Transcriptomic definition of molecular subgroups of small round cell sarcomas

Sarah Watson, Virginie Perrin, Delphine Guillemot, Stéphanie Reynaud, Jean-Michel Coindre, Marie Karanian, Jean-Marc Guinebretière, Paul Fréneaux, Francois Le Loarer, Megane Bouvet, et al.

Title: Transcriptomic definition of molecular subgroups of small round cell sarcomas

Running Title: Molecular classification of sarcoma subtypes

Authors
Sarah Watson[1,2], Virginie Perrin[1,2], Delphine Guillemot[3], Stephanie Reynaud[3], Jean-Michel Coindre[4,5], Marie Karanian[6], Jean-Marc Guinebretière[7], Paul Freneaux[8], François Le Loarer[4,5], Megane Bouvet[3], Louise Galmiche-Rolland[9,10], Frédérique Larousserie[11], Elisabeth Longchampt[12], Dominique Ranchere-Vince[6], Gaelle Pierron[3]*, Olivier Delattre[1,2,3,13]*, Franck Tirode[1,2,14]*
*Co-senior authors

Affiliations
1 INSERM U830, Laboratory of Genetics and Biology of Cancer, F-75005, Paris, France
2 Institut Curie, Paris Sciences et Lettres, F-75005, Paris, France
3 Institut Curie, Unité de génétique somatique, F-75005, Paris, France
4 Institut Bergonié, Department of Pathology, F-33000, Bordeaux, France.
5 Université Bordeaux 2, F-33000, Bordeaux, France.
6 Centre Leon Bérard, Department of Pathology, F-69008, Lyon, France.
7 Service de Pathologie, Hôpital René-Huguenin, Institut Curie, F-92210, Saint-Cloud, France.
8 Département de Biologie des Tumeurs, Institut Curie, Service d'anatomie pathologique, F-75005, Paris, France
9 Service d'Anatomie Pathologique, Hôpital Necker Enfants malades, F-75015, Paris, France
10 Université Paris Descartes, F-75006, Paris, France
11 Service d'Anatomie Pathologique, Hôpital Cochin, F-75014, Paris, France
12 Service d'Anatomie et de Cytologie Pathologiques, Hôpital Foch, F-92151, Suresnes, France
13 Ligue Contre le Cancer, Equipe labellisée
14 Univ Lyon, Université Claude Bernard Lyon 1, INSERM 1052, CNRS 5286, Cancer Research Center of Lyon, Centre Léon Bérard, F-69008, Lyon, France

Corresponding authors:
Franck Tirode, Ph.D.
Biology of Rare Sarcomas Group
Cancer Research Center of Lyon, INSERM U1052
Centre Léon Bérard, 28 Rue Laënnec, 69373 Lyon Cedex 08
Tel.: +33 4 69 85 61 45
Email: franck.tirode@lyon.unicancer.fr

Olivier Delattre, M.D., Ph.D.
INSERM U830 "Genetics and Biology of Cancers"
Institut Curie Research Centre
26 rue d'Ulm, 75248 Paris cedex 05, France
Tel.: +33 1 56 24 66 79
Fax.: +33 1 56 24 66 30
Email: olivier.delattre@curie.fr

Conflict of Interest statement:

Authors have nothing to disclose

Word count: 3929

Abstract

Sarcoma represents a highly heterogeneous group of tumours. We report here the first unbiased and systematic search for gene fusions combined with unsupervised expression analysis of a series of 184 small round cell sarcomas. Fusion genes were detected in 59% percent of samples, with half of them being observed recurrently. We identified biologically homogeneous groups of tumours such as the *CIC*-fused (to *DUX4*, *FOXO4* or *NUTM1*) and *BCOR*-rearranged (*BCOR-CCNB3*, *BCOR-MAML3, ZC3H7B-BCOR* and *BCOR* internal duplication) tumour groups. *VGLL2*-fused tumours represented a more biologically and pathologically heterogeneous group. This study also refined the characteristics of some entities such as *EWSR1-PATZ1* spindle cell sarcoma or *FUS-NFATC2* bone tumours that are different from *EWSR1-NFATC2* tumours and transcriptionally resemble *CIC*-fused tumour entities. We also describe a completely novel group of epithelioid and spindle-cell rhabdomyosarcomas characterized by *EWSR1*- or *FUS-TFCP2* fusions. Finally, expression data identified some potentially new therapeutic targets or pathways.

Keywords

Sarcoma, Fusion genes, FET-*TFCP2*, *FUS-NFATC2*, *EWSR1-PATZ1*, *VGLL2-NCOA2*, *BCOR*-Rearranged, *CIC*-Fused, RNAseq.

Introduction

Small round cell sarcoma is a heterogeneous group of tumours mostly affecting children and young adults and characterized by an overall poor prognosis [1]. They remain challenging for

pathologists since those tumours share overlapping morphological and immunophenotypical features. The identification of specific fusion transcripts has considerably improved their diagnosis as many subtypes are characterized by a specific fusion gene, such as *EWSR1/FUS-ETS, SS18-SSX and PAX-FOXO1* fusions in Ewing sarcoma [2], synovial sarcoma [3] and aRMS [4], respectively, or the more recently described *CIC-DUX4* [5], *EWSR1-NFATc2* [6] and *BCOR-CCNB3* [7] in "Ewing-like" tumours. As a result, diagnostic evaluation of specific chromosome translocations by FISH or of the resulting fusion transcripts by RT-PCR has become an asset for the molecular assessment of small round cell sarcoma. However, a number of sarcomas remains uncharacterised, raising both diagnostic and therapeutic issues.

In the present study, we report RNA-sequencing of 184 small round cell sarcoma samples from three different cohorts. By applying several fusion algorithms and performing unsupervised clustering analysis, we were able to show that recurrent fusions define homogeneous groups of tumours based on pathological features and expression profile analyses, including *FUS-NFATC2*-positive, *EWSR1-PATZ1*-positive, *CIC*-fused or *BCOR*-rearranged tumours. We also identified a new epithelioid rhabdomyosarcoma group characterized by a *EWSR1*- or *FUS-TFCP2* fusion gene. Finally, we propose genes and pathways specifically expressed in a variety of different entities that may serve as biomarkers or potential therapeutic targets.

Materials and Methods

Tumour samples. Tumour samples were chosen based on their pathological diagnosis either as unclassified sarcoma or as known sarcoma that did not carry pathognomonic genetic aberrations or suspicious for sarcoma with simple genetic or monomorphic features

(supplementary material, **Figure S1**). Samples were used in accordance with the French Biobank legislation.

Paired-end RNA sequencing. Total RNAs were isolated from crushed frozen tumours using a Trizol reagent kit (Life technologies). Library constructions were performed following the TruSeq Stranded mRNA LS protocol (Illumina). Sequencing were performed on either HiSeq 2500 (100nt paired-end) or NextSeq 500 (150nt paired-end) Illumina sequencing machines. All fastq files were deposited on the European Genome-phenome Archive (https://www.ebi.ac.uk/ega/studies/EGAS00001002189). Raw sequencing files of SMARCA4-DTS and MRT control samples were previously deposited in the Sequence Read Archive (#SRP052896).

Bioinformatics analyses. For fusion gene discovery, sequencing reads were injected into both the deFuse tools [8] and the FusionMap tool [9] with hg19 genome as reference. Only fusion transcripts supported by both tools with at least 2 split reads (and 3 spanning pairs for defuse) were considered. Gene expression values were extracted by Kallisto v0.42.5 [10] with GRCh38 release 79 genome annotation. Clustering, BGA, PCA and t-SNE were computed using R packages Cluster v2.0.3, made4 version 1.44.0, FactoMiner v1.31.4 and Rtsne version 0.10, respectively. Gene ontology analyses were performed online (https://david.ncifcrf.gov/) with DAVID v6.7 tool [11].

Fusion validation: PCR amplification of the fusion point using specifically designed primers and 50ng of cDNA were then performed with AmpliTaq Gold® DNA Polymerase with Buffer II (Thermo Fisher Scientific, Waltham, MA USA 02451) prior to Sanger sequencing on an Applied 3500XL Genetic Analyzer.

Immunohistochemistry. Immunohistochemistry was performed using the following antibodies and methods: NUTM1 (C52B1, Cell Signaling, dilution 1:50) and ALK (5A4, Cell Signaling, dilution 1:50) IHC were performed on a Ventana BenchMark platform with Cell Conditioning Solution 1 pre-treatment. NUTM1 was incubated for 56min and ALK for 80min prior to be revealed with the ultraView Universal DAB Detection Kit. ETV4 (clone 16, Santa Cruz Biotechnology, dilution 1:15) IHC was performed overnight at 4°C, following a pre-heating at 98°C for 30 min in Trisbuffer pH9, and revealed with REAL EnVision kit (Dako, Glostrup, Denmark).

Results

In molecular diagnostic routine, primitive small round or spindle cell sarcomas are systematically tested for the most common fusion genes using multiplexed Q-PCR (supplementary material, Table S1). In the present study, we performed whole transcriptome sequencing to investigate its efficacy for proper tumour classification. Our investigation cohort was composed of 94 samples randomly selected from about 700 retrospective samples in which no common fusion gene could be identified by standard diagnostic investigation (see the general scheme of the analysis in supplementary material, Figure S1). We first investigated the presence of gene fusions and the retained fusion candidates were subsequently validated and searched for within the remaining available samples (around 600 cases) using specific RT-PCR assay. The new cases thus identified were also RNA-sequenced and generated our follow-up cohort (12 samples). Basic clinical data, initial diagnosis and results of molecular analyses for all the 184 cases sequenced in this study are summarized in the supplementary material, Table S2. We next performed expression profile analyses that we compared using t-Distributed Stochastic Neighbour Embedding [12] and unsupervised clustering analyses (Figure 1A and B, supplementary material, Figure S2) to a

subset of 78 well defined cases composing our control cohort. For each group, differential expression analyses against each of all other tumour types were performed. Top variant genes are reported in the supplementary material, Table S3 together with ontology enrichment and gene set enrichment analyses (GSEA).

A fusion gene could be identified in almost 3/5 of the investigation cohort samples

We identified fusion genes in 55 out of the 94 tumours of the investigation cohort. Except for one case (SARC036) for which numerous fusions were found on chromosome 1, suggestive of chromothripsis, the mean number of fusion events detected per sample was 1.8. Forty samples carried a single fusion gene, while multiple (2 to 9) gene fusions were detected in 14 samples (supplementary material, Table S2). All fusion genes were subsequently confirmed by RT-PCR and Sanger sequencing.

Most of the RNAseq fusion-positive samples (26/55) expressed previously described fusion genes that were not included in our RT-PCR routine procedure, namely single cases of each of the following fusions: *EWSR1-PBX1*, *ACTB-GLI1*, *PAX3-MAML3*, *COL1A1-PDGFB*, *VGLL2-CITED2, FUS-NFATc2, BCOR-MAML3, ZC3H7B-BCOR, TPR-NTRK1*, *BRD3-NUTM1, KIAA1549-BRAF*; two cases with *EWSR1-PATZ1, TPM3-NTRK1, EML4-ALK* or *NAB2-STAT6* fusions; and three cases with *VGLL2-NCOA2* or *CIC-NUTM1* fusions. We also identified a variant of a *SS18-SSX2* fusion with an atypical *SS18* gene breakpoint. RT-PCR screening of these fusions in the initial cohort led to the identification of eleven additional cases: 3 *EWSR1-PATZ1,* 1 *EML4-ALK*, 2 *FUS-NFATC2*, 3 *VGLL2-NCOA2* and 2 *CIC-NUTM1* (supplementary material, Table S2).

