{T2}Rosa26-YFP* and *Apc$^{fl/fl}$Smad4$^{fl/fl}$Villin-CreER^{T2}Rosa26-YFP* genotypes, were cultured in organoid media containing Respondin-1, Noggin and EGF. Intestinal organoids that formed a normal crypt-villous morphology were exposed to 4-OHT followed by withdrawal of R-spondin1. This resulted in adenoma organoid formation in both *Apc$^{Δ/Δ}$Smad4$^{+/+}$* and *Apc$^{Δ/Δ}$Smad4$^{Δ/Δ}$* (Supplementary Fig. 3). Allele recombination was confirmed by PCR product of 259bp and 512bp for *Apc* and *Smad4 alleles*, respectively. Both the *Apc$^{fl/fl}$* and *Smad4$^{fl/fl}$* alleles recombined at the same time and rate following 4-OHT treatment, and were detected as early as day 2 post treatment (Supplementary Fig. 3). Possible alternative mechanisms included either differential responses to microenvironmental conditions or differences in progenitor cell differentiation between the small intestine and caecum in *Apc$^{fl/fl}$Smad4$^{fl/fl}$Lgr5CreER^{T2}*.

Deletion of MH1 domain of Smad4 following Cre-mediated recombination.

When we validated the expression of *Smad4* mRNA in *Smad4$^{Δ/Δ}$* and *Smad4$^{+/+}$ in vitro* using bulk RNA-Seq of adenoma organoids, we detected two different *Smad4* mRNA transcripts. *Smad4* gene tracking using Cuffdiff revealed transcripts from the intact *Smad4* gene, and a second transcript from the *Smad4* gene with exon 2 excised following loxP mediated recombination (Supplementary Fig. 1). Immunoblots confirmed expression of Smad4 full-length protein (61kDa) in both *Smad4$^{+/+}$* and *Smad4$^{+/Δ}$* adenomas, with a smaller band of (43kDa) detected in protein extracts from *Smad4$^{+/Δ}$* and *Smad4$^{Δ/Δ}$* adenomas (Supplementary Fig. 1). *In silico* analysis revealed an open reading frame in exon 5 of the *Smad4* sequence and mapping with ExPASy revealed that exon 2 deletion results in a truncated protein in which the *Smad4* MH1 domain is deleted. The predicted size of the truncated protein was 43kDa, comprising the linker and Smad4 MH2 domain (Supplementary Fig. 1). The full-length Smad4 protein in *Smad4$^{+/+}$* adenoma showed nuclear localisation following exposure of 390pM TGF-β1 for 2 hours. Under the same conditions, no nuclear localisation of Smad4 (43kDa) was detected in *Smad4$^{Δ/Δ}$* adenomas (Supplementary Fig. 1). Taken together, Cre-mediated recombination of the Smad4 gene resulted in selective MH1 deletion and a truncated protein that did not undergo nuclear localisation following TGF-β1 stimulation.

TGF-β1 dose dependency in Smad4$^{Δ/Δ}$ adenoma organoids

We next evaluated the dose and timing effects of exogenous addition of TGF-β1 to adenoma organoid cultures and the subsequent localisation of R-Smads. TGF-β1 induced dose-dependent growth inhibition of both the *Smad4*$^{+/+}$ and *Smad4*$^{\Delta/\Delta}$ adenomas, but with a 20-fold differential sensitivity (Fig. 2a). *Smad4*$^{+/+}$ adenomas were more sensitive to TGF-β1, with an IC_{50} of 24pM and threshold of 0.39pM. For *Smad4*$^{\Delta/\Delta}$, the TGF-β1 IC_{50} was 534pM with a threshold between 39-390pM. The decrease in adenoma viability with TGF-β1 dosing was prevented when the adenoma cultures were pre-incubated with SB-505124 (Fig. 2a). For subsequent experiments, a high TGF-β1 concentration of 390pM was used to differentiate the TGF-β1 context-dependent sensitivity effects between *Smad4*$^{+/+}$ (100 fold higher than the threshold) and *Smad4*$^{\Delta/\Delta}$ (at threshold but below IC_{50}) adenomas. TGF-β1 addition at 390pM resulted in the detection of phosphorylated p-Smad2 and p-Smad3 in both the *Smad4*$^{+/+}$ and *Smad4*$^{\Delta/\Delta}$ adenomas (Fig. 2b). The kinetics of response to TGF-β1 stimulation were similar between genotypes, with p-Smad2 maximally detected after 1-2 hours before subsiding with associated nuclear localisation (Fig. 2c, d). Within 24 hours of TGF-β1 administration, >90% of the *Smad4*$^{+/+}$ adenoma cells either had either morphological changes of apoptosis or changed to a marked mesenchymal phenotype (Supplementary Movies 1-4). These features appeared delayed to 72 hours in *Smad4*$^{\Delta/\Delta}$ adenomas but were still detectable (Fig. 2e). The reduced viability and altered morphology were accompanied by a significant decreased in proliferation and increased apoptosis in *Smad4*$^{+/+}$ adenoma when quantified by EdU and cleaved caspase 3 (CC3) labelling, respectively (Fig. 2f, g).

TGF-β1 and Smad4-dependent Id1 and Spp1 gene expression in adenoma organoids

We next sought to interrogate the transcriptome response to TGF-β1 that might inform differences between *Smad4*$^{+/+}$ and *Smad4*$^{\Delta/\Delta}$ adenoma organoids. Bulk paired-end RNA-Seq detected 16,143 gene transcripts for the analysis of differentially expressed genes (DEG). Three biological replicates per genotype were harvested at t=0, t=1 and t=12 hours after exposure to a 390pM TGF-β1 (Fig. 3a). Principle component analysis (PCA) of gene expression showed a clear separation of the samples based on genotype and duration of TGF-β1 exposure (Fig. 3b). Comparison of DEG between *Smad4*$^{\Delta/\Delta}$ and *Smad4*$^{+/+}$ at t=0 showed 111 upregulated genes (log_2FC >1.5) and 51 downregulated genes (log_2FC < -1.5) (Fig. 3c, Supplementary Table 1). The most important observation was the magnitude of the genotype dependent expression of a group of genes that included *Id1* (Inhibitor of DNA binding 1)*, Ddit4, Serpine1, Skil, mucins, Myc, JunB and E2f target gene* (Fig. 3c, d, Supplementary Fig. 4). Gene ontology (GO) analysis of biological processes showed enrichment for epithelial

cell proliferation and response to TGF-β1 (Supplementary Fig. 4, Supplementary Table 2). Gene set enrichment analysis (GSEA) using hallmark gene sets with "tmod" in *Smad4*$^{\Delta/\Delta}$ showed downregulation of *Myc* target genes, differential regulation of TGF-β signalling genes, mainly by upregulation, and enrichment of other pathways including IL2-Stat5 and TNF signalling (AUC >0.5, padj<0.0001, Fig. 3e, Supplementary Table 2). Importantly, mucins in *Smad4*$^{\Delta/\Delta}$ adenomas were among the DEGs with *Muc4, Muc20, Muc2* being upregulated, with *Muc6* being downregulated (padj p<0.05, Supplementary Table 1, supplementary Fig. 5). Mucin 2 immunolabeling confirmed overexpression in *Smad4*$^{\Delta/\Delta}$ adenoma tissue sections (Supplementary Fig. 5). For early timepoints, DEG were only observed in *Smad4*$^{+/+}$ adenoma organoids and included *Skil*, *Junb*, *Serpine 1*, *Ddit4* and *Id1* (Fig. 3d, Supplementary Fig. 4, Supplementary Table 1). The differential suppression of *Id1* in *Smad4*$^{\Delta/\Delta}$ compared to *Smad4*$^{+/+}$ after 1 hour achieved a 2.83 $\log_2$FC difference. By 12hr, *Id1* maintained suppression (1.2 log2FC, Fig. 3d, Supplementary Table 1). Despite the repression of gene expression, Id1 protein abundance and localisation did not appear to be significantly altered between adenoma genotypes (Supplementary Fig. 6). We also observed that *Spp1* (Secreted phosphoprotein 1), an extracellular matrix gene, was the only gene upregulated following TGF-β1 exposure for 12 hours in *Smad4*$^{\Delta/\Delta}$ ($\log_2$FC >1.5 threshold, Fig. 3c, d., Supplementary Table 3). Later time points displayed further significant time dependent differences in gene expression. For example, 35 DEGs were observed in *Smad4*$^{\Delta/\Delta}$ compared to 689 in *Smad4*$^{+/+}$ adenomas at t=12 hours vs t=0 hours (Fig.3d, Supplementary Table 1). Pathway enrichment analysis using tmod also showed that E2f, G2M, Myc and mTORC targets were more significantly downregulated in *Smad4*$^{+/+}$ adenomas compared to *Smad4*$^{\Delta/\Delta}$ adenomas (Fig. 3e, Supplementary Table 2). Time series analysis using likelihood ratios showed differences in expression of genes including *Fos*, *Lama3, Aldh1a3, Col7a1* and *Doks* associated with increased induction by t=12 hours, whereas *E2F8, Exo1, Clspn, Kif14* and *Kif12* all were decreased in *Smad4*$^{+/+}$ adenomas. Overall, the suppression of cell cycle genes was evident in *Smad4*$^{+/+}$ compared to *Smad4*$^{\Delta/\Delta}$ adenomas, as shown by Metascape overrepresentation (Fig.3f, g, Supplementary Table 3). Finally, TRRUST (Transcriptional Regulatory Relationship Unravelled by Sentence based Text mining) enrichment analysis implemented in Metascape also suggested potential regulation by transcription factors, with *Smad4*-dependent downregulated genes being co-regulated by Sp1, Smad3 and Jun, and downregulated genes by Rbl2, Tp53 and E2f1 (Supplementary Fig.4)

TGF-β1 dependent gene expression and protein correlations

Smad4 LOF did not fundamentally alter TGF-β1 phenotypic responses, but did change overall magnitude and relative gene expression consistent with epithelial to mesenchymal transition (EMT). TGF-β1 administration in *Smad4*$^{+/+}$ adenomas resulted in a >4.5 Log_2FC relative increase in expression *of Fn1* (fibronectin 1, Supplementary Fig. 7) and reduced E-cadherin immunofluorescence in both genotypes (Supplementary Fig. 7). In addition, TGF-β1 exposure resulted in a 0.25 log_2FC reduction in *Survivin* expression (12hr) in *Smad4*$^{\Delta/\Delta}$ compared to a 2.55 log_2FC reduction in *Smad4*$^{+/+}$ adenoma. Survivin western blots and nuclear localisation also suggested reduction in protein levels in *Smad4*$^{+/+}$ adenoma (Supplementary Fig. 8). The *Wnt* dependent intestinal stem cell marker *Lgr5* was downregulated in *Smad4*$^{+/+}$ following TGF-β1 (12hr) compared to *Smad4*$^{\Delta/\Delta}$ adenoma (2 log_2FC, Supplementary Table 1). E-cadherin and β-catenin proteins were both immunolocalised at the adheren junctions and nuclei in both adenoma genotypes as expected (Supplementary Fig. 9). Immunoblots of adenomas from both genotypes showed multiple MW bands for β-catenin and E-cadherin despite disruption of the β-catenin destruction complex by LOF of *Apc*$^{\Delta/\Delta}$ (Supplementary Fig. 9). Gene expression for *Myc* were not markedly different between genotypes, yet c-Myc protein levels appeared slightly increased in *Smad4*$^{\Delta/\Delta}$ (12-18hr) compared to *Smad4*$^{+/+}$ adenomas (Supplementary Fig. 9, Supplementary Table 1). Overall, there appeared no consistent correlation between changes in gene expression and protein localisation and abundance.

