Transposon-activated \textit{POU5F1B} promotes colorectal cancer growth and metastasis

Laia Simó-Riudalbas\textsuperscript{1,}\textsuperscript{a}, Sandra Offner\textsuperscript{1}, Evarist Planet\textsuperscript{1}, Julien Duc\textsuperscript{1}, Laurence Abrami\textsuperscript{1}, Sagane Dind\textsuperscript{1}, Alexandre Coudray\textsuperscript{1}, Mairene Coto-Llerena\textsuperscript{2,3}, Caner Ercan\textsuperscript{2,3}, Salvatore Piscuoglio\textsuperscript{2,3}, Claus Lindbjerg Andersen\textsuperscript{4}, Jesper Bertram Bramsen\textsuperscript{4} & Didier Trono\textsuperscript{1,}\textsuperscript{b,}\textsuperscript{c,}\textsuperscript{d,}\textsuperscript{e,}\textsuperscript{f,}\textsuperscript{g,}\textsuperscript{h,}\textsuperscript{i,}\textsuperscript{j,}\textsuperscript{k,}\textsuperscript{l,}\textsuperscript{m,}\textsuperscript{n,}\textsuperscript{o,}\textsuperscript{p,}\textsuperscript{q,}\textsuperscript{r,}\textsuperscript{s,}\textsuperscript{t,}\textsuperscript{u,}\textsuperscript{v,}\textsuperscript{w,}\textsuperscript{x,}\textsuperscript{y,}\textsuperscript{z,}\textsuperscript{ab,}\textsuperscript{ac,}\textsuperscript{ad,}\textsuperscript{ae,}\textsuperscript{af,}\textsuperscript{ag,}\textsuperscript{ah,}\textsuperscript{ai,}\textsuperscript{aj,}\textsuperscript{ak,}\textsuperscript{al,}\textsuperscript{am,}\textsuperscript{an,}\textsuperscript{ao,}\textsuperscript{ap,}\textsuperscript{aq,}\textsuperscript{ar,}\textsuperscript{as,}\textsuperscript{at,}\textsuperscript{au,}\textsuperscript{av,}\textsuperscript{aw,}\textsuperscript{ax,}\textsuperscript{ay,}\textsuperscript{az,}\textsuperscript{ba,}\textsuperscript{bc,}\textsuperscript{bd,}\textsuperscript{be,}\textsuperscript{bf,}\textsuperscript{bg,}\textsuperscript{bh,}\textsuperscript{bi,}\textsuperscript{bj,}\textsuperscript{bk,}\textsuperscript{bl,}\textsuperscript{bm,}\textsuperscript{bn,}\textsuperscript{bo,}\textsuperscript{bp,}\textsuperscript{bq,}\textsuperscript{br,}\textsuperscript{bs,}\textsuperscript{bt,}\textsuperscript{bu,}\textsuperscript{bv,}\textsuperscript{bw,}\textsuperscript{bx,}\textsuperscript{by,}\textsuperscript{bz,}\textsuperscript{ca,}\textsuperscript{cb,}\textsuperscript{cc,}\textsuperscript{cd,}\textsuperscript{ce,}\textsuperscript{cf,}\textsuperscript{cg,}\textsuperscript{ch,}\textsuperscript{ci,}\textsuperscript{cj,}\textsuperscript{ck,}\textsuperscript{cl,}\textsuperscript{cm,}\textsuperscript{cn,}\textsuperscript{co,}\textsuperscript{cp,}\textsuperscript{cq,}\textsuperscript{cr,}\textsuperscript{cs,}\textsuperscript{ct,}\textsuperscript{cu,}\textsuperscript{cv,}\textsuperscript{cw,}\textsuperscript{cx,}\textsuperscript{cy,}\textsuperscript{cz,}\textsuperscript{da,}\textsuperscript{db,}\textsuperscript{dc,}\textsuperscript{dd,}\textsuperscript{de,}\textsuperscript{df,}\textsuperscript{dg,}\textsuperscript{dh,}\textsuperscript{di,}\textsuperscript{dj,}\textsuperscript{dk,}\textsuperscript{dl,}\textsuperscript{dm,}\textsuperscript{dn,}\textsuperscript{do,}\textsuperscript{dp,}\textsuperscript{dq,}\textsuperscript{dr,}\textsuperscript{ds,}\textsuperscript{dt,}\textsuperscript{du,}\textsuperscript{dv,}\textsuperscript{dw,}\textsuperscript{dx,}\textsuperscript{dy,}\textsuperscript{dz}.

The treatment of colorectal cancer (CRC) is an unmet medical need in absence of early diagnosis. Here, upon characterizing cancer-specific transposable element-driven transpochimeric gene transcripts (TcGTs) produced by this tumor in the SYSCOL cohort, we find that expression of the hominid-restricted retrogene \textit{POU5F1B} through aberrant activation of a primate-specific endogenous retroviral promoter is a strong negative prognostic biomarker. Correlating this observation, we demonstrate that \textit{POU5F1B} fosters the proliferation and metastatic potential of CRC cells. We further determine that \textit{POU5F1B}, in spite of its phylogenetic relationship with the \textit{POUSF1}/OCT4 transcription factor, is a membrane-enriched protein that associates with protein kinases and known targets or interactors as well as with cytoskeleton-related molecules, and induces intracellular signaling events and the release of \textit{trans}-acting factors involved in cell growth and cell adhesion. As \textit{POU5F1B} is an apparently non-essential gene only lowly expressed in normal tissues, and as \textit{POU5F1B}-containing TcGTs are detected in other tumors besides CRC, our data provide interesting leads for the development of cancer therapies.

The human genome contains more than 4.5 million inserts derived from transposable elements (TEs). This so-called endovirome is a major motor of genome evolution due its gene-disruptive and recombinogenic potential and because it harbors a high density of transcription factors binding sites\textsuperscript{1,2}. TE-embedded regulatory sequences (TEeRS) can act as promoters, enhancers, repressors, terminators or insulators and, as such, exert profound influences on human development and physiology, from embryonic genome activation to metabolic control, and from brain development to innate immunity\textsuperscript{3,4}. Furthermore, the TE-based regulome confers a high degree of species specificity to the conduct of biological events due to the rapid evolutionary turnover of both its TE constituents and their sequence-specific \textit{trans}-acting controllers\textsuperscript{5,6}.

Alterations of this regulatory system can lead to cancer, as was immediately suggested by the phylogenetic relationship between endogenous retroviruses (ERV) and their exogenous relatives, the long-called RNA tumor viruses that led to the discovery of oncogenes\textsuperscript{7}. Since then, TEs have been linked to human cancer through a variety of observations and mechanisms\textsuperscript{7}, including the promotion of chromosomal rearrangements owing to their repetitiveness\textsuperscript{8}, the inactivation of tumor suppressor genes through insertional mutagenesis\textsuperscript{9}, and the production of long non-coding RNAs or chimeric gene transcripts from aberrantly activated TEeRS\textsuperscript{10,11}. A wide range of oncogene-encoding TE-driven transpochimeric gene transcripts (TcGTs) have been documented in a recent survey of cancer databases\textsuperscript{12}. However, a causal role for these products in oncogenesis has so far seldom been demonstrated.

Colorectal cancer is the second most frequently encountered non-skin malignancy in both women and men. This tumor is increasingly diagnosed in young people, and its treatment is largely unmet...
clinical need unless it is detected at an early stage (https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html). Facilitating molecular explorations, its step-wise development has been well documented, and its surgical resection commonly provides an abundance of clearly distinct normal, tumoral and sometimes metastatic tissue for in-depth molecular analyzes. Here, we used analytical pipelines tailored for the transposcriptome, or sum of TE-derived RNA transscripts present in a cell, to RNA sequencing data obtained from a large cohort of CRC patients. This led us to identify the aberrant TE-driven expression of *POUSF1B*, a hominid-restricted retrogene, as a strong negative predictor of the clinical course of this malignancy and to determine that its *POUSF1B* product is responsible for fostering the growth and metastatic potential of colorectal cancer cells through a combination of cis- and trans-effects.

**Results**

**Cancer-specific TcGTs in CRC**

We sought potentially oncogenic TcGTs through the in-depth analysis of previously published RNA-seq data from primary tumors and matched normal colon of 301 patients diagnosed with colorectal adenoma (11%) or adenocarcinoma (89%), referred to hereafter as the SYSCOL cohort. High-quality sequencing and histopathological data were available for 296 of the 301 patients. Ninety-five TcGTs involving 39 genes were detected in more than 20% of tumors and less than 10% of normal tissue samples (Table 1). About half (44/95) of these TcGTs were predicted to encode the canonical protein and the rest to yield truncated (43) or out-of-frame (8) derivatives (Table 1 and Supplementary Data 1). TcGTs involving only three of these 39 genes (*BMP7*, *RNF43*, and *SLCO1B3*) had previously been identified in cancer13,15.

