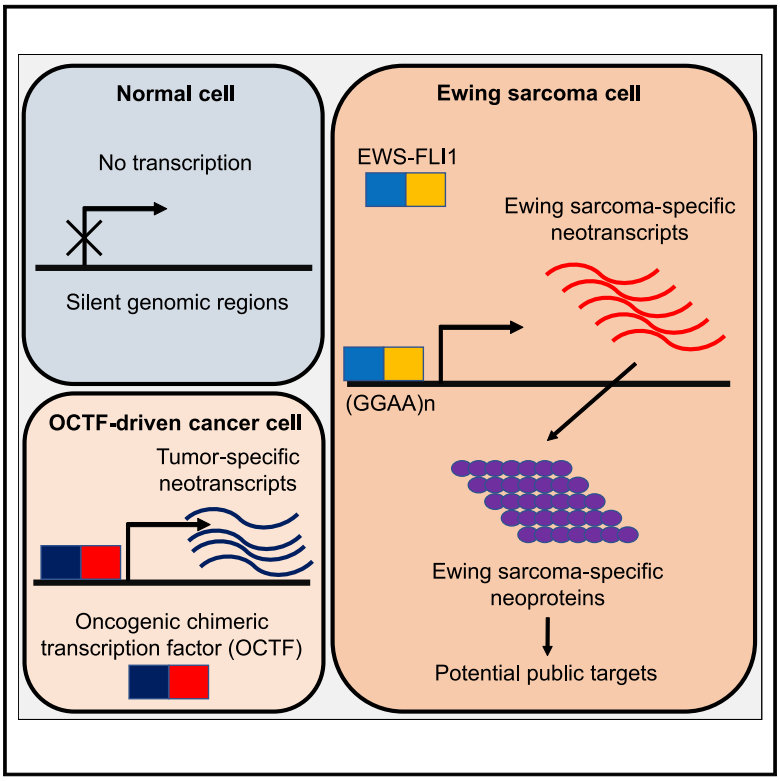
Molecular Cell

Article

Oncogenic chimeric transcription factors drive tumor-specific transcription, processing, and translation of silent genomic regions

Graphical abstract



Highlights

- EWS::FLI1 induces transcription of multiple EwS-specific novel genes
- Other oncogenic TFs also induce tumor-specific neotranscripts
- A subset of neotranscripts are translated into tumor-specific neoproteins
- Neoproteins may be novel targets in cancers with mutated TFs

Authors

Julien Vibert, Olivier Saulnier, Céline Collin, ..., Antoine Coulon, Joshua J. Waterfall, Olivier Delattre

Correspondence

joshua.waterfall@curie.fr (J.J.W.), olivier.delattre@curie.fr (O.D.)

In brief

Vibert and Saulnier et al. identify a novel activity of oncogenic chimeric transcription factors such as EWS::FLI1 in Ewing sarcoma, i.e., inducing transcription and processing of RNA transcripts from otherwise silent genomic regions. Neotranscripts are shared across patients and a subset are translated into tumor-specific neoproteins, creating new potential therapeutic targets.

Vibert et al., 2022, Molecular Cell *82*, 2458–2471
July 7, 2022
https://doi.org/10.1016/j.molcel.2022.04.019

CellPress

CellPress

Molecular Cell

Article

Oncogenic chimeric transcription factors drive tumor-specific transcription, processing, and translation of silent genomic regions

Julien Vibert,[1,2,3,22] Olivier Saulnier,[1,19,22] Céline Collin,[1] Floriane Petit,[1] Kyra J.E. Borgman,[4,5] Jérômine Vigneau,[1] Maud Gautier,[1] Sakina Zaidi,[1] Gaëlle Pierron,[6] Sarah Watson,[1,7] Nadège Gruel,[1,3] Clémence Hénon,[8] Sophie Postel-Vinay,[8,9] Marc Deloger,[10,20] Virginie Raynal,[11] Sylvain Baulande,[11] Karine Laud-Duval,[1] Véronique Hill,[1] Sandrine Grossetête,[1] Florent Dingli,[12] Damarys Loew,[12] Jacob Torrejon,[13,14] Olivier Ayrault,[13,14] Martin F. Orth,[15] Thomas G.P. Grünewald,[16,17,18] Didier Surdez,[1,21] Antoine Coulon,[4,5] Joshua J. Waterfall,[2,3,23,*] and Olivier Delattre[1,4,23,24,*]

[1]INSERM U830, Équipe Labellisée LNCC, Diversity and Plasticity of Childhood Tumors Lab, PSL Research University, SIREDO Oncology Center, Institut Curie Research Center, Paris, France
[2]INSERM U830, Integrative Functional Genomics of Cancer Lab, PSL Research University, Institut Curie Research Center, Paris, France
[3]Department of Translational Research, PSL Research University, Institut Curie Research Center, Paris, France
[4]Institut Curie, PSL Research University, Sorbonne Université, CNRS UMR 3664, Laboratoire Dynamique du Noyau, 75005 Paris, France
[5]Institut Curie, PSL Research University, Sorbonne Université, CNRS UMR168, Laboratoire Physico Chimie Curie, 75005 Paris, France
[6]Unité de Génétique Somatique, Service d'oncogénétique, Institut Curie, Centre Hospitalier, Paris, France
[7]Medical Oncology Department, PSL Research University, Institut Curie Hospital, Paris, France
[8]ATIP-Avenir group, Inserm Unit U981, Gustave Roussy, Villejuif, France
[9]Drug Development Department, DITEP, Gustave Roussy, Villejuif, France
[10]Bioinformatics and Computational Systems Biology of Cancer, PSL Research University, Mines Paris Tech, INSERM U900, Paris, France
[11]Institut Curie Genomics of Excellence (ICGex) Platform, PSL Research University, Institut Curie Research Center, Paris, France
[12]Laboratoire de Spectrométrie de Masse Protéomique, PSL Research University, Institut Curie Research Center, Paris, France
[13]Institut Curie, CNRS UMR3347, INSERM, PSL Research University, Orsay, France
[14]CNRS UMR 3347, INSERM U1021, Université Paris Sud, Université Paris-Saclay, Orsay, France
[15]Max-Eder Research Group for Pediatric Sarcoma Biology, Institute of Pathology, Faculty of Medicine, LMU Munich, Munich, Germany
[16]Division of Translational Pediatric Sarcoma Research, German Cancer Research Center (DKFZ), German Cancer Consortium (DKTK), Heidelberg, Germany
[17]Hopp-Children's Cancer Center (KiTZ), Heidelberg, Germany
[18]Institute of Pathology, Heidelberg University Hospital, Heidelberg, Germany
[19]Present address: The Arthur and Sonia Labatt Brain Tumor Research Center and Developmental & Stem Cell Biology Program, The Hospital for Sick Children, Toronto, ON M5G 0A4, Canada
[20]Present address: INSERM US23, CNRS UMS 3655, Gustave Roussy Cancer Campus, Villejuif, France
[21]Present address: Balgrist University Hospital, University of Zurich, Zurich, Switzerland
[22]These authors contributed equally
[23]Senior author
[24]Lead contact
*Correspondence: joshua.waterfall@curie.fr (J.J.W.), olivier.delattre@curie.fr (O.D.)
https://doi.org/10.1016/j.molcel.2022.04.019

SUMMARY

Many cancers are characterized by gene fusions encoding oncogenic chimeric transcription factors (TFs) such as EWS::FLI1 in Ewing sarcoma (EwS). Here, we find that EWS::FLI1 induces the robust expression of a specific set of novel spliced and polyadenylated transcripts within otherwise transcriptionally silent regions of the genome. These neogenes (NGs) are virtually undetectable in large collections of normal tissues or non-EwS tumors and can be silenced by CRISPR interference at regulatory EWS::FLI1-bound microsatellites. Ribosome profiling and proteomics further show that some NGs are translated into highly EwS-specific peptides. More generally, we show that hundreds of NGs can be detected in diverse cancers characterized by chimeric TFs. Altogether, this study identifies the transcription, processing, and translation of novel, specific, highly expressed multi-exonic transcripts from otherwise silent regions of the genome as a new activity of aberrant TFs in cancer.

INTRODUCTION

Many cancers harbor abnormal transcription factors (TFs) as a result of point mutation or gene fusion (Bushweller, 2019; Lee and Young, 2013). This latter mechanism is particularly frequent in sarcomas, i.e., bone and soft tissue cancers, where it can give rise to an oncogenic chimeric transcription factor (OCTF) (Mertens et al., 2009; Perry et al., 2019). Ewing sarcoma (EwS) is paradigmatic of the oncogenic role of such OCTFs. It is an aggressive cancer, most commonly occurring in adolescents and young adults, whose prognosis remains dismal particularly in metastatic or relapsed forms with a 5-year overall survival of less than 30%. It is characterized by specific gene fusions between members of the FET (FUS, EWSR1, TAF15) family of RNA-binding proteins and of the ETS (E-twenty-six) family of TFs, the most frequent fusion being between *EWSR1* (*EWS*) and *FLI1* in 85% of cases (Delattre et al., 1992; Grünewald et al., 2018; Riggi et al., 2021). Compared with wild-type ETS TFs, EWS::FLI1 has gain-of-function activities through its unique ability to act as a pioneer factor and activate neighboring genes by generating neo-enhancers upon binding GGAA microsatellite sequences (Boulay et al., 2017; Gangwal et al., 2008; Guillon et al., 2009; Riggi et al., 2014; Sheffield et al., 2017; Tomazou et al., 2015).

RESULTS

Long-read sequencing identifies EWS::FLI1-induced novel transcripts

We initially performed long-read RNA sequencing (RNA-seq) using the PacBio Iso-Seq protocol to investigate the full-length transcriptomic profile of the A673 EwS cell line. A total of 15,576,646 raw reads (40.3 Gbp) were generated and 56,051 high-quality circular consensus sequences were kept for downstream analysis (Figure S1A). We found 145 (0.25%) of these high-quality sequences that aligned to the human genome but had no match in the RefSeq reference transcriptome. Manual inspection of these allowed identification of 80 candidate novel transcripts, other sequences being classified as mis-annotated genes (e.g., readthrough transcription) or having low coverage support (Figure S1B).

We hypothesized that a subset of these might be direct OCTF targets and therefore investigated the A673/TR/shEF EwS cell line, which expresses a doxycycline (DOX)-inducible short hairpin RNA (shRNA) against EWS::FLI1 (Carrillo et al., 2007). This confirmed that four of the candidate novel transcripts (*Ew_NG1-4*) could be regulated by EWS::FLI1 (Figures 1 and S2–S4).

Chromatin immunoprecipitation sequencing (ChIP-seq) experiments (Aynaud et al., 2020) identified EWS::FLI1-bound GGAA microsatellites in close proximity (<5 kb) to each transcription start site (TSS). EWS::FLI1 peaks were associated with the presence of both H3K27ac and H3K4me3 activation marks, with all peaks being considerably decreased upon shRNA-mediated depletion of *EWS::FLI1* by DOX treatment (Figures 1A and S2–S4). Quantitative RT-PCR (qRT-PCR) assays in EWS::FLI1-knockdown EwS cells confirmed downregulation of these transcripts and further showed that they were induced in mesenchymal stem cells (MSCs), the putative cell of origin of EwS, engineered by CRISPR-Cas9 to express the *EWS::FLI1* fusion (Sole et al., 2021). Altogether, these data showed that EWS::FLI1 was both necessary and sufficient for expression of these transcripts (Figures 1B and 1C). Finally, we thoroughly investigated expression of these transcripts across 21,349 RNA-seq samples (10,522 normal tissues and 10,827 cancers) from Genotype-Tissue Expression (GTEx), Human Protein Atlas (HPA) (Uhlén et al., 2015), The Cancer Genome Atlas (TCGA), and our institutional collection including 132 EwS samples. We found high expression of these four transcripts in most EwS cancers. By contrast, expression was virtually undetectable in normal tissues and non-EwS cancers, with only very few samples showing some reads in the corresponding genomic regions (Figures 1D, 1E, and S2–S4). Altogether, this shows that a set of EwS-specific novel transcripts are induced by the neomorphic ability of EWS::FLI1 to bind GGAA microsatellites within transcriptionally silent genomic regions. We thus refer to these novel transcripts as "neotranscripts" encoded by "neogenes" (NGs).

Genome-guided assembly retrieves additional EwS-specific neogenes in tumor samples

Since long-read RNA-seq is technically challenging for clinical samples, we designed a strategy to further explore the existence of EwS-specific neotranscripts based on genome-guided assembly of human tumor short-read RNA-seq (Shao and Kingsford, 2017). We first identified assembled transcripts with no overlap to annotations in the GENCODE reference transcriptome that were detected in multiple EwS but not in a set of non-EwS tumors. We then explored their expression pattern across normal and tumor tissues (Figure S1C). Sixty-one neotranscripts with EwS-specific expression were thus identified and assigned to splice variants of 25 EwS-specific neogenes (*Ew_NG*s) (Figure 2A).

This approach retrieved the four initial neotranscripts except *Ew_NG3*, which was filtered out in this procedure as it was already annotated in GENCODE (AC073135.3), but not in RefSeq that was used for the long-read approach. Interestingly, this GENCODE annotation is based on five expressed sequence tags (ESTs) derived from EwS cell lines (Kimura et al., 2006). *Ew_NG3* was therefore considered EwS-specific and kept in our final list of 26 *Ew_NG*s.

