19-303.R1 Poulsom Edited USA English ORIGINAL PAPER

Undifferentiated pleomorphic sarcomas with *PRDM10* fusions have a distinct gene expression profile

Jakob Hofvander[1*], Florian Puls[2], Nischalan Pillay[3], Christopher D Steele[4], Adrienne M Flanagan[3,4], Linda Magnusson[1], Jenny Nilsson[1], Fredrik Mertens[1,5]

[1]Division of Clinical Genetics, Department of Laboratory Medicine, Lund University, Lund, Sweden;

[2]Department of Pathology and Clinical Genetics, Sahlgrenska University Hospital, Gothenburg, Sweden;

[3]Department of Histopathology, Royal National Orthopaedic Hospital, Stanmore, UK;

[4]UCL Cancer Institute, London, UK;

[5]Department of Clinical Genetics, Office for Medical Services, Division of Laboratory Medicine, 221 85 Lund, Sweden

***Correspondence to:** Jakob Hofvander, Division of Clinical Genetics, Department of Laboratory Medicine, Lund University, SE-221 84 Lund, Sweden. E-mail: jakob.hofvander@med.lu.se

Running title: *PRDM10*-rearranged sarcomas

Conflicts of interest: The authors declare no potential conflicts of interest.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/path.5326

Location of raw data: RNA-seq and SNP array data are available for academic purposes by contacting the corresponding author, as the patient consent does not cover depositing data that can be used for large-scale determination of germline variants.

Word count: 3996

ABSTRACT

Undifferentiated pleomorphic sarcoma (UPS) is a highly aggressive soft tissue tumor. A subset of UPS is characterized by a *CITED2–PRDM10* or a *MED12–PRDM10* gene fusion. Preliminary data suggest that these so-called *PRDM10*-rearranged tumors (PRT) are clinically more indolent than classical high-grade UPS, and hence important to recognize. Here, we assessed the spectrum of accompanying mutations and the gene expression profile in PRT using genomic arrays and sequencing of the genome (WGS) and transcriptome (RNA-seq). The fusion protein's function was further investigated by conditional expression of the *CITED2–PRDM10* fusion in a fibroblast cell line, followed by RNA-seq and an assay for transposase-accessible chromatin (ATAC-seq). The *CADM3* gene was found to be differentially up-regulated in PRT and cell lines and was also evaluated for expression at the protein level using immunohistochemistry (IHC). The genomic analyses identified few and non-recurrent mutations in addition to the structural variants giving rise to the gene fusions, strongly indicating that the *PRDM10*-fusions represent the critical driver mutations. RNA-seq of tumors showed a distinct gene expression profile, separating PRT from high-grade UPS and other soft tissue tumors. *CADM3* was among the genes that was consistently and highly expressed in both PRT and fibroblasts expressing *CITED2-PRDM10*, suggesting that it is a direct target of the PRDM10 transcription factor. This conclusion is in line with sequencing data from ATAC-seq, showing enrichment of PRDM10 binding sites, suggesting that the amino-terminal fusion partner contributes by making the DNA more accessible to PRDM10 binding.

Keywords: Sarcoma; PRDM10; CITED2; Fusion; Expression; Chromatin

INTRODUCTION

Undifferentiated pleomorphic sarcoma (UPS) of soft tissues is a high-grade tumor lacking any defined line of differentiation [1]. Although morphologically heterogeneous and overlapping with other sarcomas, such as myxofibrosarcoma, all cases share a marked cellular pleomorphism often admixed with spindle cells and bizarre multinucleated giant tumor cells. UPS accounts for up to 20% of sarcomas in adults [1] and is associated with a poor prognosis and a high metastatic rate. Currently, there is no specific treatment available for patients with UPS [2]. UPS is a diagnosis of exclusion and is more likely to represent a common morphological state, possibly of multiple sarcoma subtypes, rather than a distinct tumor entity [3]. In line with this hypothesis, UPS appears to be genetically heterogeneous, although most cases display highly complex genomes with extensive chromosomal rearrangements and multiple copy number alterations [1]. The spectrum of mutations in UPS was recently studied by deep sequencing, confirming an extremely high level of structural variants but a relative paucity of single nucleotide variants [4,5].

We have previously described a subset of UPS with recurrent gene fusions involving the transcription factor-encoding gene *PRDM10*, fused as the 3'-partner to *MED12* or *CITED2*; preliminary data suggested that these tumors had less complex genomic rearrangements than classical UPS cases, and that they might be associated with better outcome [6]. Indeed, in a recent study of the morphological features of *PRDM10*-rearranged tumors (PRT), we could show that they are consistently associated with a low mitotic count and a better prognosis [7].