The tumor-restricted TcGT most frequently detected in our cohort initiated within a primate-specific *LTR66* endogenous retroviral promoter on chromosome 7q24, spliced into several other TE inserts (*AluSx1*, *LTR33*, *L2b*, and *MLTH1*) and ended in *POUSF1B*, a paralog of the *POUSF1/OC74* pluripotency gene. *POUSF1B*-containing TcGTs, about two-thirds starting at *LTR66*, were found in 186 (65%) CRC primary tumors versus 11 (3.8%) samples labeled as normal colon (Fig. 1a ‘cohort 1’. Table 1, Supplementary Fig. 1a, b). Other forms of *POUSF1B* TcGTs starting from some of the intervening TEs were also noted in some tumors, but all were distinct from the annotated *POUSF1B* mono-exonic transcript driven by a promoter partly overlapping with the *MLTH1* sequence (Supplementary Fig. 1b and Supplementary Data 1). We validated these results with another publicly available dataset of 18 CRC patients (*GSE50760*)16, where the *LTR66-POUSF1B* TcGT was found in 0/18 normal colon, 10/18 primary tumors, and 12/18 liver metastases (Fig. 1a ‘cohort 2’). A re-analysis of single-cell RNA-seq (scRNA-seq) data from primary tumors and nearby normal mucosa in 11 CRC patients17 further revealed that *LTR66* and *POUSF1B* were expressed in epithelial cells and only exceptionally detected in other cell types (Supplementary Fig. 1c). Confirming these data, a RACE (5’-rapid amplification of cDNA ends) analysis documented the presence of a non-canonical ~2,500 bp *POUSF1B* TcGT starting within *LTR66* in several CRC cell lines (Supplementary Fig. 1d). In CRC RNA-seq datasets from the SYSCOL cohort (Fig. 1b), the CRC TCGA collection and 54 CRC cell lines from EGAD000010000725 (Supplementary Fig. 1e), high levels of *LTR66- and POU5F1B*-derived RNAs systematically correlated, indicating that they likely corresponded in their majority to *LTR66-POUSF1B* TcGTs.

**Table 1 | Thirty-nine TcGT-transcribed genes were detected in more than 20% of CRC tumors and less than 10% of normal colons**

| Gene       | nN | nT | %N | %T | Protein product |
|------------|----|----|----|----|----------------|
| POU5F1B    | 11 | 186| 3.8| 65.0| canon.         |
| HTR1D      | 9  | 175| 3.1| 61.2| canon.         |
| SLCO1B3    | 6  | 159| 2.1| 55.6| N-trunc.       |
| CMTR1      | 2  | 149| 0.7| 52.1| C-trunc.       |
| ABUIM2     | 1  | 111| 0.3| 38.8| N-ext.         |
| NCAED      | 28 | 124| 9.8| 43.4| canon.         |
| LTR66A3    | 25 | 119| 8.7| 41.6| N-trunc.; N-C-trunc. |
| PMPC8      | 7  | 94 | 2.4| 32.9| N-C-trunc.; out-frame; C-trunc. |
| CAPN12     | 2  | 75 | 0.7| 26.2| canon.; C-trunc. |
| FGGY       | 8  | 80 | 2.8| 28.0| canon.; N-trunc. |
| TRAF2      | 8  | 80 | 2.8| 28.0| canon.; N-ext. |
| ATP5J2     | 13 | 85 | 4.5| 29.7| canon.         |
| RNF43      | 3  | 71 | 1.0| 24.8| canon.; N-ext. |
| EDAR       | 0  | 66 | 0.0| 23.1| N-ext.         |
| TM9SF2     | 4  | 70 | 1.4| 24.5| canon.         |
| CEP72      | 25 | 91 | 8.7| 31.8| N-trunc.; N-C-trunc. |
| TSSC1      | 9  | 73 | 3.1| 25.5| N-trunc.; C-ext. |
| KRT8       | 7  | 67 | 2.4| 23.4| N-ext.         |
| WDFC3      | 24 | 83 | 8.4| 29.0| N-trunc.       |
| PALD1      | 16 | 74 | 5.6| 25.9| canon.; C-trunc.; out-frame |
| SFB2       | 12 | 69 | 4.2| 24.1| out-frame; N-trunc. |
| SPAC3A3    | 3  | 58 | 1.0| 20.3| N-ext.         |
| MCCC2      | 6  | 61 | 2.1| 21.3| out-frame; N-C-trunc. |
| POMZP3     | 20 | 70 | 7.0| 24.5| canon.; C-trunc. |
| TNRC1B     | 22 | 72 | 7.7| 25.2| N-trunc.; N-C-trunc. |
| BMP7       | 12 | 62 | 4.2| 21.7| out-frame |
| ABHD2      | 13 | 63 | 4.5| 22.0| canon.         |
| TGF2       | 14 | 63 | 4.9| 22.0| canon.         |
| ST3GAL2    | 23 | 70 | 8.0| 24.5| canon.         |
| ZNF283     | 19 | 66 | 6.6| 23.1| canon.; out-frame |
| ADAP1      | 16 | 63 | 5.6| 22.0| canon.         |
| CLSTN3     | 23 | 66 | 8.0| 23.1| canon.         |
| CLDN4      | 16 | 58 | 5.6| 20.3| canon.         |
| FAM10A     | 17 | 59 | 5.9| 20.6| canon.         |
| DAP3       | 27 | 67 | 9.4| 23.4| N-trunc.; N-C-trunc. |
| PTPDC1     | 27 | 67 | 9.4| 23.4| canon.; N-trunc. |
| PMFB1      | 20 | 58 | 7.0| 20.3| canon.; C-trunc.; N-C-trunc. |
| TSLN4      | 26 | 63 | 9.1| 22.0| N-C-trunc. |
| ZNF710     | 26 | 60 | 9.1| 21.0| N-ext.         |

*nN number of normal samples, nT number of tumors, %N percentage in normal colon, %T percentage in tumors, canon. canonical protein, N-trunc. N-terminal truncated protein, C-trunc. C-terminal truncated protein, N-C-trunc. N-C-terminal truncated protein, N-ext. N-terminal extended protein, out.frame out of frame protein.

**LTR66-POUSF1B** is a negative CRC prognostic marker in the SYSCOL cohort

We noted that increased expression of *LTR66-POUSF1B* in the SYSCOL cohort coincided with the transition from normal tissue to adenoma and from adenoma to carcinoma (Fig. 1c), which strongly suggested that *LTR66-POUSF1B* was a marker of advanced disease. Confirming this hypothesis, a Kaplan-Meier representation of overall survival and a Cox Proportional Hazards survival analysis of the SYSCOL cohort revealed that this TcGT was associated with lower overall survival (*P* = 0.0007; hazard ratio (HR) = 2.52, 95% CI = 1.41–4.51) (Fig. 1d) and shorter relapse-free survival (*P* = 0.038; HR = 8.61, 95% CI = 1.15–64.71) (Fig. 1e) in stage II and III patients. The negative prognostic value of *POUSF1B* was also significant when stage I and IV patients were
included ($P = 0.004$; HR = 1.32, 95% CI = 1.08–1.61) but their numbers ($n = 42$ and $n = 19$, respectively) were insufficient to conclude that POU5F1B was also influential at these stages. Importantly, POU5F1B-encoding TcGTs were expressed in both microsatellite stable (MSS) and microsatellite unstable (MSI) tumors, and in all four CRC consensus molecular subtypes (CMS) $^{19}$ (Supplementary Fig. 1f), although they were found more frequently in the MSS and CMS2 subsets. CMS2 tumors are characterized by marked WNT and MYC signaling activation $^{19}$, and the MYC coding sequence is located some 300 kb downstream of POU5F1B. Yet, while there was a significant correlation between POU5F1B and MYC expression in CRC cell lines and tumor samples, the presence of many outliers indicated that the two genes were not systematically coregulated (Supplementary Fig. 1g). Furthermore, while TcGT-driven POU5F1B overexpression was associated...
with shorter survival in stages II and III patients, this was not the case for MYC according to either univariate or POUSF1B/MYC multivariate analyses (Supplementary Fig. 1h). Of note, whereas these data established the LTR66-POUSF1B TcGT as a negative CRC prognostic marker in the SYSCOL cohort, we did not observe this correlation to be significant in the TCGA dataset (Supplementary Fig. II).