EWS::FLI1 directly regulates expression of *Ew_NGs*

We then explored the suspected role of EWS::FLI1 in the regulation of these *Ew_NG*s using nine EwS cell lines upon EWS::FLI1 inhibition by small interfering RNA (siRNA) or shRNA knockdown. Most *Ew_NG*s were consistently expressed in cell lines and downregulated by EWS::FLI1 depletion. Moreover, even though *Ew_NG*s were not expressed in MSCs, most were strongly induced in *EWS::FLI1*-expressing MSCs (Sole et al., 2021; Figure 2B). Conversely, tumors and normal tissues expressing high levels of wild-type *FLI1* such as acute myeloid leukemia and hematopoietic tissues did not express *Ew_NGs* (Figures 1D, 1E, and 2A). ChIP-seq (Aynaud et al., 2020) showed that GGAA microsatellite-bound EWS::FLI1 peaks were strongly enriched around the TSS of *Ew_NG*s (14/26 with a distance between TSS and nearest EWS::FLI1 microsatellite peak <10 kb, a much smaller distance than for most GENCODE transcripts,

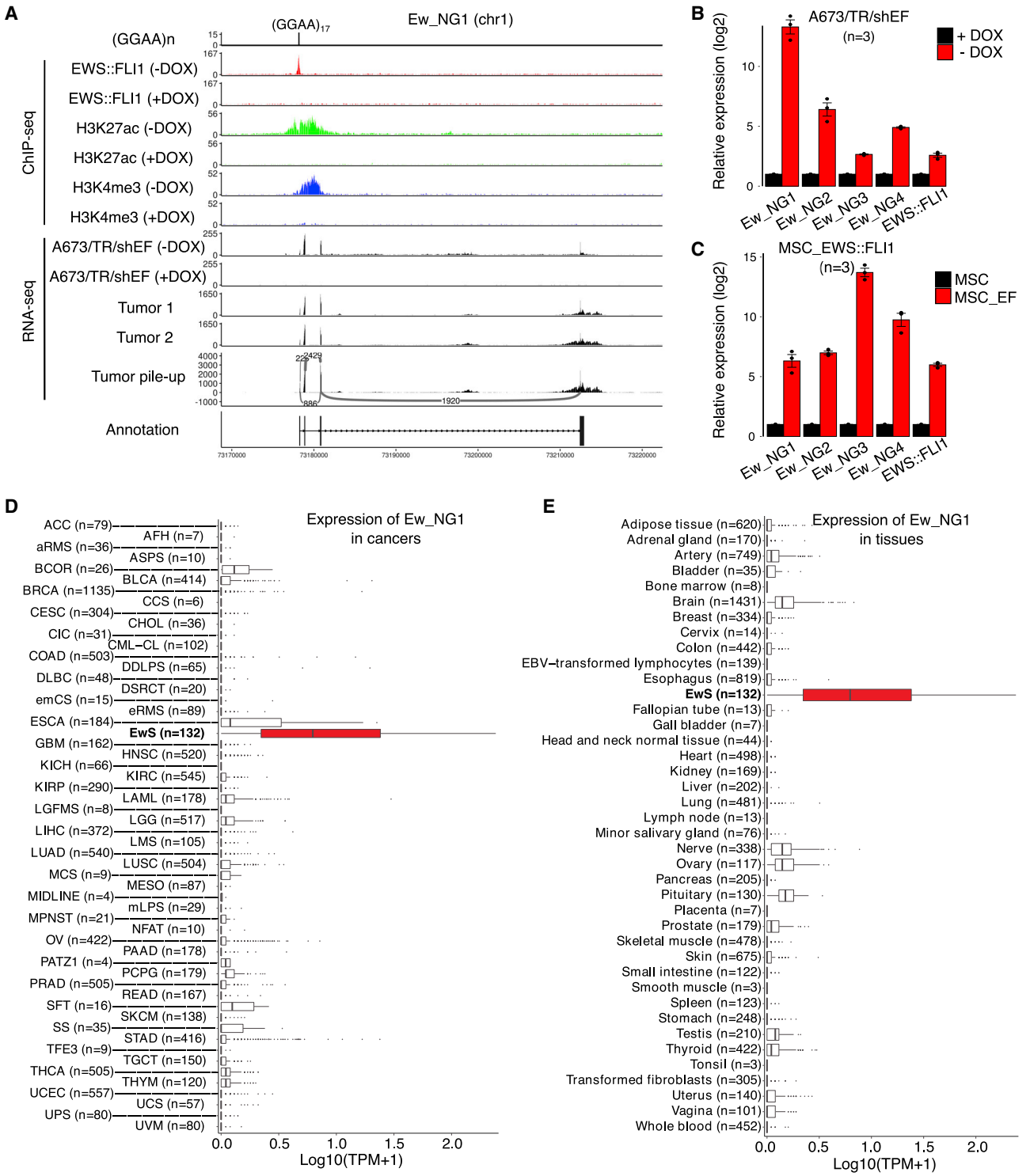


Figure 1. EWS::FLI1 regulates EwS-specific transcripts

(A) Genomic region of *Ew_NG1*. From top to bottom: GGAA microsatellites; ChIP-seq of EWS::FLI1, H3K27ac, and H3K4me3; RNA-seq of A673/TR/shEF cells in −DOX and +DOX conditions, RNA-seq of two individual tumors and of a pile-up of ten tumors (Sashimi plot shows junctions with >200 reads); and manual annotation of main spliced transcript.

(B and C) qRT-PCR analysis of *Ew_NG1-4* and *EWS::FLI1* in A673/TR/shEF cells in −DOX and +DOX conditions (B) and in wild type and *EWS::FLI1* (EF)-expressing MSCs (C); barplot represents the mean ± SEM of individual replicates (dots).

(legend continued on next page)

p value = 3.8e−15 by Wilcoxon's two-tailed test) (Figure 2C). The 26 Ew_NGs were distributed across 15 chromosomes without proximity toward centromeres or telomeres. Visual inspection of the genomic structure of *Ew_NG*s revealed that some were bidirectionally transcribed from the same EWS::FLI1-binding region and that most showed complex splicing patterns (Figure 2D). H3K27ac HiChIP, a protein-centric chromatin conformation mapping method (Mumbach et al., 2016), also identified 10 *Ew_NG*s containing EWS::FLI1-bound microsatellites within enhancer-promoter chains (Surdez et al., 2021), including 3 without direct binding of EWS::FLI1 near the TSS (Figure S5; Table S1A). To further demonstrate tumor-specific expression of *Ew_NG*s, we performed single-cell RNA-seq (scRNA-seq) on a EwS tumor and a desmoplastic small round cell tumor (DSRCT) and confirmed specific expression of *Ew_NG*s in EwS cells but neither in cells of the microenvironment nor in DSRCT cells, as shown on a uniform manifold approximation and projection (UMAP) plot of both samples (Figure 2E). Finally, CRISPR interference (CRISPRi) with three different sgRNAs targeting DNA sequences flanking EWS::FLI1-bound microsatellites upstream of six *Ew_NG*s, demonstrated a dramatic downregulation of all tested *Ew_NG*s (Figure 2F). Interestingly, CRISPRi targeting of the two microsatellites included in the enhancer-promoter chains of *Ew_NG*17 was necessary for a full downregulation of this NG (Figures 2G and S5). We noted no obvious phenotypic change of EwS cells after CRISPRi downregulation of NGs. Altogether, these experiments indicate that the expression of most *Ew_NG*s is a direct transcriptional consequence of EWS::FLI1 binding at GGAA microsatellites localized either in their promoter regions or in their enhancer-promoter chains. Notably, *Ew_NGs* were also found in EwS with non-EWS::FLI1 fusions, such as EWS::ERG and EWS::FEV, confirming the potential of these rarer EwS-defining variant fusions to induce NGs.

*Ew_NG*s encode EwS-specific neopeptides

We hypothesized that some of these neotranscripts could be translated into peptides. qRT-PCR after subcellular fractionation showed that all tested neotranscripts (NG1, 2, 3, 4, 8, and 17) were present in both nucleus and cytoplasm. Single-molecule RNA fluorescence *in situ* hybridization (smRNA FISH) more directly assessed the presence of transcripts in the nucleus and in the cytoplasm of EwS cells and their depletion upon EWS::FLI1 knockdown (Figures 2H, 2I, and S6). Using ribosome profiling (Ribo-seq) (Ingolia et al., 2009) in 2 EwS cell lines (A673 and EW7; Figure S7), we found that 16 *Ew_NG*s showed evidence of ribosome-protected fragments (RPFs) with a level of at least 0.1 transcripts per million (TPM), whereas no RPFs could be detected at similar levels at these loci in Ribo-seq of non-EwS cell lines (K562 and HepG2) (Calviello et al., 2020; Figure 3A).

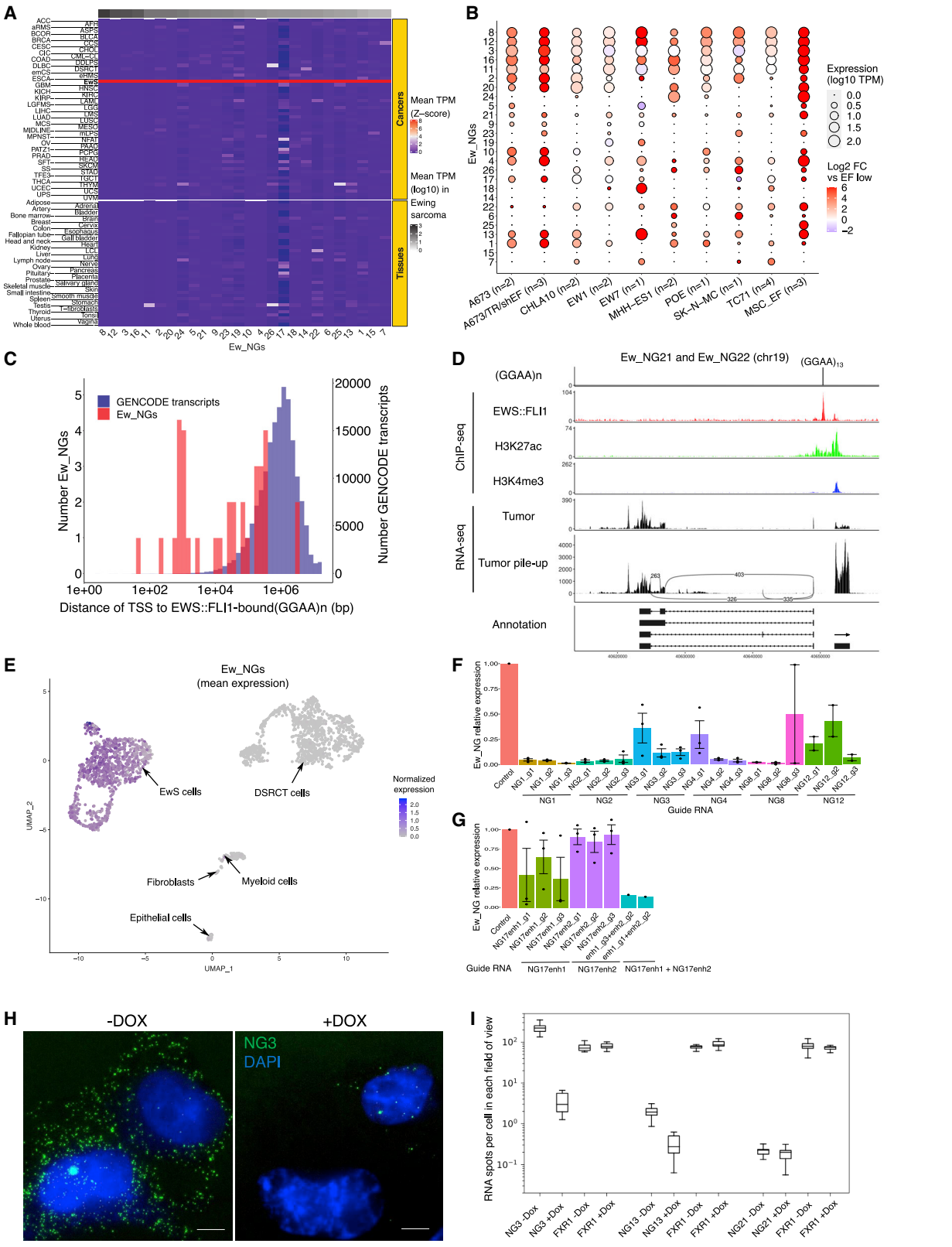
Intriguingly, a low level of RPFs (0.7 as compared with >400 TPM in EW7) mapping to *Ew_NG3* could be detected in K562, a chronic myelogenous leukemia (CML)-derived cell line, consistent with CML being the only non-EwS samples to harbor low (<2 TPM) but consistently non-zero expression of this NG in RNA-seq (Figure S3). Rigorous mapping of open reading frames (ORFs) based on Ribo-seq coverage, periodicity, and sequence characteristics identified 17 ORFs derived from 8 *Ew_NG*s (Table S2A). Quantitative label-free mass spectrometry analysis of 10 EwS cell lines and a deep proteomic workflow after high-pH reversed-phase peptide (HpH) fractionation of 2 EwS cell lines identified respectively 209 and 55 peptides, predicted to be derived from *Ew_NG*s (Tables S2B and S2C). Their hydrophobicity indices and retention times were similar to peptides from Uniprot-annotated known human proteins. Five of them matched ORFs predicted from Ribo-seq, including three—found in all replicates of all ten cell lines—that matched different parts of a predicted 88-amino-acid-long peptide encoded by an ORF detected in Ribo-seq for *Ew_NG3*, the most abundant *Ew_NG* in RNA-seq and Ribo-seq (Figures 3B and 3C). In contrast, no peptides matching these ORFs were identified by mass spectrometry of eight non-EwS medulloblastoma tumors investigated with a similar approach using deep proteomic workflow after HpH fractionation. Further characterization of one Ew_NG3 peptide (Figure 3D), using parallel reaction monitoring (PRM) experiments, showed that the mass spectrometry spectrum of peptides detected in EwS cells lines matched that of a similar synthetic peptide labeled with a heavy isotope; this was not the case in EwS cells with Ew_NG3 knockout (A673 KO NG3), with knockdown of EWS::FLI1 (ASP14 +DOX) and in non-Ewing cells (Figures 3E and 3F). This spectrum match was also present in MSCs expressing EWS::FLI1 (MSC-7-BJEF1) but not in parental MSCs (MSC-7-BJ) (Figure 3F). Thus, this peptide is unambiguously present in EwS cells, and it is dependent upon EWS::FLI1. Altogether, these data show that *Ew_NG*s can harbor actively translated ORFs and encode novel tumor-specific peptides ("neopeptides").

EWS::WT1 generates neotranscripts in DSRCT

These results in EwS prompted us to investigate other sarcomas characterized by OCTFs. DSRCT is a soft tissue sarcoma characterized by *EWS::WT1* fusions. Using the same strategy of genome-guided assembly of short-read RNA-seq from clinical samples, we identified 37 DSRCT-specific neogenes (DSRCT_NGs) corresponding to 105 DSRCT-specific neotranscripts (Figure 4A).

Similarly, we explored the potential role of EWS::WT1 in their expression. Using published RNA-seq data (Gedminas et al., 2020), we found that most DSRCT_NGs were consistently expressed in DSRCT cell lines and strongly downregulated upon EWS::WT1 knockdown (Figure 4B). Using published ChIP-seq data in JN-DSRCT-1 cells (Hingorani et al., 2020), we observed a strong enrichment of EWS::WT1 peaks around TSS of DSRCT_NGs (24/37 with distance between TSS and nearest EWS::WT1 peak <10 kb, p value = 1.3e−10, Wilcoxon's two-tailed test compared with GENCODE transcripts). EWS::WT1-bound sites were associated with RNA polymerase II (RNA Pol

(D and E) Expression of *Ew_NG1* in cancers (D) and in normal tissues (E). TPM, transcripts per million. Box represents the interquartile range, upper and lower whiskers the largest and smallest values within 1.5 times the interquartile range from the ends of the box. Abbreviations for in-house datasets are identical to Table 1. Other abbreviations are those used in the TCGA dataset.
See also Figures S1A S1B, and S2–S4.



(legend on next page)

II) binding by ChIP-seq (Figures 4C and 4D). Integration of scRNA-seq from DSRCT and EwS tumor samples also highlighted specific expression of DSRCT_NGs within DSRCT cells but not in EwS or microenvironment cells (Figure 4E).