Single cell RNA-seq of Smad4$^{fl/fl}$ ***caecal adenoma are enriched with Pak3***^{Hi} ***epithelial progenitor cells.***

In view of the changes in bulk RNA-Seq data we next sought to quantify mRNA expression at the single cell level in adenoma directly harvested from the caecum. Epithelial cells (Epcam cluster) consisted of the *Lgr5* stem cell (*Lgr5* ISC) and seven stem cell-like (SC-L) clusters based on differential expression of *Lgr5* and other stem-cell markers such as *Lrig1, Smoc2, Hopx* as previously defined [32-35] (Fig. 4a, b. Supplementary Fig. 10, Supplementary Table 4). Differential enrichment was observed between the two genotypes, consisting of 61.8% of the total cell population in *Smad4*$^{fl/fl}$ adenoma compared to 47.3% in *Smad4*$^{+/+}$ (Figure 4c, d, Supplementary Table 3). A *Pak3*^{Hi} cluster (*Pak3*; P21 (Rac1/CDC42) activated kinase 3) was the most enriched in *Smad4*$^{fl/fl}$ (10.5%) compared to *Smad4*$^{+/+}$ (2.5%) adenoma. Pathway enrichment analysis for DEGs between *Smad4*$^{fl/fl}$ and *Smad4*$^{+/+}$ for the *Pak3*^{Hi} cluster showed an enrichment for the Wnt pathway and pluripotency markers (Supplementary Fig.

11). Comparison between genotypes and all timepoints post TGF-β1 in adenoma organoid culture showed a non-significant difference in expression of *Pak3* (Supplementary Table 1) . *Slc14a1*^{Hi} cluster was also enriched in *Smad4*$^{fl/fl}$ (7%) compared to *Smad4*$^{+/+}$ (3.5%), where expression of the stem cell marker *Ly6a* was enriched (*Sac1*). Conversely, the *Lgr5* ISC cluster was significantly enriched in *Smad4*$^{+/+}$ (5.8%) compared to *Smad4*$^{fl/fl}$ (2.7%) (Figure 4d). *Fmnl2*^{Hi}, *Cachd1*^{Hi} and *Slc12a2*^{Hi} showed similar proportions between in *Smad4*$^{fl/fl}$ and *Smad4*$^{+/+}$. *Smad4*$^{fl/fl}$ showed significantly lower proportions of enterocytes (EC) as a consequence (7.8% *Smad4*$^{fl/fl}$ vs 23.4% *Smad4*$^{+/+}$). The 2.2 resolution cluster tree identified two groups of cells; the enterocyte early precursor (ECepr) expressing *Apoa1 Dmbt1, Reg4,* and *Reg3g,* and the enterocyte late precursor (EClpr) expressing *Reg3b, Mt1,* and *Mt2* [35]. In *Smad4*$^{fl/fl}$, EClpr (0.26% vs 4.1%) and ECepr (1.3% vs 6.5%) were depleted compared to *Smad4*$^{+/+}$ adenoma. The mature enterocyte clusters, EC-1 and EC-1, showed expression of *Krt19, Krt8, Lgals3, Lypd8, Emp1,* and *Krt20* with differences in *Mxd1, Cdkn1a* and *Fam3*b expression. Significantly, 2.2% of *Smad4*$^{fl/fl}$ were EC-1 cells compared to 5.4% for *Smad*$^{+/+}$ (Fig. 4a, b, c, d). The SC-L clusters and the *Lgr5* ISC originate from the same node when viewed with cluster tree at 0.2 resolution (Fig. 4a). The *Tacstd2*^{Hi} cluster, which expresses the murine foetal intestinal progenitor gene *Tacstd2* [36], and the *Clu*^{Hi} cluster, that expresses the *Clusterin* marker of the revival stem cell [37], are SC-L clusters with a similar proportions and originate from the same node (0 at 0.2 resolution). For other cell types, transient amplifying (TA) lacked stem cell expression but increased ribosomal genes [35], secretory precursor (secretory pr) showed a mixture of Paneth cells marker *Pnliprp1* (pancreatic lipase-related protein 1) [32] and enteroendocrine precursor *Gadd45g* [35] in addition to the expression of *Mt3* and *Pnliprp2*, which have not been reported. Thus, scRNA-Seq identified *Pak3*^{Hi} and *Slc14a1*^{Hi} as genes with increased expression and relatively differentially enriched in progenitor epithelial cells in *Smad4*$^{fl/fl}$ caecal adenoma. Moreover, *ID1* appeared the invariant and most significant DEG irrespective of any underlying additional contextual 'pathway' mutations of common cancer pathways. Moreover, *Id1* and *Sema5a* are the uniquely downregulated genes that are common between organoid bulk RNA-Seq and the primary caecal adenoma scRNA-Seq data (Supplementary Fig. 12).

Smad4 dependent ID1, SPP1 and PAK3 expression in human colorectal cancer.

Having identified reduced *Id1* and increased *Spp1* gene expression as marked and consistent TGFβ-1dependent bulk RNA expression gene biomarkers in *Smad4*$^{\Delta/\Delta}$ murine adenoma organoids, and that high *PAK3* expressing cells are enriched in primary caecal intestinal

adenoma, we considered whether these experimentally determined biomarkers have prognostic relevance in human colorectal cancer (Fig. 5, Supplementary Fig. 1, Supplementary Table 5). We utilised data derived from 382 individual cases from the Cancer Genome Atlas project (TCGA; COAD, Colon Adenocarcinoma, READ, rectal adenocarcinoma) who had mutational profiles in c-BioPortal and RNA-seq data. We first determined whether the co-existing mutational context alters the expression of *ID1* and key regulatory gene expression pathways by assembling subgroups based on WNT-APC-β-catenin, RAS-MAPK, PI3K, and TP53 pathway mutations (Fig. 5a). This matrix resulted in four groups for each cancer pathway (Group 1: *SMAD4* WT-No 'pathway' mutation, Group 2: *SMAD4* WT: 'pathway' mutation: Group 3: SMAD4 mutated-No 'pathway' mutation, Group 4: *SMAD4* mutated- 'pathway' mutation). Similar to when we compared gene expression signatures using the top DEGs (84 genes) derived from our murine adenoma organoids (*Smad4*$^{\Delta/\Delta}$ versus *Smad4*$^{+/+}$), the only gene which showed significant differences in expression between *SMAD4* mutated and *SMAD4* wild-type tumors in the TCGA cohorts was *ID1* ($\log_2$FC 3.09-4.91) (Fig. 5). Analysis of the TCGA cohort with respect to either the presence or absence of any additional *TGF-β/BMP* pathway variants showed that *ID1* expression was also significantly downregulated with co-existing *SMAD4* and *TGF*-β pathway mutations compared to normal control tissues (Fig. 5b). Compared to non-SMAD4 associated TGF-β/BMP pathway mutations (e.g. in Type 1 receptors, other SMADs or in the BMP pathway), *SMAD4* mutations appeared to have the greatest impact on *ID1* expression (Fig. 5b, $p<0.0001$). *Smad4* expression was associated with poorer survival in the combined cohort ($p<0.0019$), although this was non-significant in the READ cohort. Lower *ID1* expression also appeared significantly associated with poor progression-free (PFS) survival in the colorectal TCGA cohorts combined (n=325, p=0.0062, Kaplan Meier), and the READ cohort separately (n=80, p=0.0045) with a similar but non-significant trend in the COAD cohort (n=245, p=0.076) (Fig. 5c). *SPP1* expression was not significantly associated with SMAD4 mutation in these cohorts despite the significant over-expression data in the *Smad4*$^{\Delta/\Delta}$ adenoma organoids. In the context of combined TGF-β/BMP pathway mutations, however, *SPP1* mRNA appeared significantly higher in these patients compared to patients without mutations in the TGF-β and BMP pathway (p=0.044, respectively, Dunn test, Fig. 5c). High *SPP1* expression also appeared significantly associated with poor PFS in the TCGA cohorts combined (p=0.0072, Fig. 5, b, c) and the READ cohort separately (n=80, p=0.0014). For PAK3, no significant overall expression differences were observed between *Smad4* mutational groups, with PFS associated with a worse outcome when

highly expressed (p=0.012, combined COAD and READ cohorts). This presumably reflects the relative proportion of *PAK3*^{Hi} expressing sub-populations of cells.

Discussion

We sought to evaluate the specific context dependent effects of TGF-β1 dose and *Smad4* LOF on gene expression in murine conditional models. We discovered that generation of homozygous *Smad4* with *Apc* LOF alleles was associated with a discordant intestinal phenotype, with adenoma suppression in the small intestine and adenoma progression in the caecum. By deriving adenoma organoids and comparing wild-type *Smad4* (*Smad*$^{+/+}$) with *Smad4* LOF (*Smad4*$^{\Delta/\Delta}$), we discovered that TGF-β1 induces similar EMT and cell death phenotypes in a dose dependent manner, but with relative resistance in *Smad4*$^{\Delta/\Delta}$ adenoma. Following TGF-β1 exposure, we identified genes with extremes of differential expression by genotype, including reduced *Id1* and increased *Spp1* gene expression as consistent features in *Smad4*$^{\Delta/\Delta}$ murine adenoma. Single cell RNA sequencing of the primary caecal adenomas detected enrichment of *Pak3*^{Hi} expressing progenitor epithelial cells *Smad4*$^{\Delta/\Delta}$ as the main underlying difference to account for discordant progression of adenoma in the caecum. Finally, having determined the gene-pathway dependent gene expression biomarkers, we evaluated whether *SMAD4*, *ID1, SPP1* and *PAK3* mRNA expression have prognostic implications for colorectal cancer as functional readouts of *SMAD4* LOF. We provide evidence these genes may require future prospective evaluation in separate cohorts as molecular prognostic-predictive classifiers in intestinal cancers with poorer prognosis, invasion and metastasis.