Here, we have characterized the genetic features of a series of morphologically characterized [7] PRT using high-resolution single nucleotide polymorphism (SNP) arrays,

RNA sequencing (RNA-seq) for global gene expression profiling, and whole genome sequencing (WGS). In addition, the impact of one of the two recurrent fusions, *CITED2–PRDM10*, was investigated in fibroblasts using the Tet-On 3G inducible gene expression system. This allowed us to compare changes in gene expression *in vitro* with those *in vivo*, identifying several down-stream targets of the fusion. Furthermore, we could compare the changes in gene expression with changes in chromatin accessibility by performing an assay for transposase-accessible chromatin (ATAC-seq).

Materials and methods

Patients and tumors

The study included tumors from eight patients with PRT (supplementary material, Table S1). PRT were diagnosed as described [7]. The gene fusion status and the clinicopathological features have been reported [6,7]. The global gene expression profiles of six fresh frozen PRT were compared with those in high-grade UPS (n=17), myxofibrosarcomas (MFS, n=7), myxoinflammatory fibroblastic sarcomas (MIFS, n=7), dermatofibrosarcoma protuberans (DFSP, n=10), and benign fibrous histiocytoma (BFH, n=4). RNA-seq data from four formalin-fixed paraffin-embedded (FFPE) PRT were used to study the expression levels of selected genes and were compared to corresponding data from other tumors of presumed fibroblastic origin: five angiofibromas (AF), six calcifying aponeurotic fibromas (CAF), and 18 sclerosing epithelioid fibrosarcomas (SEF). All tumors used for comparison with PRT were diagnosed according to established criteria [8].

All samples were obtained after informed consent and the study was approved by the institutional review boards of the participating sarcoma centers.

Cell lines

The TERT-immortalized Bj5ta human fibroblast cell line had previously been transduced with the pLVX-Tet3G vector encoding the regulator Tet-On 3G protein [9]. Cells were further transduced with one of four different response plasmids; C-P encoding the CITED2–PRDM10 fusion found in case 2 in Hofvander *et al* (2015), WT encoding wild type PRDM10 (NM_020228.3), ZFO encoding a truncated version of PRDM10 containing only the part of the gene that is included in the fusion, or EV containing an empty vector; the size of *MED12–PRDM10* exceeded the packaging capacity of the lentiviral assay. Transductions were performed using the RetroNectin-bound virus (RBV) infection method according to the manufacturer's instructions (Takara/Clontech, Gothenburg, Sweden).

Response plasmids were selected for by adding puromycin at a final concentration of 0.5 μg/ml. Transcription of the inserted gene constructs was turned on by adding doxycycline (dox) to the culture medium at a final concentration of 1 μg/ml for 48 h; the expression of the *CITED2–PRDM10* construct was assessed by RT-qPCR. Cells were harvested and pelleted in RLT buffer (Qiagen, Valencia, California) with 1% mercaptoethanol and RNA was extracted for RNA-seq using an RNeasy micro kit (Qiagen).

RNA-seq

RNA was extracted from fresh frozen samples and from the TERT-immortalized Bj5ta fibroblast cell lines as described [9]. Libraries of cDNA were prepared from poly-A selected RNA using a TruSeq RNA Sample Preparation Kit v2 (Illumina, San Diego, USA) according to the manufacturer's instructions. Paired-end 151 bp reads were generated from the cDNA libraries using the NextSeq 500 platform (Illumina).

RNA of sufficient quality, i.e., mRNA with DV_{200} values ≥30, was extracted from FFPE samples using Qiagen's RNeasy FFPE Kit (Qiagen); cDNA libraries were prepared from 20 to 400 ng of RNA, depending on the DV_{200} value, using the capturing chemistry of the TruSeq RNA Access Library Prep Kit (Illumina). Paired-end 85 nt reads were generated from the cDNA libraries on a NextSeq 500 (Illumina). Sectioning, RNA extraction, library preparation, and sequencing were performed as described previously [9,10].

In addition to the locally obtained samples, 42 UPS and 17 MFS samples from the TCGA-SARC project were included. Fastq files were downloaded from the GDC Legacy archive (https://portal.gdc.cancer.gov/legacy-archive).

ChimeraScan 0.4.5 and FusionCatcher 1.0 with default settings and STAR-Fusion 1.4.0 with parameters --min_junction_reads 0 --FusionInspector validate, were used to identify candidate fusion transcripts from the sequence data [11–13].

The raw unfiltered RNA-seq reads were aligned to human reference genome hg19 using STAR 2.5.0a [14]. Gene expression values were calculated as fragments per kilobase of transcript per million reads (FPKM) using Cufflinks 2.2.1 [15].