POUSF1B enhances the growth and metastatic potential of CRC cells

To investigate the functional consequences of POUSF1B upregulation in colorectal cancer, we used a combination of in vitro and in vivo studies. Stable transduction with a lentiviral vector expressing an HA-tagged form of POUSF1B (Fig. 2a) stimulated the colony forming and proliferation abilities of SW480 (Fig. 2b, c), HT29, and SW620 (Supplementary Fig. 2a, b) cells. Moving to xenotransplantation experiments, we found that subcutaneous injection of luciferase-containing POUSF1B-overexpressing SW480 cells in NOD/SCID gamma (NSG) immunodeficient mice induced the formation of tumors that were bigger and heavier than with GFP-overexpressing controls (Fig. 2d). When equal-size fragments of these tumors were then implanted in the cecum of another set of NSG mice, total body bioluminescence monitoring yielded higher signals in SW480-POUSF1B versus SW480-GFP-engrafted animals (Fig. 2e). The transplanted tumor fragments themselves presented roughly similar masses and volumes at the implantation site (Supplementary Fig. 2c), suggesting that the fitness of well-constituted tumors may be less critically dependent on POUSF1B than their initial establishment and growth. However, SW480-POUSF1B recipient mice presented greater numbers of macroscopically visible metastases in the liver and of additional tumor-invaded organs (Fig. 2f), which explained their overall higher level of bioluminescence (Fig. 2e). Finally, when POUSF1B-overexpressing SW480 cells were injected in the spleen of NSG mice and these animals splenectomized rapidly thereafter, they induced markedly higher numbers of liver metastases than their GFP-expressing controls (Fig. 2g). To extend these data, we repeated the intra-splenic injection experiment with similarly modified SW620 CRC cells, and observed that POUSF1B-overexpressing cells yielded higher densities of liver metastases and greater numbers of extra-hepatic metastases than control cells (Supplementary Fig. 2d).

We then turned to in vivo loss-of-function experiments using POUSF1B-expressing CRC cell lines. Downregulating POUSF1B by dox-inducible lentivector-mediated RNA interference impaired the growth of LS1034 (Fig. 2h, Supplementary Fig. 2e), HT55 and LS174T CRC cells (Supplementary Fig. 2f, g), all of which express this TcGT at baseline. Mice kept on doxycycline for one week beforehand were then injected subcutaneously with dox-treated control or POUSF1B knockdown LS1034 cells. Tumor growth was markedly decreased after injection of POUSF1B-depleted cells, confirming the pro-oncogenic effect of this factor (Fig. 2i). Furthermore, following intrasplenic injection and splenectomy, LTR66-POUSF1B knockdown LS1034 cells also displayed a decreased metastatic phenotype both in and out of the liver, compared with control cells (Supplementary Fig. 2h).

LTR66-POUSF1B TcGTs result from aberrant activation of an intronic enhancer

To determine what triggered the production of LTR66-POUSF1B TcGTs, we analyzed molecular changes occurring at their source locus, located in a gene desert that harbors previously identified risk loci for several epithelial cancers such as the GWAS prostate and colorectal cancer-associated SNP rs6983267. This SNP resides 15 kb upstream of POUSF1B in a regulatory sequence previously described as an enhancer for MYC. For samples for which sequencing depth allowed an assessment of the allelic mode of expression of LTR66-POUSF1B, it was found to be biallelic in a significant fraction (Supplementary Fig. 3a). In addition, detection of this TcGT correlated neither with the rs6983267 genotype in the SYSCOL tumors nor with copy number variations (CNVs) in the TCGA CRC database (Supplementary Fig. 3b, c). We thus turned to chromatin analyzes. By exploring publicly available data, we found that the rs6983267-containing locus physically interacts with LTR66 in three CRC cell lines and that both regions are characterized by ATAC-seq (transposase-accessible chromatin sequencing) peaks in CRC tumor samples from TCGA (Fig. 3a and Supplementary Fig. 3d). This suggested that the rs6983267-containing sequence acts as an LTR66-POUSF1B intronic enhancer. Accordingly, this locus was enriched in the activation marks H3K4me1 and H3K27ac in LS1034 and HT55, two CRC cell lines expressing the TcGT, but not in SW480, where LTR66-POUSF1B is not detected (Fig. 3c and Supplementary Fig. 1e). Correspondingly, the neighboring LTR66 was adorned with the active promoter marks H3K4me3 and H3K27ac in the first two but not the third of these cell lines (Fig. 3d). In HT29 cells, where RNA-seq detected low levels of LTR66 and POUSF1B reads (Supplementary Fig. 1e), some H3K4me3 was detected at LTR66 whereas the intronic enhancer region was more enriched in H3K4me1 and H3K27ac (Fig. 3c, b). We then manipulated the chromatin status at LTR66 and its putative enhancer through gRNA-targeting of a dCas9-KRAB fusion protein (CRISPRi), which induces deposition of the repressive mark H3K9me3, or of its dCas9-VRPR counterpart (CRISPRa), which activates transcription (Fig. 3d). Expression of the LTR66-POUSF1B TcGT was decreased in LS1034 and HT55 cells when CRISPRi was targeted either to its LTR66 promoter or to its putative intronic enhancer (Fig. 3e, f, upper plots). Inversely, it was induced by targeting CRISPRa to LTR66 and, modestly, to the TcGT internal enhancer in HT29 cells (Fig. 3e, f, lower plots).

POUSF1B is a hominoïd-restricted OCT4 paralogs localizing to cytoplasm and enriched in membranes

POUSF1B results from the retrotransposition of POUSFI/OCT4, the stem cell pluripotency gene, in the last common ancestor of modern Hominidae (Supplementary Fig. 4a). In human, the products of these two paralogs differ at 15 out of 359/360 amino acid positions (Supplementary Fig. 4b). However, immunofluorescence microscopy revealed that, contrasting with the exclusively nuclear OCT4, POUSF1B was detected only at low levels in this compartment and instead strongly accumulated in the cytoplasm of overexpressing SW480 cells (Fig. 4a and Supplementary Fig. 4c). Of note, the cytoplasmic enrichment of POUSF1B was not cell type-specific, as it was also apparent in transfected 293 T cells (Supplementary Fig. 4d). This preferential subcellular localization was confirmed by immunohistochemistry analysis of endogenous POUSF1B in primary CRC biopsy samples, where a positive signal strictly correlated with the detection of POUSF1B transcripts by RT-PCR was detected in 4 out of 5 examined tumors but none of their matching normal tissue counterparts (Fig. 4b and Supplementary Fig. 4e, g). Cytoplasmic endogenous POUSF1B was also detected in LS1034 and HT55 cells (Supplementary Fig. 4f). Furthermore, the subcellular fractionation of SW480 cells overexpressing HA-tagged POUSF1B demonstrated that it was cytoplasmic and enriched in membranes, whereas OCT4 displayed a preference for the nucleus (Fig. 4f). Accordingly, chromatin immunoprecipitation studies failed to document significant genomic recruitment for POUSF1B (not illustrated). Isolation of detergent-resistant membranes (DRMs) from POUSF1B- and GFP-overexpressing SW480 cells further revealed that POUSF1B was enriched in this fraction (F2) (Fig. 4d), which interestingly comprises cholesterol-sphingolipid rafts where signaling complexes are commonly assembled. Of note, a fraction of transferrin receptor was also found to relocate partly to DRMs in POUSF1B-overexpressing cells, suggesting that the TcGT product perhaps exerted some restructuring influence on membranes.
Multi-omic characterization of POU5F1B-induced molecular changes

To define the impact of POU5F1B activation, we first analyzed RNA-seq data from GFP- and POU5F1B-overexpressing HT29 and SW480 cells, from control and POU5F1B-knockdown LS1034 cells, and from stages I, II, and III SYSCOL CRC primary tumors (Supplementary Fig. 5a, b). Of note, the transcriptional changes associated with POU5F1B expression were globally mild when compared with those triggered by POU5F1/OCT4, as illustrated by a side-by-side comparison of HT29 cells modified to overexpress either one of these two paralogs (Supplementary Fig. 5a). For the 3 settings (SW480, LS1034, SYSCOL) with enough upregulated genes in the presence of POU5F1B, delineation of GO terms using the Functional Annotation Chart from DAVID bioinformatics resources 6.8 consistently pointed to the term ‘extracellular space’ (Supplementary Fig. 5b). Several genes stood out, albeit not systematically in all 4 settings, which encoded for secreted proteins known to stimulate cell proliferation in trans, such as AREG, NTS, EREG, TGFBI, IGF2, and IL33.
Next, we characterized the protein interactome of POU5F1B by affinity purification and mass spectrometry (AP-MS). Sixty proteins were found highly enriched in POU5F1B-specific immuno-precipitates (Fig. 5a and Supplementary Data 2), several known to associate in complexes endowed with signaling functions (Fig. 5a, blue balls) and complexes endowed with signaling functions (Fig. 5a, red lines). To probe for a broader role of POU5F1B in human cancer, we sought to identify novel targets (Fig. 5a, red lines).