Novel fusions were detected in 29 samples (supplementary material, Table S2). Two cases presented variants of known fusions: *UXT-TFE3*, a variant of the *ASPSCR1-* or *SFPQ-TFE3* fusions found in alveolar soft part sarcomas and renal cell carcinoma; and *IKBKG-ALK*, a

new variant among the numerous *ALK* fusions found in inflammatory myofibroblastic tumours. In two cases, we identified a completely new fusion between *EWSR1* or *FUS* and *TFCP2*. Eleven samples harboured previously undescribed fusions, involving genes relevant for tumorigenesis, but for which no additional sample could be identified during the RT-PCR screen. These fusions were then considered as private to their respective tumour. Finally, in the remaining 14 fusion-positive cases, we were unable to propose a definitive driver event due to our lack of knowledge on the implication in cancer of the fused genes.

Emerging *VGLL2-NCOA2/CITED*, *CIC-fused* and *BCOR*-rearranged sarcoma subtypes

A total of 7 samples harbouring a *VGLL2* fusion with either *NCOA2* (n=6) or *CITED* (n=1) were identified (supplementary material, Figure S3A) with two samples forming a primary/relapse couple (SARC070_Primary and SARC070_Relapse). These fusion genes characterized tumours occurring in very young children (below 5yo), as previously described [13]. Centralized review of cases highlighted two subtypes with specific pathological aspects. In three cases (SARC061, SARC065 and SARC070_primary), the tumours were composed of a large amount of fibrous stroma with scarce tumour cells (Figure 2A). Tumour cells presented neither atypia nor abnormal mitotic figures. They were negative for desmin, whereas few cells (below 1%) showed a nuclear positivity for myogenin. Fewer than 25% of cells were Ki67 positive. In contrast, in the four other cases (SARC070 at relapse, SARC085, SARC088 and SARC102), the cellular mass was far denser with cells presenting numerous atypia and mitoses. Myogenin staining was strongly positive in around 10% of cells and desmin and Ki67 immuno-reactivity was displayed by over 30% of tumour cells (Figure 2B). Consistently with this histological heterogeneity, t-SNE analysis and unsupervised clustering revealed a relatively disparate group of tumours (Figure 1A and supplementary material, Figure S2A) which samples could further be separated according to the histology, when

different clustering conditions were applied (supplementary material, Figure S2C-D). When considering these two histological subtypes separately it appeared that the “fibrous” subtype was enriched in immune/inflammatory response genes, while the “dense” subtype was enriched in cell cycle/proliferation genes (supplementary material, Table S3). Supervised group comparisons and between group analyses (BGA) indicated that *VGLL2*-fused tumours expressed numerous muscle-related genes as well as genes involved in extracellular matrix and in the epithelial-mesenchymal transition process. Despite clear positivity for muscle differentiation markers, none of the two subtypes clustered with rhabdomyosarcoma samples.

A total of five *CIC-NUTM1* (supplementary material, Figure S3B) fusions were retrieved. All but one case were observed in young children, at various locations (supplementary material, Table S2). Tumours harbouring *CIC-NUTM1* fusion clustered tightly with the other *CIC-DUX4*- or -*FOXO4*-positive samples (Figure 1). All these tumours overexpressed ETS family members of the *PEA3* type [5,14,15] as well as a variety of secreted and matrix protein genes such as *VGF*, *BMP2*, *glypican 3/4*, *NRG1*, *PTX3*, *pleiotrophin* and *spondins*. GSEA revealed enrichment in genes of extracellular matrix but also in genes overlapping proliferation/response to drug/immune response categories (supplementary material, Table S3).

The seven samples presenting a rearrangement of *BCOR*, including *BCOR-MAML3* or *ZC3H7B-BCOR* fusions [16] and internal tandem duplication (ITD) [17–19] (supplementary material, Figure S3C), grouped together with the *BCOR-CCNB3*-positive samples from the control cohort, describing a very-well defined cluster. These samples also shared some pathological features, as previously described [17]. Among the most significantly enriched gene ontologies for this *BCOR*-rearranged group were “developmental protein” and

"homeobox" (supplementary material, Table S3) with a strong overexpression of *HOX-A*, *-B*, *-C* and *-D* family genes as well as *HMX1*, *PITX1*, *ALX4* or *DLX1*. More specifically, we observed very strong gene sets and ontology enrichments for the morphogenesis, development and differentiation of neurons and for the skeletal system development. Finally, different genes encoding membrane receptors including *RET, FGFR2/3, EGFR, PDGFRA, NTRK3, KIT* or *NGFR* were also highly and constantly overexpressed in these tumours and may hence constitute actionable target genes.

Identification of a *FUS-NFATC2* fusion gene in a new bone tumour entity transcriptionally distinct from *EWSR1-NFATC2*-positive tumours

We identified *FUS-NFATC2* fusion genes in three tumours of the femur of adult patients (median age 38.3yo), like the only case described in the literature (17). Tumours displayed areas composed of round tumour cells arranged in sheets or short fascicles. Tumour cells were embedded in a variably myxoid stroma (Figure 3A). Focal hemangiopericytic vascular network was present in all cases (Figure 3B). More distinctively, these tumours harboured focal myxohyaline foci raising suspicion of cartilaginous differentiation (Figure 3C). All tumours displayed brisk mitotic activity and necrosis. Altogether, these tumours were histologically reminiscent of *CIC*-fused sarcomas. Clustering analyses indicated that *FUS-NFATC2*-positive tumours grouped together and were clearly distinct from all other *FET*-fused samples, including *EWSR1-NFATC2*-positive tumours (Figure 1). While the *EWSR1-NFATC2* tumours were strongly enriched in genes associated with inflammatory and immune responses (supplementary material, Tables S3 and S4), the *FUS-NFATC2* tumours were enriched in proliferation and drug resistance signatures. In accordance with the potential areas of cartilaginous differentiation, genes involved in the extracellular matrix of cartilaginous tissues (like *ACAN*, *COL9A2*, *MATN3*, *COMP* or *CILP2*) or encoding secreted

proteins (*pleiotrophin*, *DKK3*, *ANGPTL2*, *SBSPON* or *WNT5B*) were preferentially expressed in *FUS-NFATC2*-positive tumours.

EWSR1-PATZ1 **fusion gene is found in a new tumour group unrelated to any other** ***EWSR1*****-fused tumour.**

While only single cases of tumour carrying an *EWSR1-PATZ1* fusion have been previously reported [21–23], we identified in our cohorts a total of 5 cases (Figure 4A and supplementary material, Figure S3D). All tumours were from soft tissues and occurred across a very broad age range (from 0.9 to 68.5yo). Centralized pathological review of three cases indicated three different morphological aspects (Figure 4B): i) A first tumour was composed of bundles of relatively monomorphic spindle cells with an eosinophilic cytoplasm and enlarged hyperchromatic nuclei; ii) A second case consisted of a diffuse proliferation with several vessels forming a “haemangioma-like” vasculature. Tumour cells were mostly round, and focally spindle, with scant cytoplasm and an enlarged nucleus containing one nucleolus; iii) The last case resided in a diffuse proliferation of spindle cells displaying an eosinophilic cytoplasm, oval nuclei with an irregular chromatin distribution with one or several nucleoli. Some nests of epithelioid cells with abundant, eosinophilic or clear, cytoplasm and atypical nuclei could be seen. Nevertheless, two features were observed consistently: a fibrous stroma and the presence of at least one component of spindle-shaped cells. All cases were negative for EMA and AE1/AE3 and inconsistent for PS100, cytoplasmic CD99 and vimentin staining. Ki67 was high (between 20% and 70%). In agreement with this heterogeneous histological presentation, the proposed diagnoses for *EWSR1-PATZ1*-positive tumours were quite variable including unclassified spindle cell sarcoma, myxoid liposarcoma, Ewing-like sarcoma, or unclassified malignant neuroectodermal tumours. Nevertheless, all five tumours tightly clustered together and away from other *EWSR1*-fused

tumours, indicative of a transcriptionally different entity. GSEA identified enrichment of genes correlated with SMARCA2 expression in prostate cancer (supplementary material, Table S3).

A new "epithelioid rhabdomyosarcoma" entity characterized by *EWSR1/FUS-TFCP2* fusion

New fusions involving either *EWSR1* (exon 5) or *FUS* (exon 6) and *TFCP2* (exon2) were identified in three cases (Figure 5A and supplementary material, Figure S4A), linking part of the low complexity domains of EWSR1/FUS to the CP2 DNA binding and the SAM/pointed domains of TFCP2. The three tumour samples formed a discrete cluster away from any other *EWSR1/FUS*-fused tumour samples (Figure 1 and supplementary material, Figure S2). *EWSR1/FUS-TFCP2*-positive tumours arose in young adult females (age range 16-38yo) and developed in either the pelvic region, chest wall or the sphenoid bone. All three tumours were extremely aggressive, since patient survival did not exceed 5 months. Upon review, all three tumours were composed of an epithelioid proliferation arranged in small sheets or short fascicles. Tumour cells presented monotonous round nuclei with high grade features and prominent nucleoli (Figure 5B). They were associated with variable amounts of fibrous stroma with focal sclerosing areas that were present in all cases. The three tumours stained positive for desmin, MYOD1 and myogenin. Gene expression profile comparison confirmed a strong expression of *MYOD1* and *DES* as well as an impressive overexpression of *TERT* and *ALK* (supplementary material, Figure S4B), the latter being heterogeneously expressed with both a cytoplasmic and membranous staining (Figure 5B). *In silico* functional analyses highlighted T cell immune response as well as keratin intermediate filament enrichment (supplementary material, Table S3). Altogether the observation of positivity for both muscle and epithelial markers suggests that this *EWSR1/FUS-TFCP2* fusion defines a new aggressive "epithelioid rhabdomyosarcoma" entity.

Expression profiling reveals distinct transcriptomic patterns and potential biomarkers

Supervised group comparisons and between group analyses (BGA) highlighted genes specifically expressed in the molecularly defined entities (Figure 6; supplementary material, Table S3). More specifically, crossing pairwise differential analysis and BGA analysis, we identified genes that were specifically expressed in each tumour type including *VGLL2*-fused tumours (*LANCL2*), *CIC*-fused tumours (*ETV4*), *BCOR*-rearranged (*HES7*), *FUS-NFATC2*-positive sarcomas (*CD8B*), *EWSR1-PATZ1*-positive tumours (*GPR12*) and *FET-TFCP2*-positive tumours (*REQL*), but also for almost all of the other tumour entities (supplementary material, Figure S5).