Based on our observations in EwS and DSRCT, we propose the concept of "OCTF-driven" NGs that demonstrate the following: (1) specific expression in a given OCTF-driven cancer type, (2) expression regulation by the OCTF in cell lines, and (3) OCTF binding near the TSS or within an enhancer-promoter regulatory chain. Using these criteria, 16 (out of 26) *Ew_NG*s and 19 (out of 37) DSRCT_NGs, i.e., most of these cancer type-specific NGs, are considered OCTF-driven (Table S1). These numbers may still be underestimated due to low ChIP-seq sensitivity or unmapped long-range regulatory interactions. Conversely, expression of some cancer type-specific NGs appears to be OCTF-independent and could potentially be rather linked to the specific tumor cell of origin.

Specific neogenes characterize other OCTF-driven cancers

We further hypothesized that such NGs could exist in other OCTF-driven malignancies. Using our institutional database of clinical RNA-seq and the same assembly-based strategy, we studied 16 other OCTF-driven cancers, all being sarcomas except midline carcinoma and TFE3-translocated renal cell carcinoma. Cancer type-specific neotranscripts were identified in every malignancy (Figure 5A).

Overall, we found 398 NGs corresponding to 807 neotranscripts across 18 cancer types (Tables 1 and S3). Most neotranscripts were multi-exonic (58.7%) and used consensus splice sites (>99%) (Figure 5B). Evolutionary sequence conservation analysis showed much lower constraint than for protein-coding genes, as the sequence conservation of neotranscripts was similar to that of previously characterized and annotated long intergenic non-coding RNAs (lincRNAs), most demonstrating little to no evolutionary constraint at the sequence level (Figure 5C, Hezroni et al., 2015).

For the additional cancer types, we could not distinguish between OCTF-driven and OCTF-independent NGs since only clinical samples were available. However, we note that both angiomatoid fibrous histiocytoma (AFH) and clear cell sarcoma (CCS), two cancers with very different clinico-pathological features but that share an identical OCTF (EWS::ATF1/CREB1), express many common neotranscripts. A similar pattern of sharing is also observed between alveolar soft part sarcoma (ASPS) and TFE3-translocated renal cell carcinoma that express the same ASPSCR1::TFE3 OCTF (Figure 5A). This suggests that the common NGs are directly regulated by the shared OCTF. Interestingly, a small number of the NGs found in alveolar rhabdomyosarcoma (aRMS), characterized by the aberrant PAX3/7::FOXO1 TF, are also expressed in PAX3/7::FOXO1-negative embryonal rhabdomyosarcoma, suggesting conversely that these shared NGs are rather related to a common tissue of origin.

DISCUSSION

Altogether, our results show the existence of exquisitely specific novel transcription units (*NGs*) in many OCTF-driven cancers, of which a significant fraction are directly induced (OCTF-driven) by the OCTF. Whereas mutated TFs are known to affect many steps of mRNA transcription and processing in diverse cancers, the generation of such neotranscripts from otherwise silent chromatin regions is a qualitatively different neomorphic activity of this well-studied class of oncogenes. These polyadenylated transcripts of more than 200 nucleotides long, mostly multi-exonic and without long ORFs, can be considered as lincRNAs (Statello et al., 2021). Accordingly, they show lincRNA-type sequence conservation scores. Though many cancer-specific lincRNAs have been described (Iyer et al., 2015), their regulatory mechanisms remain elusive. Our data indicate that one such mechanism relies on gain-of-function properties of OCTFs. In EwS, the current knowledge on the mechanism of action of EWS::FLI1 enables the proposition of a conceptual framework. Specifically, EWS::FLI1 presents the neomorphic ability to bind GGAA microsatellites and to recruit chromatin remodeling complexes and histone modifying enzymes, as well as the transcription machinery, possibly through phase transition mechanisms mediated by the low-complexity EWS domain (Boulay et al., 2017). This leads to localized transcription of unprocessed enhancer RNAs (eRNAs) at many sites (Boulay et al., 2018) but also, as shown here, to highly abundant, spliced, transcripts

Figure 2. Features of EwS-specific neogenes

(A) Heatmap of *Ew_NG*s expression levels in cancers and normal tissues. Expression levels are in TPM. Main heatmap reports *Z* scores scaled by neogene. *Ew_NG*s are ordered by mean expression in EwS shown in the top heatmap (abbreviations are as in Figure 1).

(B) Modulation of *Ew_NG*s expression by EWS::FLI1. *Ew_NG*s are ordered as in (A). Dot size shows mean expression level in EwS cell lines and MSCs transformed by EWS::FLI1 (capped at 100). Color represents $\log_2$-fold change (capped at 6) as compared respectively with EWS::FLI1 knockdown conditions and parental MSCs (EF low). Data from GSE133228, GSE164373, and GSE150783 were used in this figure.

(C) Distances between TSS and nearest EWS::FLI1-bound GGAA microsatellite for *Ew_NG*s and GENCODE transcripts.

(D) Genomic region of two bidirectional *Ew_NG*s. From top to bottom: GGAA microsatellites; ChIP-seq of EWS::FLI1, H3K27ac, and H3K4me3 in A673/TR/shEF cells; RNA-seq of one tumor and of a pile-up of ten tumors (Sashimi plot shows junctions with >200 reads); and predicted annotation.

(E) UMAP plot of scRNA-seq of EwS and DSRCT tumor samples showing mean expression level of *Ew_NG*s.

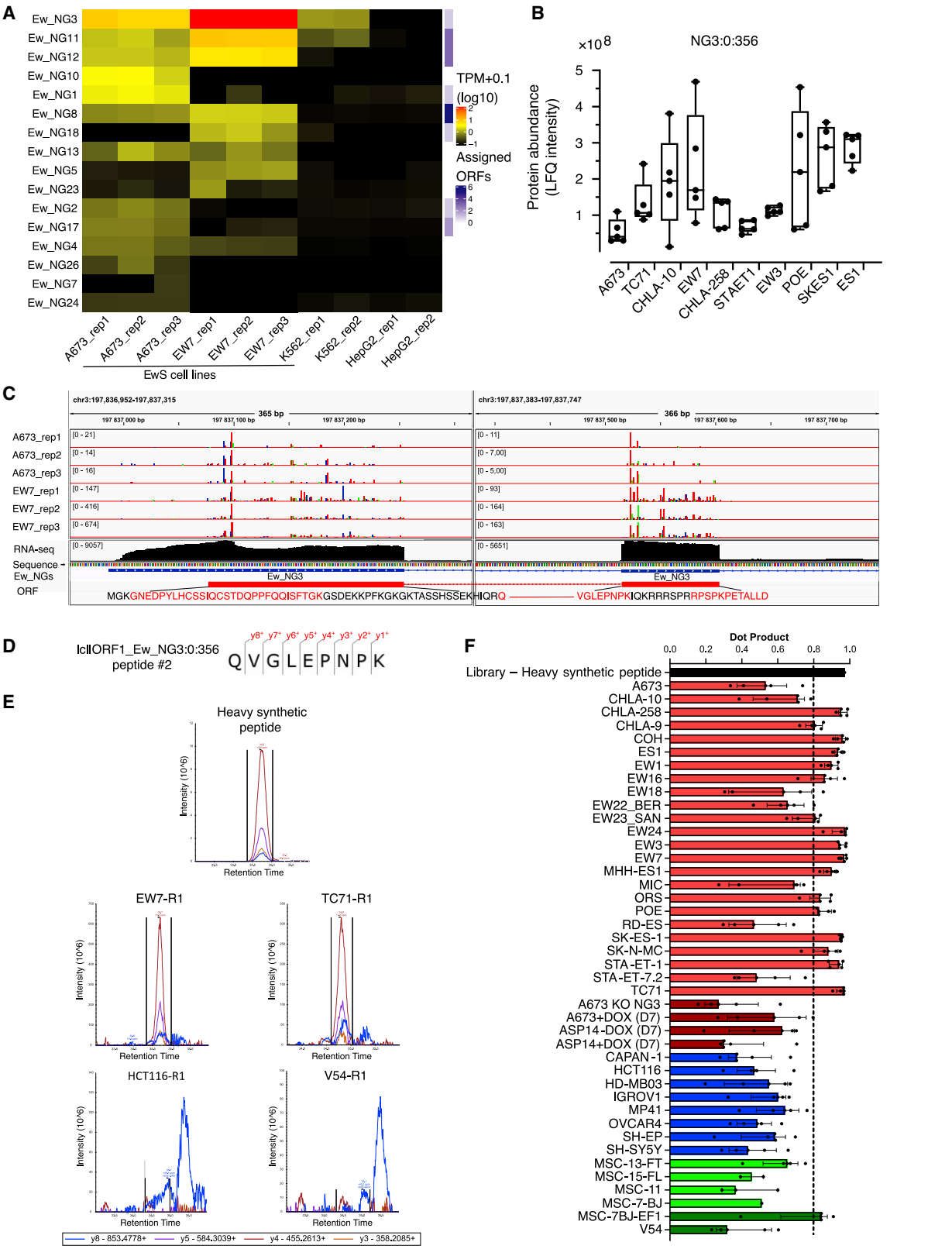
(F) CRISPRi experiments with three different guide RNAs for NG1, 2, 3, 4, 8, and 12.

(G) CRISPRi experiment targeting the two EWS::FLI1-bound microsatellites regulating NG_17 (enh1 and enh2). Barplots in (F) and (G) show mean ± SEM of individual replicates (dots).

(H) Single-molecule RNA FISH with NG3 transcript. Probes for NG3 were labeled with Quasar570 (green spots) and nuclei with DAPI. +DOX refers to inhibition of EWS::FLI1 with an inducible shRNA. Scale bars, 10 μm.

(I) Quantification of smRNA FISH on three different *Ew_NGs* upon EWS::FLI1 expression (−DOX) or knockdown (+DOX). The FXR1 transcript, which is not regulated by EWS::FLI1, is used as an internal control. Box represents the interquartile range, upper and lower whiskers the largest and smallest values within 1.5 times the interquartile range from the ends of the box.

See also Figures S1C S5, and S6 and Tables S1 and S4.



(legend on next page)

from otherwise silent regions of the genome (Figure 5D). The emergence of this latter phenomenon only at a subset of EWS::FLI1-bound microsatellites may depend on local sequence features (Core et al., 2014), chromatin architecture, and three-dimensional structure of the genome. Additionally, we acknowledge that our search may not be exhaustive since we restricted our search to highly expressed transcripts. Additional neotranscripts with lower expression levels may exist at other EWS::FLI1 binding sites. The same model is also expected to apply to DSRCT, though EWS::WT1-specific binding sites have not been precisely defined yet (Hingorani et al., 2020). The observation that different cancer types expressing the same OCTF share a set of neotranscripts also supports a similar mechanism in other tumors.

These results open several research perspectives and translational opportunities. Molecular studies could shed light on the precise mechanisms from OCTF DNA binding to transcription activation and explore their potential functional roles. Indeed, lincRNAs can exhibit diverse functions (Statello et al., 2021), and some NGs may play a role in oncogenesis. However, the absence of sequence conservation across species and the complexity of their splicing patterns do not support a strong selection pressure (Palazzo and Koonin, 2020) and rather suggest that many NGs may be obligatory, but non-functional, "by-products" of general OCTF activity. The tumor specificity and high recurrence across patients of these neotranscripts make them attractive diagnostic markers. Additionally, our demonstration of translation by both Ribo-seq and proteomics in EwS, together with recent studies indicating an unexpected potential of lincRNAs to generate HLA class I-associated peptides (Chen et al., 2020; Chong et al., 2020; Laumont et al., 2016, 2018; Ouspenskaia et al., 2022; Ruiz Cuevas et al., 2021), suggest that peptides from these NGs may be presented on the surface as exquisitely tumor-specific antigens. EwS and other OCTF-driven sarcomas are usually poorly responsive to immune checkpoint therapies and are generally defined as "immune cold," suggesting that the recognition of the putative neoantigens by the immune system may be impaired. Different hypotheses may account for this apparent discrepancy. Clonal deletion of tumor-responsive T cells, T cell exclusion from tumors resulting from immunosuppressive cells—including tumor cells themselves, macrophages, myeloid-derived cells or cancer-associated fibroblasts—as well as tumor vasculature barriers and unique metabolic environments are some of the mechanisms currently proposed to account for T cell depletion in the tumor microenvironment (Lanitis et al., 2017). In synovial sarcoma (SS), a tumor driven by the SS18::SSX OCTF, recent data showed the potential therapeutic interest of co-targeting the interplay between immune evasion and oncogenic processes (Jerby-Arnon et al., 2021). In neuroblastoma, another "immune cold" tumor, a recent paper showed that peptide-centric chimeric antigen receptor T cells can target the cancer cells *in vivo* (Yarmarkovich et al., 2021). Finally, preliminary results from the Ewing cell vaccination Vigil study suggest that Ewing cells are indeed capable of generating an immune response (Ghisoli et al., 2016). Such peptides, expressed by all tumors expressing a given OCTF, may therefore constitute a source of public neoantigens of considerable interest for immunotherapy approaches.

Finally, our observations on OCTFs raise the hypothesis that other types of gain-of-function mutations of TFs, being cancer-associated or germline, may also lead to new DNA binding properties leading to aberrant transcription and splicing of otherwise silent intergenic regions. It is tempting to speculate that, across evolution, such mutations may generate new transcription units that could subsequently acquire specific functions (Van Oss and Carvunis, 2019; Palazzo and Koonin, 2020).

Limitations of the study

Although we identified tumor-specific transcripts in many OCTF-driven cancers, we only showed their direct regulation by the OCTF in EwS and DSRCT, as we did not have functional data to explore this mechanism in other less well-studied OCTF-driven cancers. Nonetheless, we showed indirect evidence in favor of this hypothesis, and we believe that future functional experiments may confirm it. Moreover, some of the tumor-specific transcripts in EwS and DSRCT did not show evidence of being directly driven by the OCTF. Other mechanisms may also be considered to account for this specificity such as tissue-specific transcripts—as observed for some common transcripts between alveolar and embryonal rhabdomyosarcoma. Indirect regulation by the OCTF or other unidentified tumor-specific mechanisms can also be hypothesized. Our search for neotranscripts was not necessarily exhaustive, as our pipeline was designed to recover very specific, broadly, and highly expressed transcripts; enlarging the search over more tumor samples and loosening filtering criteria may

Figure 3. EwS-specific neotranscripts can be translated into neopeptides

(A) Heatmap of ribosome-protected fragments (RPFs) mapping to *Ew_NG*s in EwS and non-EwS cell lines. Levels are in $\log_{10}$(TPM + 0.1). *Ew_NG*s are ordered from top to bottom by maximum RPF levels. Right heatmap reports number of computationally predicted ORFs in corresponding *Ew_NG*. rep: replicate.

(B) Mass spectrometry label-free quantification (LFQ) analysis of a neopeptide derived from *Ew_NG3* (NG3:0:356) in ten EwS cell lines performed from five replicates (each dot is a replicate). Box represents the interquartile range, upper and lower whiskers the largest and smallest values within 1.5 times the interquartile range from the ends of the box.