Mouse models of *Smad4* loss of function in the intestine have shown that *Smad4*$^{+/-}$ mice spontaneously developed juvenile polyposis (JP) at 1.5 years and invasive adenocarcinoma when crossed with *Apc*$^{\Delta716/+}$ [38-41]. In addition, *Apc*$^{1638N/+}$*Smad4*$^{fl/fl}$*K19-CreERT2* showed an increased tumour burden [42] and CRISPR/Cas9-mediated genome editing of human intestinal cultures introduced mutations in *APC*, *TP53*, *KRAS* and *SMAD4* which, when combined and selected, developed aggressive invasive carcinomas[25]. Moreover, *SMAD4* LOF blocked differentiation during tumour progression in an orthotopic organoid transplantation model of intestinal adenoma-carcinoma sequence [43]. These results inform the selection of deletions of chromosome arms 18p and 18q (including *SMAD4*) in 66% of colorectal tumours [44]. The pattern of adenoma development in *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$ suggests that the colon is more prone to adenoma initiation and progression when both genes are disrupted. One could hypothesize that this may be due to the combined inflammatory microenvironment and the

local TGF-β-BMP ligand supply. RNA-Seq showed downregulation of *Pla2g2a* (2.5$\log_2$FC) in *Smad4*$^{\Delta/\Delta}$ adenoma. *Pla2g2a* is considered to be a tumour suppressor independent of the Wnt/β-catenin signalling pathway, and intraluminal secretion of Pla2g2a is a host defence mechanism during the active phase of ulcerative colitis [45]. Furthermore, Reg3*γ* and *Reg3*β were also downregulated (1.2$\log_2$FC) in *Smad4*$^{\Delta/\Delta}$ adenomas and as both genes are important for epithelial defence against bacterial invasion [46,47]. Modifier alleles co-segregating with loxP alleles can also occur despite backcrossing to >10 generations of C57Bl/6J [30], best demonstrated for the context of β-catenin where "just-right'' signalling is modified [42]. For example, targeted deletion of *Bmpr1a* leads to crypt expansion with increased proliferation, fused villi and adenoma formation due to an increase in β-catenin independent of LOF of *Apc*[48]. Overall, *Smad4* LOF may create the conditions for local TGF-β1 ligand supply, differentiation arrest, chronic inflammation and impaired epithelial defence mechanisms.

Time-dependent TGF-β1 (390pM) induced adenoma mRNA gene expression was *Smad4* dependent with respect to the decreased expression of *Id1. Id1* codes for a non-DNA binding inhibitor of basic-helix-loop-helix (bHLH) transcription factors (e.g. GATA4) to inhibit multiple cell processes including differentiation. *Id1* showed an early transient upregulation with TGF-β1 treatment in wild-type Smad4$^{+/+}$adenomas followed by reduction, whereas *Id1* was decreased throughout in *Smad4*$^{\Delta/\Delta}$ adenomas. These effects can be explained by ChiP-Seq assays that have shown that Smad3 and Smad4 can bind the *Id1* promoter to directly stimulate early induction following TGF-β1 exposure [49]. Later *Id1* repression is also thought to be pSmad3 dependent through activation of *ATF3*, that then represses *Id1* expression[50]. *Id1* over-expression in pancreatic ductal adenocarcinoma in mouse and human is associated with more invasive and undifferentiated cells, a mechanism proposed to be dependent on EIF2 signalling [51] and independent of the TGF-β pathway[52]. Here, we found that *SMAD4* mutation in CRC patients was associated with low *ID1* mRNA expression and poor survival whereas this association has been also been conflicting [53]. Recently, a Nur77 based molecular switch has also been proposed where stabilisation of TGF-β induced Smad3 degradation leads to an increase in *Id1* expression, whereas in the absence of TGF-β, Nur77 enhances Smad3 degradation and reduces *Id1* expression[54]. *Pak3* is a well-known downstream effector of Cdc42 and Rac1 GTPases that mediates motility and MAPK activation associated with cancer invasion and metastasis [55-57]. In an adenovirus Cre-induced *Kras*^{G12D}, *p53*$^{fl/fl,}$ *Smad4*$^{fl/fl}$ model of lung cancer, increased metastasis was associated with *Smad4* LOF and *Pak3* activation was identified as a downstream effector of *Smad4* via the PAK3-JNK-Jun pathway [57]. Interestingly the mechanism appeared due to attenuation of *Smad4* dependent

transcription of miR-495 and miR-543 that directly bind to the PAK3 3'UTR. Overexpression of PAK family members are also associated with CRC outcome, transition of adenoma to carcinoma and stabilisation of β-catenin [58-64]. It will be important to investigate *ID1, SPP1* and *PAK3* as gene-pathway specific biomarkers in CRC particularly in the context of *SMAD4* LOF, their functional role in the anti-tumour function of the TGF-β pathway and as future components of the improvement in pathway classification of colorectal cancers [65].

Materials and methods

Conditional Cre intestinal adenoma model

Work was performed under a UK Home Office licence, and approved by University of Oxford ethics committee. The mouse lines were as reported except for *Smad4*$^{fl/fl}$ (*Smad4*^{tm1Rob}) (Supplementary Material and Methods). Genotyping PCRs were briefly (5'-3', forward-F and reverse-R): *Apc* F-GTTCTGTATCATGGAAAGATAGGTGGTC and R- CAC-TCAAAACGCTTTTGAGGGTTG (GAGTACGGGGTCTCTGTCTCAGTGAA for recombined product), *Smad4* F-CTTTTATTTTCAGATTCAGGGGTTC (F-AAAATGGGAA- AACCAACGAG for recombined product) and R- TACAAGTGTATGTCTTCAGCG. Genotype combinations were generated by experimental breeding. *Smad4*$^{fl/fl}$*RosaYFP*$^{fl/fl}$ and *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$*RosaYFP*$^{fl/fl}$ were crossed with either *Lgr5-CreER*^{T2} or *Vil-CreER*^{T2} mice. *Apc*$^{fl/fl}$*Smad4*$^{+/+}$*RosaYFP*$^{fl/fl}$ *Lgr5-CreER*^{T2}/*Vil-CreER*^{T2} animals were used as controls. Mice bred with *Vil-CreER*^{T2} mice were used for an *in vitro* experiment only. For *Lgr5-CreER*^{T2} activation, adult mice age ≥ 8 weeks had an intraperitoneal injection of 5mg (200 mg.kg^{-1}) tamoxifen dissolved in corn oil. Mice were monitored daily, small intestinal and colonic adenomas were scored for number, diameter, distribution and weight as described[30].

Intestinal adenoma organoids

Adenoma formed *in vivo* from *Smad*$^{fl/fl}$ or *Smad4*$^{+/+}$ Lgr5 mice were cultured as described[30,66] . For crypt culture, the small intestines were collected from *Smad4*$^{fl/fl}$*RosaYFP*$^{fl/fl}$ and *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$*RosaYFP*$^{fl/}$ *Vil-CreER*^{T2} mice and processed as described [66]. Briefly, crypts were cultured in Matrigel and supplemented with ADF2 and crypt growth factors, ENR (8.1 nM EGF (E), 500ng ml-1 R-spondin1 (R), and 2.16nM Noggin (N)) and incubated at 37°C, 5% as described. At D3, crypts generated from transgenic mice were incubated with 500nM 4-hydroxy-tamoxifen (4-OHT) (Millipore, 579002) dissolved in 100% EtOH (test), vehicle (negative control) with 5µl EtOH for 12hr at 37°C. The test and the vehicle treated organoids were supplemented with EN only (no R) and the normal crypts with ENR. Proteins were extracted from organoids as described [67].

Cell labelling and imaging

Neutral buffered formalin fixed tissues were paraffin embedded and 5µm sections were processed either by immunohistochemistry (IHC) or immunofluorescence (IF). The antibodies that used in this study are listed (Supplementary Materials and Methods). Organoids growing on coverslips in 24 well plates were used for IF. Organoid proliferation used the thymidine analogue 5-ethynyl-2'-deoxyuridine (EdU) for 2 hrs using Click-iT EdU Alexa Fluor-555 Imaging Kit (Invitrogen, C10338). H&E and IHC slides were viewed by brightfield-

fluorescent microscopy (Olympus BX60). Images were acquired by CCD camera and Nuance 2.10 software (CRI, Massachusetts). Confocal images were captured with an Olympus FluoView FV1000 using 40x 1.3NA lens and associated software. Labelled slides (EdU, IHC or IF) were analysed using ImageJ 1.48 software. AlamarBlue (AB) (AbD Serotec, BUF012B), was used to quantify adenoma growth. Absorbance after 24hr was read at 570nm and 600nm using a Spectramax M5 plate reader (MTX Lab Systems). Live cell imaging was conducted with a Zeiss Axiovert 200 inverted microscope with a BSI Prime CMOS camera with a Neofluar 10x 0.3NA lensfor adenoma organoids treated with TGF-β for 24hrs. MetaMorph software (v 7.8.12.0) multidimensional Z-stack acquisition with 100 milliseconds exposure acquired at regular intervals of 15 minutes over an extended duration of 6-18 hrs culture at 37.5 °C, and 5% CO_2 in air.

TGF-β1 induced gene expression bulk RNASeq.

Early passage adenomas seeded at 9,000 cells per 25µl well were used for RNA-Seq and miRNA-Seq. Adenomas were grown in ADF2 medium supplemented with 8.1nM EGF, 10µM Y-27632 and 1µM SB-505142. The medium was changed on D2, the cells were washed, EGF and 390pM TGF-β1 added and samples extracted from three different time points (t=0, t=1, t=12 hours post TGF-β1 treatment). Three biological replicates from each genotype and RNA was pooled from 8 wells per treatment. RNA extraction was performed using Quick-RNA MiniPrep (Zymo Research, R1054), assessed using a Nanodrop-1000 spectrophotometer and RNA Clean & concentrator™-5 (Zymo Research, R1015). RNA-Seq and miRNA-Seq were performed at the Wellcome Trust Center for Human Genetics, University of Oxford. Briefly, directional multiplexed mRNA and miRNA Libraries were prepared. All mRNA was ribo-depleted and converted to cDNA. Paired-end sequencing (100nt) were run on a HiSeq 2500 (Illumina) for mRNA-Seq.

Transcript abundances from RNA-Seq data were quantified using the Tuxedo pipeline[68]. Fastq reads were aligned to the mouse genome (mm10/GRCm38) with the TopHat-Bowtie2 aligner, versions 2.1.0 and 2.3.4, respectively and expression of transcripts was quantified with Cufflinks 2.2.1, as fragments per kilobase per million mapped reads (FPKM). Counts were also generated using the feature 'Counts' from the Subread package. Pairwise differential expression analysis between groups was performed in R using DESeq2 version 1.28[69]. Genes that have less than 5 reads in at least 3 samples were removed. Genes with adjusted p-values (FDR) < 0.05 were considered significant. Shrunk log2fold change (log2FC) was calculated using as a shrinkage estimator. Genes considered as differentially regulated between 2 conditions have ±1.5log2FC. Time-series analysis was performed using the

likelihood ratio test implemented in DESeq2. Gene ontology term enrichment analysis for the significant DEG genes performed using "clusterProfiler" where we evaluated biological processes and molecular functions[70]. For easier visualisation, we used "tmod"package for enrichment analysis (https://peerj.com/preprints/2420v1). In tmod, we applied CERNOtest, Coincident Extreme Ranks in Numerical Observations (CERNO) after converting mouse genes to human orthologs. We performed the enrichment analysis using Hallmark gene sets from MSigDB v7.2 (Broad Institute) and the blood signature from LI modules in tmod, gene sets with the area under the curve (AUC) ≥ 0.5 and padj<0.05 were considered positive. Metascape website was utilised for pathway analysis of certain gene lists, and to identify the transcription regulatory network based on TRRUST ontology[71].