SNP array analysis

SNP array analysis was performed for two cases using the Affymetrix Cytoscan HD array

(Affymetrix, Santa Clara, USA) and from two cases using the Illumina HumanOmni-Quad version 1.0 array (Illumina), as described [16]. From one additional case, DNA was extracted from FFPE blocks and prepared for SNP array analysis using the Oncoscan CNV array (Affymetrix), as described [9]. The position of the SNPs was based on the GRCh37/hg19 sequence assembly.

Whole genome sequencing (WGS)

WGS was performed on two frozen tumor samples, 4a and 5a. DNA was extracted using an automated magnetic bead extraction and purification system according to the manufacturer's protocols (Prepito DNA Tissue10 Kit, Perkin Elmer Ltd, Beaconsfield, UK). DNA from matching blood was obtained using a column-based system (QIAamp DNA Blood Maxi kit; Qiagen). Library preparation and sequencing were performed on the XTen instrument (Illumina) according to the manufacturer's protocol, using 150 bp paired-end libraries with a PCR-free workflow. The average coverage of tumors was at least 70X, and of normal DNA at least 30X.

Single nucleotide variants (SNVs) were called using CavEman [17] and indels using cgpPindel [18]. Only mutations that had median assembly score (ASMD) ≥140 and median clipped bases (CLPM)=0 were considered reliable mutations.

Structural rearrangements were called using BRASS [19] and allele specific copy number and ploidy were called using ASCAT NGS [20].

Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)

ATAC-seq was performed on biological duplicates or triplicates of the Bj5ta cell line with C-P, WT, ZFO or EV constructs after 48 h of dox treatment. Libraries were prepared from 50,000-

75,000 cells (per replicate) according to the Omni-ATAC protocol [21] with minor adjustments. In order to remove primer-dimers and large fragments, the final spin column purification step was replaced with double-sided (0.5X and 1.3X) AMPure XP bead purification. Libraries were sequenced using a NextSeq 500 (Illumina) with paired-end reads of 80 bp.

The first step in the analysis of the ATAC-seq data was to remove remaining adapter sequences from the FASTQ files using Trim-galore (v0.4.1). The trimmed reads were then aligned to the human reference genome hg19 using BWA-MEM (v0.7.10) and duplicate reads were removed using Picard (v2.2.4). Bam files were filtered using SAMTools (v1.3). Reads from mitochondrial DNA and the Y chromosome were removed and only properly paired reads with high mapping quality (-q 30) was used for further analysis. An average of 104 million reads per sample was retained after filtering and the fragment-size distribution for the individual bam files was used for quality control (supplementary material Figure S1). Coverage tracks for visualization were generated using deepTools bamCoverage with parameters -bs=1 --normalizeUsingRPKM. Peak calling was performed with Genrich using the parameter –j (ATAC-seq mode) and Encode Blacklisted regions were excluded. Peak files from all samples were combined and overlapping peaks were merged using bedtools. The number of reads per peak was counted separately for each sample using featureCounts [22] and library size normalization was performed using DEseq2 [23].

De novo motif discovery was performed with HOMER [24], using peaks found across all samples as background sequences. HOMER was also used for annotation of peak regions.

Evaluation of RNA-seq and ATAC-seq data

Correlation-based principal component analysis (PCA) and hierarchical clustering analysis were performed using the Qlucore Omics Explorer version 3.3 (Qlucore AB, Lund, Sweden). FPKM and DEseq2 values were log2 transformed and the data were normalized to a mean of 0 and a variance of 1. Variables were filtered based on variance and the projection score was used to determine the optimal filtering threshold [25]. Hierarchical clustering of both samples and variables was performed using Euclidean distance and average linkage.

Two-sided *t*-tests, corrected for multiple testing by the Benjamini–Hochberg method, were used to identify statistically significant differences in gene expression and chromatin accessibility between groups. Fold-change ≥2 and FDR ≤0.05 were used as cutoffs unless otherwise specified.

Gene ontology (GO) enrichment was performed using metascape [26] with the multiple gene list option.

Immunohistochemistry (IHC)

Cultured Bj5ta cells were scraped off culture tissue flasks and fixed in 10% buffered formalin for 1 h. Cells were spun down and fixed further in PreservCyt® Solution (Hologic, Toronto, Canada) before paraffin embedding using the Cellient Automated Cell Block System (Hologic). CADM3 IHC on cell blocks and formalin-fixed paraffin-embedded (FFPE) tissue from 3 PRT and 14 other soft tissue sarcomas (3 UPS, 5 MIFS, 3 solitary fibrous tumors, and 3 DFSP) was performed using a mouse monoclonal antibody raised against recombinant human SynCAM3 (Pro21-Tyr329, R&D Systems, Minneapolis, MN, USA; clone #730004, dilution 1:60)

following heat induced epitope retrieval using Novocastra Epitope Retrieval Solution pH9 in a Dako PTLink processor (Dako-Agilent, Kista, Sweden). The EnVision Dual Link system (Dako-Agilent) was used for visualization. Surplus surgical tissue from cortical dysplasia including neurons and glial tissue was used as positive control.