(C) Genomic view of first two exons of *Ew_NG3*, derived ORF predicted by Ribo-seq and corresponding peptides identified in mass spectrometry (NG3:0:356). From top to bottom: Ribo-seq P-sites for EwS cell lines, colored by frame (0: red; 1: green; 2:blue); RNA-seq in A673; nucleotide sequence; transcript annotation; and predicted ORF. Peptides highlighted in red are identified in mass spectrometry (quantification in [B]).

(D) Sequence of tryptic neopeptide lcl|ORF1_Ew_NG3:0:356 with the observed ions identified by tandem mass spectrometry (MS/MS).

(E) Extracted ion chromatogram examples of heavy synthetic peptide and endogenous peptides measured by PRM in Ewing cell lines (EW7, TC71) and non-Ewing cell lines (HCT116, V54). R1: replicate 1.

(F) Dot product values in all tested cell lines; Ewing (red) n = 24, Ewing cellular models (dark red) n = 4, other cancers cell lines (blue) n = 8, MSCs (green) n = 4, and MSCs models (dark green) n = 2. Dot product value indicates the degree of the match between spectral library MS/MS and the extracted ion chromatograms of the corresponding transitions. High dot product (>0.8) indicates the absence of interfering signals. Barplots show mean ± SEM of individual replicates (dots). See also Figure S7 and Tables S2 and S5.

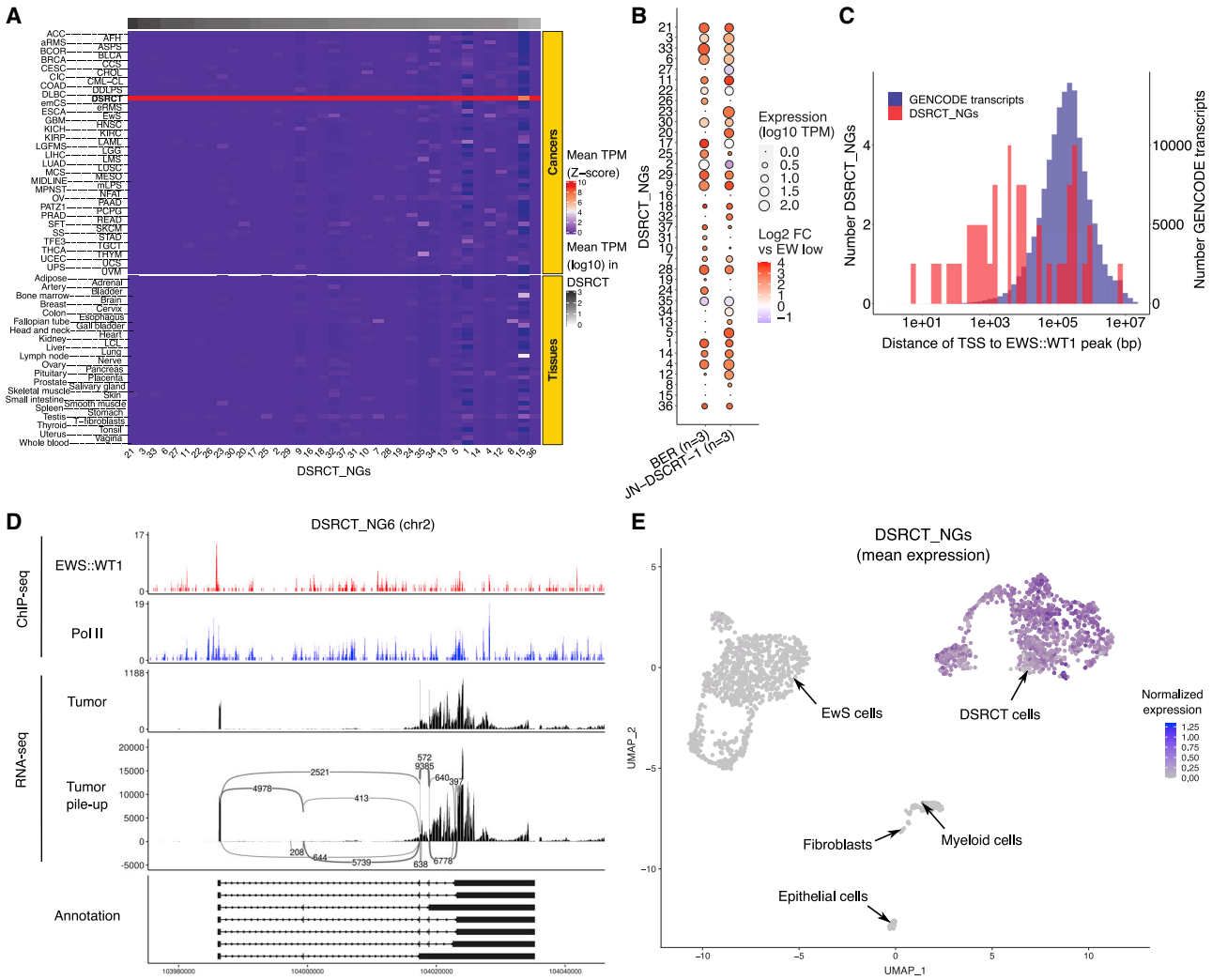


Figure 4. Features of DSRCT-specific neogenes

(A) Heatmap of DSRCT_NGs expression levels in cancers and normal tissues. Expression levels are in TPM. Main heatmap reports *Z* scores scaled by neogene. DSRCT_NGs are ordered by mean expression in DSRCT shown in the top heatmap (abbreviations are as in Figure 1).

(B) Modulation of DSRCT_NGs expression by EWS::WT1. DSRCT_NGs are ordered as in (A). Dot size shows mean expression level in DSRCT cell lines. Color represents $\log_2$-fold change as compared with EWS::WT1 knockdown conditions (EW low) (Gedminas et al., 2020).

(C) Distances between TSS and nearest EWS::WT1 peak for DSRCT_NGs and GENCODE transcripts.

(D) Genomic region of a representative DSRCT neogene. From top to bottom: ChIP-seq of EWS::WT1 and RNA Pol II in JN-DSRCT-1 cells (Hingorani et al., 2020); RNA-seq of one tumor and of a pile-up of ten tumors (Sashimi plot shows junctions with >200 reads); and predicted annotation.

(E) UMAP plot of scRNA-seq of DSRCT and EwS tumor samples showing mean expression level of DSRCT_NGs.

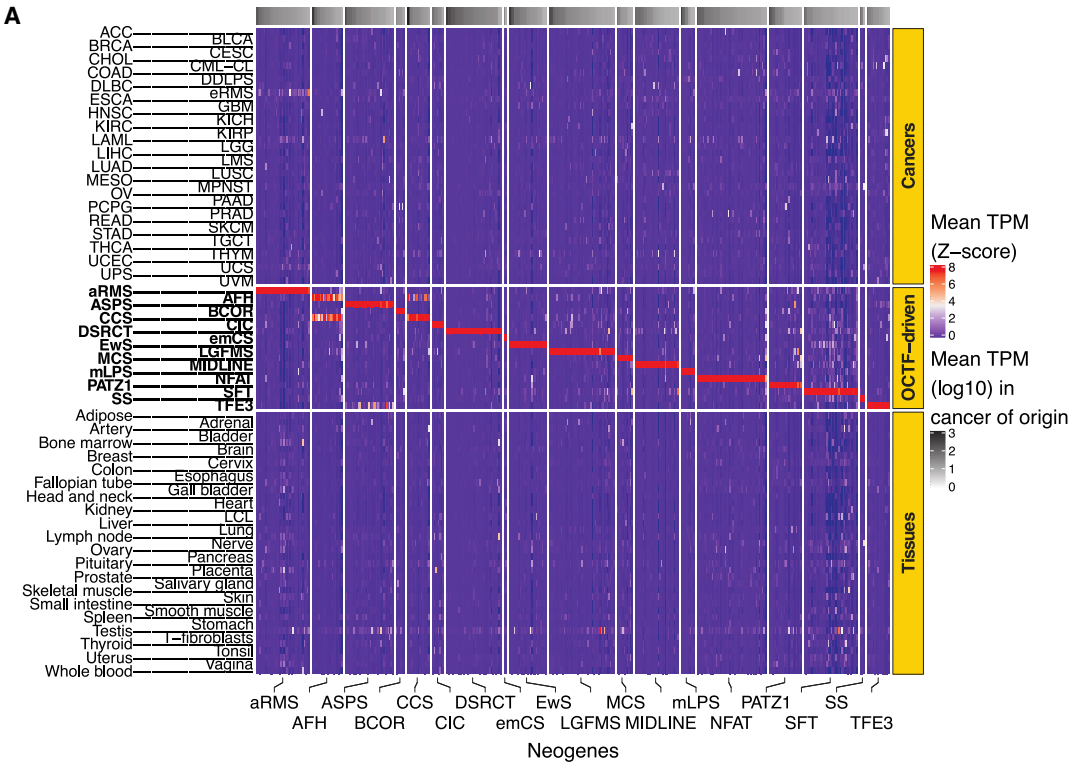
See also Table S1.

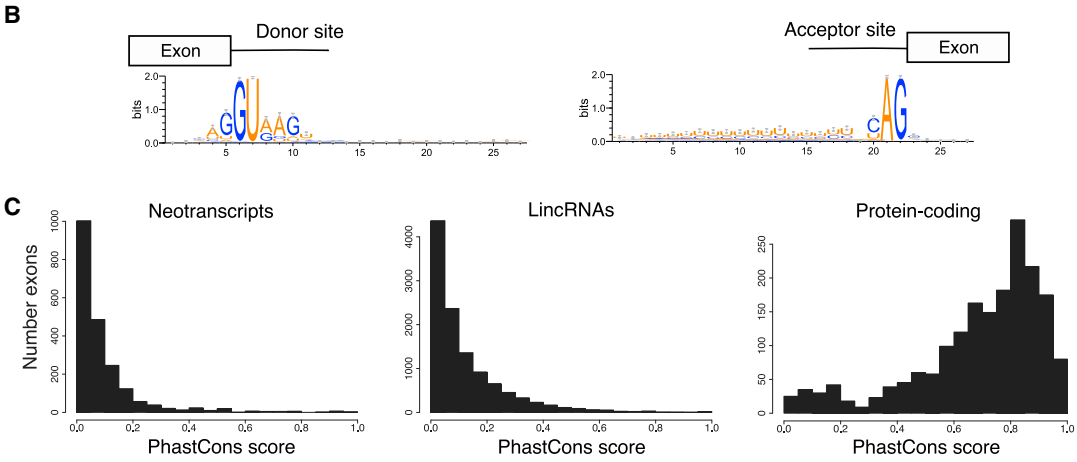
identify many other neotranscripts with lower and less widespread expression across patients. Future experiments are needed to explain the exact mechanisms leading an OCTF to activate transcription in closed chromatin regions and to investigate whether some neotranscripts have any functional roles. Although we showed evidence that some EwS neotranscripts are translated into peptides, we do not as yet have data for other OCTF-driven cancers. Finally, the neoantigenic potential of these tumor-specific peptides remains to be demonstrated, though their exquisite specificity and recurrent expression in patients make them ideal targets for off-the-shelf immunotherapies.

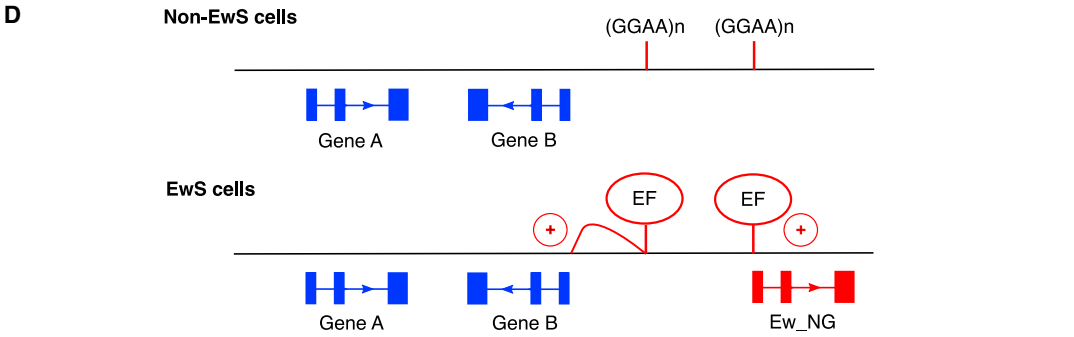
STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS







(legend on next page)

Table 1. List of the 18 OCTF-driven cancers studied, associated OCTF, and numbers of neogenes and neotranscripts discovered for each

Cancer type	Abbreviation	OCTF	Number of neogenes	Number of neotranscripts
Angiomatoid fibrous histiocytoma	AFH	EWSR1::ATF1/CREB1*	20	28
Alveolar rhabdomyosarcoma	aRMS	PAX3*/PAX7::FOXO1	36	72
Alveolar soft part sarcoma	ASPS	ASPSCR1::TFE3	33	103
BCOR-rearranged sarcoma	BCOR	BCOR::CCNB3	6	6
Clear cell sarcoma	CCS	EWSR1::ATF1*/CREB1	15	24
CIC-fused sarcoma	CIC	CIC::DUX4*/NUTM1	8	16
Desmoplastic small round cell tumor	DSRCT	EWSR1::WT1	37	105
Extraskeletal myxoid chondrosarcoma	emCS	EWSR1::NR4A3	2	2
Ewing sarcoma	EwS	EWSR1::FLI1*/ERG	26	62
Low-grade fibromyxoid sarcoma	LGFMS	FUS::CREB3L2	44	70
Mesenchymal chondrosarcoma	MCS	HEY::NCOA2	11	21
Midline carcinoma	MIDLINE	BRD::NUT	29	64
Myxoid liposarcoma	mLPS	FUS::DDIT3	9	14
EWSR1::NFATC2 sarcoma	NFAT	EWSR1::NFATC2	47	88
EWSR1::PATZ1 sarcoma	PATZ1	EWSR1::PATZ1	22	30
Solitary fibrous tumor	SFT	NAB2::STAT6	35	54
Synovial sarcoma	SS	SS18::SSX1*/SSX2	3	16
TFE3-translocated renal cell carcinoma	TFE3	ASPSCR1::TFE3	15	32

For cancers with alternative fusion partners, the specific gene found in the index patient used for neotranscript identification is indicated with an asterisk (*). See also Table S3.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.molcel.2022.04.019.