Early TGF-β target genes were classified based on log2FC ±1.5 expression and adjusted p <0.05 after 1h compared to t=0 in *Smad4*$^{+/+}$ adenomas. Transient early upregulated/ down-regulated genes were the genes with opposite directions to their initial regulation when compared to late time points (t=12 vs t=0 and/ or t=12 vs t=1 hour) in *Smad4*$^{+/+}$ adenomas. Late TGF-β upregulated/downregulated genes were classified according to their upregulation/downregulation at 12 hours in *Smad4*$^{+/+}$ adenomas compared to t=0 and/or t=1 hour with log2FC ±1.5 and adjusted p <0.05. Genes were considered to be *Smad4*-independent if they were significantly expressed in both *Smad4*$^{+/+}$ and *Smad4*$^{\Delta/\Delta}$ adenomas with TGF-β1 treatment compared to their corresponding untreated baseline. However, if genes were differentially expressed in an opposite direction to that of TGF-β target genes in *Smad4*$^{\Delta/\Delta}$ vs *Smad4*$^{+/+}$ at t=0, they were classified as Smad4-dependent genes and excluded from the *Smad4*$^{\Delta/\Delta}$ vs *Smad4*$^{+/+}$ t=0 list. TGF-β target genes that were only upregulated in *Smad4*$^{+/+}$ adenomas were classified as *Smad4*-dependent genes. Genes that were expressed in *Smad4*$^{\Delta/\Delta}$ vs *Smad4*$^{+/+}$ at t=0 hour and were not among TGF-β target genes and considered for baseline comparisons.

Caecal adenoma single cell mRNA sequencing

Pooled single-cell suspensions from caecal tumours from three *Apc*$^{fl/fl}$,*Smad4*$^{+/+}$, *Lgr5Cre*^{ERT2} and five *Apc*$^{fl/fl}$, *Smad4*$^{fl/fl}$, *Lgr5Cre*^{ERT2} mice underwent library preparation using the Chromium Single Cell 3' kit (10X Genomics) and sequencing using Illumina HiSeq 2x150bp (Azenta, USA), with approximately 25,000 reads per cell. Sample demultiplexing, barcode processing and single-cell 5' unique molecular identifier (UMI) counting (Cell Ranger Software Suite v7.0.1) with FASTQ alignment to the mouse reference transcriptome mm10-3.0.0. The following criteria were then applied to each cell: gene number between 250 and 7,000, UMI count >500 and mitochondrial gene percentage <0.15, log10 Genes Per UMI >0.80.

Doublets were removed using the scDblFinder library. Genes with ten read counts or less were removed. A total of 20886 cells were analysed (11212 for *Smad4*$^{fl/fl}$ and 9674 for *Smad4*$^{+/+}$). Dimensionality reduction and clustering used the filtered gene-barcode matrix normalised using the SCTransform with Seurat (v4.3.0) [72]. Variables including mitochondria Ratio, nUMI, nGene and the cell cycle gene difference were regressed and the top 3000 variable genes were used. The filtered gene-barcode matrix of all samples were then integrated, k nearest neighbour (k-NN) generated using FindNeighbors (Seurat) using the top 40 principal components and clusters identified using FindCluster function (Louvain algorithm). We used resolution 1.2 to classify the different cell types and resolution 2.2 to determine the enterocyte precursors based on clustree (v0.5.1) [73]. Data was visualised using Uniform Manifold Approximation and Projection (UMAP) dimensional reduction and plots were generated using Seurat and the scCustomize library. Each cluster was determined using FindAllMarkers function in Seurat (Wilcox test). The log fold change threshold was set to 0.25. Manual cell labelling utilised gene clusters in previous studies [32-35], the Lgr5 intestinal stem cell signature [34] and the CellMarker 2.0 database [74]. The significant differences in cell type proportion used the permutation test implemented in scProportionTest package in R which calculate a p-value for each cluster, and a confidence interval for the magnitude difference via bootstrapping [75]. All analysis was performed in R V4.2.3. Differential gene expression analysis for each cluster between *Smad4*$^{fl/fl}$ and *Smad4+/+* were performed using MAST [76] with FindAllMarkers function in Seurat. Differentially expressed genes were selected based on log fold change threshold and expressed in at least in 25% in one of the groups. Pathway analysis was performed using Metascape [71].

TCGA colorectal cancer analysis

Normalised gene expression from RNA-seq (illumine hiseq Level 3 RSEM normalized genes) retrieved from http://firebrowse.org/?cohort=COADREAD accessed October 2019. Genes and pathways that are frequently mutated in colorectal cancer, retrieved from Cancer Genome Atlas project (TCGA)[44], include Wnt/β-catenin pathway (APC, CTNNB1, DKK1, DKK2, DKK3, DKK4, LRP5, FZD10, FAM123B:AMER1, AXIN2, TCF7L2, FBXW7, and ARID1A). TP53 pathway (ATM, and TP53). PI3K (IGF2, IRS2, PTEN, PIK3R1, and PIK3CA). RTK-RAS pathway (NRAS, KRAS, BRAF, ERBB2, and ERBB3). TGF-β pathway (ACVR1B, ACVR2A, SMAD2, SMAD3, SMAD4, TGFBR1, and TGFBR2 and BMP pathway (SMAD1, SMAD9, SMAD5, BMPR1A, BMPR1B and ACVR1). Genomic alteration (point mutation, HOMDEl or amplification or fusion) in these genes for COAD and READ TCGA cohorts retrieved from cBioPortal (https://www.cbioportal.org), accessed December 2019. In COAD and READ TCGA cohorts

the mutation status was available for 56 patients with SMAD4 mutations. 98.1% of these patients have alterations in the Wnt/β-catenin pathway. 59.3% have alteration in the P53 pathway, 46.3% have alteration in the PI3K pathway and 64.8% have alteration in the RTK-RAS pathway. Normalized RNA-seq was available for 51 normal tissues and 326 patient tumours with mutational data. In TCGA COAD-READ cohorts, 47 patients with probably damaging, and possibly damaging SMAD4 mutations with no other mutations in the TGF-β pathway have RNA-seq data. One patient did not have an alteration in the Wnt pathway and thus was excluded from the analysis. Patients with known mutations in the TGF-β pathway were excluded from the control group. Using this patient group we validated the top DEG in the mouse adenomas, genes with ±2log2FC and padj<0.05 from $Smad4^{\Delta/\Delta}$ vs $Smad4^{+/+}$ t=0 hour, in the presence or absence of mutations in the PI3K, P53, PTK-RAS pathways in the context of SMAD4 mutation. Statistical analysis was performed using Kruskal-Wallis test followed by FDR adjustment. Genes with p<0.05 following FDR adjustment were considered significant. Multiple comparisons were performed using Dunn test followed by Holm adjustment to find SMAD4-dependent genes. The expression between 2 groups were considered positive if adjusted p<0.05. We considered genes to be SMAD4 dependent genes based on three criteria. First, there is a significant difference in gene expression between Group 1-Group 3, and Group 1-Group 4, Group 2-Group 3 and Group 2-Group 4 in the context of the three altered pathways. Second, no difference in gene expression between Group 1-Group 2, and between Group 3 -Group 4. Third, the change in expression followed the same direction as the mouse adenomas. 84 genes differentially regulated at t=0 with absolute log2FC >=2 and adjusted p<= 0.05. Of those, 76 genes had human orthologs, and 73 genes had RNA data in the TCGA cohorts. Survival data for the COAD-READ TCGA cohort were retrieved from Liu et al.,2018 Table S1 Tab TCGA-CDR, accessed May 2019[77]. The cut-off point between high and low mRNA expression levels for each gene was computed based on the maximally selected rank statistics using the "survminer" package in R. mRNA levels with values above the cut-off point were considered high and below were considered low.

Statistical analysis

Statistical and graphical data analyses were performed using either Graph-Pad Prism version 6.0. software or R version 4.0.2. Images were analysed using ImageJ 1.48 software. Final Figures were generated using Photoshop (CS5). Immunoblot analysis was performed as follow; loading control and test from scanned immunoblots was aligned horizontally on Photoshop (CS5) and analyzed using ImageJ 1.48 software.

Acknowledgements

We thank Elisabeth Robertson for the conditional Smad4 model. Linda Randall and Nicky Hamp, for technical assistance.

Disclosure of potential conflicts of interest

None.

Author Contributions

Experiment conceived by ABH. Mouse breeding and phenotypic analysis by MS and JM, imaging by MS, SB, and CB, bulk gene expression MS and DB, genotyping by MS, EC, and JH, histopathology HM, evaluation of cancer cell model systems, MS, PM and CJI, single cell sequencing analysis MS, TCGA analysis, MS, manuscript was written by ABH, MS and DB. All authors approved the manuscript.

Grant Support

We thank Cancer Research UK (A429- ABH), Clarendon-Arab Bank scholarship programme (MS), Rhodes Trust (JM) and Dunn School (ABH) for funding.

References

1 Batlle, E. & Massague, J. Transforming Growth Factor-beta Signaling in Immunity and Cancer. *Immunity* **50**, 924-940, doi:10.1016/j.immuni.2019.03.024 (2019).

2 David, C. J. & Massague, J. Contextual determinants of TGFbeta action in development, immunity and cancer. *Nat Rev Mol Cell Biol* **19**, 419-435, doi:10.1038/s41580-018-0007-0 (2018).

3 Massague, J. TGFbeta signalling in context. *Nat Rev Mol Cell Biol* **13**, 616-630, doi:10.1038/nrm3434 (2012).

4 Zhang, Y. & Que, J. BMP Signaling in Development, Stem Cells, and Diseases of the Gastrointestinal Tract. *Annu Rev Physiol* **82**, 251-273, doi:10.1146/annurev-physiol-021119-034500 (2020).

5 Fleming, N. I. *et al.* SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. *Cancer Res* **73**, 725-735, doi:10.1158/0008-5472.CAN-12-2706 (2013).

6 Massague, J. & Sheppard, D. TGF-beta signaling in health and disease. *Cell* **186**, 4007-4037, doi:10.1016/j.cell.2023.07.036 (2023).

7 Bramsen, J. B. *et al.* Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer. *Cell Rep* **19**, 1268-1280, doi:10.1016/j.celrep.2017.04.045 (2017).