Discussion

We have used RNA sequencing to investigate unclassified small round or spindle cell sarcomas. We first confirm that sarcomas, like haematological malignancies [24], demonstrate a high incidence of gene fusions as compared to carcinomas. In this respect, it is noteworthy that the mesenchyme, from which sarcomas are derived, and haematopoiesis, from which arise leukaemias and lymphomas, both originate from the mesoderm. One hypothesis may rely on the activity of specific recombinases in mesodermal derived tissues. In this respect we can mention the role of the RAG1 recombinase in the generation of lymphoma-specific translocations [25] and the recently suspected role of the PGBD5 recombinase in malignant rhabdoid tumours [26].

Combining fusion gene discovery and expression profiling, our analysis resulted in the delineation of biologically homogeneous groups of tumours. While the *CIC-NUTM1* fusion was discovered as a new brain tumor entities [14], we show here that they are encompassed in an homogeneous *CIC*-fused group of tumors together with *CIC-DUX4*- and *CIC-FOXO4*-positive samples [27]. The overexpression of genes of the *PEA3* type of the *ETS* family

observed in all *CIC*-fused samples is known to be a consequence of a loss of function of *CIC* [28], which was also involved in resistance to MAPK inhibitors [28,29] or in the promotion of metastasis [30]. Further analyses should confirm whether a dominant negative effect on wild type *CIC* is a consequence of *CIC*-fusions and if it may be correlated with the aggressiveness of *CIC*-fused tumors. The *BCOR*-rearranged group of tumors includes various *BCOR*-fusion genes and *BCOR*-ITD. *BCOR*-ITD were described in CCSK [19], in a new brain tumor entity CNS HGNET-BCOR [14] and in soft tissue undifferentiated round cell sarcoma of infancy sharing strong similarities with CCSK [17]. This homogeneous biological entity may also present clinical specificities depending on the type of genetic lesions. Indeed, while *BCOR-CCNB3* was exclusively observed in bone, *BCOR*-ITD was only observed in soft tissues. *BCOR* rearrangements may lead to an abnormal activity of the non-conventional polycomb repressive BCOR (or PRC1.1) complex [31]. In this regard, the observation that BCOR-rearrangements are associated with increased expression of genes correlated with *SMARCA2* expression (**Figure 6A**) may suggest that impairment of PRC1.1 activity lead to an increased activity of the antagonist SWI/SNF complex. Thanks to a unique collection, we found several samples carrying *FUS-NFATC2* or *EWSR1-PATZ1* fusions, which formed tight groups. *FUS-NFATC2*-positive tumours present different morphologies but with features rather reminiscent of *CIC*-fused tumours and are definitively different from *EWSR1-NFATC2*-positive tumours. *EWSR1-PATZ1* tumours present relatively divergent cell morphologies, rendering their pathological identification challenging, though areas of spindle cells found in all three samples may guide pathologists during diagnosis. The identification of more cases is nevertheless mandatory to improve the characterization of these ultra-rare FET-fused sarcomas.

We also describe here a new entity carrying an *EWSR1- or FUS-TFCP2* fusion gene. TFCP2 (also known as LSF or LBP1) does not resemble any of the usual *FET*-fusion partners. It was first described as an activator of the late SV40 promoter and later found to bind globins and HIV-1 promoters [32,33]. TFCP2 is ubiquitously expressed and plays determinant roles in lineage-specific gene expression or cell cycle regulation [34]. Potential involvement of TFCP2 in oncogenesis has been suggested and may depend on tumor type: In hepatocellular carcinomas, TFCP2 is overexpressed [35] and its inhibition by small molecules leads to cell cycle arrest [36], whereas in skin melanoma TFCP2 is under-expressed, and its re-expression induces growth arrest [37]. Interestingly, in addition to the expression of *MYOD1* (supplementary material, Table S3), a number of up-regulated genes in these *FUS/EWSR1-TFCP2*-positive tumours are involved in muscle biology (such as *Cholinergic Receptor Nicotinic subunit alpha1*, *delta* and *gamma* or *sarcoglycan alpha*). This may suggest either a muscle cellular origin of these tumours or a muscle differentiation program triggered by the fusion protein. Consistently with the genes expressed, pathological review confirmed that *TFCP2*-rearranged tumours were an epithelioid variant of rhabdomyosarcoma, although the morphological spectrum of these tumours needs to be assessed on a larger scale. Despite a common muscle phenotype, these samples do not cluster together with other rhabdomyosarcomas. Considering potential therapeutic targets, it is noteworthy that *ALK* and *TERT* are highly expressed in these tumours. Moreover, recently developed small molecules impeding the DNA-binding activity of TFCP2 [36,38], which remains in the fusion protein, might also be effective in hindering EWSR1-TFCP2 binding and possibly in inhibiting the development of these very aggressive tumours.

Samples carrying *VGLL2-NCOA2* or *VGLL2-CITED2* fusion genes demonstrate some transcriptome heterogeneity related to two different histological subtypes: one presenting

only few tumour cells surrounded by fibrous stroma, rather suggestive of relatively benign tumours, the other with a spindle cell rhabdomyosarcoma-like morphology similar to that reported [13] and exhibiting expression profiles enriched in cell cycle genes. Interestingly, one fibrous tumour and one rhabdomyosarcoma-like tumour represented the primary and the relapse tumours from the same patient, respectively. It is therefore likely that the rhabdomyosarcoma-like tumour represents a more aggressive evolution of the fibrous tumour. Further analyses including exome-sequencing of the two tumours may enable the identification of additional genetic mutations accounting for this evolution.

In addition to the fundamental role of expert surgical pathology, molecular analysis is now becoming an integral part of the diagnosis of small round cell tumours, in particular for resolving frequent diagnostic dilemmas. With the increase of sarcoma subclasses, it becomes difficult for the pathologist to ascertain a precise diagnosis in a reliable and cost effective way. In this respect, and depending on the expertise and technical availabilities at each pathology department, two molecular approaches may be proposed. One may rely on the design of a specific panel of genes, based on our selection of genes that are specifically expressed in each tumour group, tested in a simple assay (p.e., using Nanostring technology. Alternatively, thanks to the constant decrease of whole transcriptome costs and the availability of robust bioinformatics tools, we strongly believe that RNA-sequencing is a key approach. Indeed, our work indicates that the convergent information from gene fusion detection and expression profiles-based clustering is extremely powerful to identify biologically homogeneous groups of tumours. Hence, in the short term, RNA-sequencing is expected to constitute a mandatory methodology for an efficient molecular diagnosis of sarcoma, a requirement recently proposed in the GENSARC study [39]. Such a precise subgrouping of sarcomas is essential i) to investigate the clinical characteristics of each

subgroup, particularly regarding the risk of evolution and response to current treatment options, ii) to identify new therapeutic opportunities, as potentially *ALK* overexpression in *EWSR1/FUS-TFCP2* sarcomas, iii) to construct relevant animal models to design robust pre-clinical studies, and iv) to design highly specific assays to monitor circulating cell and tumour DNA during treatment. Further increasing the number of samples, within large international consortia, is essential to identify groups of tumours of sufficient size to enable robust conclusions and to allow the collection of otherwise “orphan” samples.

Acknowledgments

The authors wish to thank pathologists who provided tumour material: E. Angot, H. Antoine-Poirel, S. Aubert, A. Babik, J.P. Barbet, C. Bastien, A.M. Bergemer-Fouquet, D. Berrebi, L. Boccon-Gibod, C. Bossard, S. Boudjemaa, E. Cassagnau, J. Champigneulle, M.A. Chrestian, A. Clemenson, S. Collardeau-Frachon, S. Corby, J.F. Cote, A. Coulomb-Lhermine, A. Croue, C. Daniliuc, A. De Muret , A. Dhouibi, C. Douchet, F. Dujardin, J.M. Dumollard, H. Duval, M. Fabre, C. Fernandez, S. Fraitag, F. Galateau-Salle, L. Gibault, A. Gomez-Brouchet, C. Guettier, M.F. Heymann, J.F. Ikoli, J.F. Jazeron, C. Jeanne-Pasquier, A. Jouvet, R. Kaci, M. Karanian-Philippe, O. Kerdraon, J. Klijanienko, C. Labit-Bouvier, M. Lae, T. Lazure, F. Le Pessot , F. Lemoine, A. Liprandi, F. Llamas - Gutierrez, M.C. Machet, A. Maran-Gonzalez, L. Marcellin, P. Marcorelles, C. Marin, A. Maues De Paula, C.A. Maurage, A. Moreau, A. Neuville, H. Perrochia, J.M. Picquenot, C. Renard, V. Rigau, J. Riviere, C. Rouleau, A. Rouquette, H.Sartelet, I. Serre, N. Stock, H. Szabo, P. Terrier, M. Terrier-Lacombe, V. Thomas De Montpreville, J. Tran Van Nhieu, E. Uro-Coste, P. Validire, I. Valo, V. Verkarre, J.M. Vignaud, M.O. Vilain, M.L. Wassef, D. Zachar, L. Zemoura and the tumorothèque Necker-Enfants Malades. We also thank Brigitte Manship for the careful

reading of the manuscript. This work was supported by the Institut National de la Santé et de la Recherche Médicale, the Institut Curie, the Ligue National Contre Le Cancer, the Institut National du Cancer and la Direction générale de l'offre de soins (INCa-DGOS_5716), the European PROVABES (ERA- 649 NET TRANSCAN JTC-2011), ASSET (FP7-HEALTH-2010-259348), and EEC (HEALTH-F2-2013-602856) projects. U830 is also indebted to the Société Française des Cancers de l'Enfant, Enfants et Santé, Courir pour Mathieu, Dans les pas du Géant, La course de l'espoir du mont Valérien, Au nom d'Andréa, Association Abigaël, Association Marabout de Ficelle, Les Bagouz à Manon, Les amis de Claire, the association Adam. SW was supported by a grant from the Fondation Nuovo-Soldati. High-throughput sequencing has been performed by the NGS platform of Institut Curie, supported by the grants ANR-10-EQPX-03 and ANR10-INBS-09-08 from the Agence Nationale de la Recherche (investissements d'avenir) and by the Canceropôle Ile-de-France.

Author contributions

SW, GP, OD and FT designed the study and wrote the manuscript. SW, VP, DG, SR and MB performed all the experiments. SW, GP and FT, acquired and analysed the data. JMC, MK, JMG, PF, DRV and FLL provided samples and were the reference senior pathologists. LGR, FL and EL recruited patients and provided numerous samples and clinical information.