We went on to document POU5F1B-induced changes in the cell proteome by conducting SILAC (stable isotope labeling by amino acids in cell culture) coupled to LC-MS/MS and computational analysis of lysates and supernatants of two independent pairs of GFP- and POU5F1B-overexpressing SW480 cells (Supplementary Fig. 5a and Supplementary Data 1). We also sought novel targets (Fig. 5a, red lines).}

POU5F1B-overexpressing SW480 cell lines were used to identify novel targets (Fig. 5a, red lines). We went on to document POU5F1B-induced changes in the cell proteome by conducting SILAC (stable isotope labeling by amino acids in cell culture) coupled to LC-MS/MS and computational analysis of lysates and supernatants of two independent pairs of GFP- and POU5F1B-overexpressing SW480 cells (Supplementary Fig. 5a and Supplementary Data 1). We also sought novel targets (Fig. 5a, red lines).
POU5F1B was lowly expressed and POU5F1B TcGTs were undetectable in more than 8,800 samples from 29 different normal tissues of the Genotype–Tissue Expression (GTEx) dataset (Fig. 6a). POU5F1B RNA levels were similarly low in human embryonic stem cells (hESC), at least 1,000-fold below those of its paralog OCT4 (Supplementary Fig. 6c). In all tumors where high levels of POU5F1B transcripts were documented, they originated from some upstream TE, POU5F1B RNAs devoid of sequences 5' of the annotated TSS being only exceptionally detected. Together, these data demonstrate that the tumor-preferential overexpression of POU5F1B through the onco-exaptation of normally silenced TE promoters is a widespread phenomenon in human cancer.
Fig. 3 | The rs6983267-containing enhancer regulates LTR66-POU5F1B expression. a The LTR66-POU5F1B genomic locus, with depiction of TcGT (black boxes linked by broken dashed lines), rs6983267-containing enhancer from UCSC GeneHancer tracks (turquoise box), long-range chromatin interaction in CRC cell lines from the literature (gray line), and ATAC-seq peaks in CRC tumor samples from TCGA (gray boxes). b, c ChIP-PCR analyzes of LTR66-POU5F1B-expressing (LS1034 and HT55) and non-expressing (SW480 and HT29) CRC cell lines for indicated chromatin marks (n = 3 independent experiments; from left to right b: ***P = 6.70e-08, **P = 9.64e-03, **P = 5.58e-07, n.s. = 0.68; c: ***P = 6.45e-04, **P = 4.81e-02, **P = 1.01e-05, ***P = 6.38e-03, respectively, by two-sided t-test). d Schematic representation of gRNAs (g) and primers (P) used to target CRISPRi or CRISPRa to either LTR66 or the rs6983267-containing region and measure LTR66-POU5F1B transcripts by qRT-PCR. e, f Impact of indicated manipulations on LTR66-POU5F1B RNA levels (n = 3 independent experiments; e: LS1034 P1 **P = 4.21e-03, P2 **P = 1.33e-03, P3 **P = 4.67e-03; HT55 P1 ***P = 6.79e-05, P2 **P = 6.03e-05, P3 **P = 7e-05; HT29 P1 ***P = 6.30e-03, P2 P = 4.93e-02, P3 **P = 4.32e-02; f: LS1034 P1 ***P = 2.96e-03, P2 *P = 2.94e-02, P3 ***P = 9.62e-03; HT55 P1 *P = 2.03e-02, P2 **P = 3.92e-03, P3 ***P = 3.69e-05; HT29 P1 ns P = 8.64e-02, P2 nsP = 0.18, P3 ***P = 6.62e-04, respectively, by two-sided t-test). Data in b–f are presented as mean ± s.e.m., with single values as circles. Ns, not significant; *P < 0.05, **P < 0.01, ***P < 0.001. Source data are provided as a Source Data file.

Fig. 4 | POU5F1B differs from OCT4. a Representative immunofluorescence–confocal microscopy of POU5F1B-HA overexpressing SW480 cells, with DAPI in blue, HA in pink, and the nuclear membrane marker lamin-B1 in green. Two fields out of eight are shown. b Immunohistochemistry for endogenous POU5F1B in adjacent normal colon and primary colon adenocarcinoma samples from CRC patient 1. Two levels of magnification from a representative field out of three are shown. Two patients are depicted in Supplementary Fig 4e. c Subcellular fractionation into the cytoplasm (cyto), membrane (mem), and nuclear (nuc) compartments from OCT4-, POU5F1B- and GFP-overexpressing SW480 cells, with total cell extracts on right. Calnexin, lamin-B1, and beta-tubulin are used as controls for membrane, nucleus, and cytoplasm, respectively (representative blot out of three independent experiments). d Top, isolation of detergent-resistant membranes (DRMs) from POU5F1B- and GFP-overexpressing SW480 cells. F1–F3 correspond to insoluble fractions, F4–F6 to soluble fractions, DRMs being traditionally found in fraction F2. Caveolin1 and transferrin receptors are used as controls for insoluble and soluble fractions, respectively (representative blot out of three independent experiments). Bottom, western blot quantification indicating the percentage of protein in each fraction (n = 3 independent experiments; P = 2.1e-02 by two-sided t-test). Source data are provided as a Source Data file.
Discussion

The POU5F1B retrogene arose by retrotransposition of OCT4 in the last common ancestor of great apes, hence is absent in the mouse, the animal model most commonly used to study human cancers. Furthermore, POU5F1B displays biological properties fundamentally different from those of its highly conserved OCT4 relative. As such, POU5F1B is a dually novel oncogene. The chromosomal region between LTR66 and POU5F1B hosts cancer risk loci for breast, colon, and prostate cancer\(^5\). Here, we demonstrate that production of the LTR66-POU5F1B TcGT is stimulated by an intronic enhancer previously found to display a state of open chromatin in several human cancers, including colon, breast, and prostate\(^5\). This sequence was initially
identified as a MYC enhancer, but evidence indicates that it also acts on the LTR66 integrant situated upstream, including i) a physical interaction between the two elements documented in several CRC cell lines\(^2\); ii) a concordance of their ATAC-seq profiles in CRC tumors from the TCGA cohort\(^3\), and iii) our functional data. Of note, we found that the production of LTR66-POUSF1B TcGTs is not linked to a polymorphism previously mapped to this enhancer and also that it can be bi-allelic. Thus, induction of LTR66-POUSF1B might result from trans-acting influences or from cis-acting mutations situated at greater distances and affecting both alleles. However, the finding that in other cancers, POU5F1B-encoding TcGTs can arise from TE integrants situated downstream of LTR66 argues for a strong selective pressure for expression of this protein irrespective of the underlying mechanism.

We found POU5F1B overexpression to be a negative CRC prognostic marker in the SYSCOL cohort. This corroborates similar observations in hepatocellular carcinoma and gastric cancer\(^4\). Furthermore, POU5F1B-encoding transcripts have been detected in circulating but not primary tumor cells from pancreatic ductal adenocarcinoma (PDAC), where they have been found to be associated with a more rapid clinical deterioration\(^5\), suggesting that POU5F1B contributes to conferring PDAC cells with the phenotype of circulating and, ultimately, of metastasis-initiating cells (MICs)\(^6\). Accordingly, the apparent absence of prognostic value for POU5F1B in the TCGA dataset is intriguing and could stem from differences in patient populations and therapeutic regimens compared with the SYSCOL cohort.

Our protein-centered analyzes suggest a prominent role in the activation of intracellular signaling events and cytoskeletal rearrangements in the pro-oncogenic effects of POU5F1B. First, the retrogene product could be co-immunoprecipitated in CRC cells with a number of signaling molecules, including protein kinases such as MSTIR, PKRCA, ERBB2 and several of its known interactors, as well as modulators of GTPase activation such as the RAS effector protein RIN1. Second, POU5F1B expression induced the upregulation or increased phosphorylation of several downstream mediators of protein kinase-mediated signaling. Third, POU5F1B was enriched in sphingolipid/cholesterol-enriched membrane subdomains or lipid rafts, where signaling effectors are commonly concentrated and was associated with increased levels of several components of caveolae, which are biochemically closely related to lipid rafts and play an important role in the trafficking of signaling receptors such as ERBB2, EGFR, insulin-R, TGFα and PDGF\(^7\). Caveolae are also associated with the activation of the contractile actin cytoskeleton, which favors cell migration\(^8\), and we relatedly found POU5F1B to interact with or to induce the upregulation and/or phosphorylation of several mediators of cytoskeletal rearrangements. This suggests that modifications of cell architecture and motility may underlie the pro-metastatic effect of POU5F1B documented in our xenotransplantation experiments.