8 Dunne, P. D. & Arends, M. J. Molecular pathological classification of colorectal cancer-an update. *Virchows Arch*, doi:10.1007/s00428-024-03746-3 (2024).

9 Isella, C. *et al.* Stromal contribution to the colorectal cancer transcriptome. *Nat Genet* **47**, 312-319, doi:10.1038/ng.3224 (2015).

10 Heldin, C.-H. & Moustakas, A. Role of Smads in TGFβ signaling. *Cell and tissue research* **347**, 21–36 (2012).

11 Macias, M. J., Martin-Malpartida, P. & Massague, J. Structural determinants of Smad function in TGF-beta signaling. *Trends Biochem Sci* **40**, 296-308, doi:10.1016/j.tibs.2015.03.012 (2015).

12 Inman, G. J., Nicolas, F. J. & Hill, C. S. Nucleocytoplasmic shuttling of Smads 2, 3, and 4 permits sensing of TGF-beta receptor activity. *Mol Cell* **10**, 283-294, doi:10.1016/s1097-2765(02)00585-3 (2002).

13 Guerrero-Martinez, J. A., Ceballos-Chavez, M., Koehler, F., Peiro, S. & Reyes, J. C. TGFbeta promotes widespread enhancer chromatin opening and operates on genomic regulatory domains. *Nat Commun* **11**, 6196, doi:10.1038/s41467-020-19877-5 (2020).

14 Inman, G. J. & Hill, C. S. Stoichiometry of active smad-transcription factor complexes on DNA. *J Biol Chem* **277**, 51008-51016, doi:10.1074/jbc.M208532200 (2002).

15 Gaarenstroom, T. & Hill, C. S. 107–118 (Elsevier).

16 Massagué, J., Seoane, J. & Wotton, D. Smad transcription factors. *Genes & development* **19**, 2783–2810 (2005).

17 Gehart, H. & Clevers, H. Tales from the crypt: new insights into intestinal stem cells. *Nat Rev Gastroenterol Hepatol* **16**, 19-34, doi:10.1038/s41575-018-0081-y (2019).

18 Sato, T. *et al.* Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature* **469**, 415-418, doi:10.1038/nature09637 (2011).

19 Tan, S. H. *et al.* A constant pool of Lgr5(+) intestinal stem cells is required for intestinal homeostasis. *Cell Rep* **34**, 108633, doi:10.1016/j.celrep.2020.108633 (2021).

20 Beumer, J. *et al.* Enteroendocrine cells switch hormone expression along the crypt-to-villus BMP signalling gradient. *Nat Cell Biol* **20**, 909-916, doi:10.1038/s41556-018-0143-y (2018).

21 Beumer, J. & Clevers, H. Cell fate specification and differentiation in the adult mammalian intestine. *Nat Rev Mol Cell Biol* **22**, 39-53, doi:10.1038/s41580-020-0278-0 (2021).

22 Haramis, A.-P. G. *et al.* De novo crypt formation and juvenile polyposis on BMP inhibition in mouse intestine. *Science* **303**, 1684–1686 (2004).

23 Winesett, M. P., Ramsey, G. W. & Barnard, J. A. Type II TGF(beta) receptor expression in intestinal cell lines and in the intestinal tract. *Carcinogenesis* **17**, 989-995, doi:10.1093/carcin/17.5.989 (1996).

24 Hahm, K. B. *et al.* Loss of transforming growth factor beta signalling in the intestine contributes to tissue injury in inflammatory bowel disease. *Gut* **49**, 190-198, doi:10.1136/gut.49.2.190 (2001).

25 Drost, J. *et al.* Sequential cancer mutations in cultured human intestinal stem cells. *Nature* **521**, 43-47, doi:10.1038/nature14415 (2015).

26 Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457**, 608-611, doi:10.1038/nature07602 (2009).

27 Chu, G. C., Dunn, N. R., Anderson, D. C., Oxburgh, L. & Robertson, E. J. Differential requirements for Smad4 in TGFbeta-dependent patterning of the early mouse embryo. *Development* **131**, 3501-3512, doi:10.1242/dev.01248 (2004).

28 Costello, I., Biondi, C. A., Taylor, J. M., Bikoff, E. K. & Robertson, E. J. Smad4-dependent pathways control basement membrane deposition and endodermal cell migration at early stages of mouse development. *BMC Dev Biol* **9**, 54, doi:10.1186/1471-213X-9-54 (2009).

29 Hassan, A. B. & Howell, J. A. Insulin-like growth factor II supply modifies growth of intestinal adenoma in Apc(Min/+) mice. *Cancer Res* **60**, 1070-1076 (2000).

30 Matheson, J. *et al.* Epithelial-mesenchymal transition and nuclear beta-catenin induced by conditional intestinal disruption of Cdh1 with Apc is E-cadherin EC1 domain dependent. *Oncotarget* **7**, 69883-69902, doi:10.18632/oncotarget.11513 (2016).

31 Sato, T. *et al.* Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* **141**, 1762-1772, doi:10.1053/j.gastro.2011.07.050 (2011).

32 Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333-339, doi:10.1038/nature24489 (2017).

33 Yan, K. S. *et al.* Non-equivalence of Wnt and R-spondin ligands during Lgr5(+) intestinal stem-cell self-renewal. *Nature* **545**, 238-242, doi:10.1038/nature22313 (2017).

34 Munoz, J. *et al.* The Lgr5 intestinal stem cell signature: robust expression of proposed quiescent '+4' cell markers. *EMBO J* **31**, 3079-3091, doi:10.1038/emboj.2012.166 (2012).

35 Grun, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251-255, doi:10.1038/nature14966 (2015).

36 Mustata, R. C. *et al.* Identification of Lgr5-independent spheroid-generating progenitors of the mouse fetal intestinal epithelium. *Cell Rep* **5**, 421-432, doi:10.1016/j.celrep.2013.09.005 (2013).

37 Ayyaz, A. *et al.* Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. *Nature* **569**, 121-125, doi:10.1038/s41586-019-1154-y (2019).

38 Kitamura, T. *et al.* SMAD4-deficient intestinal tumors recruit CCR1+ myeloid cells that promote invasion. *Nat Genet* **39**, 467-475, doi:10.1038/ng1997 (2007).

39 Takaku, K. *et al.* Gastric and duodenal polyps in Smad4 (Dpc4) knockout mice. *Cancer Res* **59**, 6113-6117 (1999).

40 Takaku, K. *et al.* Intestinal tumorigenesis in compound mutant mice of both Dpc4 (Smad4) and Apc genes. *Cell* **92**, 645-656, doi:10.1016/s0092-8674(00)81132-0 (1998).

41 Takaku, K., Wrana, J. L., Robertson, E. J. & Taketo, M. M. No effects of Smad2 (madh2) null mutation on malignant progression of intestinal polyps in Apc(delta716) knockout mice. *Cancer Res* **62**, 4558-4561 (2002).

42 Freeman, T. J. *et al.* Smad4-mediated signaling inhibits intestinal neoplasia by inhibiting expression of beta-catenin. *Gastroenterology* **142**, 562-571 e562, doi:10.1053/j.gastro.2011.11.026 (2012).

43 Fumagalli, A. *et al.* Genetic dissection of colorectal cancer progression by orthotopic transplantation of engineered cancer organoids. *Proc Natl Acad Sci U S A* **114**, E2357-E2364, doi:10.1073/pnas.1701219114 (2017).

44 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).

45 Fijneman, R. J. *et al.* Expression of Pla2g2a prevents carcinogenesis in Muc2-deficient mice. *Cancer Sci* **99**, 2113-2119, doi:10.1111/j.1349-7006.2008.00924.x (2008).

46 Dessein, R. *et al.* Toll-like receptor 2 is critical for induction of Reg3 beta expression and intestinal clearance of Yersinia pseudotuberculosis. *Gut* **58**, 771-776, doi:10.1136/gut.2008.168443 (2009).

47 Haapamaki, M. M. *et al.* Gene expression of group II phospholipase A2 in intestine in ulcerative colitis. *Gut* **40**, 95-101, doi:10.1136/gut.40.1.95 (1997).

48 He, X. C. *et al.* BMP signaling inhibits intestinal stem cell self-renewal through suppression of Wnt-beta-catenin signaling. *Nat Genet* **36**, 1117-1121, doi:10.1038/ng1430 (2004).

49 Liang, Y. Y., Brunicardi, F. C. & Lin, X. Smad3 mediates immediate early induction of Id1 by TGF-beta. *Cell Res* **19**, 140-148, doi:10.1038/cr.2008.321 (2009).

50 Kang, Y., Chen, C. R. & Massague, J. A self-enabling TGFbeta response coupled to stress signaling: Smad engages stress response factor ATF3 for Id1 repression in epithelial cells. *Mol Cell* **11**, 915-926, doi:10.1016/s1097-2765(03)00109-6 (2003).

51 Tang, Y., Zhang, S., Li, J., Wu, C. & Fan, Q. ID1 marks the tumorigenesis of pancreatic ductal adenocarcinoma in mouse and human. *Sci Rep* **12**, 13555, doi:10.1038/s41598-022-17827-3 (2022).

52 Huang, Y. H. *et al.* ID1 Mediates Escape from TGFbeta Tumor Suppression in Pancreatic Cancer. *Cancer Discov* **10**, 142-157, doi:10.1158/2159-8290.CD-19-0529 (2020).

53 Roschger, C. & Cabrele, C. The Id-protein family in developmental and cancer-associated pathways. *Cell Commun Signal* **15**, 7, doi:10.1186/s12964-016-0161-y (2017).

54 Niu, B. *et al.* Interplay between transforming growth factor-beta and Nur77 in dual regulations of inhibitor of differentiation 1 for colonic tumorigenesis. *Nat Commun* **12**, 2809, doi:10.1038/s41467-021-23048-5 (2021).

55 Gao, Z. *et al.* PAK3 promotes the metastasis of hepatocellular carcinoma by regulating EMT process. *J Cancer* **13**, 153-161, doi:10.7150/jca.61918 (2022).

56 Ma, T., Li, H., Yang, W., Liu, Q. & Yan, H. Over-expression of miR-193a-3p regulates the apoptosis of colorectal cancer cells by targeting PAK3. *Am J Transl Res* **14**, 1361-1375 (2022).

57 Tan, X. *et al.* Loss of Smad4 promotes aggressive lung cancer metastasis by de-repression of PAK3 via miRNA regulation. *Nat Commun* **12**, 4853, doi:10.1038/s41467-021-24898-9 (2021).

58 Carter, J. H. *et al.* Pak-1 expression increases with progression of colorectal carcinomas to metastasis. *Clin Cancer Res* **10**, 3448-3456, doi:10.1158/1078-0432.CCR-03-0210 (2004).

59 Gong, W. *et al.* P21-activated kinase 5 is overexpressed during colorectal cancer progression and regulates colorectal carcinoma cell adhesion and migration. *Int J Cancer* **125**, 548-555, doi:10.1002/ijc.24428 (2009).