References

1 Antonescu C. Round cell sarcomas beyond Ewing: emerging entities. *Histopathology* 2014; **64**: 26-37

2 Delattre O, Zucman J, Melot T, *et al.* The Ewing family of tumors – A subgroup of small-round-cell tumors defined by specific chimeric transcripts. *N Engl J Med* 1994; **331**: 294-299

3 Clark J, Rocques PJ, Crew AJ, *et al.* Identification of novel genes, SYT and SSX, involved in the t(X;18)(p11.2;q11.2) translocation found in human synovial sarcoma. *Nat Genet* 1994; **7**: 502-508

4 Galili N, Davis RJ, Fredericks WJ, *et al.* Fusion of a fork head domain gene to PAX3 in the solid tumour alveolar rhabdomyosarcoma. *Nat Genet* 1993; **5**: 230-235

5 Kawamura-Saito M, Yamazaki Y, Kaneko K, *et al.* Fusion between CIC and DUX4 up-regulates PEA3 family genes in Ewing-like sarcomas with t(4;19)(q35;q13) translocation. *Hum Mol Genet* 2006; **15**: 2125-2137

6 Szuhai K, IJszenga M, de Jong D, *et al.* The NFATc2 gene is involved in a novel cloned translocation in a Ewing sarcoma variant that couples its function in immunology to oncology. *Clin Cancer Res* 2009; **15**: 2259-2268

7 Pierron G, Tirode F, Lucchesi C, *et al.* A new subtype of bone sarcoma defined by BCOR-CCNB3 gene fusion. *Nat Genet* 2012; **44**: 461-466

8 McPherson A, Hormozdiari F, Zayed A, *et al.* deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data. *PLOS Comput Biol* 2011; **7**: e1001138

9 Ge H, Liu K, Juan T, *et al.* FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinforma Oxf Engl* 2011; **27**: 1922-1928

10 Bray NL, Pimentel H, Melsted P, *et al.* Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016; **34**: 525-527

11 Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009; **37**: 1-13

12 Van der Maaten L, Hinton G. Visualizing High-Dimensional Data Using t-SNE. *J Mach Learn Res* 2008; **9**: 2579-2605

13 Alaggio R, Zhang L, Sung Y-S, *et al.* A Molecular Study of Pediatric Spindle and Sclerosing Rhabdomyosarcoma: Identification of Novel and Recurrent VGLL2-Related Fusions in Infantile Cases. *Am J Surg Pathol* 2016; **40**: 224-235

14 Sturm D, Orr BA, Toprak UH, *et al.* New Brain Tumor Entities Emerge from Molecular Classification of CNS-PNETs. *Cell* 2016; **164**: 1060-1072

15 Le Guellec S, Velasco V, Pérot G, *et al.* ETV4 is a useful marker for the diagnosis of CIC-rearranged undifferentiated round-cell sarcomas: a study of 127 cases including mimicking lesions. *Mod Pathol* 2016; **29**: 1523-1531

16 Specht K, Zhang L, Sung Y-S, *et al.* Novel BCOR-MAML3 and ZC3H7B-BCOR Gene Fusions in Undifferentiated Small Blue Round Cell Sarcomas. *Am J Surg Pathol* January 2016

17 Kao Y-C, Sung Y-S, Zhang L, *et al.* Recurrent BCOR Internal Tandem Duplication and YWHAE-NUTM2B Fusions in Soft Tissue Undifferentiated Round Cell Sarcoma of Infancy: Overlapping Genetic Features With Clear Cell Sarcoma of Kidney. *Am J Surg Pathol* 2016; **40**: 1009-1020

18 Kao Y-C, Sung Y-S, Zhang L, *et al.* BCOR Overexpression Is a Highly Sensitive Marker in Round Cell Sarcomas With BCOR Genetic Abnormalities: *Am J Surg Pathol* 2016; **40**: 1670-1678

19 Roy A, Kumar V, Zorman B, *et al.* Recurrent internal tandem duplications of BCOR in clear cell sarcoma of the kidney. *Nat Commun* 2015; **6**: 8891

20 Brohl AS, Solomon DA, Chang W, *et al.* The genomic landscape of the Ewing sarcoma family of tumors reveals recurrent STAG2 mutation. *PLoS Genet* 2014; **10**: e1004475

21 Mastrangelo T, Modena P, Tornielli S, *et al.* A novel zinc finger gene is fused to EWS in small round cell tumor. *Oncogene* 2000; **19**: 3799-3804

22 Qaddoumi I, Orisme W, Wen J, *et al.* Genetic alterations in uncommon low-grade neuroepithelial tumors: BRAF, FGFR1, and MYB mutations occur at high frequency and align with morphology. *Acta Neuropathol (Berl)* 2016; **131**: 833-845

23 Johnson A, Severson E, Gay L, *et al.* Comprehensive Genomic Profiling of 282 Pediatric Low- and High-Grade Gliomas Reveals Genomic Drivers, Tumor Mutational Burden, and Hypermutation Signatures. *The Oncologist* September 2017: theoncologist.2017-0242

24 Lieber MR. Mechanisms of human lymphoid chromosomal translocations. *Nat Rev Cancer* 2016; **16**: 387-398

25 Papaemmanuil E, Rapado I, Li Y, *et al.* RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nat Genet* 2014; **46**: 116-125

26 Kentsis A, Eisenberg A, Blackford AN, *et al.* PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nat Genet* 2017; **49**: 1005-1014

27 Sugita S, Arai Y, Tonooka A, *et al.* A novel CIC-FOXO4 gene fusion in undifferentiated small round cell sarcoma: a genetically distinct variant of Ewing-like sarcoma. *Am J Surg Pathol* 2014; **38**: 1571-1576

28 Dissanayake K, Toth R, Blakey J, *et al.* ERK/p90RSK/14-3-3 signalling has an impact on expression of PEA3 Ets transcription factors via the transcriptional repressor capicúa. *Biochem J* 2011; **433**: 515-525

29 Wang B, Krall EB, Aguirre AJ, *et al.* ATXN1L, CIC, and ETS Transcription Factors Modulate Sensitivity to MAPK Pathway Inhibition. *Cell Rep* 2017; **18**: 1543-1557

30 Okimoto RA, Breitenbuecher F, Olivas VR, *et al.* Inactivation of Capicua drives cancer metastasis. *Nat Genet* 2017; **49**: 87-96

31 Gearhart MD, Corcoran CM, Wamstad JA, *et al.* Polycomb Group and SCF Ubiquitin Ligases Are Found in a Novel BCOR Complex That Is Recruited to BCL6 Targets. *Mol Cell Biol* 2006; **26**: 6880-6889

32 Swendeman SL, Spielholz C, Jenkins NA, *et al.* Characterization of the genomic structure, chromosomal location, promoter, and development expression of the alpha-globin transcription factor CP2. *J Biol Chem* 1994; **269**: 11663-11671

33 Yoon JB, Li G, Roeder RG. Characterization of a family of related cellular transcription factors which can modulate human immunodeficiency virus type 1 transcription in vitro. *Mol Cell Biol* 1994; **14**: 1776-1785

34 Veljkovic J, Hansen U. Lineage-specific and ubiquitous biological roles of the mammalian transcription factor LSF. *Gene* 2004; **343**: 23-40

35 Yoo BK, Emdad L, Gredler R, *et al.* Transcription factor Late SV40 Factor (LSF) functions as an oncogene in hepatocellular carcinoma. *Proc Natl Acad Sci U S A* 2010; **107**: 8357-8362

36 Rajasekaran D, Siddiq A, Willoughby JLS, *et al.* Small molecule inhibitors of Late SV40 Factor (LSF) abrogate hepatocellular carcinoma (HCC): Evaluation using an endogenous HCC model. *Oncotarget* 2015; **6**: 26266-26277

37 Goto Y, Yajima I, Kumasaka M, *et al.* Transcription factor LSF (TFCP2) inhibits melanoma growth. *Oncotarget* 2016; **7**: 2379-2390

38 Grant TJ, Bishop JA, Christadore LM, *et al.* Antiproliferative small-molecule inhibitors of transcription factor LSF reveal oncogene addiction to LSF in hepatocellular carcinoma. *Proc Natl Acad Sci* 2012; **109**: 4503-4508

39 Italiano A, Di Mauro I, Rapp J, *et al.* Clinical effect of molecular methods in sarcoma diagnosis (GENSARC): a prospective, multicentre, observational study. *Lancet Oncol* 2016; **17**: 532-538

Figure Legend

Figure 1: Fusion genes define tumour groups.

Kallisto-extracted expression values were used for all genes having a CDS annotation in Ensembl GRCh38p5. A. t-Distributed Stochastic Neighbour Embedding (t-SNE) analysis. Tumour samples (coloured spheres) carrying identical or similar fusion genes are linked to the centroid of the group (small grey sphere). Fusion genes defining major groups are indicated. Unclassified sarcomas are not represented here for the sake of clarity but were taken into account in the analysis. An enlarged view of the centre of the figure is presented in

the supplementary material, Figure S2A. B. Hierarchical clustering. 1-pearson correlation and Ward's method were used as distance and clustering method, respectively; identified recurrent fusion genes are indicated. An enlarged view of *TFE3*-fused, *CIC*-fused and BCOR-rearranged samples are presented below (see supplementary material, Figure S2B for the full annotated cluster).

Figure 2: Subtypes of *VGLL2-NCOA2*-positive samples

A. Representative images of fibrous tumour samples. SARC070_primary and SARC065 tumour samples contained abundant fibrous stroma, displayed low Ki67 and were mostly negative for myogenin and desmin. B. Representative images of dense tumour samples. SARC070_relapse and SARC102 cells were packed and strongly positive for Ki67, with numerous cells being positive for myogenin and desmin. Scale bar: 100 µm

Figure 3: *FUS-NFATC2*-positive tumours

Morphology (hematoxylin/eosin staining) of *FUS-NFATC2* tumours. A. Proliferation of round tumour cells embedded in a variable myxoid stroma. B. Round to spindle tumour cells arranged in sheets associated with haemangioperycitic vessels. C. Tumour cells were focally embedded within myxohyaline stroma reminiscent of cartilaginous differentiation. Scale bars: 100 µm

Figure 4: *EWSR1-PATZ1*-positive tumours

A. Schematic diagram of native EWSR1 and PATZ1 proteins indicating fusion point of *EWSR1-PATZ1* fusion gene. LC: Low complexity domain; RRM: RNA recognition motif; RGG: Arg-Gly-Gly rich region; Zn: Zinc Finger, RanBP2-type; BTB/POZ: Broad-Complex, Tramtrack and Bric a brac / POxvirus and Zinc finger; H: AT hook; C2H2: Cys2-His2 Zinc-finger. Exon number based on the indicated RefSeq accession number is indicated below

each protein scheme together with amino acid length scale. B. H & E staining of three *EWSR1-PATZ1*-positive samples demonstrating a variety of morphologies (scale bar: 100 µm).

Figure 5: *FET-TFCP2* fusion

A. Schematic diagram of native proteins and fusion proteins. EWSR1-TFCP2 fusion was identified in SARC049 while FUS-TFCP2 was identified in RNA009_16_062 and RNA020_16_136 samples. LC: Low complexity domain; RRM: RNA recognition motif; RGG: Arg-Gly-Gly rich region; Zn: Zinc Finger, RanBP2-type; SAM: Sterile alpha motif / pointed domain. Exon number based on the indicated RefSeq accession number is indicated below each protein scheme. Amino acid length scale is presented at the bottom. B. Morphological spectrum of *FET-TFCP2* tumours. Hematoxylin/eosin staining (left panels) illustrating epithelioid (top), spindling (middle) and sclerosing (bottom) areas and immunohistochemistry (right panels) for MYOD1 (top), desmin (middle), and ALK (bottom) seen in all cases are illustrated in *EWSR1-TFCP2*-positive sample SARC049. ALK staining was cytoplasmic-positive with nuclear and membrane exclusion. Morphological aspects were evocative of an epithelioid and spindled variant of rhabdomyosarcoma, with numerous mitotic figures. Scale bar: 40 µm

Figure 6: Expression profiles functional analyses

Heatmap representing the specificity of the top 100 genes (detailed in the supplementary material, Table S3) identified in the between group analysis for each 24 tumour groups containing at least two samples. Data were centred and scaled (in row direction) prior analysis. Colour key of the scaled expression values is shown.