Our results have important medical implications. First, they suggest that the detection of POU5F1B RNA or protein in a CRC biopsy warrants aggressive management of the underlying tumor. Second, although a physiological role for the recently emerged POU5F1B is yet to be identified, this gene is highly tolerant to loss-of-function mutations (pLI = 0), and can be bi-allelically inactivated in some individuals (https://gnomad.broadinstitute.org\(^9\)), and displays little expression in normal tissues. Accordingly, its product is an attractive target for the development of novel cancer therapies. Third, blocking the trans-acting mediators released by POU5F1B-expressing cells could also be of benefit, as these molecules are predicted to increase the oncogenic properties of neighboring cells and of the tumor microenvironment. Finally, the presence of POU5F1B-encoding TcGTs in several other malignancies, including breast, prostate, stomach, and uterus, the previous identification of the POU5F1B locus as an integration hotspot for human papillomavirus and the coincidence of its transcripts with more advanced histological grades in cervical cancer\(^10\),\(^11\), as well as the documentation of its amplification in some gastric tumors\(^12\) and growth-promoting effect on hepatocellular carcinoma cell lines\(^13\), all indicate that our findings have a relevance that likely extends well beyond colorectal cancer alone.

### Methods

#### RNA sequencing

Reads were mapped to the human genome (hg19) using hisat2 (2.1.0)\(^14\). Samtools (1.4) were used to manipulate the alignments. Counts on genes and TEs were generated using featureCounts (from subread 1.5.2)\(^15\). To avoid read assignment ambiguity between genes and TEs, a gtf file containing both was provided to featureCounts. For repetitive sequences, an in-house curated version of the RepeatMasker database (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013–2015; http://www.repeatmasker.org\(^16\)) was used, where fragmented LTR elements were fused together and flanking LTR elements were grouped with their internal counterparts. Only uniquely mapped reads were used for counting on genes and TEs. Library sizes used in the CPMs’ calculation were computed with the TMM method as implemented in the limma package of Bioconductor\(^17\). Library sizes for genes were used for both genes and TEs. Differential gene expression analysis was performed as explained in the ‘David enrichment analysis’ methods chapter.

#### Transspochimeric gene transcripts analysis

First, a per sample transcriptome was computed from the RNA-seq bam file using Stringtie (1.3.4c)\(^18\) with parameters –p 1 –c 1. For TCGA, samples were downsampled down to 17mio reads in order to match the average sequencing depth of the SYSCOL dataset. Each transcriptome was then crossed using BEDTools (2.30.0)\(^19\) to both the ensemble hg19 coding exons and curated RepeatMasker to extract TcGTs for each sample. Second, a custom program was used to annotate and aggregate the sample level TcGTs into counts per groups (normal, tumor, tissue, or cancer type depending on the dataset). In brief, for each dataset, a gtf file containing all annotated TcGTs was created and TcGTs having their first exon overlapping an annotated gene or TSS not overlapping a TE were discarded. From this filtered file, TcGTs associated with the same gene and having a TSS 100 bp within each other were aggregated. Finally, for each aggregate, its occurrence per group was computed.
Fig. 6 | POU5F1B-encoding TcGTs are detected in several other cancers. a Left, heatmap depicting the percentage of patients with POU5F1B TcGTs for indicated TCGA cancers and normal tissues of the Genotype–Tissue Expression (GTEx) dataset. Locations of TEs acting as TSS are indicated on top, with the intensity of gray shade proportional to usage frequency. Right, levels of expression of POU5F1B transcripts in 9,566 samples from 32 cancer types of the TCGA dataset and in 8,878 samples from 29 normal tissue types of the GTEx dataset. For each TCGA and GTEx category, expression of TcGT samples was compared to non-TcGT samples with a two-sided t-test. Significance levels are shown as stars (*** pval < 0.001, ** pval < 0.01, * pval < 0.05). Right margin, percentage of samples with (red) or without TcGT (blue -cancer-, gray -normal tissue-). Abbreviations: ACC adrenocortical carcinoma; BLCA bladder urothelial carcinoma; BRCA breast carcinoma; CESC cervical squamous cell carcinoma; CHOL cholangiocarcinoma; COAD colon adenocarcinoma; DLBC diffuse large B-cell lymphoma; ESCA esophageal carcinoma; GBM glioblastoma multiforme; HNSC head, and neck squamous cell carcinoma; KICH kidney chromophobe; KIRC kidney renal clear cell carcinoma; KIRP kidney renal papillary cell carcinoma; LAML acute myeloid leukemia; LUAD lung adenocarcinoma; LUSC lung squamous cell carcinoma; MESO mesothelioma; OV ovarian cancer; PAAD pancreatic adenocarcinoma; PCPG pheochromocytoma, and paraganglioma; PRAD prostate adenocarcinoma; READ rectal adenocarcinoma; SARC sarcoma; SKCM skin cutaneous melanoma; STAD stomach adenocarcinoma; THCA thyroid carcinoma; THYM thymoma; UCEC uterine corpus endometrial carcinoma; UCS uterine carcinosarcoma; UVM uveal melanoma.
Analysis of public datasets
Raw RNA sequencing reads were downloaded from NCBI’s dbGaP for both the TCGA and the GTEX datasets. Raw single-cell RNA-seq data was downloaded from the European Genome-phenome Archive (EGAD00001002727). Raw RNA-seq data from cancer cell lines were downloaded from the EGAD0000100725 repository. Raw RNA-seq data from eighteen CRC patients were downloaded from GEO (GSE60945, GSE83765) and from the CMS classification repository. Long-range chromatin interactions in three CRC cell lines were obtained from Jäger, R. et al. TCGA ATAC-seq peaks were downloaded from https://gdc.cancer.gov/about-data/publications/ATACseq-AWG.

CMS classification
The CMS classification was obtained as implemented in the CMSclassifier software (https://github.com/dFenix7/CMSclassifier).

5′-rapid amplification of the cDNA ends (5′-RACE)
We applied SMARTer 5′-RACE technique in LS1034, HCT116, LS174T, NCH508 and LoVo CRC cell lines, according to the manufacturer’s protocol (Clontech) using the gene-specific primer:

5′- GATTATGATTTAAAGATCACCTGTG -3′
5′- TGCTGTTGA -3′

RACE products were loaded in an agarose gel and a fragment around 2,500 bp was excised for DNA purification and In-Fusion cloning (Clontech). Ten transformed colonies per condition were analyzed by restriction digestion and only those with the RACE insert were Sanger sequenced. A scheme of the sequenced mRNA species and their frequency is shown.

Cell culture
LS1034, NCH508, and DLD1 cell lines were obtained from the American Type Culture Collection (ATCC) and maintained in RPMI 1640 medium (Gibco) supplemented with 10% FCS (Bioconcent 2-01F36-I); LS174T (ATCC) and HT55 (Sigma) cells were cultured in EMEM supplemented with 10% FBS, HT55 being complemented with 2 mM glutamine (Gibco), and 1% of non-essential aminoacids (Sigma). SW480 and SW620 (ATCC) were cultured in L15 medium (Sigma); HT29 (ATCC) in McCoy’s 5 A (Thermo Fisher); LoVo (ATCC) in Ham’s F12K (ThermoFisher); and HCT116 (ATCC) in DMEM (Gibco), being all supplemented with 10% FCS. For lentiviral vector production, 293 T cells were cultured in DMEM supplemented with 10% FBS, HT55 being complemented with 2 mM glutamine (Gibco) and 1% of non-essential aminoacids (Sigma). SW480 and SW620 were cultured in DMEM supplemented with 10% FCS, HT55 being complemented with 2 mM glutamine. 25 μg ml−1 penicillin, 100 μg ml−1 streptomycin, and 26 μg ml−1 glutamine (Corning 30-009-C) at 37°C in a humidified atmosphere of 5% CO2. All cells tested negative for mycoplasma.