Results

Gene fusion status in tumors

Five PRT harbored the *MED12–PRDM10* fusion and three the *CITED2–PRDM10* fusion. At RNA-seq, the fusions between *MED12* and *PRDM10* varied slightly, juxtaposing exons 42 or 43 with exons 13 or 14, respectively, while the fusion point in *CITED2* always was in exon 2 (supplementary material, Table S1). Thus, in all cases, the part of PRDM10 predicted to be retained in the chimeric protein would lack the PR domain but include 5-9 of the 10 zinc finger domains. It could be noted that in spite of a large number of uniquely mapped reads (mean 25.2 million) among the fusion-positive samples, multiple fusion finding algorithms were needed to identify them. **Gene expression profile of *PRDM10*-rearranged tumors**

The gene expression profiles of six PRT were compared with those of 45 other soft tissue tumors, including 17 high-grade UPS and seven MFS. By unsupervised hierarchical clustering of 1737 variables, the PRT formed a distinct expression cluster; furthermore, the three PRT with *CITED2-PRDM10* and the three PRT with *MED12-PRDM10* formed separate sub-clusters (Figure 1A). A total of 404 coding genes were identified as being significantly differentially expressed between PRT and the other tumor types (FDR<0.1; supplementary material, Table S2).

PRDM10 was not among the significantly differentially expressed genes, but it could be

noted that it was expressed at higher levels in tumors with the *CITED2–PRDM10* compared to those with *MED12–PRDM10* (supplementary material, Figure S2). The two most strongly differentially-expressed genes were for the surface receptors transmembrane serine protease 4 (*TMPRSS4*) and cell adhesion molecule 3 (*CADM3*) (Figure 1B,C). *TMPRSS4* and *CADM3* were also significantly over-expressed in the four FFPE PRT samples compared to the 29 control samples (supplementary material, Figure S3A,B).

No fusion transcripts involving *PRDM10* could be identified in the UPS and MFS samples from the TCGA-SARC cohort. The gene expression of these tumors was analyzed together with the local UPS, MFS and PRT samples using unsupervised principal component analysis (PCA). The PRT tumors formed a separate group while UPS and MFS were largely intermixed (supplementary material, Figure S4A). The inability to separate UPS from MFS samples by gene expression analysis is in agreement with previous data [4]. As for the initial cohort, the gene expression levels of *CADM3* and *TMPRSS4* were consistently higher in PRT compared to UPS and MFS samples (supplementary material, Figure S4B,C), highlighting their potential as diagnostic markers. All 3 PRT showed strong expression of CADM3 by IHC, whereas the other 14 soft tissue tumors were negative (Figure 2).

Gene expression profile and ATAC-seq findings in cell lines

Large changes in gene expression were observed between the cells expressing the *CITED2–PRDM10* (C-P) fusion (n=3) and those with an empty vector (EV; n=4); 16% of the coding genes were considered differentially expressed. Among the 242 genes found to be upregulated in PRT, 77 overlapped with the overexpressed genes in the fusion expressing cell line; including

CADM3 and *TMPRSS4*. We confirmed the increased expression of CADM3 using IHC, finding strong expression in Bj5ta cells with the *CITED2–PRDM10* fusion, whereas Bj5ta cells with empty vector were negative (Figure 2).

Expression of the fusion gene also resulted in major changes in chromatin accessibility, as evaluated by ATAC-seq, an assay for evaluating the genomic distribution of open chromatin. ATAC-seq showed that 41% of the peaks were differentially open between C-P and EV. *De novo* motif discovery in the peaks found to be differentially open in C-P identified PRDM10 as the most enriched transcription factor binding motif (supplementary material, Table S3) suggesting that the *CITED2-PRDM10* fusion product still interacts with PRDM10 binding sites and that it plays an important role in the observed changes in chromatin accessibility. To investigate the impact of the fusion protein on biological processes, GO term enrichment of upregulated genes and coding genes harboring an open PRDM10 peak within the gene body or promoter region was performed. The top co-enriched GO terms suggested that the fusion affects several processes involved in development and differentiation (supplementary material, Figure S5).

Among the total peaks (n=145,364), 670 were reported to be differentially open in promoter regions in the fusion expressing cells and 122 of these harbored the PRDM10 motif.