Supporting information:

Supplementary Figure S1: Constitution of the cohorts

Supplementary Figure S2: Details of the unsupervised analyses

Supplementary Figure S3: Details of VGLL2-NCOA2/CITED2, CIC-NUTM1, BCOR-ITD and EWSR1-PATZ1 fusion points.

Supplementary Figure S4: Sequence of the fusion points and genes expressed in FET-*TFCP2*-positive tumors

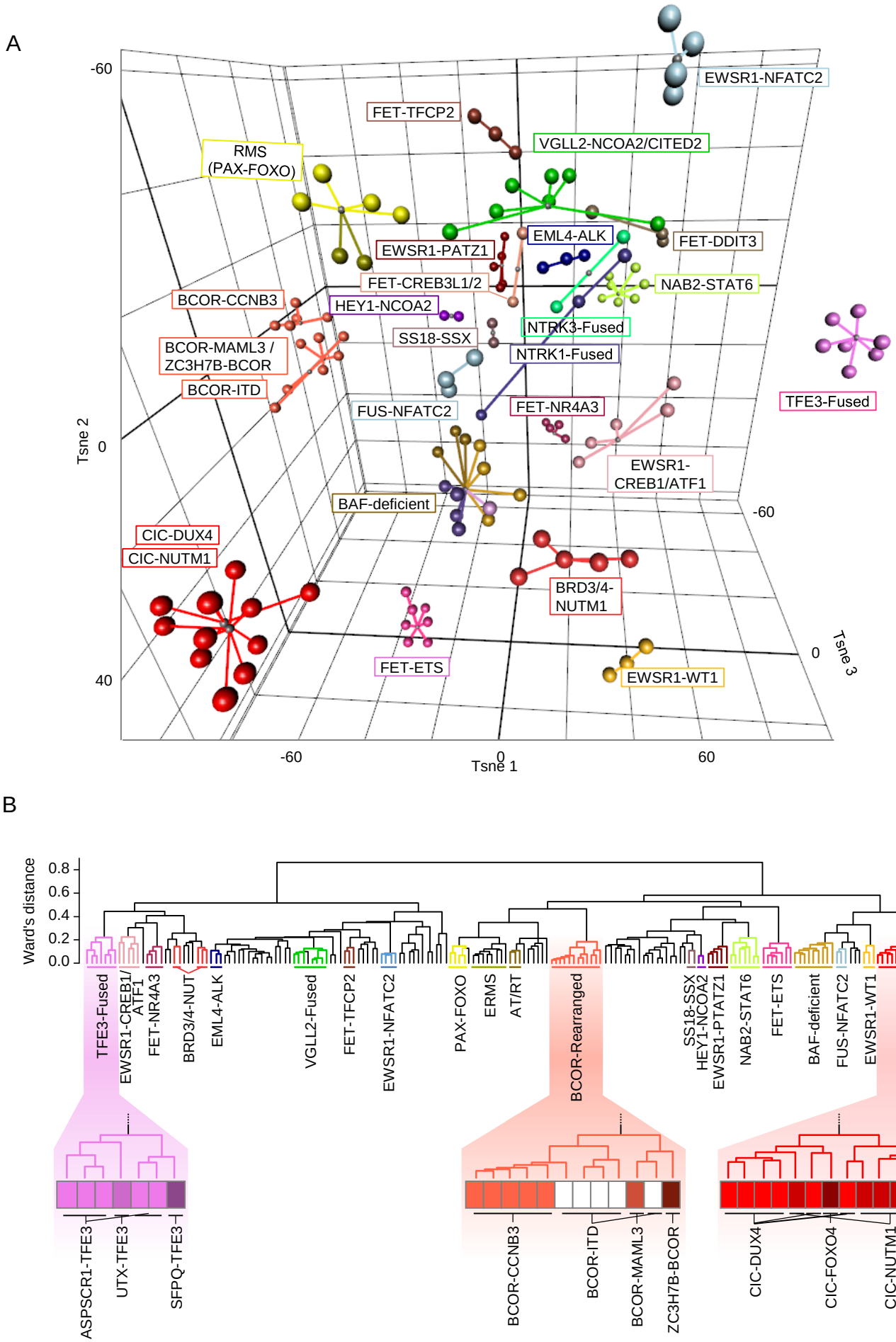
Supplementary Figure S5: Expression of the most specific gene for each tumor entity (with more than 2 samples)

Supplementary Table S1: List of the fusion genes routinely tested by RT-PCR at the Institut Curie Unité de Génétique Somatique.

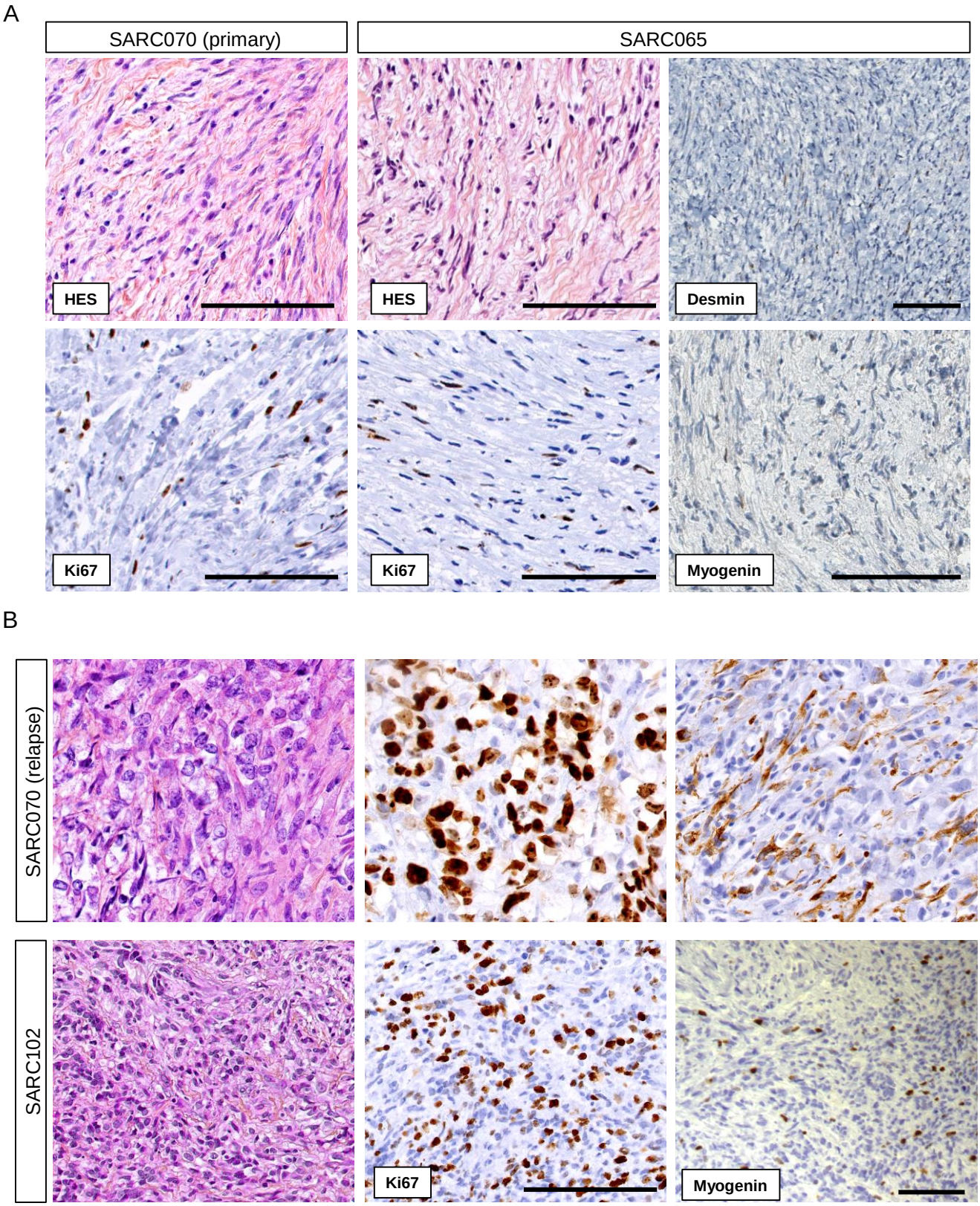
Supplementary Table S2: Description of the samples from the three cohorts

Supplementary Table S3: Differentially expressed genes in the BGA and supervised analyses and Gene ontology / GSEA enrichment for all sarcomas subtypes

Supplementary Table S4: Differentially expressed genes and gene ontology enrichment for FUS-NFATC2- vs EWSR1-NFATC2-positive tumors