Overexpression and shRNA vectors
For the POU5F1B-expressing vector, genomic DNA from DL1 cell lines (ATCC® CCL-221™) was PCR amplified with the POU5F1 primers 5′- CACATGCGGATACCTGTCGATTCTC-3′; 5′- CGTTGATCATGGGAGGAGCCAG-3′, cloned into a pENTR TOPO donor vector (Thermo Fisher), that was recombined with the doxycycline-inducible lentiviral destination vector pSIN-ΔTRE-3xHA-puro (Thermo Fisher), that was used in the i-Score Designer system, and adapted to the miR-E shRNA structure and cloned into the LT3GEP pIRL backbone following Fellmann C et al. instructions10. shRNA3 5′- TCTGTTCTGACATGGGACGCACAGGTATTAGTTTTAAGTATGGAAAGCCACATGGATTTAATATCTATACTAATGCAGTTGGTACCTCATGCTGTTGG-3′; shRNA5 5′- TCTGCTGGA CAGTGTCGATCTGCTGTTGG-3′ with a scrambled shRNA serving as negative control 5′- TGGT TTCTGTTGACATGGGACGCACAGGTATTAGTTTTAAGTATGGAAAGCCACATGGATTTAATATCTATACTAATGCAGTTGGTACCTCATGCTGTTGG-3′ with a scrambled shRNA serving as negative control 5′- TCTGTTGACATGGGACGCACAGGTATTAGTTTTAAGTATGGAAAGCCACATGGATTTAATATCTATACTAATGCAGTTGGTACCTCATGCTGTTGG-3′ with a scrambled shRNA serving as negative control 5′- TCTGTTGACATGGGACGCACAGGTATTAGTTTTAAGTATGGAAAGCCACATGGATTTAATATCTATACTAATGCAGTTGGTACCTCATGCTGTTGG-3′ with a scrambled shRNA serving as negative control 5′- TCTGTTGACATGGGACGCACAGGTATTAGTTTTAAGTATGGAAAGCCACATGGATTTAATATCTATACTAATGCAGTTGGTACCTCATGCTGTTGG-3′ with a scrambled shRNA serving as negative control 5′- TCTGTTGACATGGGACGCACAGGTATTAGTTTTAAGTATGGAAAGCCACATGGATTTAATATCTATACTAATGCAGTTGGTACCTCATGCTGTTGG-3′ with a scrambled shRNA serving as negative control 5′- TCTGTTGACATGGGACGCACAGGTATTAGTTTTAAGTATGGAAAGCCACATGGATTTAATATCTATACTAATGCAGTTGGTACCTCATGCTGTTGG-3′ with a scrambled shRNA serving as negative control 5′- TCTGTTGACATGGGACGCACAGGTATTAGTTTTAAGTATGGAAAGCCACATGGATTTAATATCTATACTAATGCAGTTGGTACCTCATGCTGTTGG-3′ with a scrambled shRNA serving as negative control.
microscopically following H&E tissue staining. Mice were maintained on doxycycline food pellets (0.625 g kg⁻¹; SAFE E8404 version 0002). Subcutaneous and intrasplenic injection experiments with LS1034 cells were performed in 3 groups (sh3, sh5, scramble) of 8 animals each, which were fed with doxycycline food pellets from one week before injection to the experimental endpoint. LS1034 cells were treated with doxycycline 72 h before injection.

Micrometastases quantification
Images were acquired from Hematoxylin&Eosin stains on an Olympus VSI20 Whole Slide Scanner, using a 20x objective (UPLSAPO, NA. 0.75) and a color camera (Pike F505 Color) with an image pixel size of 0.345 microns. Obtained images were analyzed using the software QuPath (0.3.2) using groovy scripts, making use of a pixel classifier to segment and measure cancer cell clusters. For more details, please access the following Zenodo project (https://doi.org/10.5281/zenodo.6523649).

Chromatin immunoprecipitation
Chromatin was prepared as described previously⁵. Pellets were lysed by resuspension in LBI for 10 min (50 mM HEPES-KOH pH 7.4, 140 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 10% Glycerol, 0.5% N Pep, 0.25% TritonX100, protease inhibitors), centrifuged, resuspended in LBI2 for 10 min (10 mM Tris pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA and protease inhibitors), centrifuged, and resuspended in LB3 (10 mM Tris pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5% SDS, 0.1% NaDOC, 0.1% SDS and protease inhibitors). LB incubations were done in a rotating wheel at 4 °C and centrifugations at 1,700 g for 5 min at 4 °C.

Subcellular fractionation
Twenty million SW480 GFP- and POU5F1B-expressing cells were washed three times with cold PBS and harvested upon scraping in 0.4 ml cold buffer A (10 mM KCl, 2 mM MgOAc, 20 mM HEPES pH 7.2, 0.5 mM DTT, 0.01% digitonin) containing protease inhibitor cocktail (Roche). After centrifuging at 300 g 4 °C for 5 min, supernatant was collected as cytoplasmic fraction and pellet was resuspended in 200 µl buffer C (10 mM KCl, 2 mM MgOAc, 20 mM HEPES pH 7.2, 0.5% SDS, 20 mM EDTA, 0.5 mM NaF, 2 mM benzamidine) with protease inhibitors. After centrifuging at 300 g 4 °C for 10 min, supernatant was collected as membrane fraction and nuclear pellet was resuspended in 0.4 ml of buffer C (1% NP-40, 500 mM Tris-HCl pH 7.4, 0.05% SDS, 20 mM EDTA, 10 mM NaF, 2 mM benzamidine) with protease inhibitors. After incubating 10 min on ice and centrifuging at 1,000 rpm 4 °C 10 min, supernatant was collected as nuclear fraction. Equal volumes of cytoplasm, membrane, and nuclear fractions were submitted to SDS-PAGE and analyzed by immunoblotting using anti-beta tubulin (Sigma T4026, 1:1,000), HRP-conjugated anti-HA (clone 3F10 Roche 12103819001, 1:1,000), calnexin (Bethyl A303-696A, 1:2,000), lamin B1 (Abcam ab16048, 1:1,000), HRP-conjugated anti-rabbit (Santa Cruz sc-2004, 1:100,000), and HRP-conjugated anti-mouse (GE Healthcare 2004, 1:10,000, and HRP-conjugated anti-mouse (GE Healthcare 2004, 1:10,000).
NA93IV, 1:10,000) antibodies. Equal volumes of the three fractions were mixed and loaded as total cell extract. Uncropped blots are shown in the Source Data File.

Isolation of detergent-resistant membranes (DRMs)
Approximately 1 × 10^9 cells were resuspended in 0.5 mL cold TNE buffer (25 mM Tris-HCl, pH 7.5, 150 mM NaCl, 5 mM EDTA, and 1% Triton X-100; Surface Amps, ThermoFisher) with a tablet of protease inhibitors (Roche). Membranes were solubilized in a rotating wheel at 4 °C for 30 min. DRMs were isolated using an Optiprep™ gradient; the cell lysate was adjusted to 40% Optiprep™, loaded at the bottom of a TLS.55 Beckman tube, overlaid with 600 μL of 30% Optiprep™ and 600 μL of TNE, and centrifuged for 0.5 h at 259,000 g at 4 °C. Six fractions of 400 μL were collected from top to bottom. DRMs were found in fraction 2. Equal volumes from each fraction were analyzed by SDS-PAGE and western blot analysis using HRP-conjugated anti-HA, caveloin (Santa Cruz sc-894, 1:500) and transferrin receptor (Thermo Fisher 13-6800, 1:10,000) antibodies.

Interactome of POU5F1B
Three technical replicates of 80 million HA-tagged POU5F1B and GFP overexpressing LS1034 and HT29 cells were used in this experiment (a total of 12 samples). To assess affinity purification, protein expression was induced with 500 μg/mL doxycycline for three days. Once subconfluent, cells were harvested in PBS 1 mM EDTA. Dry pellets were lyzed in HNN lysis buffer (0.5% NP40, 50 mM HEPES pH 7.5, 150 mM NaCl, 50 mM NaF, 200 mM Na3VO4, 1 mM EDTA, 1 μg/mL PMSF, and protease inhibitors) and fixed with 3 mM DSP for 40 min. Reactive DSP was quenched with 100 mM Tris pH 7.5. The lysates were subjected to 250 μL U1 benzamidase (Merck, 72103) for 30 min at 37 °C. Lysate was then centrifuged for 15 min at 17,000 g in order to remove insoluble material. Supernatant was then incubated with 100 μL of pre-washed anti-HA agarose beads (Sigma, A2095) for 2 h on a rotating wheel at 4 °C. Immunoprecipitates were washed three times with 2 mL and twice 1 mL HNN lysis buffer, and three times with 2 mL and twice 1 mL HNN buffer (50 mM HEPES pH 7.5, 150 mM NaCl, 50 mM NaF). Proteins were then eluted with 3 × 100 μL of 0.2 M glycine pH 2.5. Samples were neutralized and denatured with 500 μL 0.5 M NH4HCO3 pH 8.8, 6 M urea, reduced with 1 mM PMSF, and protease inhibitors (Roche). Samples were then digested with 1 μg trypsin (Promega, V5113) overnight at 37 °C in the dark. The next day, trypsin digestion was stopped by lowering the pH with the addition of 50 μL of formic acid (AppliChem, A398S.0500) and peptides were purified and prepared for mass spectrometry injection at the EPFL proteomics facility as previously described. All samples selected for further analyses had to display more than 10 bait POU5F1B spectral counts in the three technical replicates in order to ensure proper bait protein levels. Only proteotypic, unique spectral counts were used. The CRAPome matrix reduced list of proteins was used to subtract unspecific protein-protein interactions. Significance between bait-prey interactions was computed with the R package lsmeans. The most significant interactions were defined as having an adjusted p value lower than 0.01 and fold change enrichment over control bigger than 5 (Supplementary Data 2). A POU5F1B interactome was established with Cytoscape software, using the fold change to draw force-directed edges between proteins. Previously described protein-protein interactions amongst some detected prey were found in the BioGRID website or the literature.