Genes with open promoter regions included *CADM3* (Figure 3) but not *TMPRSS4*, however; here, a peak in intron 7 was found to be differentially open. At closer inspection, the C-P cells did not express a full length *TMPRSS4* transcript but instead a truncated version starting in exon 8, just downstream of the identified peak (Figure 3) and the same truncated version was found to be expressed in PRT (supplementary material, Figure S6).

Because many of the differentially expressed genes appeared to be direct targets of PRDM10, harboring an open peak with the PRDM10 motif, we investigated whether similar alterations in gene expression and chromatin accessibility could be achieved by overexpression of the WT or ZFO constructs. WT and ZFO showed fewer differentially expressed genes than C-P when compared to EV, and approximately half of these genes overlapped with the genes found in C-P versus EV (Figure 4; supplementary material, Table S4). Similar observations were seen in the ATAC-seq data, as WT and ZFO resulted in smaller changes in chromatin accessibility than C-P – 20% and 21% of peaks, respectively – when compared to EV. Among the differentially open promoter regions in WT and ZFO, 42% and 59%, respectively, overlapped with C-P (Figure 3).

The gene expression profiles were also analyzed by unsupervised PCA at which the C-P and EV cells formed distinct clusters while WT and ZFO formed a mixed cluster (Figure 4). Overexpression of the WT or ZFO constructs appeared to have similar effects on the cell lines and though they resulted in changes in the global gene expression, the effects were smaller than in C-P. This was further supported by unsupervised PCA of peaks in promoter regions. Though WT and ZFO formed separate clusters, they were more similar to each other than to EV or C-P and the largest changes in chromatin accessibility of promoter regions was seen between C-P and EV (Figure 3). Thus, both the RNA-seq and ATAC-seq data showed that overexpression of the *CITED2–PRDM10* fusion gene resulted in greater changes in gene expression and chromatin accessibility than WT and ZFO. Though the latter constructs appeared to share a substantial amount of their targets with the fusion gene, the observed effect on these targets was more

severe, regarding both expression and chromatin accessibility, in the fusion-expressing cells (Figures 3 and 4).

Genomic features

From the five cases with SNP-array information, three showed no imbalances, one had a del(11)(q14q22), a del(12)(p11p13), and -14, and one had trisomies 7 and 16 as the only changes.

The number of SNVs was low in the two *PRDM10*-rearranged tumors subjected to WGS: 633 SNVs and 171 indels in sample 4a and 1030 SNVs, and 349 indels in sample 5a. Out of these variants, only 8 were amino acid altering (missense) and two were reported to affect splice sites (supplementary material, Table S5). None of the genes affected was involved in both tumors, nor did they overlap with the commonly mutated genes in high-grade UPS.

There were 16 and 95 structural rearrangements, respectively, in the two cases. Translocations correlating to the expected breakpoints in *MED12*, *CITED2*, and *PRDM10* were identified. Both fusions had multiple breaks in each gene, suggesting that more than two DNA double-strand breaks are involved in the fusion process (supplementary material, Figure S7).

DISCUSSION

UPS is one of the most common sarcoma subtypes and its high genetic complexity has been demonstrated using several different techniques, including WGS [1,4,5]. However, marked heterogeneity in both clinical outcome and genomic complexity provide compelling arguments for more comprehensive attempts to delineate genetic subgroups of UPS. We have previously

identified a subset of UPS showing gene fusions involving the transcription factor PRDM10. These tumors appear to be important to recognize, as the metastasis rate for PRT seems to be lower than for high-grade UPS; ~30% of the latter metastasize [1], compared to none of the PRT.

Here, we have studied a larger series of PRT, showing they share a distinct gene expression profile that is easily distinguishable from those of regular UPS and MFS as well as other morphologically similar tumors including MIFS, BFH and DFSP. We also identified several highly expressed genes, including *TMPRSS4* and *CADM3*, that could potentially be useful as diagnostic markers; strong expression of CADM3 was verified also at the protein level.

PRT and high-grade UPS and MFS differ also at the genomic level. G-banding, SNP array and WGS analyses have identified few structural aberrations and copy number changes in PRT in addition to the structural variants resulting in the gene fusions. Furthermore, the number of SNVs and indels were also substantially lower in the two WGS-analyzed cases in comparison to what has been reported for regular UPS and myxofibrosarcoma [4,5]. The relative lack of secondary aberrations, and the fact that none of them was recurrent, strongly suggests that the PRDM10 fusions are strong driver mutations.