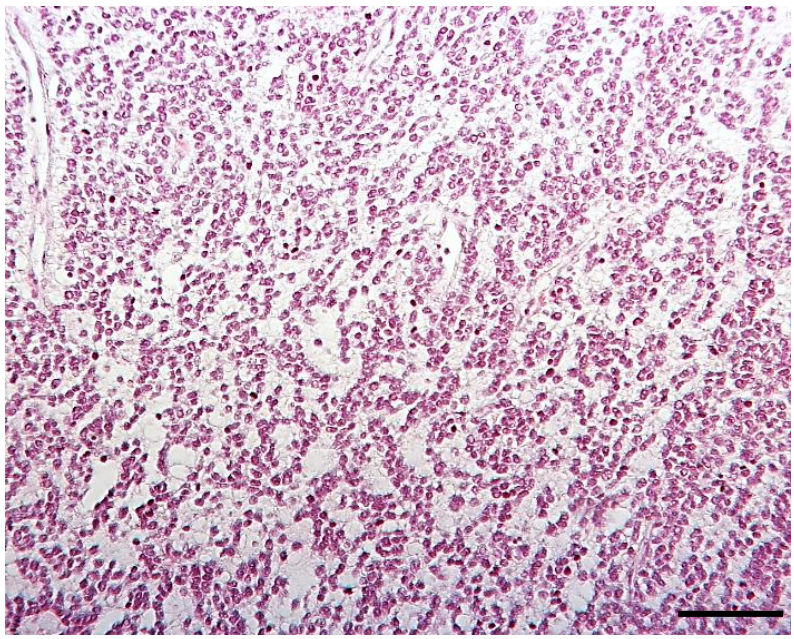


Watson et al. Figure 1

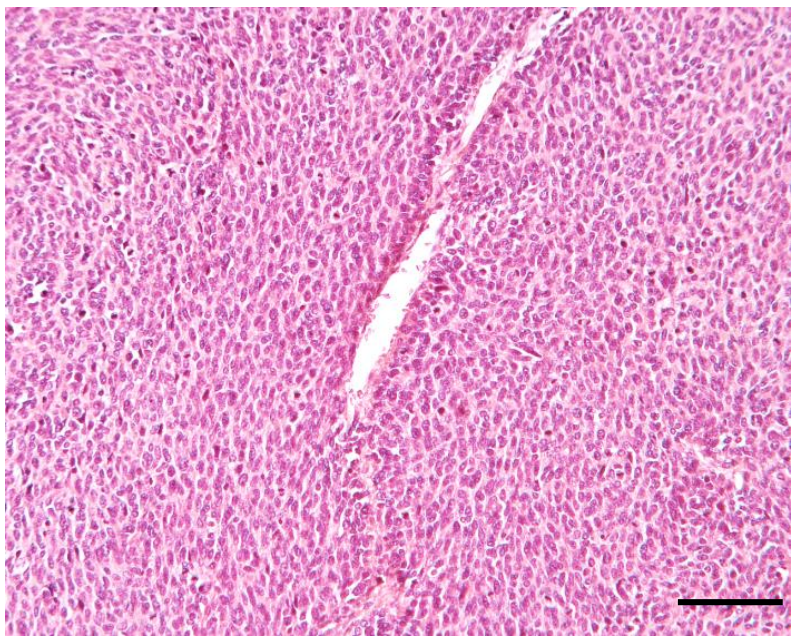


Watson et al. Figure 2

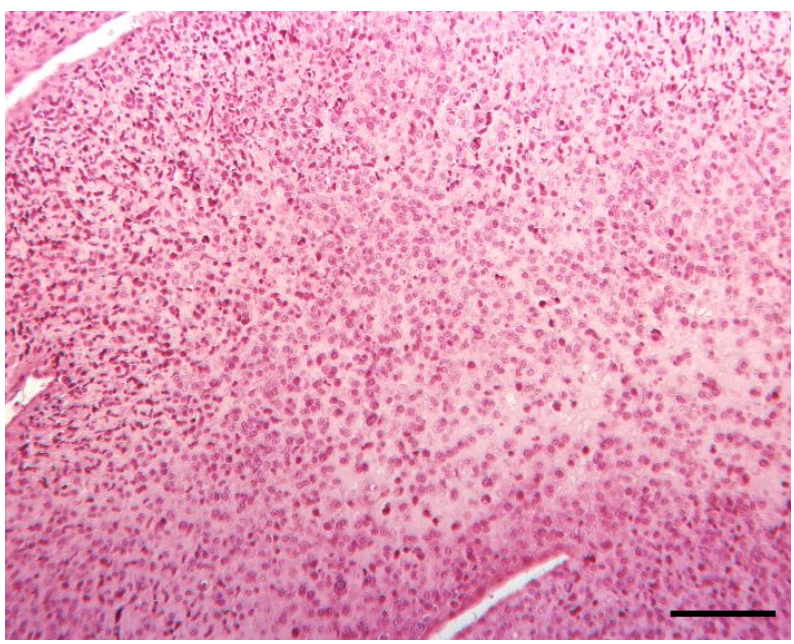
A



B

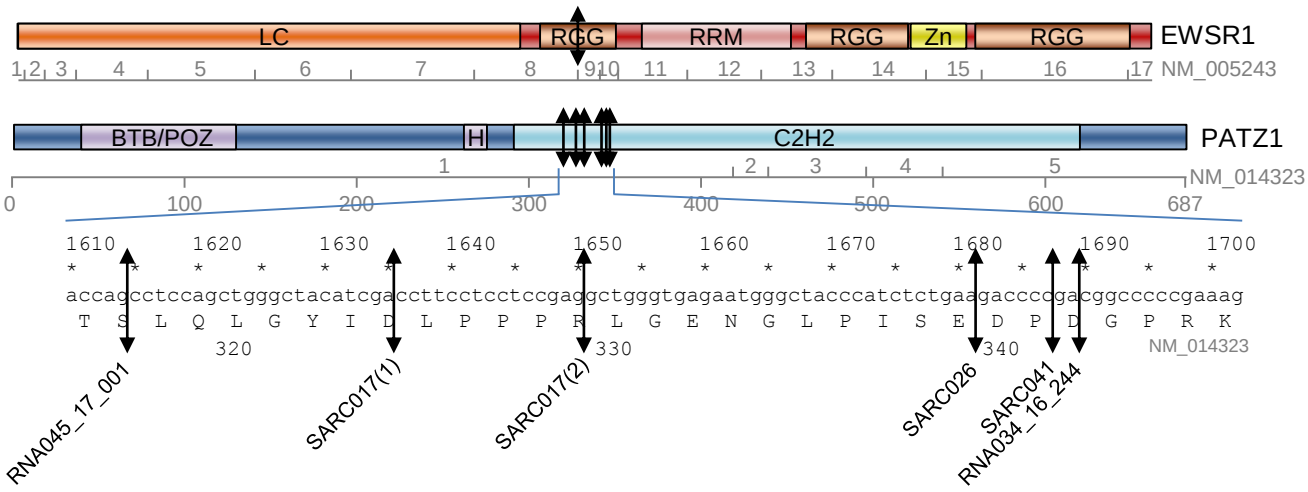


C

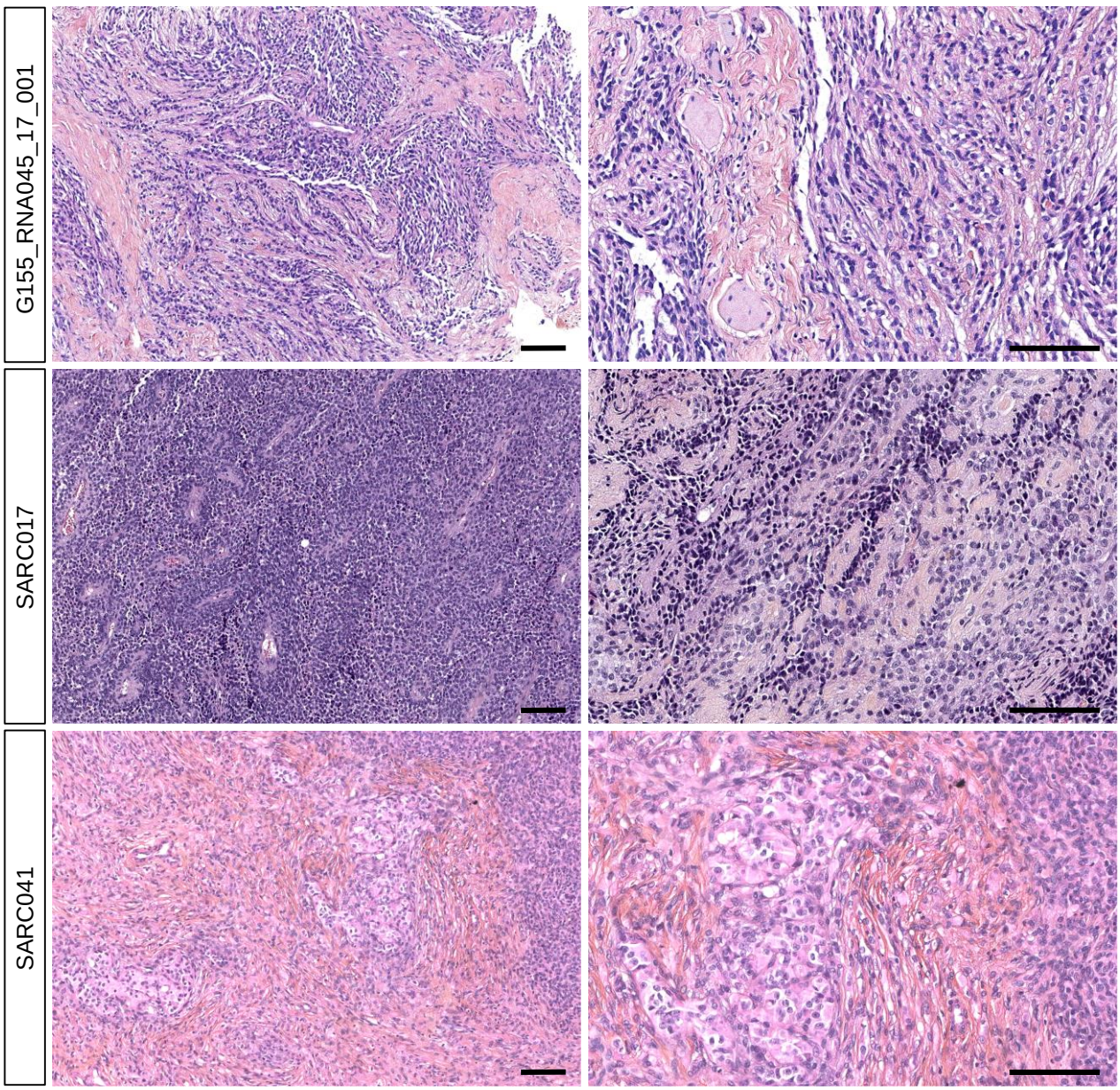


Watson et al. Figure 3

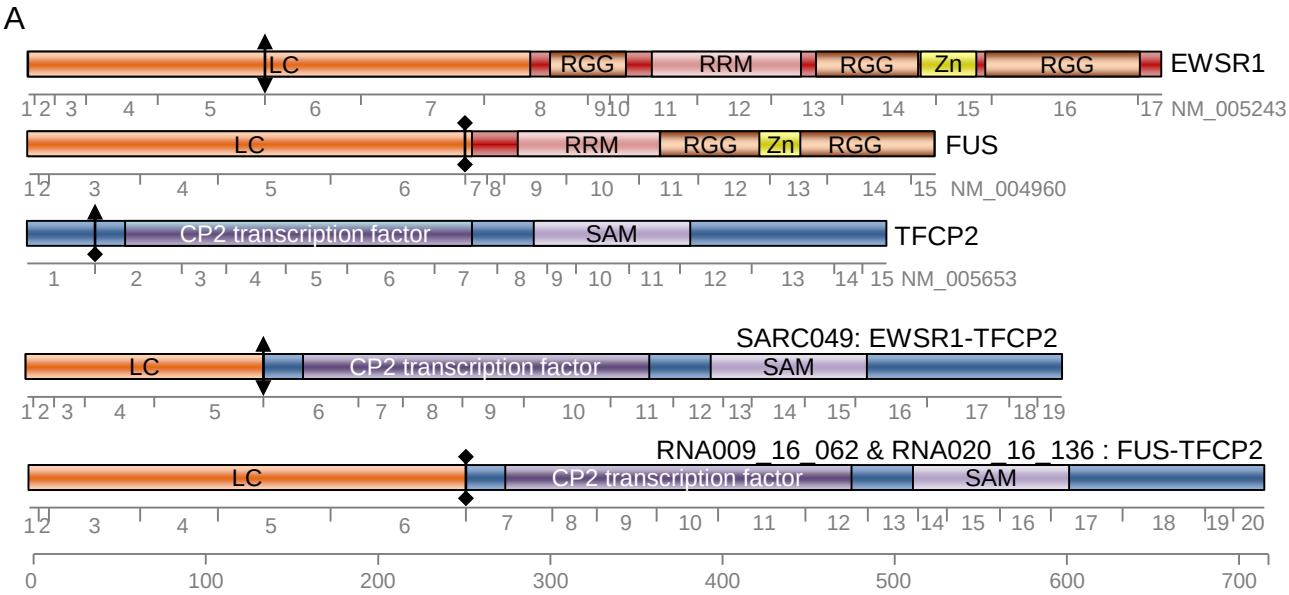
A

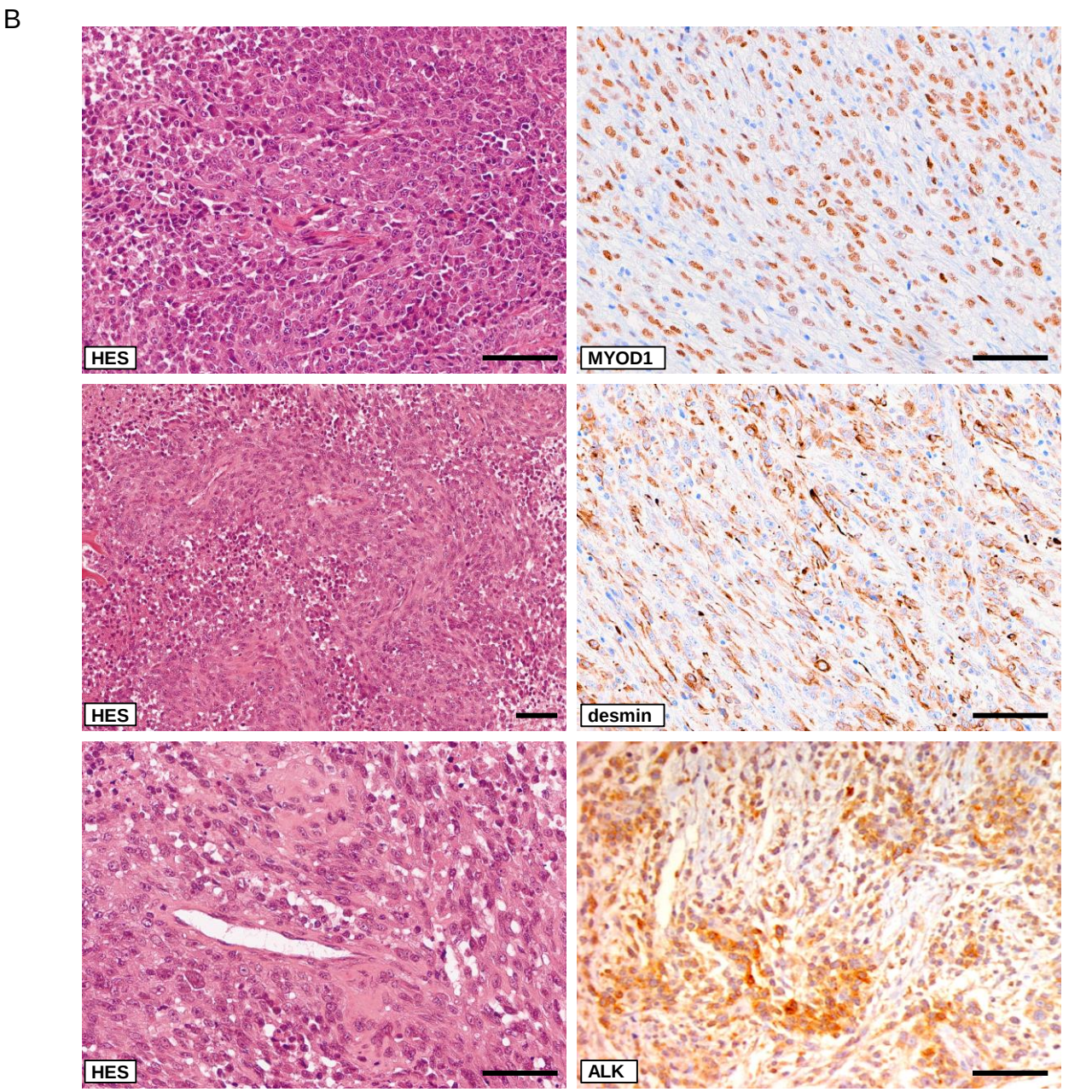