60 Huynh, N. *et al.* Depletion of p21-activated kinase 1 up-regulates the immune system of APC(Δ14/+) mice and inhibits intestinal tumorigenesis. *BMC Cancer* **17**, 431, doi:10.1186/s12885-017-3432-0 (2017).

61 King, H., Nicholas, N. S. & Wells, C. M. Role of p-21-activated kinases in cancer progression. *Int Rev Cell Mol Biol* **309**, 347-387, doi:10.1016/B978-0-12-800255-1.00007-7 (2014).

62 Li, C. *et al.* High-level expression of P21-Cdc/Rac-activated kinase 7 is closely related to metastatic potential and poor prognosis of colon carcinoma. *Oncotarget* **7**, 46042-46055, doi:10.18632/oncotarget.10017 (2016).

63 Li, X. & Li, F. p21-Activated Kinase: Role in Gastrointestinal Cancer and Beyond. *Cancers (Basel)* **14**, doi:10.3390/cancers14194736 (2022).

64 Song, B., Wang, W., Zheng, Y., Yang, J. & Xu, Z. P21-activated kinase 1 and 4 were associated with colorectal cancer metastasis and infiltration. *J Surg Res* **196**, 130-135, doi:10.1016/j.jss.2015.02.035 (2015).

65 Malla, S. B. *et al.* Pathway level subtyping identifies a slow-cycling biological phenotype associated with poor clinical outcomes in colorectal cancer. *Nat Genet*, doi:10.1038/s41588-024-01654-5 (2024).

66 Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature* **459**, 262-265, doi:10.1038/nature07935 (2009).

67 Lee, G. Y., Kenny, P. A., Lee, E. H. & Bissell, M. J. Three-dimensional culture models of normal and malignant breast epithelial cells. *Nat Methods* **4**, 359-365, doi:10.1038/nmeth1015 (2007).

68 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).

69 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

70 Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).

71 Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523, doi:10.1038/s41467-019-09234-6 (2019).

72 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).

73 Zappia, L. & Oshlack, A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience* **7**, doi:10.1093/gigascience/giy083 (2018).

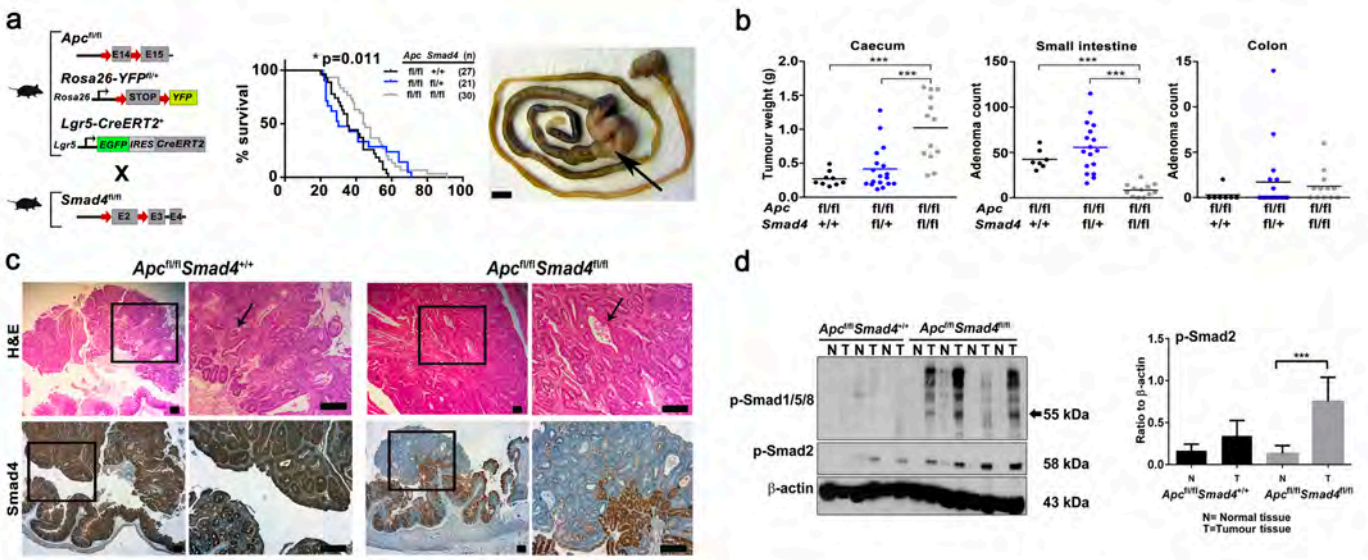
74 Hu, C. *et al.* CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res* **51**, D870-D876, doi:10.1093/nar/gkac947 (2023).

75 Miller, S. A. *et al.* LSD1 and Aberrant DNA Methylation Mediate Persistence of Enteroendocrine Progenitors That Support BRAF-Mutant Colorectal Cancer. *Cancer Res* **81**, 3791-3805, doi:10.1158/0008-5472.CAN-20-3562 (2021).

76 Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278, doi:10.1186/s13059-015-0844-5 (2015).

77 Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416 e411, doi:10.1016/j.cell.2018.02.052 (2018).

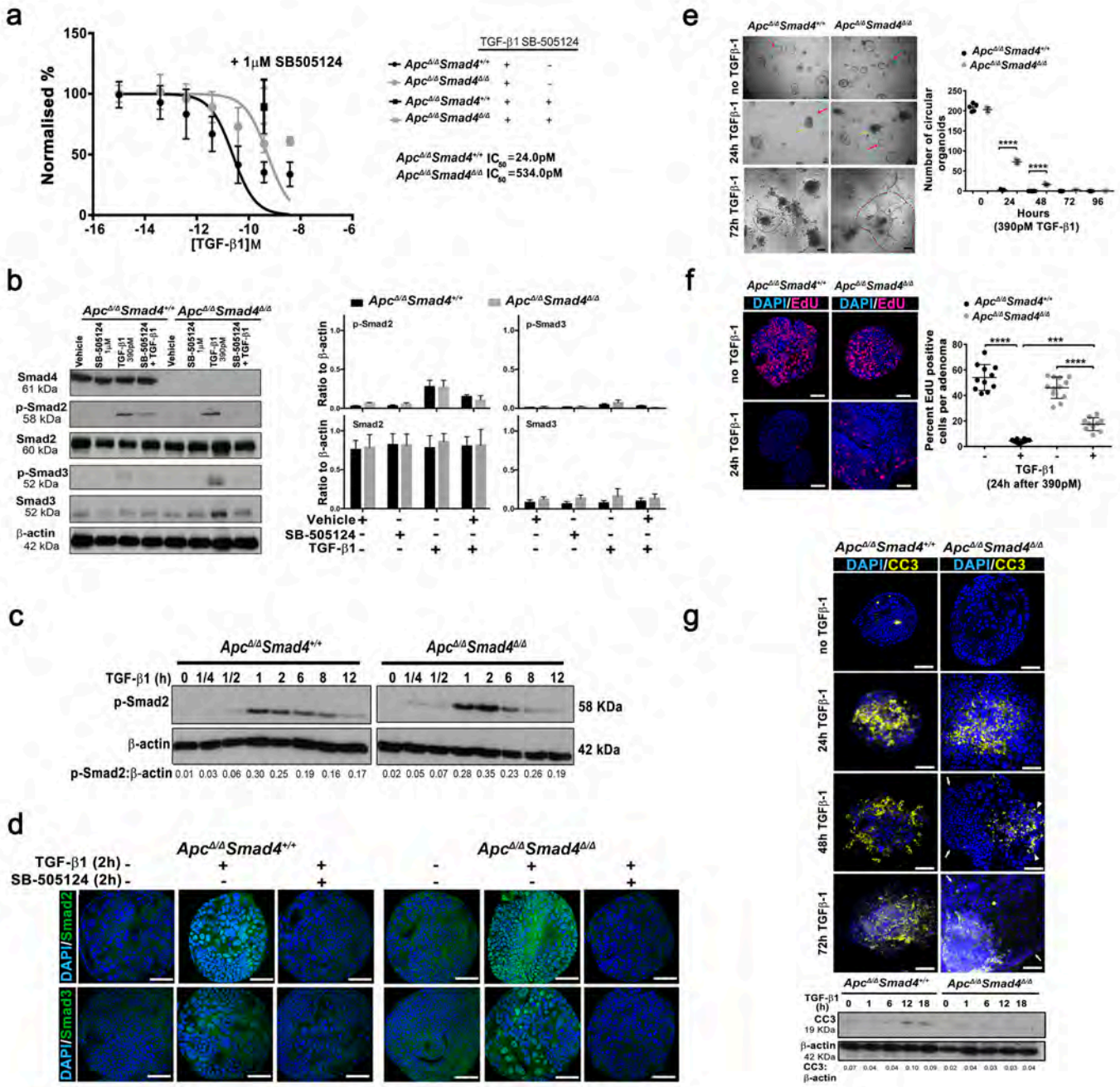
Fig. 1



Conditional disruption of *Apc* and *Smad4* using *Lgr5-CreER^{T2}* results in discordant adenoma growth in the caecum and small intestine.

a. Breeding strategy to combine loxP conditional alleles with conditional *Lgr5-CreER^{T2}* prior to tamoxifen injection in adult mice. Kaplan-Meier survival of *Apc$^{fl/fl}$Smad4$^{+/+}$*, *Apc$^{fl/fl}$Smad4$^{fl/+}$* and *Apc$^{fl/fl}$Smad4$^{fl/fl}$* mice post-Cre induction. Comparison between *Apc$^{fl/fl}$Smad4$^{+/+}$* and *Apc$^{fl/fl}$Smad4$^{fl/fl}$* genotypes, *p=0.011, Log-rank test. Caecal enlargement (arrow) from *Apc$^{fl/fl}$Smad4$^{fl/fl}$* mice. Bar 1cm. **b.** Caecal (adenoma) weight, small intestine and colon total adenoma from *Apc$^{fl/fl}$Smad4$^{+/+}$* (n=9), *Apc$^{fl/fl}$Smad4$^{fl/+}$* (n=17) and *Apc$^{fl/fl}$Smad4$^{fl/fl}$* (n=13). Horizontal bars; mean. Statistical analysis one-way ANOVA with Bonferroni's multiple comparison test. ***p<0.001. **c**. Haematoxylin and Eosin (H&E) stained histological sections (H&E) and Smad4 antibody immunohistochemistry localisation in *Apc$^{fl/fl}$Smad4$^{+/+}$* and *Apc$^{fl/fl}$Smad4$^{fl/fl}$* caecal adenoma. Arrows show crypt abscesses, upper panel. Smad4 localisation is lost in adenoma and confined to adjacent normal colonic crypt cells in *Apc$^{fl/fl}$Smad4$^{fl/fl}$* mice. Scale bar 100µm. **b.** Western blot of adenoma (T) and normal tissue (N) from *Apc$^{fl/fl}$Smad4$^{+/+}$* and *Apc$^{fl/fl}$Smad4$^{f/fl}$* with anti- phospho-Smad2 (p-Smad2) and anti-phospho-Smad1/5/8 antibodies. Densitometric analysis for p-Smad2 represented as ratio to β-actin (for *Apc$^{fl/fl}$Smad4$^{+/+}$* n=3 and for *Apc$^{fl/fl}$Smad4$^{fl/fl}$* n=4 experimental replicates). Significant increase in p-Smad2 in adenoma (T) compared to normal tissues (N) in *Apc$^{fl/fl}$Smad4$^{fl/fl}$* mice. ***p<0.001. Non-parametric one-way ANOVA with Dunn's multiple comparison test. Mean ± S.D.