ACKNOWLEDGMENTS

We thank all members of the Diversity and Plasticity of Pediatric Tumors and Integrative Functional Genomics of Cancer laboratories for helpful discussions. We thank H. Kovar and the Childhood Cancer Repository for providing EwS cell lines and A. Solé and E. Brunet for providing RNA-seq from MSCs and EWS::FLI1-expressing MSCs. We are grateful to C. Pierre-Eugène for her experimental assistance as well as P. Legoix, V. Marsaud, V. Laigle, C. Kamoun, and E. Barillot for bioinformatic or technical assistance. We thank S. van Heesch for advice on ribosome profiling. The results shown here are in

Figure 5. Cancers with gene fusions involving a transcription factor express specific neogenes

(A) Heatmap of expression levels in cancers and normal tissues of all neogenes found in OCTF-driven cancers. Expression levels are in TPM. Main heatmap reports *Z* scores scaled by neogene. Top heatmap shows mean expression of neogenes in corresponding tumors (abbreviations are as in Figure 1).

(B) Sequence logo for splicing junctions of neogenes.

(C) Sequence conservation scores for neotranscripts, lincRNAs, and a subset of protein-coding transcripts in GENCODE. Conservation scores are calculated as the mean of all PhastCons base scores for each exon.

(D) Model of induction of specific neogenes by EWS::FLI1. EWS::FLI1 binding sites such as GGAA microsatellites are in closed chromatin conformation in non-EwS cells (top). Upon binding of EWS::FLI1 (EF), these sequences are transformed into neo-enhancers able to activate neighboring known target genes (Gene B) but also to induce transcription of *Ew_NGs*.

part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The GTEx project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx database through dbGaP. With regard to funding, this work was supported by Institut Curie, Institut national de la santé et de la recherche médicale (INSERM), Ligue Nationale Contre le Cancer (Equipe labellisée and program Ligue EAC 2020), Institut National du Cancer (PLBIO19-192), Agence Nationale de la Recherche (ANR-10-EQPX-03), Institut Curie Génomique d'Excellence (ICGex), and Société française de lutte contre les leucémies et cancers de l'enfant et de l'adolescent. This project also received support from European funding as follows: ERA-NET TRANSCAN JTC-2011 (01KT1310), TRANSCAN JTC-2014 (TRAN201501238), TRANSCAN JTC-2017 (TRANS201801292), EEC (HEALTH-F2-2013-602856), H2020-IMI2-JTI-201 5-07 (116064—ITCC P4), and H2020-SC1-DTH-2018-1 (SEP-210506374—iPC). We are indebted to the following associations for providing essential support: L'Etoile de Martin, la Course de l'Espoir, M la vie avec Lisa, ADAM, Couleur Jade, Dans les pas du Géant, Courir pour Mathieu, Marabout de Ficelle, Olivier Chape, Les Bagouzamanon, Enfants et Santé, Les Amis de Claire, Un Elan pour Lucas, and Amarape. J.V. is supported by a Ligue Nationale Contre le Cancer PhD fellowship and Institut Curie. O.S. was supported by a fellowship from the French Ministry of Higher Education and Research and Institut Curie. C.H. is supported by a Gustave Roussy Philanthropia fellowship. D.S. is supported by the Institut Curie-SIRIC (Site de Recherche Intégrée en Cancérologie) program. The laboratory of T.G.P.G. is supported by the Matthias-Lackas Foundation, the Dr. Leopold and Carmen Ellinger Foundation, the Boehringer-Ingelheim Foundation, the Dr. Rolf M. Schwiete Foundation, the German Cancer Aid (DKH-70112257 and DKH-70114278), the Gert und Susanna Mayer Foundation, the Barbara und Wilfried Mohr Foundation, the SMARCB1 association, and the Deutsche Forschungsgemeinschaft (DFG-458891500). Finally, the laboratory of A.C. is supported by the LabEx DEEP (ANR-11-LABX-0044, ANR-10-IDEX-0001-02) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 757956).

AUTHOR CONTRIBUTIONS

Conceptualization, J. Vibert, O.S., J.J.W., and O.D.; methodology, J. Vibert, O.S., J.J.W., and O.D.; investigation, J. Vibert, O.S., C.C., F.P., K.J.E.B., J. Vigneau, S.Z., N.G., K.L.-D., F.D., and D.S.; data curation, J. Vibert, O.S., G.P., and S.W.; resources, C.H., S.P.-V., V.R., S.B., F.D., D.L., J.T., O.A., M.F.O., T.G.P.G., and A.C.; formal analysis, J. Vibert, O.S., M.G., M.D., V.H., S.G., and J.T.; software, J. Vibert, O.S., M.G., M.D., V.H., and S.G.; visualization, J. Vibert, O.S., C.C., F.P., K.J.E.B., J. Vigneau, M.G., V.H., D.L., A.C., J.J.W., and O.D.; funding acquisition, J.J.W. and O.D.; project administration, J.J.W. and O.D.; supervision, J.J.W. and O.D.; writing—original draft, J. Vibert, O.S., J.J.W., and O.D.; writing—review & editing, all authors.

DECLARATION OF INTERESTS

J. Vibert, O.S., J. Vigneau, M.G., C.C., J.J.W., and O.D. are named inventors of a patent application on neotranscripts. All other authors declare that they have no competing interests.

Received: September 15, 2021
Revised: February 20, 2022
Accepted: April 14, 2022
Published: May 11, 2022

REFERENCES

Aynaud, M.-M., Mirabeau, O., Gruel, N., Grossetête, S., Boeva, V., Durand, S., Surdez, D., Saulnier, O., Zaïdi, S., Gribkova, S., et al. (2020). Transcriptional programs define intratumoral heterogeneity of Ewing sarcoma at single-cell resolution. Cell Rep. *30*, 1767–1779.e6.

Boulay, G., Sandoval, G.J., Riggi, N., Iyer, S., Buisson, R., Naigles, B., Awad, M.E., Rengarajan, S., Volorio, A., McBride, M.J., et al. (2017). Cancer-specific retargeting of BAF complexes by a prion-like domain. Cell *171*, 163–178.e19.

Boulay, G., Volorio, A., Iyer, S., Broye, L.C., Stamenkovic, I., Riggi, N., and Rivera, M.N. (2018). Epigenome editing of microsatellite repeats defines tumor-specific enhancer functions and dependencies. Genes Dev. *32*, 1008–1019.

Bushweller, J.H. (2019). Targeting transcription factors in cancer—from undruggable to reality. Nat. Rev. Cancer *19*, 611–624.

Calviello, L., Hirsekorn, A., and Ohler, U. (2020). Quantification of translation uncovers the functions of the alternative transcriptome. Nat. Struct. Mol. Biol. *27*, 717–725.

Calviello, L., Sydow, D., Harnett, D., and Ohler, U. (2019). Ribo-seQC: comprehensive analysis of cytoplasmic and organellar ribosome profiling data. Preprint at bioRxiv. https://doi.org/10.1101/601468.

Carrillo, J., García-Aragoncillo, E., Azorín, D., Agra, N., Sastre, A., González-Mediero, I., García-Miguel, P., Pestaña, A., Gallego, S., Segura, D., et al. (2007). Cholecystokinin down-regulation by RNA interference impairs Ewing tumor growth. Clin. Cancer Res. *13*, 2429–2440.

Chen, J., Brunner, A.-D., Cogan, J.Z., Nuñez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann, M., Leonetti, M.D., et al. (2020). Pervasive functional translation of noncanonical human open reading frames. Science *367*, 1140–1146.

Chong, C., Müller, M., Pak, H., Harnett, D., Huber, F., Grun, D., Leleu, M., Auger, A., Arnaud, M., Stevenson, B.J., et al. (2020). Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. Nat. Commun. *11*, 1293.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. *46*, 1311–1320.

Coulon, A., Ferguson, M.L., de Turris, V., Palangat, M., Chow, C.C., and Larson, D.R. (2014). Kinetic competition during the transcription cycle results in stochastic RNA processing. Elife *3*, e03939.

Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol. Cell. Proteomics *13*, 2513–2526.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

Delattre, O., Zucman, J., Plougastel, B., Desmaze, C., Melot, T., Peter, M., Kovar, H., Joubert, I., de Jong, P., Rouleau, G., et al. (1992). Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. Nature *359*, 162–165.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinform. Oxf. Engl. *29*, 15–21.

Edelstein, A., Amodaj, N., Hoover, K., Vale, R., and Stuurman, N. (2010). Computer control of microscopes using μManager. Curr. Protoc. Mol. Biol. *92*, 14.20.1–14.20.17.

Gangwal, K., Sankar, S., Hollenhorst, P.C., Kinsey, M., Haroldsen, S.C., Shah, A.A., Boucher, K.M., Watkins, W.S., Jorde, L.B., Graves, B.J., et al. (2008). Microsatellites as EWS/FLI response elements in Ewing's sarcoma. Proc. Natl. Acad. Sci. USA *105*, 10149–10154.

Gedminas, J.M., Chasse, M.H., McBrairty, M., Beddows, I., Kitchen-Goosen, S.M., and Grohar, P.J. (2020). Desmoplastic small round cell tumor is dependent on the EWS-WT1 transcription factor. Oncogenesis *9*, 41.

Ghisoli, M., Barve, M., Mennel, R., Lenarsky, C., Horvath, S., Wallraven, G., Pappen, B.O., Whiting, S., Rao, D., Senzer, N., et al. (2016). Three-year follow up of GMCSF/bi-shRNA(furin) DNA-transfected autologous tumor immunotherapy (Vigil) in metastatic advanced Ewing's sarcoma. Mol. Ther. J. Am. Soc. Gene Ther. *24*, 1478–1483.

Grünewald, T.G.P., Cidre-Aranaz, F., Surdez, D., Tomazou, E.M., de Álava, E., Kovar, H., Sorensen, P.H., Delattre, O., and Dirksen, U. (2018). Ewing sarcoma. Nat. Rev. Dis. Primers *4*, 5.

Guillon, N., Tirode, F., Boeva, V., Zynovyev, A., Barillot, E., and Delattre, O. (2009). The oncogenic EWS-FLI1 protein binds *in vivo* GGAA microsatellite sequences with potential transcriptional activation function. PLoS One *4*, e4932.

Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296.

Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P., and Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep. *11*, 1110–1122.

Hingorani, P., Dinu, V., Zhang, X., Lei, H., Shern, J.F., Park, J., Steel, J., Rauf, F., Parham, D., Gastier-Foster, J., et al. (2020). Transcriptome analysis of desmoplastic small round cell tumors identifies actionable therapeutic targets: a report from the Children's Oncology Group. Sci. Rep. *10*, 12318.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. *47*, 199–208.

Jerby-Arnon, L., Neftel, C., Shore, M.E., Weisman, H.R., Mathewson, N.D., McBride, M.J., Haas, B., Izar, B., Volorio, A., Boulay, G., et al. (2021). Opposing immune and genetic mechanisms shape oncogenic programs in synovial sarcoma. Nat. Med. *27*, 289–300.

Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H., et al. (2006). Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. Genome Res. *16*, 55–65.

Kovar, H., Jug, G., Aryee, D.N., Zoubek, A., Ambros, P., Gruber, B., Windhager, R., and Gadner, H. (1997). Among genes involved in the RB dependent cell cycle regulatory cascade, the p16 tumor suppressor gene is frequently lost in the Ewing family of tumors. Oncogene *15*, 2225–2232.

Krokhin, O.V., Craig, R., Spicer, V., Ens, W., Standing, K.G., Beavis, R.C., and Wilkins, J.A. (2004). An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. Mol. Cell. Proteomics *3*, 908–919.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lanitis, E., Dangaj, D., Irving, M., and Coukos, G. (2017). Mechanisms regulating T-cell infiltration and activity in solid tumors. Ann. Oncol. *28*, xii18–xii32.

Laumont, C.M., Daouda, T., Laverdure, J.-P., Bonneil, É., Caron-Lizotte, O., Hardy, M.-P., Granados, D.P., Durette, C., Lemieux, S., Thibault, P., et al. (2016). Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat. Commun. *7*, 10238.

Laumont, C.M., Vincent, K., Hesnard, L., Audemard, É., Bonneil, É., Laverdure, J.P., Gendron, P., Courcelles, M., Hardy, M.-P., Côté, C., et al. (2018). Noncoding regions are the main source of targetable tumor-specific antigens. Sci. Transl. Med. *10*, eaau5516.

Lee, T.I., and Young, R.A. (2013). Transcriptional regulation and its misregulation in disease. Cell *152*, 1237–1251.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. arXiv:180203426.

Mertens, F., Antonescu, C.R., Hohenberger, P., Ladanyi, M., Modena, P., D'Incalci, M., Casali, P.G., Aglietta, M., and Alvegård, T. (2009). Translocation-related sarcomas. Semin. Oncol. *36*, 312–323.

Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., and Chang, H.Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat. Methods *13*, 919–922.

Ouspenskaia, T., Law, T., Clauser, K.R., Klaeger, S., Sarkizova, S., Aguet, F., Li, B., Christian, E., Knisbacher, B.A., Le, P.M., et al. (2022). Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. Nat. Biotechnol. *40*, 209–217.

Palazzo, A.F., and Koonin, E.V. (2020). Functional long non-coding RNAs evolve from junk transcripts. Cell *183*, 1151–1161.

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., et al. (2019). The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. *47*, D442–D450.

Perry, J.A., Seong, B.K.A., and Stegmaier, K. (2019). Biology and therapy of dominant fusion oncoproteins involving transcription factor and chromatin regulators in sarcomas. Annu. Rev. Cancer Biol. *3*, 299–321.

Pertea, G., and Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. F1000Res *9*, 304.

Poullet, P., Carpentier, S., and Barillot, E. (2007). myProMS, a web server for management and validation of mass spectrometry-based proteomic data. Proteomics *7*, 2553–2556.

R Core Team. (2017). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing). https://www.R-project.org/.

Riggi, N., Knoechel, B., Gillespie, S.M., Rheinbay, E., Boulay, G., Suvà, M.L., Rossetti, N.E., Boonseng, W.E., Oksuz, O., Cook, E.B., et al. (2014). EWS-FLI1 utilizes divergent chromatin remodeling mechanisms to directly activate or repress enhancer elements in Ewing sarcoma. Cancer Cell *26*, 668–681.

Riggi, N., Suvà, M.L., and Stamenkovic, I. (2021). Ewing's sarcoma. N. Engl. J. Med. *384*, 154–164.

Ruiz Cuevas, M.V., Hardy, M.-P., Hollý, J., Bonneil, É., Durette, C., Courcelles, M., Lanoix, J., Côté, C., Staudt, L.M., Lemieux, S., et al. (2021). Most non-canonical proteins uniquely populate the proteome or immunopeptidome. Cell Rep. *34*, 108815.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. *16*, 259.

Shao, M., and Kingsford, C. (2017). Accurate assembly of transcripts through phase-preserving graph decomposition. Nat. Biotechnol. *35*, 1167–1169.

Sheffield, N.C., Pierron, G., Klughammer, J., Datlinger, P., Schönegger, A., Schuster, M., Hadler, J., Surdez, D., Guillemot, D., Lapouble, E., et al. (2017). DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma. Nat. Med. *23*, 386–395.

Sole, A., Grossetête, S., Heintzé, M., Babin, L., Zaïdi, S., Revy, P., Renouf, B., De Cian, A.D., Giovannangeli, C., Pierre-Eugène, C., et al. (2021). Unraveling Ewing sarcoma tumorigenesis originating from patient-derived mesenchymal stem cells. Cancer Res. *81*, 4994–5006.

Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nat. Rev. Mol. Cell Biol. *22*, 96–118.

Surdez, D., Zaidi, S., Grossetête, S., Laud-Duval, K., Ferre, A.S., Mous, L., Vourc'h, T., Tirode, F., Pierron, G., Raynal, V., et al. (2021). STAG2 mutations alter CTCF-anchored loop extrusion, reduce cis-regulatory interactions and EWSR1-FLI1 activity in Ewing sarcoma. Cancer Cell *39*, 810–826.e9.

The, M., MacCoss, M.J., Noble, W.S., and Käll, L. (2016). Fast and accurate protein false discovery rates on large-scale proteomics data sets with Percolator 3.0. J. Am. Soc. Mass Spectrom. *27*, 1719–1727.

Thompson, R.E., Larson, D.R., and Webb, W.W. (2002). Precise nanometer localization analysis for individual fluorescent probes. Biophys. J. *82*, 2775–2783.

Tomazou, E.M., Sheffield, N.C., Schmidl, C., Schuster, M., Schönegger, A., Datlinger, P., Kubicek, S., Bock, C., and Kovar, H. (2015). Epigenome mapping reveals distinct modes of gene regulation and widespread enhancer reprogramming by the oncogenic fusion protein EWS-FLI1. Cell Rep. *10*, 1082–1095.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419.

Valot, B., Langella, O., Nano, E., and Zivy, M. (2011). MassChroQ: a versatile tool for mass spectrometry quantification. Proteomics *11*, 3572–3577.

Van Oss, S.B.V., and Carvunis, A.-R. (2019). De novo gene birth. PLoS Genet. *15*, e1008160.

Waszak, S.M., Robinson, G.W., Gudenas, B.L., Smith, K.S., Forget, A., Kojic, M., Garcia-Lopez, J., Hadley, J., Hamilton, K.V., Indersie, E., et al. (2020). Germline Elongator mutations in Sonic Hedgehog medulloblastoma. Nature *580*, 396–401.

Watson, S., Perrin, V., Guillemot, D., Reynaud, S., Coindre, J.-M., Karanian, M., Guinebretière, J.-M., Freneaux, P., Le Loarer, F.L., Bouvet, M., et al. (2018). Transcriptomic definition of molecular subgroups of small round cell sarcomas. J. Pathol. *245*, 29–40.

Yarmarkovich, M., Marshall, Q.F., Warrington, J.M., Premaratne, R., Farrel, A., Groff, D., Li, W., di Marco, M., Runbeck, E., Truong, H., et al. (2021). Cross-HLA targeting of intracellular oncoproteins with peptide-centric CARs. Nature *599*, 477–484.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-seq (MACS). Genome Biol. *9*, R137.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Frozen human tumor samples	Curie Institute, UGS	N/A
Fresh human tumor samples	Curie Institute and Gustave Roussy	N/A
Chemicals, Peptides, and Recombinant Proteins		
Penicillin/streptomycin	Gibco	Cat#15140122
Doxycycline	Invitrogen	Cat#BP26535
RPMI-1640 Medium	Sigma-Aldrich	Cat#R8758
Minimum Essential Medium Alpha, no nucleosides	Gibco	Cat#12561056
L-glutamine	Gibco	Cat#25030081
IMDM	Gibco	Cat#12440053
Insuline-Transferrin-Selenium	Gibco	Cat#41400045
DMEM/High glucose with 4.0mM L-glutamine, with sodium pyruvate	HyClone	Cat#SH30243.LS
Fetal bovine serum	Eurobio	Cat#S182H-500
CO_2 Independent Medium	Gibco	Cat#18045088
Liberase	Roche	Cat#5401020001
DNase I	Sigma-Aldrich	Cat#AMPD1-1KT
Debris Removal Solution	Miltenyi Biotec	Cat#130-109-398
Opti-MEM	Gibco	Cat#31985062
Urea	Euromedex	Cat#EU0014-B
Ammonium Bicarbonate	Fisher Scientific	Cat#A643-500
Dithiothreitol	Sigma-Aldrich	Cat#D0632
Iodoacetamide	Sigma-Aldrich	Cat#I6125
Trypsin/Lys-C	Promega	Cat#PAV5072
Trifluoroacetic Acid	Thermo Scientific	Cat#85183
Sep-Pak C18 Classic Cartridge	Waters	Cat#WAT051910
Formamide	Ambion	Cat#AM9342
ProLong Gold Antifade Mountant	Thermo Fisher	Cat#P10144
Critical Commercial Assays		
Power SYBR Green PCR Master	Applied Biosystems	Cat#4367659
Mycoplasma detection kit	Minerva Biolabs	Cat#11-9250
NucleoSpin RNA kit	Macherey-Nagel	Cat#740955.50
High-Capacity cDNA Reverse Transcription kit	Applied Biosystems	Cat#4374967
AmpliTaqGold DNA Polymerase kit with Gold Buffer and MgCl2	Applied Biosystems	Cat#4311806
TruSeq Stranded mRNA Library preparation kit	Illumina	Cat#20020594
SMARTer PCR cDNA Synthesis Kit	Clontech	Cat#634925
PrimeSTAR GXL DNA Polymerase	Clontech	Cat#R050A
SMRTbell Template Prep Kit	Pacific Biosciences	Cat#100-259-100
Chromium Next GEM Single Cell 3' GEM, Library & Gel Bead Kit v3.1	10x Genomics	Cat#PN-1000121
Ribosome profiling (Ribo-seq)	Ribomaps Ltd	N/A
RNeasy Plus Mini Kit	Qiagen	Cat#74134
Pierce BCA Protein Assay Kit	Thermo Scientific	Cat#23225
Pierce High pH Reversed-Phase peptide fractionation kit	Thermo Scientific	Cat#84868

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Ewing sarcoma cell lines RNA-seq data	Surdez et al., 2021	GEO: GSE133228; GEO: GSE164373
MSCs +/- EWS::FLI1 (EWIma) RNA-seq data	Sole et al., 2021	GEO: GSE150783
DSRCT cell lines RNA-seq data	Gedminas et al., 2020	GEO: GSE137561
TCGA RNA-seq data	N/A	https://gdc.cancer.gov/
Ribo-seq data for K562 and HepG2 cell lines	Calviello et al., 2020	GEO: GSE129061
GTEx RNA-seq data	N/A	https://gtexportal.org/home/protectedDataAccess
Human Protein Atlas RNA-seq data	Uhlén et al., 2015	http://www.proteinatlas.org
Ewing sarcoma cell lines HiChIP data	Surdez et al., 2021	GEO: GSE133227
Ewing sarcoma cell lines mass spectrometry data	This paper	PRIDE: PXD027309
Ewing sarcoma cell line ChIP-seq data	Aynaud et al., 2020	GEO: GSE129155
DSRCT cell line ChIP-seq	Hingorani et al., 2020	GEO: GSE156277
Clinical RNA-seq data	Watson et al., 2018	EGA: EGAS00001002189
Experimental Models: Cell Lines		
A673	ATCC	ATCC CRL-1598
A673/TR/shEF	Carrillo et al., 2007	N/A
TC-71	DSMZ	ACC 516
EW7	IARC	N/A
STA-ET-1	Kovar et al., 1997	N/A
POE	Curie Institute	N/A
SK-ES-1	ATCC	N/A
MHH-ES1	DSMZ	N/A
EW3	IARC	N/A
CHLA-10	COG Repository	N/A
CHLA-258	COG Repository	N/A
MSCs +/- EWS::FLI1 (EWIma)	Sole et al., 2021	N/A
Oligonucleotides		
Primers for reverse transcription, see Table S4A	This paper	N/A
Guide RNAs for CRISPR interference, see Table S4B	This paper	N/A
Probes for single-molecule RNA FISH, see Table S4C	This paper	N/A
Software and Algorithms		
R (v.3.5.1)	N/A	https://cran.r-project.org
Scallop (v.0.10.4)	Shao and Kingsford, 2017	https://github.com/Kingsford-Group/scallop
Gffcompare	Pertea and Pertea, 2020	https://github.com/gpertea/gffcompare
Bowtie2 (v2.2.9)	Langmead and Salzberg, 2012	https://github.com/BenLangmead/bowtie2
MACS2 (v.2.1.1)	Zhang et al., 2008	https://github.com/macs3-project/MACS
HiC-Pro (v.2.10.1)	Servant et al., 2015	https://github.com/nservant/HiC-Pro
STAR (v.2.5.0a and v.2.7.0e)	Dobin et al., 2013	https://github.com/alexdobin/STAR
CellRanger (v.3.0.2)	10x Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation
Seurat (v.3.1.4)	N/A	https://satijalab.org/seurat
WebLogo (v.3.6)	Crooks et al., 2004	https://github.com/WebLogo/weblogo
phastCons	N/A	http://compgen.cshl.edu/phast/
Trim Galore! (v.0.6.5)	N/A	https://github.com/FelixKrueger/TrimGalore
RiboseQC (v0.99.0)	Calviello et al., 2019	https://github.com/lcalviell/Ribo-seQC

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ORFquant (v1.02.0)	Calviello et al., 2020	https://github.com/lcalviell/ORFquant
Samtools (v.1.9)	Li et al., 2009	http://samtools.sourceforge.net/
ORFfinder (v.0.4.3)	N/A	https://www.ncbi.nlm.nih.gov/orffinder/
Proteome Discoverer (v.2.4)	Thermo Scientific	https://www.thermofisher.com/fr/fr/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/multi-omics-data-analysis/proteome-discoverer-software.html
myProMS (v3.9.3)	Poullet et al., 2007	https://github.com/bioinfo-pf-curie/myproms
Percolator	The et al., 2016	https://github.com/percolator/percolator
MassChroQ (v.2.2.21)	Valot et al., 2011	http://pappso.inrae.fr/bioinfo/masschroq/
pbsmrtpipe	Pacific Biosciences	https://github.com/PacificBiosciences/pbsmrtpipe
Skyline (v.21.1.0.278)	MacCoss Lab	https://skyline.ms/project/home/software/Skyline/begin.view
Stellaris Probe designer software	LGC Biosearch Technologies	https://www.biosearchtech.com/support/tools/design-software/stellaris-probe-designer
MicroManager software	Edelstein et al., 2010	https://micro-manager.org/
MatchAnnot	N/A	https://github.com/TomSkelly/MatchAnnot
Integrative Genomics Viewer (IGV)	N/A	https://software.broadinstitute.org/software/igv

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Olivier Delattre (olivier.delattre@curie.fr).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data from TCGA, GTEx, HPA are available through their respective dedicated platforms. ChIP-seq data for the A673/TR/shEF cell line are available in the Gene Expression Omnibus (GEO) repository under GSE129155 (Aynaud et al., 2020). RNA-seq data for EwS cell lines are available in GEO under GSE133228 (A673, TC71, EW1, EW7, CHLA10, SK-N-MC, POE) (reviewer token wpqvyqoyndmbbux) and GSE164373 (A673/TR/shEF, MHH-ES1, EW24, TC71) (reviewer token wtkxmocuvdgjdmt). RNA-seq data for MSCs are available in GEO under GSE150783 (reviewer token yxgfqkgkxpqfrkn) (Sole et al., 2021). HiChIP data are available in GEO under GSE133227 (Surdez et al., 2021). ChIP-seq data for the JN-DSRCT-1 cell line are available in GEO under GSE156277 (Hingorani et al., 2020). RNA-seq data for DSRCT cell lines are available in GEO under GSE137561 (Gedminas et al., 2020). Ribo-seq data for K562 and HepG2 cell lines are available in GEO under GSE129061 (Calviello et al., 2020). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository (Perez-Riverol et al., 2019) with the dataset identifier PXD027309 (Username: reviewer_pxd027309@ebi.ac.uk, Password: gLTwS23i). Clinical sequencing data is available in the European Genome-phenome Archive (EGA) dataset EGAS00001002189 through controlled access by a Data Access Committee (DAC).
- This paper does not report original code. All custom code used for the analyses was written with existing software as detailed in the STAR Methods section and is available upon request.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

All cell lines were routinely checked by PCR for the absence of Mycoplasma. A673 and A673/TR/shEF (also called ASP14) cell lines (Carrillo et al., 2007) were cultured at 37°C, in 5% CO_2 with Dulbecco's Modified Eagle Medium (DMEM) with High Glucose, 4mM of

L-Glutamine, 4,500mg/L Glucose and sodium pyruvate (HyClone) supplemented with 10% FBS (Eurobio) and 1% antibiotics (v/v) (penicillin and streptomycin (Gibco)). Induction of EWS::FLI1 specific shRNA was performed by adding 1 μg/mL of doxycycline in the medium ex-tempo. After seven days of treatment, doxycycline was removed and cells were washed three times to stop the shRNA induction, thus enabling re-expression of EWS::FLI1. TC71, EW7, STA-ET-1 (Kovar et al., 1997), POE, SKES1, ES1 cell lines were cultured in Roswell Park Memorial Institute (RPMI) 1640 medium (Sigma), 10% FBS (Eurobio), 1% antibiotics (v/v) (penicillin and streptomycin (Gibco)). EW3 cell line was cultured in Minimum Essential Medium Alpha (MEM α), no nucleosides (Gibco), 10% FBS (Eurobio), 1% L-glutamine (Gibco). CHLA-10 and CHLA-258 cell lines were cultured in Iscove's Modified Dulbecco's Medium (IMDM) (Gibco), 20% FBS (Eurobio), 4 mM Glutamine (Gibco), 1X ITS (5 μg/mL insulin, 5 μg/mL transferrin, 5 ng/mL selenous acid) (Gibco).

Patient samples

All patient samples used in this study were stored in a tumor bank at the Institut Curie. The study was approved by the Institutional Review Board of the Institut Curie. Written informed consent was obtained.

METHOD DETAILS

RNA extraction and reverse transcription

RNA from MSCs expressing or not EWS::FLI1 was obtained from E. Brunet's lab (Sole et al., 2021). For A673/TR/shEF, total RNA was isolated using the Nucleospin II kit (Macherey-Nagel) and reversely transcribed using the High-Capacity cDNA Reverse Transcription kit (Applied Biosystems). Next, cDNA molecules were amplified by PCR performed using the AmpliTaqGold DNA Polymerase kit with Gold Buffer and MgCl2 (Applied Biosystems). One microgram of template total RNA was used for each reaction. Next, cDNA molecules were amplified by qPCR performed using SYBR Green (Applied Biosystems). Reactions were run on a 7500 QPCR instrument and analyzed using the 7500 system SDS software (Applied Biosystems). Relative quantification of neotranscripts was normalized to an endogenous control (*RPLP0*) and was performed using the comparative Ct method. Error bars indicate standard error of the mean (n=3). Oligonucleotides were purchased from MWG Eurofins Genomics (Table S4A).