B

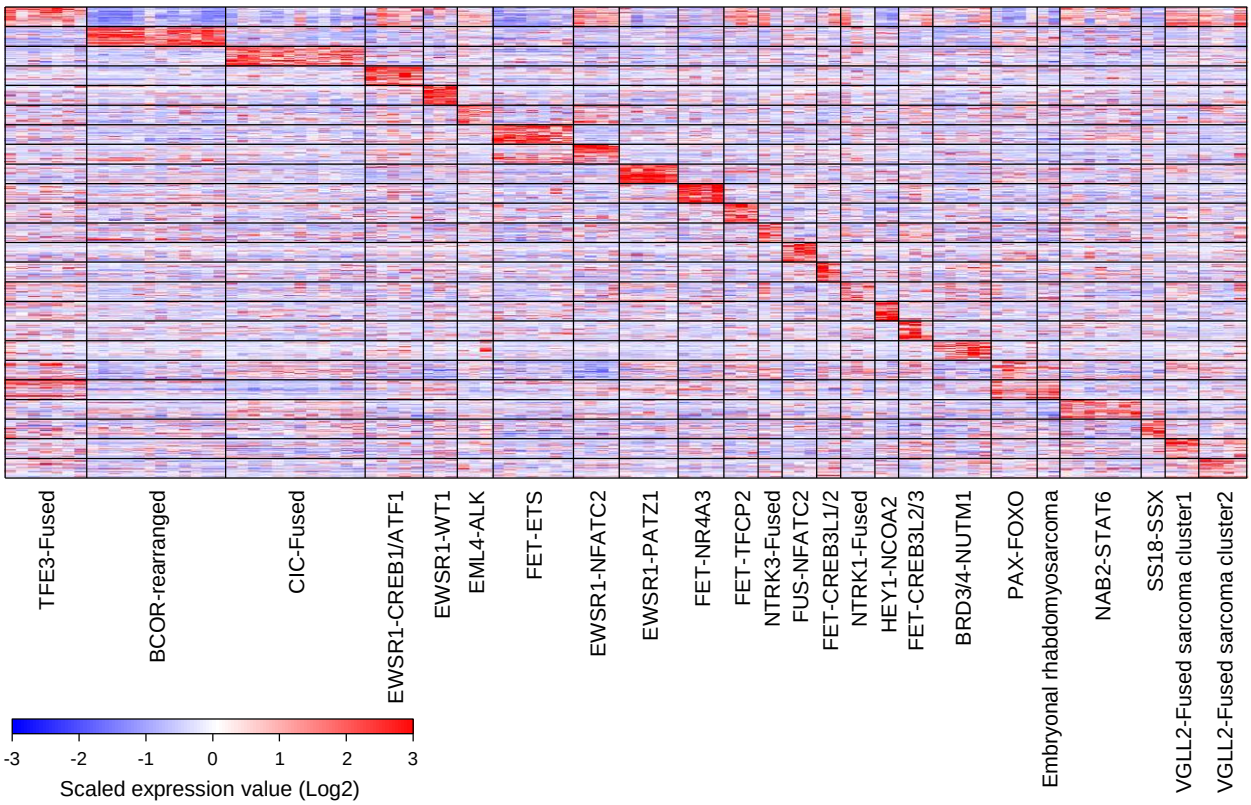


Watson et al. Figure 4

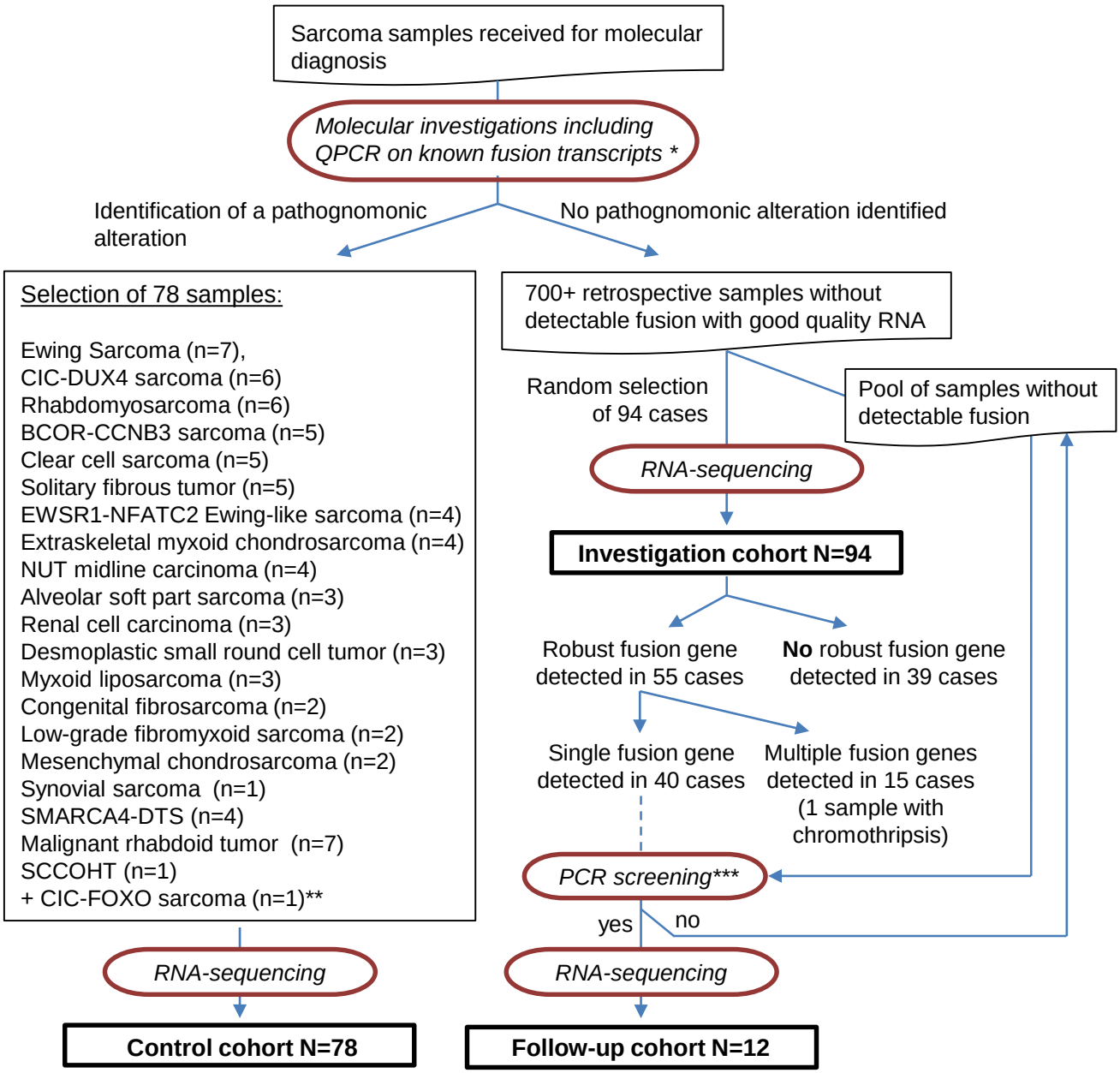




Watson et al. Figure 5



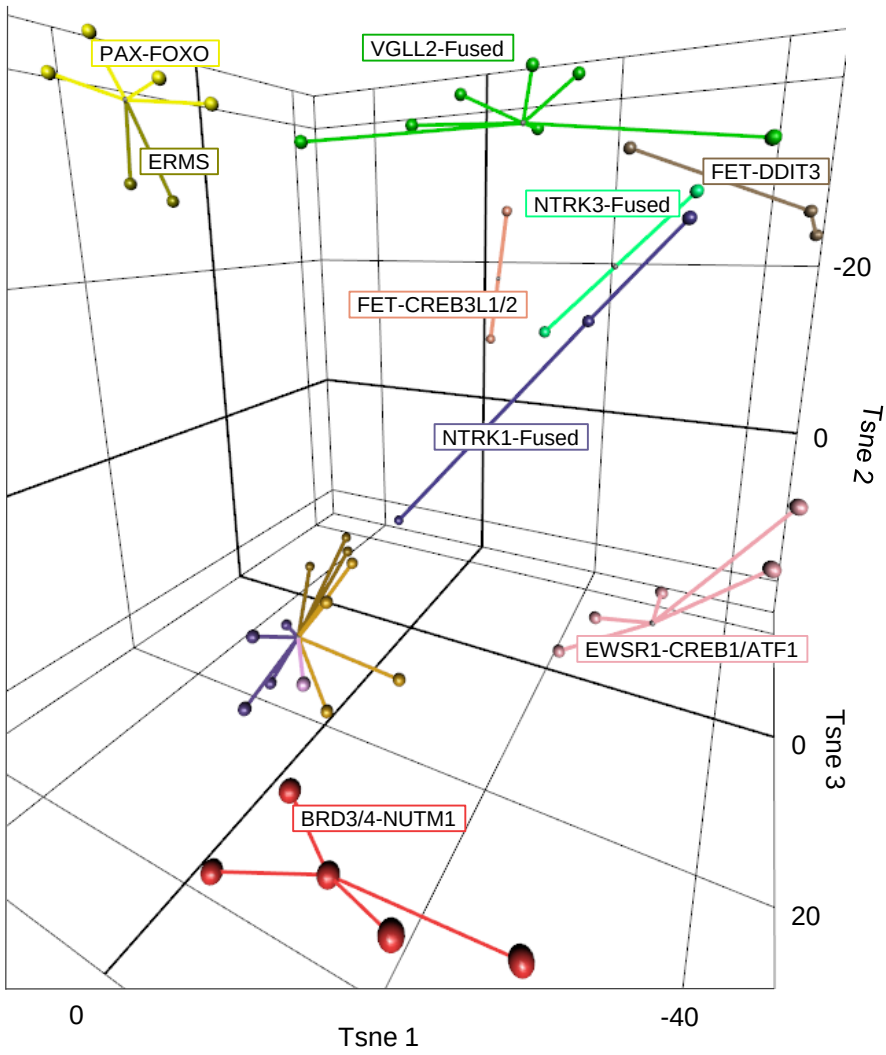
Watson et al. Figure 6



Supplementary Figure S1: Constitution of the cohorts

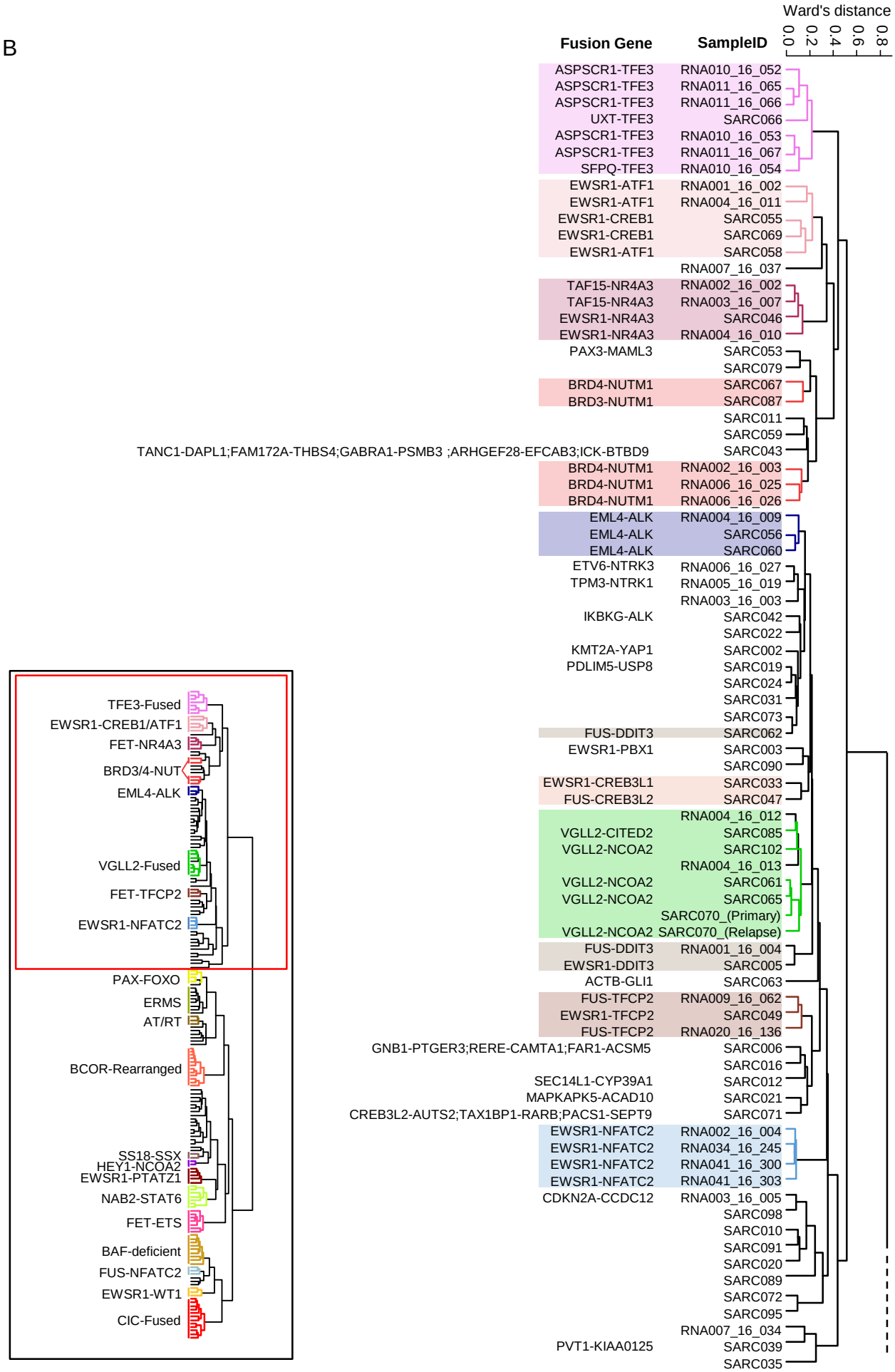
* See Supplementary table S1