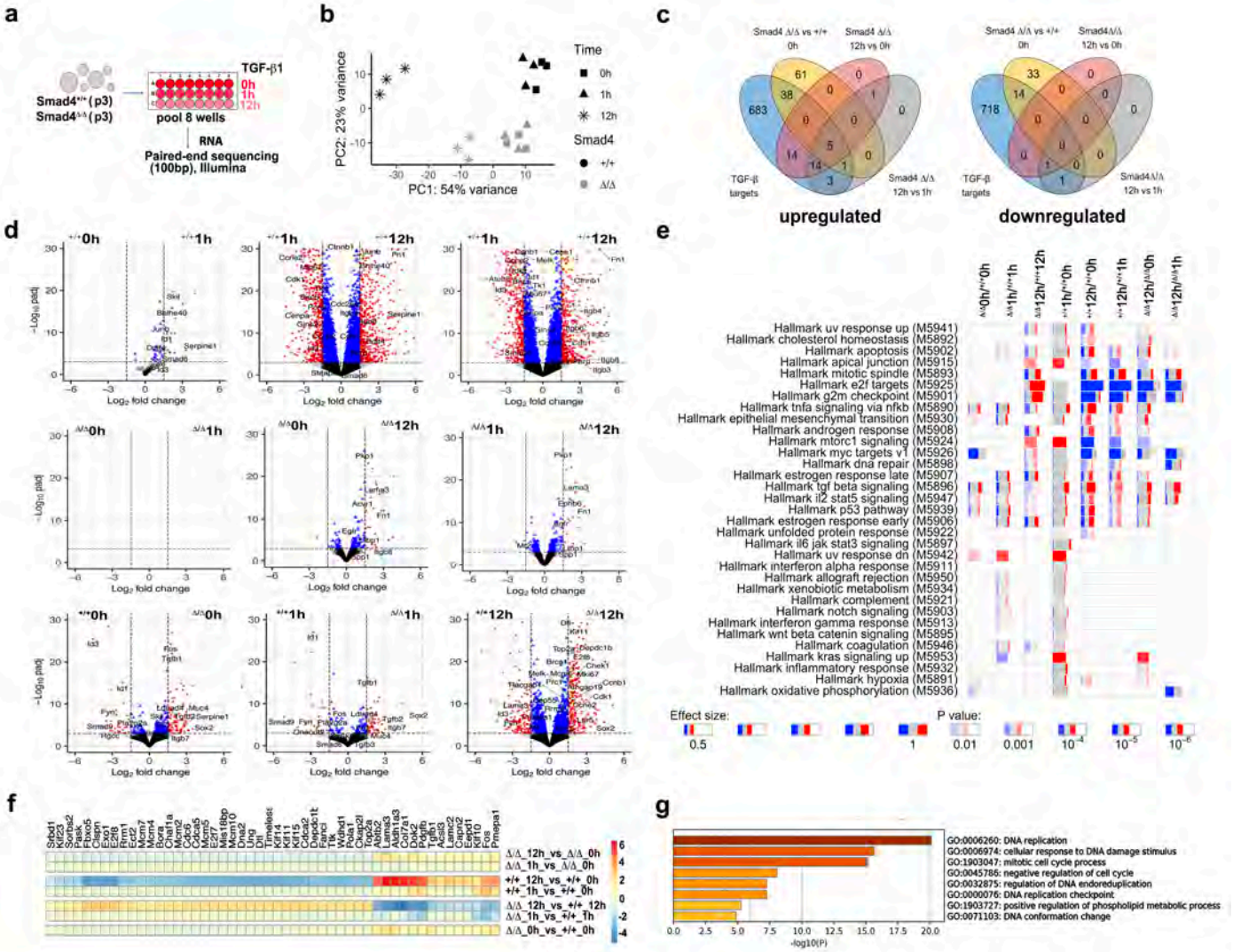
Fig. 2


a
+ 1μM SB505124
Normalised %
[TGF-β1]M
TGF-β1 SB-505124
Apc Δ/Δ Smad4 +/+
Apc Δ/Δ Smad4 Δ/Δ
IC50 =24.0pM
IC50 =534.0pM
b
Vehicle
SB-505124 1μM
TGF-β1 390pM
SB-505124 + TGF-β1
Smad4 61 kDa
p-Smad2 58 kDa
Smad2 60 kDa
p-Smad3 52 kDa
Smad3 52 kDa
β-actin 42 kDa
Ratio to β-actin
p-Smad2
p-Smad3
Smad2
Smad3
c
TGF-β1 (h) 0 1/4 1/2 1 2 6 8 12
p-Smad2
β-actin
58 KDa
42 kDa
p-Smad2:β-actin 0.01 0.03 0.06 0.30 0.25 0.19 0.16 0.17
0.02 0.05 0.07 0.28 0.35 0.23 0.26 0.19
d
TGF-β1 (2h)
SB-505124 (2h)
DAPI/Smad2
DAPI/Smad3
e
no TGFβ-1
24h TGFβ-1
72h TGFβ-1
Number of circular organoids
Hours
(390pM TGF-β1)
f
DAPI/EdU
Percent EdU positive cells per adenoma
TGF-β1
(24h after 390pM)
g
DAPI/CC3
48h TGFβ-1
TGF-β1 (h) 0 1 6 12 18
CC3 19 KDa
β-actin 42 KDa
CC3: β-actin 0.07 0.04 0.04 0.10 0.09 0.02 0.04 0.03 0.03 0.04

TGF-β1 and Smad4 dependent signalling in intestinal adenoma organoids.

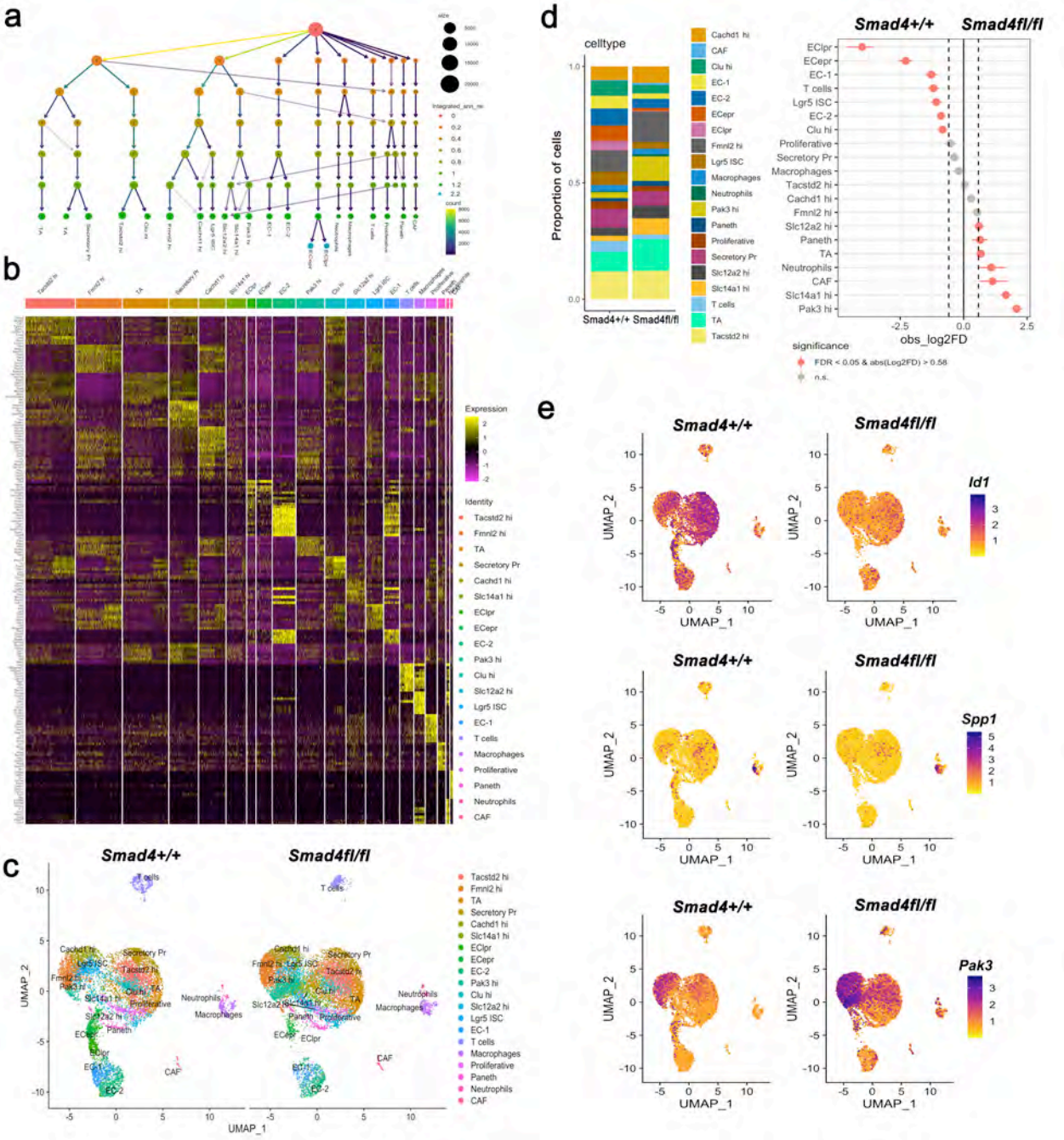
a. TGF-β1 induced growth response inhibition of adenoma organoids. Organoids (Day 2) from *Apc*$^{\Delta/\Delta}$*Smad4*$^{+/+}$ and *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$ mice were incubated with increasing doses of TGF-β1 (range: 0.039-3900pM) with growth quantified by alamarBlue (AB) measured at 24h. Data normalized to untreated control (100%) and Matrigel only (0%) (n=4 experimental replicates). Growth inhibition of 390pM TGF-β1 rescued by pre-treatment with SB-505124 (ALK4,5,7 inhibitor) for 2hr (1mM). Error bars ± S.D. **b**. Representative immunoblots of Smad4, total R-Smads and p-R-Smads of extracts from *Apc*$^{\Delta/\Delta}$*Smad4*$^{+/+}$ and *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$ adenoma organoids. Increased phosphorylation of Smad2 and Smad3 following TGF-β1 treatment decreased with SB-505124. P-Smad2/ Smad2 and p-Smad3/ Smad3 densitometry relative to β-actin (n=3 experimental replicates). Statistical analysis non-parametric one way-ANOVA with Dunn's multiple comparison. **c.** Immunoblots of p-Smad2 at different time points (h) following TGF-β1 (390pM) showing similar time course of activation (densitometry ratios of p-Smad2/β-actin shown). **d.** Confocal immunofluorescence images of Smad2 and Smad3 nuclear localisation (DAPI) following TGF-β1 (390pM) exposure and SB-505124 inhibition. Scale bar 50µm. **e**. Bright field confocal images of organoids 24-72h following TGF-β1 exposure. Red arrows indicate examples of intact spherical (circular) organoid shape, yellow arrows show organoids with shape change, red dotted line indicates organoids with marked disruption, cell spreading and cellular transformation. Quantification shows relative preservation of spheroid (circular) organoids in *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$ samples. One-way ANOVA with Sidak's multiple comparison test ****$p<0.0001$. **f**. Suppression of organoid proliferation following TGF-β1 treatment assessed with EdU incubation for 2hr. Confocal images of EdU incorporation (green) in nuclei (DAPI- blue) from *Apc*$^{\Delta/\Delta}$*Smad4*$^{+/+}$ and *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$ and quantification using ImageJ (n=10 organoids). One-way ANOVA with Sidak's multiple comparison test *** $p<0.001$, ****, $p<0.0001$, respectively. **g**. Cleaved Caspase 3 (CC3) activation following 24-72hr of TGF-β1 (390pM). CC3 (arrowheads) labelling in confocal images appeared greater in *Apc*$^{\Delta/\Delta}$*Smad4*$^{+/+}$ compared to *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$, supported by immunoblots. Arrows indicate margin of organoids. Scale bar 50µm.