RNA sequencing of cell lines (Illumina & PacBio)

Illumina: Every RNA sample was evaluated for integrity using the BioAnalyzer instrument (Agilent). All samples displayed excellent quality (RNA Integrity Number above 9). Libraries were performed using the TruSeq Stranded mRNA Library Preparation Kit. Equimolar pools of libraries were sequenced on an Illumina HiSeq 2500 machine using paired-end reads (PE, 2x101bp) and High Output run mode allowing to get around 200 millions of raw reads per sample. Raw reads were mapped on the human reference genome hg19 using the STAR aligner (v.2.5.0a) (Dobin et al., 2013). PCR-duplicated reads and low mapping quality reads (MQ<20) were removed using Picard tools and SAMtools (Li et al., 2009), respectively.

Pacbio: Libraries were prepared following the protocol from Pacific Biosciences: "Procedure & Checklist - Iso-Seq™ Template Preparation for Sequel® Systems - Version 5 (November 2017)". 1 μg of total RNA was used as input. The cDNA was synthesized with the SMRTer PCR cDNA Synthesis Kit from Clontech, following manufacturer's recommendations. The cDNA was then amplified with the PrimeSTAR GXL DNA Polymerase from Clontech with 12 cycles of PCR. This number was set up after PCR optimization, in order to obtain enough yield, and to avoid PCR bias. The amplified cDNA was then split into only 2 fractions to perform 2 different purifications using AMPure beads (ratio of 0.4x and 1x). No third fraction was isolated to perform a size-selection step using Blue Pippin system. Then an equimolar pool was made from the 2 fractions. The SMRTbell was prepared from 2.8μg of this equimolar pool of cDNA using the SMRTbell Template Prep Kit from Pacific Biosciences, following manufacturer's recommendations. The sequencing was performed on a Sequel system, using V2.1 chemistry and Magbead loading. 4 SMRTcells were used for the ASP14 sample and 3 SMRTcells for the ASP14+DOX sample. The sequencing runs were set up with a pre-extension step of 240 minutes and 10 hours of movie. We used the implemented pbsmrtpipe pipeline to perform read processing. To annotate IsoSeq reads, we used the MatchAnnot script (https://github.com/TomSkelly/MatchAnnot). MatchAnnot assigns each read to annotated transcripts using score base-pair matching. Reads that had no match on the GENCODE v19 reference were annotated as NA. We manually curated the list of NA reads (n=145) using Integrative Genomics Viewer (IGV, https://software.broadinstitute.org/software/igv/). We also used Illumina RNA-seq data, ChIP-seq data (EWS::FLI1, H3K27me3, H3K27ac, H3K4me3) and GGAA repeats track at the same time in order to identify EWS::FLI1-regulated reads from intergenic regions. After applying these filters, we found 4 clusters corresponding to four distinct expressed-intergenic regions.

Quantification of neotranscripts in RNA-seq across cancers and tissues

FASTQ files from TCGA, GTEx and HPA (http://www.proteinatlas.org) datasets were downloaded and aligned to the hg19 genome assembly using STAR (v2.7.0e) (Dobin et al., 2013). The GTF file used for alignment and quantification of gene expression was based on evidence-based annotation of the human genome (GRCh37), version 19 (Ensembl 74) provided by GENCODE, to which was added the annotation of the neotranscripts in GTF format. Gene expression was quantified using the GeneCounts procedure from STAR. Raw counts were then normalized to Transcripts Per Million (TPM).

Tumor short-read RNA-seq for discovery of neogenes

We used paired-end RNA-seq from our institutional database of fresh-frozen patient tumor tissue to search for tumor-specific neogenes. RNA sequencing was performed using established protocols on Illumina instruments as previously described (Watson et al., 2018). All diagnoses were made by pathological examination, confirmed by fusion gene detection in the case of OCTF-driven cancers and independently reviewed by an expert clinician.

RNA-seq alignment, transcript assembly and detection of unannotated transcripts

We used Scallop (Shao and Kingsford, 2017), a reference-based transcript assembler, to predict all transcript sequences based on aligned RNA-seq reads, independently of a reference transcriptome annotation. First, paired-end FASTQ files were aligned to the hg19 human reference genome using STAR (v2.7.0e) (Dobin et al., 2013). We then ran Scallop (v0.10.4) on the resulting BAM file with default parameters to assemble all expressed transcript sequences. To conserve only unannotated transcripts, we used Gffcompare (Pertea and Pertea, 2020) to compare the Scallop output GTF file with the reference GENCODE v19 GTF file, and conserved only transcripts labeled by Gffcompare as « u » (unknown, intergenic), « y » (contains a reference within its introns) and « x » (exonic overlap on the opposite strand). Finally, to remove lowly expressed transcripts and decrease the rate of false positives, we removed all transcripts with *coverage* less than 10 as output by Scallop.

Selection of tumor-specific unannotated transcripts

To discover tumor-specific neogenes and discard all other transcripts assembled by Scallop, we used three steps of selection.

1. First, we ran Scallop as described previously on one RNA-seq sample of the cancer of interest to generate a first set of candidate unannotated transcripts (*Candidate set 1*).
2. Then in order to quickly discard non tumor-specific transcripts from this first candidate set, we applied a first filtering process based on high and tumor-specific expression as compared to a limited set of other tumors: for this we quantified the expression of *Candidate set 1* neogenes on 3 samples of at least 8 different tumor types (3 samples for each tumor type) by re-aligning each sample with STAR and quantifying expression using the GeneCounts procedure with the GENCODE v19 reference GTF file to which we added the *Candidate set 1* neotranscripts. Raw counts were converted to transcripts per million (TPM) before the filtering process. To retain only tumor-specific and highly expressed candidates, we selected transcripts with:
   (i) mean expression in the disease of interest of more than 10 TPM,
   (ii) log-fold change of mean expression in samples of other diagnoses versus mean expression in disease of interest of less than -2,
   (iii) mean expression in samples of other diagnoses of less than 3 TPM,
   (iv) maximum expression in samples of other diagnoses of less than 15 TPM, resulting in a second set of candidate neogenes (*Candidate set 2*). For EwS, we also ran Scallop as described previously on one RNA-seq sample from a cell line (A673/TR/shEF) and applied the same above filters to the resulting *Candidate set 1* of neogenes. The resulting *Candidate set 2* neogenes not overlapping *Candidate set 2* neogenes from the EwS tumor sample were included in the final *Candidate set 2* neogenes used for the subsequent round of filtering.
3. Finally, a second filtering step was applied based on expression levels across a wide range of cancers and normal tissues: for this we quantified expression of *Candidate set 2* neogenes in all tumors from our institutional database, all cancer types from TCGA (either all the samples from one type or 50 samples if number of samples exceeded 50), all normal tissue samples from TCGA, all normal tissues from GTEx (either all the samples from one type or 50 samples if number of samples exceeded 50) and all normal tissue samples from the Human Protein Atlas. Every sample was re-aligned with STAR and expression quantified by the GeneCounts procedure with the use of a GTF file including GENCODE v19 and *Candidate set 2* neotranscripts. Raw counts were converted to TPM before filtering. To retain tumor-specific candidates with a relatively high expression level (accounting for potentially lower tumor content in some samples we diminished the first threshold (i) as compared to the first filter) and quasi-null expression in other cancers and normal tissues, we selected transcripts with:
   (i) mean expression in the disease of interest of more than 7.5 TPM,
   (ii) log-fold change of mean expression in other samples versus mean expression in disease of interest of less than -3,
   (iii) mean expression in other samples of less than 2 TPM,
   (iv) 99 % quantile of expression in other samples of less than 10 TPM,
   (v) maximum mean expression in another cancer or tissue of less than 10 TPM (excluding testis and placenta), resulting in a final set of tumor-specific neogenes.

We noted during this procedure that some neogenes could be moderately expressed (most less than 10 TPM) in germinal tissues (testis and placenta), reflecting known higher transcriptomic diversity and exclusivity there (e.g., for cancer-testis antigens), and therefore allowed the few genes (less than 1.5% of neogenes in this study) expressed in these tissues at more than 10 TPM to pass the filter (v) nonetheless.

During the procedure we also found that a large part of candidate neogenes were mutually and exclusively expressed in the following pairs of diagnoses: angiomatoid fibrous histiocytoma (AFH) and clear cell sarcoma (CCS), alveolar soft part sarcoma

(ASPS) and TFE3-translocated renal cell carcinoma (TFE3), alveolar rhabdomyosarcoma (aRMS) and embryonal rhabdomyosarcoma (eRMS). We did not remove those neogenes to account for neogenes driven by the same fusion gene in two different diseases (AFH/CCS, ASPS/TFE3) and disease-associated neogenes common to both types of rhabdomyosarcoma (aRMS/eRMS).

Analysis of OCTF binding and histone marks around neogenes

We analyzed ChIP-seq data to explore the epigenetic landscape of neogenes in two fusion-driven sarcomas (EwS and DSRCT). For EwS, we used ChIP-seq data from our laboratory for EWS::FLI1, H3K27ac and H3K4me3 in the A673/TR/shEF cell line (GEO accession GSE129155), generated as previously described (Aynaud et al., 2020). For DSRCT, ChIP-seq from public data was used: EWS::WT1 and PolII (GSE156277) for the JN-DSRCT-1 cell line (Hingorani et al., 2020). For public data, FASTQ files were downloaded from SRA and aligned with Bowtie2 (v2.2.9) (Langmead and Salzberg, 2012), duplicates and multi-mapped reads were removed with SAMtools (Li et al., 2009). Peaks were identified using MACS2 (Zhang et al., 2008) with default parameters and q-value < 0.05. For EWS::FLI1, only peaks associated to GGAA repeats were kept (i.e., peaks with > 4 GGAA repeats within less than 1 kb).

To document enrichment of OCTF binding sites near the TSS of neogenes, we used Wilcoxon's two-tailed test to compare the distributions of distances of TSS to nearest ChIP-seq peak (GGAA-microsatellite EWS::FLI1 for EwS, EWS::WT1 for DSRCT) between neotranscripts and all transcripts in the GENCODE reference transcriptome. NB: One gene can have multiple and sometimes identical TSSs corresponding to alternative transcripts, but we only counted unique TSSs for the analysis.

Analysis of EWS::FLI1 enhancer-promoter chains for EwS neogenes

H3K27ac HiChIP were processed with HiC-Pro (v2.10.1) (Servant et al., 2015) using at least two replicates for each experiment with a bin resolution of 5kb and all analyses were performed using valid pairs. Chains started from H3K4me3 peaks overlapping TSS (GENCODE v19 annotation of the hg19 mapping assembly) of expressed neogenes in A673 or TC71 EwS cell lines. Each 5kb promoter bin (BIN-P) overlapped with at least one of these promoter regions (i.e., several BIN-P overlapping H3K4me3 peaks were allowed). Starting from a single BIN-P, the first enhancer (BIN-E1) element of the promoter-enhancer chain was identified as a bin displaying overlap with H3K27ac peaks (in respectively A673 and TC71 ChIP-seq data) and displaying the strongest interaction (greater than 4 reads) located at least 20,000 (4-bin gap) away from BIN-P in H3K27ac HiChIP matrix of respectively A673 or TC71 data. A recursive algorithm following these rules allowed to construct promoter enhancer chains up to the 20th enhancer BIN (chains: BIN-P linked to BIN-E1 up to BIN-E20). Promoter-enhancer chains were assigned to respectively 7 and 10 neogenes in A673 and TC71, representing respectively 26.9% and 38.5% of the population of expressed neogenes in A673 and TC71 (Table S1). In Figure S5, bin size and chain thickness are proportional to respective numbers of reads.

Analysis of expression of neogenes in cell lines

For EwS and DSRCT, we quantified expression of neogenes in cell lines having normal or downregulated expression of the fusion transcript (respectively EWS::FLI1 and EWS::WT1). RNA-seq reads were aligned with STAR (v.2.7.0e) (Dobin et al., 2013) using a GTF file containing GENCODE v19 and the corresponding neogenes, quantification was done using GeneCounts. For EwS, we used RNA-seq data from our laboratory for nine cell lines having either normal expression of EWS::FLI1 at day 0 or downregulated expression by a doxycycline-inducible system or siRNA after 7 days (GSE133228 and GSE164373). We also used RNA-seq data for MSCs, some of which were induced to express EWS::FLI1 (GSE150783). For DSRCT, we used RNA-seq from public data (GSE137561) for BER and JN-DSRCT-1 cell lines treated with siRNA against EWS::WT1 for 48 hours or control siRNA (Gedminas et al., 2020). All FASTQ files from public data were downloaded from SRA.

Classification of neogenes

We classified neogenes in EwS and DSRCT according to the two following criteria: i) dependence on the OCTF, as shown by regulation of expression in cell lines with activity of the OCTF (mean $\log_2$FC > 1.5), and ii) evidence of physical binding of the OCTF near their TSS (less than 10 kb) or within an enhancer-promoter chain as defined by HiChIP.

Neogenes were thus classified as OCTF-driven (both criteria present) or not. See Table S1.

Generation of genomic-browser style figures for neogene loci

Genomic browser-style figures showing neotranscript sequences, RNA-seq read alignments, and ChIP-seq data were generated with custom scripts written in R (R Core Team, 2017). RNA-seq reads in FASTQ format were aligned to the hg19 human reference genome with STAR (Dobin et al., 2013) using a GENCODE v19 reference GTF annotation and visualized in BAM format. ChIP-seq BAM files used for visualization were generated as described previously. For EwS, GGAA repeats (EWS::FLI1 canonical binding sites) were also displayed in the same figure. For pile-up tracks, RNA-seq BAM files were merged with the SAMtools *merge* function (Li et al., 2009) for ten EwS and ten DSRCT tumors.

Single-cell RNA-seq of human fresh tumors

Two male patients of 20 and 24 years old, respectively diagnosed with EwS and DSRCT, underwent tumor surgery as part of their clinical care and provided written consent to analysis of their tumor samples for this research, as specified by ethical regulation of our institutions.

Tissues were processed within 1 hour after tumor resection, and sorted cells were loaded in a 10x Chromium instrument within 3 hours. Fresh tumor samples were cut in small pieces then dissociated 30 min at 37°C in CO2-independent medium (Gibco) containing 150 μg/ml of Liberase TL (Roche) and DNase 1 (Sigma Aldrich). Dissociated cells were then filtered with a 30 μm cell strainer and washed in PBS. Debris were removed by centrifugation (3,000 xg for 10 min at 4°C), using the debris removal solution (Miltenyi Biotech). Cells were then resuspended in PBS + 2 mM EDTA, counted and adjusted at 10^6 cells/ml in PBS-2mM EDTA. 6,000 cells were loaded on a 10X Chromium (10X Genomics) and libraries were prepared using a Single Cell 3′ Reagent Kit (NextGem kit, 10X Genomics), according to the manufacturer's protocol, targeting 3,000 recovered cells per cell type and sample. Libraries were sequenced on an Illumina NovaSeq 6000 sequencing platform.