The role of the *CITED2–PRDM10* fusion in tumor development was further evaluated in cell lines where much of the characteristic gene expression profile observed in the tumors could be recapitulated: 77 of the 242 upregulated genes in the tumors were upregulated also in the cell lines expressing the fusion. Expression of the *CITED2–PRDM10* transcript seems to have major impact also on histone regulation as major changes in chromatin accessibility were observed at ATAC-seq. Furthermore, analysis of the ATAC-seq data identified the PRDM10 motif as the

most enriched transcription factor-binding motif in open regions. These results also correlated well with the RNA-seq data as 21 % of the upregulated genes were reported to contain an open peak with the PRDM10 motif; additionally several GO terms were co-enriched between these data sets.

Very little is known about the role of PRDM10, which belongs to the PRDM family of transcription factors. The majority of PRDM proteins, including PRDM10, share a common structure including an N-terminal PR domain, with potential methyltransferase activity, and multiple C2H2-type zinc-finger domains involved in sequence-specific DNA-binding. Based on a few animal models it may be involved in the embryonic development of both mesenchymal tissues and the central nervous system [27,28], and IHC studies have revealed widespread nuclear and cytoplasmic reactivity, with particularly strong positivity in renal tubules, the placenta, and gastrointestinal organs (https://www.proteinatlas.org/ENSG00000170325-PRDM10/).

The fusions reported in this study all result in loss of the PR domain while the majority of the zinc-finger domains are included in the fusion product; this is reminiscent of the fusions involving PRDM16, also resulting in loss of the PR domain, that have been reported in acute myeloid leukemia and myelodysplasia [29]. Interestingly, the full length product of some PRDM family members has been reported to act as a tumor suppressor while shorter isoforms, lacking the PR domain, are oncogenic [30]. We could not, however, detect any major differences between full-length and truncated PRDM10; expression of WT and ZFO constructs in the Bj5ta fibroblast cell line appeared to have very similar effects on gene expression.

While the results of the present study strongly suggest that many of the differentially expressed genes are direct targets of the transcription factor PRDM10, the amino-terminal

partner in the fusion chimera could also be important for its tumorigenic properties. This assumption is based on several observations. First, unlike other sarcoma-associated gene fusions where the main function of the 5'-gene is restricted to providing a stronger and constitutively active promoter, the *MED12–PRDM10* and *CITED2–PRDM10* fusions do not result in increased expression of *PRDM10*; indeed, the fusion transcript was often difficult to detect when using fusion-calling algorithms with default settings. Second, the breakpoints in *MED12* and *CITED2* were highly clustered. Actually, all three tumors with a *CITED2–PRDM10* fusion had the exact same exonic breakpoint in *CITED2* (including at the DNA level), resulting in loss of only its last 3 aa, and the breakpoints in *MED12* were all located within a restricted region, leading to loss of 41–88 of its carboxy-terminal aa. Third, it has become increasingly apparent that many of the amino-terminal partners of transcription factors in sarcoma-associated fusions interact with a variety of protein complexes that influence the accessibility of DNA. For instance, the FUS, EWSR1 and SS18 proteins, serving as amino-terminal partners in myxoid liposarcoma, Ewing sarcoma, and synovial sarcoma, respectively, were recently shown to interact with the BRG1/BRM-associated factor (BAF) chromatin remodeling complex [31–33]. It is thus tempting to speculate that also MED12 and CITED2 contribute by affecting the ability of PRDM10 to bind to DNA and/or by redirecting PRDM10 to novel binding sites. In line with this notion, the majority of open PRDM10 peaks were unique for the *CITED2–PRDM10* expressing cells and the effect on the shared PRDM10 peaks were stronger in C-P than for WT and ZFO (supplementary material, Figure S8).

MED12 is a known component of the kinase module of the so-called Mediator complex, which is involved in the regulation of transcription in several ways, e.g., by directly linking DNA-bound transcription factors to RNA polymerase II, by facilitating the formation of DNA

loops that juxtapose non-adjacent chromosome segments, as well as by changing the chromatin architecture [34]. Of particular interest here, it has been shown in hematopoietic stem and progenitor cells (HSPC) that >80% of genomic regions associated with MED12 were located outside promoters and that these regions, based on the histone acetylation pattern, were within active chromatin corresponding to enhancers [35]. CITED2 is also an important transcriptional co-factor with an impact on chromatin configuration. Through its interaction with the acetyltransferase CBP/p300 it can both increase and repress the expression of the target genes of its associated transcription factors [36]. Intriguingly, both CBP and p300 interact also with MED12. Aranda-Orgilles and coworkers [35] showed that in HSPC, there is an extensive overlap between enhancer sites occupied by MED12 and p300. Hence, the two amino-terminal partners in *PRDM10*-rearranged tumors may well have highly similar functions in the chimeric protein, and thus be exchangeable.