Fig. 3



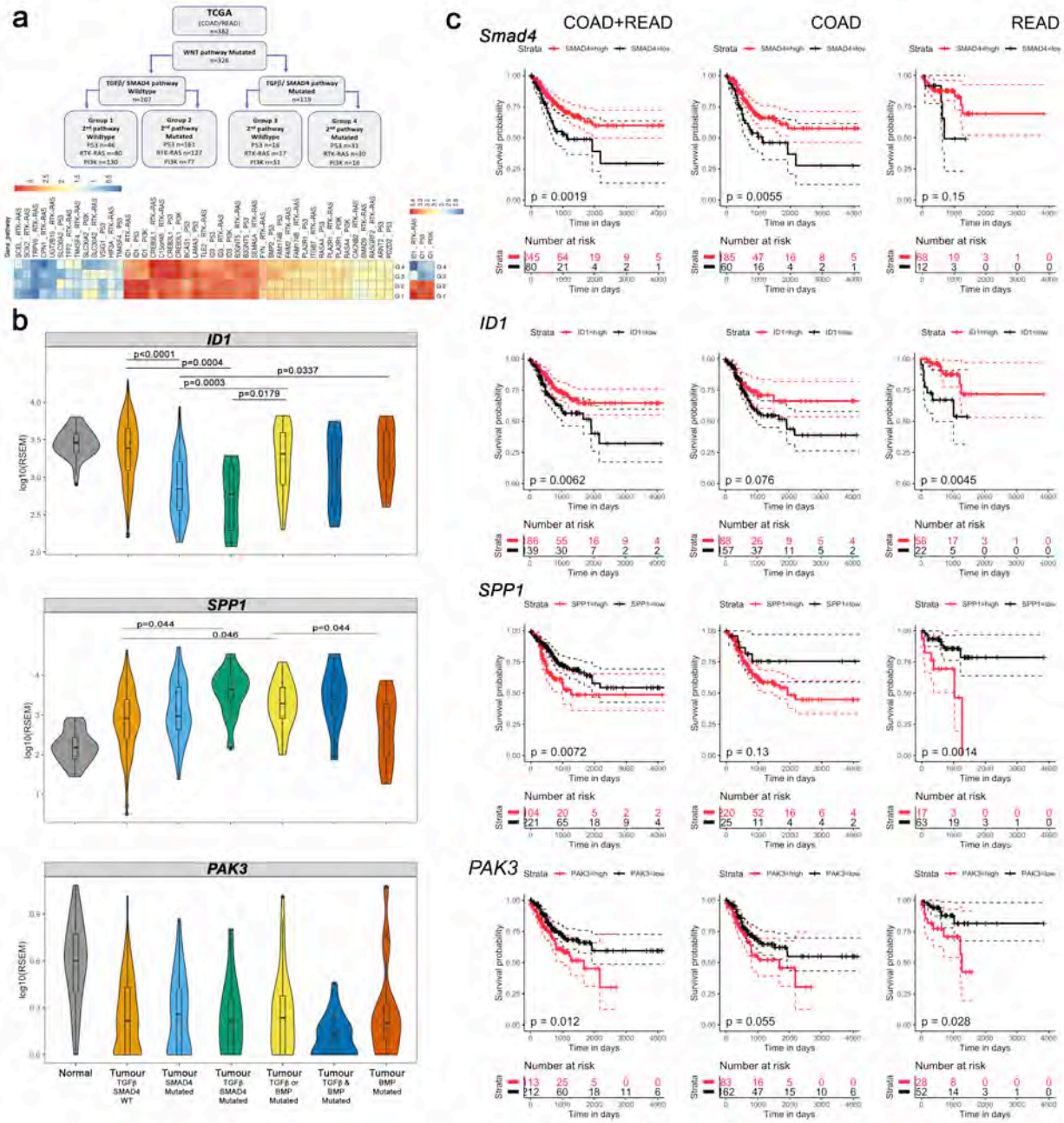
TGF-β1 and Smad4 dependent bulk gene expression signatures in adenoma organoids.
a. Adenoma organoids of *Apc*$^{\Delta/\Delta}$*Smad4*$^{+/+}$ and *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$ genotypes were exposed to TGF-β1 (390pM) for 0, 1 and 12hr, eight wells were pooled for paired-end bulk RNA-Seq and the experiment repeated three times. **b.** Principal components of transcriptional clusters according to genotype and time. Data have been vst-transformed. **c**. Venn diagram classifies TGF-β1 dependent Smad4 down-regulated and upregulated genes. Genes used have 1.5 log2FC (1.5) and adjusted p-value <0.05. **d.** Volcano plots of differentially expressed genes comparing time and genotype ($^{+/+}$=*Apc*$^{\Delta/\Delta}$*Smad4*$^{+/+}$, $^{\Delta/\Delta}$ = *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$). In all volcano plots, up to 30 genes are shown with -log(p-adj). Cut-off values for log2FC (1.5) and adjusted-p-value (1e-3) are marked with dashed lines. *Apc*$^{\Delta/\Delta}$*Smad4*$^{\Delta/\Delta}$ DEGs are much reduced in number following TGF-β1 treatment. **e**. Enriched gene modules of DEG using the hallmark gene set enrichment (GSEA with tmod package). Comparisons shown in each column. Red and blue indicate the proportion of genes in a module that are either upregulated or downregulated, respectively. Width of each box relates to its effect size, where lighter, less saturated colours indicate lower significance with p-values, enrichment using "tmod" package. **f**. Heatmap showing time-dependent changes in expression (log2FC) for the 50 genes with the lowest adjusted p-values and **g**. pathway enrichment for these genes are shown generated using Metascape.

Fig. 4



***Smad4* dependent adenoma cell populations using single cell RNA-Seq from *in vivo* caecal adenoma.**
a. Cluster tree shows the resolution from 0 to 1.2 labelled with each cluster name. The EClpr and ECepr clusters were identified at resolution 2.2 and added to the tree. **b**. The heatmap shows the scaled expression of the top 5 genes identified in each cluster (genes selected based on the biggest difference between the cluster compared to all clusters). Lgr5 ICS = Lgr5 Intestinal stem cell; EC =enterocyte; EClpr =enterocyte late precursor; ECepr =enterocyte early precursor; TA= transit amplifying; CAF =cancer associated fibroblast; Secretory Pr =secretory precursor; hi =high referring to specific genes. **c**. UMAP plot shows the cell-type clusters of single cells from *Apc*$^{fl/fl}$*Smad4*$^{+/+}$ (9674, n=3) and *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$ (11212, n=5) excised adenoma based on the expression of known marker genes. **d**. The bar graph shows the frequencies of clusters of the single cells in *Apc*$^{fl/fl}$*Smad4*$^{+/+}$ and *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$ adenoma. A point range plot shows the confidence interval for the absolute log2FC (obs_log2FC) for the different identified cell types. Significant results are labelled in red. Note the differential enrichment of *Pak3*^{high}, *Slc14a1*^{high}, CAF and neutrophil populations and the reduction in ECepr and EClpr populations in *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$ adenoma. **e**. UMAP plots comparing single cell expression of *ID1* (suppression in all cell types in *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$), *Spp1* and *Pak3* (increased expression in progenitor enterocyte population of *Apc*$^{fl/fl}$*Smad4*$^{fl/fl}$).

Fig. 5



***SMAD4, ID1, SPP1 and PAK3* expression biomarkers for Smad4-TGF-β gene-pathway and human colorectal cancer (TCGA)**

a. Flow diagram of TCGA bulk somatic DNA and RNA NGS sequencing data (COAD =colon adenocarcinoma, READ =rectal adenocarcinoma) with WNT pathway mutations (n=326) for validation of significant tumour DEGs for SMAD4 dependent genes depending on patient groups with or without TGF-β1/SMAD4 pathway mutations, co-existing common pathway mutations in either RTK-RAS, TP53 or PI3K pathways. Median RSEM expression (Log10 value) for the significant genes validated in the TCGA COAD-READ cohorts (left heatmap) based on *SMAD4* mutations in the context of P53, RTK-RAS, and PI3K pathway mutations. *ID1* median expression when patients with BMP pathway mutations excluded is shown separately (right heatmap). **b**. *ID1, SPP1 and PAK3 mRNA* expression in cases with somatic pathogenic mutations in either *SMAD4* alone, *SMAD4* with other genes in the TGF-β/BMP pathway (TGF-β *SMAD4* mutated), single gene mutation in either the TGF-β or BMP pathway (TGF-β or BMP mutated), more than one mutation in TGF-β /BMP (TGF-β and BMP mutated) or mutations of BMP pathway genes alone (BMP mutated). Statistical analysis was performed using Kruskal-Wallis test ($p < 0.0001$) followed by Dunn test multiple comparison test adjusted by Holm test (shown in the plot). Note the significant repression of *ID1* and increase in *SPP1* in *SMAD4* mutated cases. **c.** Kaplan–Meier survival curves based on high and low *SMAD4*, *ID1, SPP1 and PAK3* expression. The plots used COAD-READ cohort combined (n=326), COAD alone (n=246), and READ alone (n=80). Statistical analyses were performed using the Log-rank test and the p-values are indicated on the graphs.