Single-cell RNA-seq analysis

Single cell RNA-seq raw base call (BCL) files were demultiplexed and converted into FASTQ files by using the 10X Genomics Cell Ranger pipeline (v3.0.2) *mkfastq* command. FASTQ files were then processed with the Cell Ranger *count* command to perform quality control, barcode processing, and single-cell gene counting. Sequencing reads were aligned to the GRCh38 human reference genome (v3.0.0 Cell Ranger index). Downstream analysis was conducted using Seurat (https://satijalab.org/seurat/) (v3.1.4) in R (v3.5.1). Cells with less than 1000 features, and features expressed in less than 3 cells, were filtered out. Cells with a high proportion of mitochondrial reads (more than 10% and 20% for EwS and DSRCT, respectively) were filtered out based on quality control analyses in Seurat. After merging both tumor sample count matrices with *merge*, normalization was performed using the Seurat function *sctransform* (v0.2.0) (Hafemeister and Satija, 2019). UMAP (McInnes et al., 2020) was performed with *RunUMAP* on 50 principal components after *RunPCA*. Assignment of clusters was performed manually with inspection of marker genes. To quantify expression of neotranscripts, a custom transcriptome was produced by appending sequences of the neotranscripts to the reference transcriptome and running CellRanger *count* with this custom index. Counts for neotranscripts were log-normalized and the average log-normalized expression level was plotted with *FeaturePlot*. Other marker genes were also visualized by *FeaturePlot*.

Analysis of splicing sites of neotranscripts

The analysis used a custom script to count canonical splice sites. The sequence logo was generated with WebLogo 3.6 (Crooks et al., 2004) with splicing sites from all neotranscripts.

Analysis of evolutionary sequence conservation

Evolutionary sequence conservation was evaluated using PhastCons. Scores were calculated for each exon as the mean for all base scores, for neotranscripts, and all lincRNAs and a random subset of 2,000 protein-coding transcripts (without UTRs) in the GENCODE reference transcriptome.

Ribosome profiling

Ribosome profiling (Ribo-seq) on A673 and EW7 cells was performed by Ribomaps Ltd (https://ribomaps.com). Three independent replicas of A673 and EW7 were grown, harvested in ice-cold polysome isolation buffer supplemented with cycloheximide. Following sequencing of Ribo-seq libraries, the per base sequencing quality of each replicate passed the quality threshold.

Ribo-seq data analysis

Ribo-seq data for K562 and HepG2 cell lines in two replicates each (Calviello et al., 2020) were downloaded in FASTQ format from SRA (GEO: GSE129061). Adapters were trimmed using Trim Galore! (v.0.6.5). Ribo-seq reads were mapped with STAR (v2.7.0e) with options –outFilterMultimapNmax 1 –outFilterMismatchNmax 2 to conserve only uniquely mapping reads with a maximum of two mismatches, using a GTF file containing GENCODE v19 reference transcripts to which was added the annotation of the *Ew_NG*s. Ribo-seq quality control analyses were performed with RiboseQC (v0.99.0) (Calviello et al., 2019) using default parameters, after which P-site positions and number of reads mapping to *Ew_NG*s (raw and TPM) were extracted. Read length distribution for Ribo-seq datasets fell within the expected range of 25–35 nt, with a peak between 28 and 32 nt showing strong periodic signals and an enrichment in annotated CDSs: QC measures are displayed in Figure S6. ORF predictions were then performed with ORFquant (v1.02.0) (Calviello et al., 2020) using default parameters on RiboseQC output data.

CRISPR interference (CRISPRi)

Transfection

The A673 cell line was previously transfected using the plasmid Lenti-dCas-KRAB-blast (Addgene) with a MOI (Multiplicity Of Infection) of 3, and cultured at 37°C, in 5% CO_2 with Roswell Park Memorial Institute (RPMI) 1640 medium (Sigma) supplemented with 10% of Foetal Bovine Serum (Eurobio) and 1% Penicillin/Streptomycin. 24h after plating (50,000 cells per well in a 6-well plate), cells were incubated in a mix of OptiMEM (Gibco) and Roswell Park Memorial Institute (RPMI) 1640 medium (Sigma) supplemented with 10% of Foetal Bovine Serum (Eurobio) and transfected with guide RNAs and trackRNA (IDT DNA) each at a final concentration of 10 nM for four days. The genomic coordinates and sequences of guide RNAs are in Table S4B.

RNA extraction and qPCR

After four days, the culture media was removed and RNA was extracted using the RNeasy Plus Mini Kit (Qiagen) and the Qiacube device. RNA quantification was performed with the Nanodrop device (ThermoFisher). RT-PCR was then performed using the Applied Biosystems High-Capacity cDNA Reverse Transcription kit with RNAse Inhibitor (Fisher Scientific) and run was completed with the Applied Biosystems GeneAmp PCR System 2700. After assessment of the DNA concentration, qPCR was performed using the SYBR Green PCR Master Mix and the CFX384 Touch Real-Time PCR System. Results were analyzed with the CFX Manager Software.

Additional primers for this experiment are in Table S4A.

Proteomics and mass spectrometry analysis

Sample Preparation

Proteome cell samples were lysed in a buffer containing 8 M urea (Euromedex), 200 mM ammonium bicarbonate (ABC, FisherScientific) for 30 minutes at room temperature. Lysates were sonicated to decrease viscosity and centrifuged at 20,000 x g for 10 minutes. The protein concentration was measured using the BCA assay (Pierce). 60 μg of total protein were reduced by 5 mM dithiothreitol (DTT, Sigma) for 30 minutes at 55°C, alkylated with 10 mM iodoacetamide (IAM, Sigma) for 30 minutes in the dark. Samples were then diluted 10-fold with 200 mM ABC to obtain a final concentration of urea of 1 M before overnight digestion with Trypsin-LysC (Promega) at a 1:50 ratio at 37°C. Digested samples were acidified with 1% trifluoroacetic acid (TFA, Thermo) for 15 minutes on ice and centrifuged at 2,000 x g for 15 minutes. Peptides were purified using 50 mg Sep-Pak C18 cartridge (Waters) and dried with a SpeedVac apparatus.

Deep HpH-proteome samples were obtained by mixing 12 μg purified peptides of each five cell replicates. Peptide fractionation was carried out with the Pierce High pH Reversed-Phase peptide fractionation kit (Cat number 84868). 75μg of peptides were eluted successively using six elution buffers containing the following percentages of acetonitrile: 10, 12.5, 15, 17.5, 20 and 50%. Eluted peptides were then vacuum-concentrated to dryness and reconstituted in 0.3% TFA to a concentration of 1μg/μL prior to liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis.

LC-MS/MS Analysis

Online chromatography was performed with an RSLCnano system (Ultimate 3000, Thermo Scientific) coupled online to an Orbitrap Exploris 480 mass spectrometer (Thermo Scientific). Peptides were trapped on a C18 column (75 μm inner diameter × 2 cm; nanoViper Acclaim PepMapTM 100, Thermo Scientific) with buffer A (2/98 MeCN/H2O in 0.1% formic acid) at a flow rate of 3.0 μL/min over 4 minutes. Separation was performed on a 50 cm x 75 μm C18 column (nanoViper Acclaim PepMapTM RSLC, 2 μm, 100Å, Thermo Scientific) regulated to a temperature of 40°C with a linear gradient of 3% to 32% buffer B (100% MeCN in 0.1% formic acid) at a flow rate of 300 nL/min over 211 minutes. MS full scans were performed in the ultrahigh-field Orbitrap mass analyzer in ranges m/z 375–1500 with a resolution of 120 000 at m/z 200. For every full scan, the top 20 most intense ions were isolated and subjected to further fragmentation via high-energy collision dissociation (HCD) activation and a resolution of 15 000 with the AGC target set to 100%. We selected ions with charge state from 2+ to 6+ for screening. Normalized collision energy was set at 30 and the dynamic exclusion at 40s.

Data processing

ORFs of *Ew_NG*s were predicted computationally with ORFfinder (v.0.4.3; options: minimal length=75 nucleotides; start codon: any) to constitute a database of potential neopeptides (Table S5). For identification, the data were searched against the Homo Sapiens (UP000005640_9606) UniProt database, this neopeptide database and a database of the common contaminants using Sequest HT through Proteome Discoverer (version 2.4). Enzyme specificity was set to trypsin and a maximum of two miss cleavages sites were allowed. Oxidized methionine, Met-loss, Met-loss-Acetyl and N-terminal acetylation were set as variable modifications. Carbamidomethylation of cysteins were set as fixed modification. Maximum allowed mass deviation was set to 10 ppm for monoisotopic precursor ions and 0.02 Da for MS/MS peaks. The resulting files were further processed using myProMS v3.9.3 (Poullet et al., 2007) (https://github.com/bioinfo-pf-curie/myproms). FDR calculation used Percolator (The et al., 2016) and was set to 1% at the peptide level for the whole study. The label-free quantification was performed by peptide Extracted Ion Chromatograms (XICs) computed with MassChroQ version 2.2.21(Valot et al., 2011). For protein quantification, XICs from all proteotypic peptides, and proteotypic only, were used and missed cleavages were allowed. Median correction and variance scale normalization was applied on the total signal to correct XICs for each biological replicate. Label-free quantification (LFQ) was performed following the algorithm as described (Cox et al., 2014), with the minimum number of peptide ratios set to 1 and the large ratios stabilization feature, and the LFQ values were also normalized to correct for remaining total intensity biases. The final LFQ intensities were used as protein abundance.

Medulloblastoma dataset

We selected this proteome dataset to investigate the potential presence of EwS neopeptides in another type of tumor because protein extraction and quantification were performed by the same Proteomic Mass Spectrometry facility following similar approaches: same acquisition method (Data Dependent Acquisition, DDA) and neopeptide database (Table S5). The dataset corresponds to two replicates of High pH Reversed-Phase peptide fractionation in 24 fractions of a mix of eight primary medulloblastomas (two of each consensus medulloblastoma molecular subgroups), which were previously used to develop the spectral library for Data Independent Acquisition (DIA) quantification (Waszak et al., 2020). The search for neopeptides identification was performed as described in *Data processing*.

Targeted proteomic method for neopeptide validation

LC-MS/MS analysis was performed as previously by adding a parallel reaction monitoring (PRM) mode with an acquisition list generated from the heavy synthetic tryptic NG3 peptide QVGLEPNPK ordered from SB-peptide (Mass [491.2718]; charge 2; extracted fragments y8, y5, y4, y3; purity>95%) based on the DDA results of this heavy synthetic peptide.

PRM data analysis

All raw files were processed using Skyline (version 21.1.0.278) MacCoss Lab Software, Seattle, WA; (https://skyline.ms/project/home/software/Skyline/begin.view) for the generation of the extracted-ion chromatograms and peak integration. To identify peptides in the Skyline platform, we imposed a mass accuracy of withing 10 ppm for fragment ions. The targeted peptides were manually checked to ensure that the transitions for multiple fragment ions exhibited the same elution time in the pre-selected retention time window of the synthetic heavy peptide. The data were then processed so that the distribution of relative intensities of multiple transitions associated with the same precursor ion must be correlated with the theoretical distribution in the MS/MS spectral library entry. The assessment of MS/MS matching was performed by Skyline and Proteome Discoverer.

Single-molecule RNA FISH

To visualize the transcription of NG3, NG13 and NG21 with single-molecule RNA FISH, a single custom probe set for each neogene transcript was designed using Stellaris® Probe designer software (LGC, Biosearch Technologies) (Table S4C). As an input for the design, regions of the transcripts were chosen to be >1kb exonic regions with minimal splicing isoform variations. Probes for the three neogenes with fluorophore Quasar570 and probes for the control gene *FXR1* (exonic probe set from (Coulon et al., 2014)) with fluorophore Quasar670 were ordered from LGC, Biosearch Technologies. Single-molecule RNA FISH was performed by growing cells on 18 mm diameter polylysine-coated (10 ug/mL, Sigma Aldrich) #1.5 glass coverslips and following the Stellaris® RNA FISH protocol for adherent cells with minor modifications. Briefly, the cells were fixed with 4% formaldehyde in 1X PBS, washed twice with PBS and permeabilised using 70% ethanol. The cells were incubated with 2.5 nM of total probe concentration for both probe sets (NG of interest and control gene *FXR1*) in RNA FISH Hybridization Buffer (10% Dextran sulfate, 10% Formamide (Ambion), 2xSSC) overnight at 37°C in parafilm-sealed humified chamber. The coverslips were washed twice for 30 minutes using the wash solution (10% Formamide in 2xSSC) at 37°C the next day. Then, incubated with 5 ng/ml DAPI in 2xSSC for 5 minutes at 37° C, and finally washed twice with 2xSSC. The cells were mounted in Prolong Gold antifade mountant (ThermoFisher) and imaging was performed using a home-built widefield epifluorescence microscope composed of an Eclipse Ti2 stand (Nikon), a Spectra-X light source with a NIR module (Lumencor), a CFI Plan Apochromat Lambda 60x/1.4NA oil immersion objective (Nikon), an Orca Flash4.0 V3 sCMOS camera (Hamamatsu), a H117E1NN stage (Prior), a Nano-ZL100 stage piezo (Mad City Labs). The microscope was controlled by the MicroManager software (Edelstein et al., 2010). The following filters (Semrock) were used for multicolor fluorescence imaging: Quasar670 (FF01-636/8, FF649-Di01, FF01-680/42), Quasar570 (FF01-543/22, FF560-FDi01, FF01-575/15), DAPI (390/22, FF409-Di03-25x36, FF02-447/60-25), compensation channel (FF01-589/15, FF605-Di02, FF01-623/24). For each sample, stacks of 31 z-planes were taken for all channels for 60 fields of view. Pixel size was 0.108 μm.

Analysis of the smRNA FISH data

Raw images were first corrected for 3D chromatic shift. Broad-spectrum autofluorescent structures in the Quasar570 and Quasar670 channels were then attenuated by fluorescence compensation using the 589/15-623/24 channel. An image-wide offset was subtracted to the images of the Quasar570 channel images to compensate for the fast-bleaching homogeneous fluorescence background visible throughout each z-stack. Using custom-written Python scripts, diffraction-limited spots were identified in the compensated Quasar570 and Quasar670 images and 589/15-623/24 images by detecting local 3D maxima after band-pass filtering, followed by an iterative 3D Gaussian mask fit (Thompson et al., 2002), yielding fluorescence intensities and 3D coordinates of the spots. RNA molecules were then distinguished from the background by using a fluorescence intensity threshold and eliminating spots in the Quasar570 and Quasar670 images that co-localize with a spot in the 589/15-623/24 images. Finally, the average number of RNA molecules per cell was calculated for each field of view and each sample, for both the NGs and the control gene *FXR1*. Python scripts are available on https://github.com/CoulonLab/FISHingRod.

QUANTIFICATION AND STATISTICAL ANALYSIS

The tests used for statistical analyses are described in the main text and STAR Methods and have been performed using R v3.5.