ERBB2 immunoprecipitation
Ten million SW480, HT29, and LS1034 GFP- and POU5F1B-expressing cells were harvested, washed with PBS, resuspended in lysis buffer (400 mM NaCl, 10 mM HEPES, pH 7.5, 0.1% NP-40, protease inhibitor cocktail – Roche), and incubated 30 min at 4 °C with gentle agitation. The NaCl concentration was then brought down to 150 mM, samples were sonicated using a probe sonicator (three times 10 seconds at 30% amplitude) and then centrifuged at 17,000 rcf to remove insoluble material. The protein concentration of the lysates was measured using a BCA assay (Thermo Fisher, 23225), and equivalent protein quantities were incubated overnight on a rotating wheel at 4 °C with 50 μL of streptavidin agarose beads (Sigma S1638) previously bound to 3 μg of anti-ERBB2 antibody (ThermoFisher BMS120BT) for 3 h on a rotating wheel at 4 °C. Samples were then incubated three times for 5 min with 1 mL of BC100 (100 mM KCl, 10 mM Tris pH 7.8, 0.5 mM EDTA, 10% glycerol, 0.1 mM PMSE, 0.1 mM DTT), and five times with 1 mL of BC500 (BC100 with 500 mM KCl) on a rotating wheel at 4 °C. Immunoprecipitates were eluted twice in 50 μL 0.2 M glycine pH 2. Similar elution volumes and protein quantities for the IPs and the inputs, respectively, were submitted to SDS-PAGE and analyzed by immunoblotting using anti-ERBB2 (ThermoFisher MAS-13102, 1:100), HRP-conjugated anti-HA and HRP-conjugated anti-mouse antibody.

Stable isotope labeling by amino acids in cell culture (SILAC)
SILAC experiments were performed as described. POU5F1B- and GFP-overexpressing SW480 cells were grown in parallel in heavy or medium SILAC labeling mixes for 15 days (~9 cell divisions), performing a second independent experiment with inverted labeling. SILAC RPMI 1640 culture medium was complemented with 200 mg/L light L-proline, 150 mg/L heavy L-lysine (K5), and 50 mg/L heavy L-arginine (R10) for the heavy amino-acid labeling; and 200 mg/L light L-proline, 150 mg/L medium L-lysine (K4), and 50 mg/L medium L-arginine (R6) for the medium amino-acid labeling. Medium and heavy cells were combined and lysed as a mixed population. The day 15 time point was used for LC-MS/MS analysis of the total cell proteome after verifying that it allowed ~96% SILAC amino-acid incorporation rate.

For the secretome studies, exponentially growing SILAC-labeled cells were washed twice with PBS and incubated in 10 mL serum-free SILAC medium at 37 °C for 24 h. The cell death rate was below 2% (measured by trypsin blue staining). Conditioned medium (CM) was collected, centrifuged at 200 g for 10 min, and filtered through 0.22 μm membranes adding protease inhibitors (Roche). Medium and heavy labeled CM were combined at equal volumes and stored at −80 °C. CM from SILAC cells at day 15, corresponding to maximum amino acid incorporation rate, was used for further analysis. CM was concentrated using Amicon Ultra-4 and ~0.5 centrifugal filters with 3 kDa cutoff (Millipore) and subjected to LC-MS/MS analysis.

The same total cell extract and secretome SILAC protocols were applied to sh-scramble, shRNA3, and shRNAs LS1034 cells, labeled with light, medium, and heavy medium, respectively, with a second independent experiment inverting the labeling of shRNA3 and shRNA 5 cells.

Each sample was digested by Filter Aided Sample Preparation (FASP) with minor modifications. Dithiothreitol (DTT) was replaced by Tris (2-carboxyethyl)phosphine (TCEP) as a reducing agent and iodoacetamide by Chloroacetamide as alkylating agent. Combined proteolytic digestion was performed using Endoproteinase Lys-C and Trypsin. Peptides were desalted on SDB-RPS StageTips and dried down by vacuum centrifugation. Samples were then fractionated into 12 fractions using an Agilent OFFGEL 3100 system. The resulting fractions were desalted on SDB-RPS StageTips and dried by vacuum centrifugation. For LC MS/MS analysis, peptides were resuspended and separated by reversed-phase chromatography on a Dionex Ultimate 3000 RSLC nanoUPLC system in-line connected with an Orbitrap Lumos Fusion Mass-Spectrometer. Database search was performed
using MaxQuant (1.6.10.43) against a concatenated database consisting of the Uniprot human database (Uniprot release 2019_06, 74468 sequences) and common fetal bovine serum protein. Carbamidomethylation was set as a fixed modification, whereas oxidation (M), phosphorylation (S,T,Y), Gln to pyro-Glu and acetylation (Protein (and < intensity>1000, and average fold change between the two arrays >1.2) (Supplementary Data 7).

Midomethylation was set as a fixed modification. SILAC total cell (Supplementary Data 3 and 4) and secreteme candidates (Supplementary Data 5 and 6) were selected when both experimental replicates showed p value < 0.05 (outlier detection test as computed by MaxQuant), high intensity (above quantile 25%), and fold changes in the same direction for both replicates.

DAVID enrichment analysis
For each RNA-seq and SILAC data set, we obtained a list of differentially expressed candidates. Differential gene expression analysis was performed using voom as it has been implemented in the limma package of Bioconductor. A gene was considered to be differentially expressed when the fold change between groups was bigger than 2 and the p-value was smaller than 0.05. A moderated t-test (as implemented in the limma package of R) was used to test significance. P-values were corrected for multiple testing using the Benjamini–Hochberg’s method. Proteins enriched in the SILAC experiments were selected as detailed above. RNA-seq and SILAC data were plotted in R and python. The resulting candidate lists were subjected to a functional annotation chart using the online bioinformatics resource DAVID (6.8).

Kinexus antibody microarray preparation and analysis
Kinexus KAM-1325 antibody microarray kits (Kinexus) were used according to the manufacturer’s protocols. Each of the two arrays was loaded with protein extracts from SW480 POU5F1B- and GFP-overexpressing cells obtained from two independent lentiviral transductions. Experts at Kinexus conducted array scanning and data collection. Data analysis was performed in-house, for a final selection of candidates based on p-value < 0.05 (two-way ANOVA test), average spot intensity >1000, and average fold change between the two arrays >1.2 (and < –1.2) (Supplementary Data 7).

Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The RNA-sequencing data generated in this study have been deposited in the GEO database under accession code GSE182467. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the identifier PXD028034; PXD028035; and PXD028036. The TCGA and GTEX publicly available data used in this study are available in the dbGaP database under the accession codes phs000178.v10.p8 and phs000424.v7.p2, respectively. The raw single-cell RNA-seq and the cancer cell lines RNA-seq publicly available data used in this study are available in the European Genome–phenome Archive under accession codes EGAD00001002727 and EGAD00001000725, respectively. RNA-seq data from eighteen CRC patients were down-loaded from GEO under the accession code GSE50760. RNA-seq data from H9 cell lines and the collection of hESC clones was loaded from GEO (GSE60943, GSE83769). Long-range chromatin interactions in three CRC cell lines were obtained from Jäger, R. et al. TCGA ATAC-seq peaks were downloaded from https://gdac.broadinstitute.org/about-data/publications/ATACseq AWC. Human genome hg19. Source Data are provided with this paper. All data are available in the article file, Supplementary Information and Source Data. Source data are provided with this paper.

Code availability
Code for RNA-seq processing and Transposchimeric Gene Transcript analysis is published in the Supplementary Code file of this article: https://genome.cshlp.org/content/suppl/2021/08/16/gr.275133.120.DC1.