In conclusion, the data provide compelling evidence that PRT is a distinct tumor type, separate from classical UPS. The gene expression profile of PRT, much of which could be reproduced in fibroblasts expressing the *CITED2–PRDM10* fusion, combined with ATAC-seq data suggests that the pathogenetic effects of the chimeric protein can be attributed to both its carboxy-terminal and its amino-terminal parts. Furthermore, the consistent up-regulation of several genes, such as *CADM3*, could be exploited for developing new diagnostic biomarkers.

Acknowledgements

The Swedish Cancer Society, Governmental Funding of Clinical Research within the National Health Service, and the RNOH Research and Development department. The results published

here are in part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.

Author contributions statement

JH and FM designed research. JH, FP, NP, CDS, LM, JN, and FM performed research. FP, NP, and AMF provided clinical and histopathological data. All authors assisted with drafting and revising the manuscript.

References

1 Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F (Eds.). *WHO Classification of Tumours of Soft Tissue and Bone*. 2013. Lyon: IARC Press, 468 pp.

2 Goldblum JR. An approach to pleomorphic sarcomas: can we subclassify, and does it matter? *Mod Pathol* 2014; **27:** 39–46.

3 Fletcher CDM, Gustafson P, Rydholm A, *et al*. Clinicopathologic re-evaluation of 100 malignant fibrous histiocytomas: prognostic relevance of subclassification. *J Clin Oncol* 2001; **19:** 3045-3050.

4 TCGA: the Cancer Genome Atlas Research Network. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell* 2017; **171:** 950–965.

5 Steele CD, Tarabichi M, Oukrif D, *et al*. Undifferentiated sarcomas develop through distinct evolutionary pathways. *Cancer Cell* 2019; **35:** 441–456.

6 Hofvander J, Tayebwa J, Nilsson J, *et al.* Recurrent PRDM10 gene fusions in undifferentiated pleomorphic sarcoma. Clin Cancer Res 2015; **21:** 864–869.

7 Puls F, Pillay N, Fagman H, *et al*. *PRDM10*-rearranged soft tissue tumor: a clinicopathologic study of nine cases. *Am J Surg Pathol* 2019; **43:** 505–513.

8 Fletcher CDM, Chibon F, Mertens F. Undifferentiated/unclassified sarcomas. In: Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F (Eds.). *WHO Classification of Tumours of Soft Tissue and Bone*. 2013. Lyon: IARC Press, pp 236–238.

9 Arbajian E, Puls F, Antonescu CR, *et al*. In-depth genetic analysis of sclerosing epithelioid fibrosarcoma reveals recurrent genomic alterations and potential treatment targets. *Clin Cancer Res* 2017; **23:** 7426–7434.

10 Walther C, Hofvander J, Nilsson J, *et al*. Gene fusion detection in formalin-fixed paraffin-embedded benign fibrous histiocytomas using fluorescence in situ hybridization and RNA sequencing. *Lab Invest* 2015; **95:** 1071–1076.

11 Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* 2011; **27:** 2903–2904.

12 Nicorici D, Satalan M, Edgren H, *et al*. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv* 2014.

13 Haas B, Dobin A, Stransky N, *et al*. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv* 2017.

14 Dobin A, Davis CA, Schlesinger F, *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; **29:** 15–21.

15 Trapnell C, Williams BA, Pertea G, *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010; **28:** 511–515.

16 Walther C, Mayrhofer M, Nilsson J, *et al*. Genetic heterogeneity in rhabdomyosarcoma revealed by SNP array analysis. *Genes Chromosomes Cancer* 2016; **55:** 3–15.

17 Varela I, Tarpey P, Raine K, *et al*. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* 2011; **469:** 539–542.

18 Raine KM, Hinton J, Butler AP, *et al*. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr Protoc Bioinformatics* 2015; **52:** 15.7.1–12.

19 Nik-Zainal S, Davies H, Staaf J, *et al*. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016; **534:** 47–54.

20 Van Loo P, Nordgard SH, Lingjærde OC, *et al*. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci USA* 2010; **107:** 16910–16915.

21 Corces MR, Trevino AE, Hamilton EG, *et al*. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* 2017; **14:** 959–962.

22 Liao Y, Smyth GK, Shi W. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014; **30:** 923–930.

23 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; **15:** 550.

24 Heinz S, Benner C, Spann N, *et al*. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010; **38:** 576–589.

25 Fontes M, Soneson C. The projection score – an evaluation criterion for variable subset selection in PCA visualization. BMC *Bioinformatics* 2011; **12:** 307.

26 Zhou Y, Zhou B, Pache L, *et al*. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* 2019; **10:** 1523.

27 Siegel DA, Huang MK, Becker SF. Ectopic dendrite initiation: CNS pathogenesis as a model of CNS development. *Int J Dev Neurosci* 2002; **20:** 373–389. Erratum in: *Int J Dev Neurosci*. 2003; **21:** 169–170.