References
1. Bourque, G. et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 18, 1752–1762 (2008).
2. Trizzino, M. et al. Transposable elements are the primary source of novelty in primate gene regulation. Genome Res 27, 1623–1633 (2017).
3. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351, 1083–1087 (2016).
4. Trono, D. Transposable elements, polydactyl proteins, and the genesis of human-specific transcription networks. Cold Spring Harb. Symp. Quant. Biol. 80, 281–288 (2015).
5. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature 543, 550–554 (2017).
6. Bister, K. Discovery of oncogenes: The advent of molecular cancer research. Proc. Natl Acad. Sci. USA 112, 15259–15260 (2015).
7. Burns, K. H. Transposable elements in cancer. Nat. Rev. Cancer 17, 415–424 (2017).
8. Tomlins, S. A. et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature 448, 595–599 (2007).
9. Miki, Y. et al. Disruption of the APC gene by a retrotransposon insertion of L1 sequence in a colon cancer. Cancer Res. 52, 643–645 (1992).
10. Lamprecht, B. et al. Derepression of an endogenous long terminal repeat activates the CSFIR proto-oncogene in human lymphoma. Nat. Med. 16, 571–579 (2010).
11. Babaian, A. et al. Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. Oncogene 35, 2542–2546 (2016).
12. Wiesner, T. et al. Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. Nature 526, 453–457 (2015).
13. Jang, H. S. et al. Transposable elements drive widespread expression of oncogenes in human cancers. Nat. Genet. 51, 611–617 (2019).
14. Bramsen, J. B. et al. Molecular-Subtype–Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer. Cell Rep. 19, 1268–1280 (2017).
15. Babaian, A. & Mager, D. L. Endogenous retroviral promoter exaptation drives IRF5 expression in colorectal tumors. Mol. Oncol. 8, 1653–1666 (2014).
16. Li, H. et al. Reference component analysis of single–cell transcriptionomes elucidates cellular heterogeneity in human colorectal tumors. Nat. Genet. 49, 708–718 (2017).
17. Klijn, C. et al. A comprehensive transcriptional portrait of human cancer cell lines. Nat. Biotechnol. 33, 306–312 (2015).
18. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. Nat. Med. 21, 1350–1356 (2015).
19. Tomlins, I. et al. A genome–wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat. Genet. 39, 984–988 (2007).
20. Zanke, B. W. et al. Genome–wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat. Genet. 39, 989–994 (2007).
22. Haiman, C. A. et al. A common genetic risk factor for colorectal and prostate cancer. Nat. Genet. 39, 954–956 (2007).
23. Pomerantz, M. M. et al. The 8q24 cancer risk variant rs6983267 demonstrates long-range interaction with MYC in colorectal cancer. Nat. Genet. 41, 882–884 (2009).
24. Wright, J. B., Brown, S. J. & Cole, M. D. Upregulation of c-MYC in cis through a long downstream loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. Mol. Cell. Biol. 30, 1411–1420 (2010).
25. Jäger, R. et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nat. Commun. 6, 1–9 (2015).
26. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. Science 362, eaav1998 (2018).
27. Thakore, P. I. et al. Highly specific epigenome editing by CRISPR/Cas9 repressors for silencing of distal regulatory elements. Nat. Methods 12, 1143–1149 (2015).
28. Chavez, A. et al. Highly efficient Cas9-mediated transcriptional programming. Nat. Methods 12, 326–328 (2015).
29. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57 (2009).
30. Kim, S.-W. et al. Tid1, the human homologue of a drosophila tumor suppressor, reduces the malignant activity of ErbB-2 in carcinoma cells. Cancer Res 64, 7732–7739 (2004).
31. Li, L.-Y. et al. Nuclear ErbB2 enhances translation and cell growth by activating transcription of ribosomal RNA genes. Cancer Res 71, 4269–4279 (2011).
32. Liang, G., Fang, X., Yang, Y. & Song, Y. Silencing of CEMIP suppresses colorectal cancer dependent on COX2/PGE2. Cancer Res. 78, 9915–9926 (2018).
33. Koivunen, J., Aaltonen, V. & Peltonen, J. Protein kinase C (PKC) family in cancer progression. J. Exp. Clin. Cancer Res. 37, 168 (2018).
34. Pan, Y. et al. POU5F1B promotes hepatocellular carcinoma proliferation by activating AKT. Biomed. Pharmacother. 100, 374–380 (2018).
35. Massagué, J. & Ganesh, K. Metastasis-initiating cells and ecosystems. Cancer Disc. 11, 971–994 (2021).
36. Patel, H. H., Murray, F. & Insel, P. A. Caveolae as organizers of pharmacologically relevant signal transduction molecules. Annu. Rev. Pharmacol. Toxicol. 48, 359–391 (2008).
37. del Pozo, M. A. et al. Phospho-caveolin-1 mediates integrin-regulated membrane domain internalization. Nat. Cell Biol. 7, 901–908 (2005).
38. Pereira, P. M. R. et al. Caveolin-1 mediates cellular distribution of HER2 and affects trastuzumab binding and therapeutic efficacy. Nat. Commun. 9, 5137 (2018).
39. Hetmanski, J. H. R. et al. Membrane tension orchestrates rear retraction in matrix-directed cell migration. Dev. Cell 51, 460–475.e10 (2019).
40. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020).
41. Hu, Z. et al. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. Nat. Genet. 47, 158–163 (2015).
42. Chen, L. et al. HPV-16 E2/E6 and POU5F1B as biomarkers to determine cervical high-grade squamous lesions and more. J. Inflamm. Res. 13, 813–821 (2020).
43. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360 (2015).
44. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an ef- cient Cas9-mediated transcriptional alignment of large gene lists using DAVID bioinformatics resources. Nat. Methods 10, 326–328 (2015).
45. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80 (2004).
46. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295 (2015).
47. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
48. Takashima, Y. et al. Resetting transcription factor control circuitry toward ground-state pluripotency in human. Cell 158, 1254–1269 (2014).
Theunissen, T. W. et al. Molecular criteria for defining the naive human pluripotent state. Cell Stem Cell 19, 502–515 (2016).
67. Fellmann, C. et al. An optimized microRNA backbone for effective single-copy RNAi. Cell Rep. 5, 1704–1713 (2013).
68. Tseng, W., Leong, X. & Engleman, E. Orthotopic mouse model of colorectal cancer. J. Vis. Exp. JoVE 484, 10.3791/484 (2007).
69. Bankhead, P. et al. QutPath: open source software for digital pathology image analysis. Sci. Rep. 7, 16878 (2017).
70. Heusser, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPR. Genome Biol. 17, 148 (2016).
71. Harder, T., Scheiffele, P., Verkade, P. & Simons, K. Lipid domain structure of the plasma membrane revealed by patching of membrane components. J. Cell Biol. 141, 929–942 (1998).
72. Helleboid, P. et al. The interactome of KRAb zinc finger proteins reveals the evolutionary history of their functional diversification. EMBO J. 38, e101220 (2019).
73. Mellacheruvu, D. et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. Nat. Methods 10, 730–736 (2013).
74. Bankhead, P. et al. QuPath: open source software for digital pathology image analysis. Sci. Rep. 7, 16878 (2017).
75. Benjami, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300 (1995).
76. Ong, S.-E. & Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). Nat. Protoc. 1, 2650–2660 (2006).
77. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. Nat. Methods 6, 359–362 (2009).
78. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nat. Methods 11, 319–324 (2014).
79. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372 (2008).
80. Shin, J. et al. Use of Composite Protein Database including Search Result Sequences for Mass Spectrometric Analysis of Cell Secretome. PLOS ONE 10, e0121692 (2015).
81. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 15, R29 (2014).

Acknowledgements
We thank the members of the Trono lab and Maria-Eugenia Zaballa, Jose Vicente Sanchez-Mut, and Marco Cassano for stimulating discussions. We gratefully acknowledge the Proteomics Core Facility (PCF), the Histology Core Facility (HCF), the Bioimaging and Optics Platform (BIOP), the Centre of Phenogenomics (CPG), and the Gene Expression Core Facility (GECF) at EPFL for their support & assistance in this work. This work was supported by grants from the European Research Council (KRAbNAP, No. 206872; Transpos-X, No. 6946588), the Personalized Health and Related Technologies (PHRT-508) program, the Swiss National Science Foundation (310030_152879 and 3100308_173337), the Swiss Cancer League and the Aclon Foundation to D.T., and a Marie Sklodowska-Curie Fellowship to L.S.R.

Author contributions
L.S.R. and D.T. conceived the study, interpreted the data, and wrote the manuscript; L.S.R. designed, performed and analyzed all the experiments, for some with the expert help of S.O. and L.A.; E.P., J.D., and S.D. performed the bioinformatics analyzes, A.C. the phylogenetic study, M.C.L., C.E., and S.P. the histopathological characterization of primary CRC biopsies, and C.L.A. and J.B.B. generated the SYSCOL cohort data.

Competing interests
L.S.R., E.P., J.D., and D.T. are inventors on an international patent application (title: Transpochimeric gene transcripts (tcgts) as cancer biomarkers; identification number: US2022145395 (A1)) submitted by the École Polytechnique Fédérale de Lausanne that covers methods for transposcriptome-based biomarker discovery. The remaining authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-32649-7.

Correspondence and requests for materials should be addressed to Laia Simó-Riudalbas or Didier Trono.

Peer review information Nature Communications thanks Deepali Rathore and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022