28 Park JA, Kim KC. Expression patterns of PRDM10 during mouse embryonic development. *BMB Rep* 2010; **43:** 29–33.

29 Duhoux FP, Ameye G, Montano-Almendras CP, *et al*. PRDM16 (1p36) translocations define a distinct entity of myeloid malignancies with poor prognosis but may also occur in lymphoid malignancies. *Br J Haematol* 2012; **156:** 76–88.

30 Mzoughi S, Tan YX, Low D, *et al*. The role of PRDMs in cancer: one family, two sides. *Curr Opin Genet Dev* 2016; **36:** 83–91.

31 Boulay G, Sandoval GJ, Riggi N, *et al*. Cancer-specific retargeting of BAF complexes by a prion-like domain. *Cell* 2017; **171:** 163–178.

32 McBride MJ, Kadoch C. Disruption of mammalian SWI/SNF and polycomb complexes in human sarcomas: mechanisms and therapeutic opportunities. *J Pathol* 2018; **244:** 638–649.

33 Lindén M, Thomsen C, Grundevik P, *et al*. FET family fusion oncoproteins target the SWI/SNF chromatin remodeling complex. *EMBO Rep* 2019; e45766.

34 Allen BL, Taatjes DJ. The Mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol* 2015; **16:** 155–166.

35 Aranda-Orgilles B, Saldaña-Meyer R, Wang E, *et al*. MED12 regulates HSC-specific enhancers independently of mediator kinase activity to control hematopoiesis. *Cell Stem Cells* 2016; **19:** 784–799.

36 Mattes K, Berger G, Geugien M, *et al*. CITED2 affects leukemic cell survival by interfering with p53 activation. *Cell Death Dis* 2017; 8**:** e3132.

Figure legends

Figure 1. Gene expression analysis of *PRDM10*-rearranged tumors (PRT) and five morphologically similar tumor types (DFSP = dermatofibrosarcoma protuberans; MIFS = myxoinflammatory fibroblastic sarcoma; UPS = undifferentiated pleomorphic sarcoma; MFS = myxofibrosarcoma; BFH = benign fibrous histiocytoma). (A) Unsupervised hierarchical clustering of 1,737 genes. Colored boxes below the dendrogram indicate the tumor type of each sample. PRT forms a distinct expression cluster. The three PRT with *CITED2–PRDM10* and the three PRT with *MED12–PRDM10* formed separate sub-clusters. (B and C) Boxplots of the expression of *CADM3* and *TMPRSS4*, respectively. The boxes are defined by the end of the first and third quartiles and whiskers extend from the boxes to the highest and lowest values.

Figure 2. Immunohistochemistry for CADM3. (A and B) Whereas PRDM10-rearranged tumors with the two known variants of *PRDM10* fusions showed distinct membranous positivity, (C and D) high-grade UPS and MIFS were negative. (E) Bj5ta cells containing the empty vector did not express CADM3, (F) whereas Bj5ta cells expressing *CITED–PRDM10* showed distinct membranous positivity.

Figure 3. ATAC-seq analysis of differentially accessible promoter regions in TERT-immortalized Bj5ta human fibroblast cell line with the *CITED2–PRDM10* (C-P), wild type (WT), truncated (ZFO), or empty vector (EV) plasmids. (A) Venn diagram of differentially open promoter regions in C-P, WT and ZFO. (B) RPKM normalized ATAC-seq coverage tracks for the regions surrounding *CAMD3* exon 1 and *TMPRSS4* exon 8. The data range is 0-2500 RPKM

for the *CADM3* panel and 0-4000 RPKM for the *TMPRSS4* panel. (C) unsupervised PCA at a variance threshold of 0.13, retaining 1204 variables. All constructs formed separate groups and the largest distance was observed between C-P and EV. (D) visualization of coverage (RPKM) in differentially open promoter regions shared between C-P and WT or ZFO. Plots display a region of ±1kb from the peak centers (PC). The line plot shows the mean coverage around shared open (blue) and closed (green) regions. The heatmap displays the coverage around shared open (upper panel) and closed (lower panel) regions.

Figure 4. Gene expression analysis of TERT-immortalized Bj5ta human fibroblast cell line with the *CITED2–PRDM10* (C-P), wild type (WT), truncated (ZFO), or empty vector (EV) plasmids. (A) unsupervised PCA at a variance threshold of 0.3, retaining 835 variables. C-P and EV formed separate groups while WT and ZFO were intermixed. (B) Heatmap of the upregulated and downregulated genes shared between C-P and WT or ZFO. (C) Venn diagram of upregulated genes in C-P, WT and ZFO. (D) Venn diagram of downregulated genes in C-P, WT and ZFO.