** : RNAseq raw fastq files kindly provided by Yasuhito Arai and Tadashi Hasegawa [27]
*** Screened fusions were:
BCOR-MAML3, CDKN2A-CCDC12, CIC-NUTM1, EML4-ALK, EWSR1-PATZ1, EWSR1-TFCP2, FUS-TFCP2, FUS-NFATC2, KMT2A-YAP1 / YAP1-KMT2A, MN1-TAF3, NF1-RHOT1, PARG-BMS1, VGLL2-NCOA2 and VGLL2-CITED2.

A



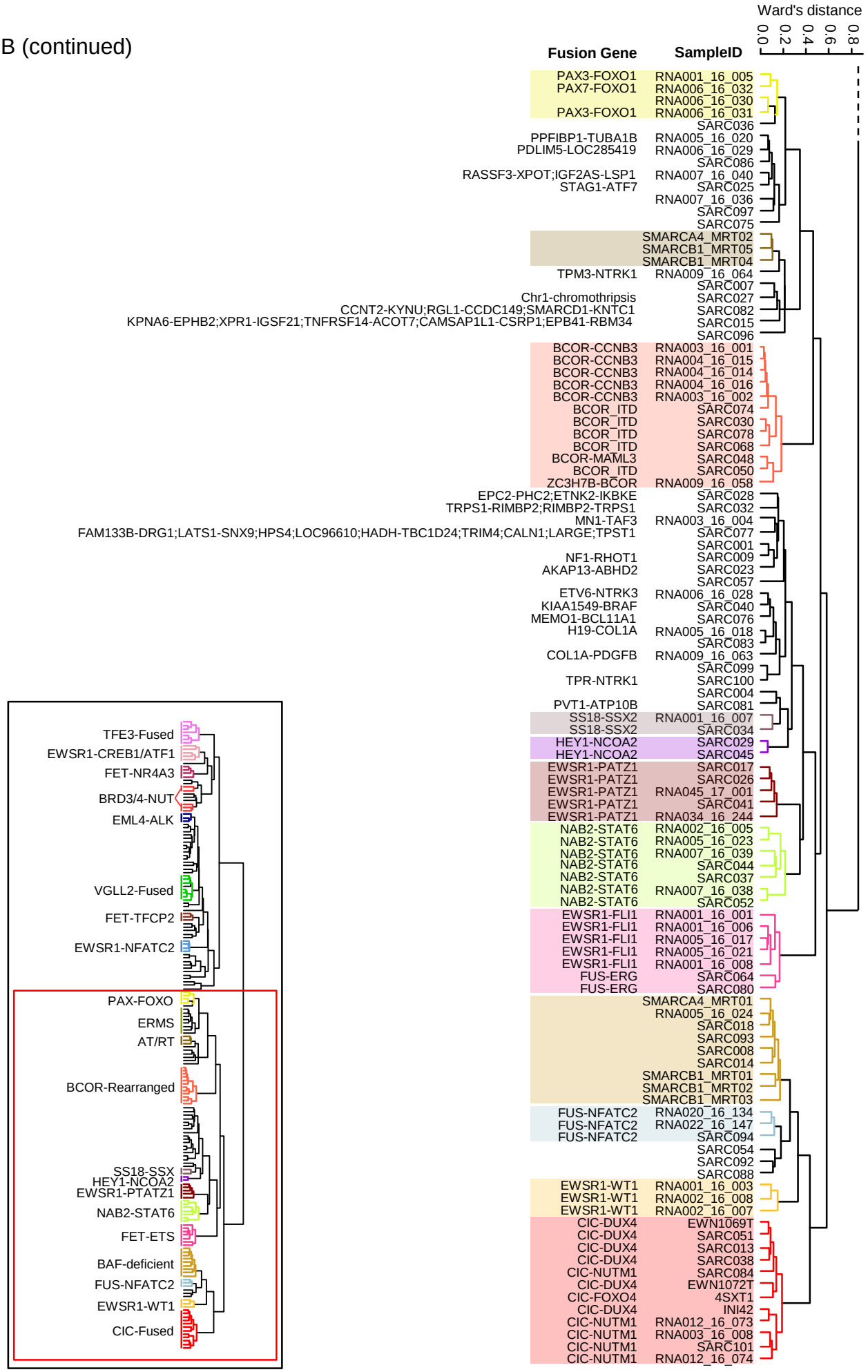
Watson et al., Supplementary Figure S2

B



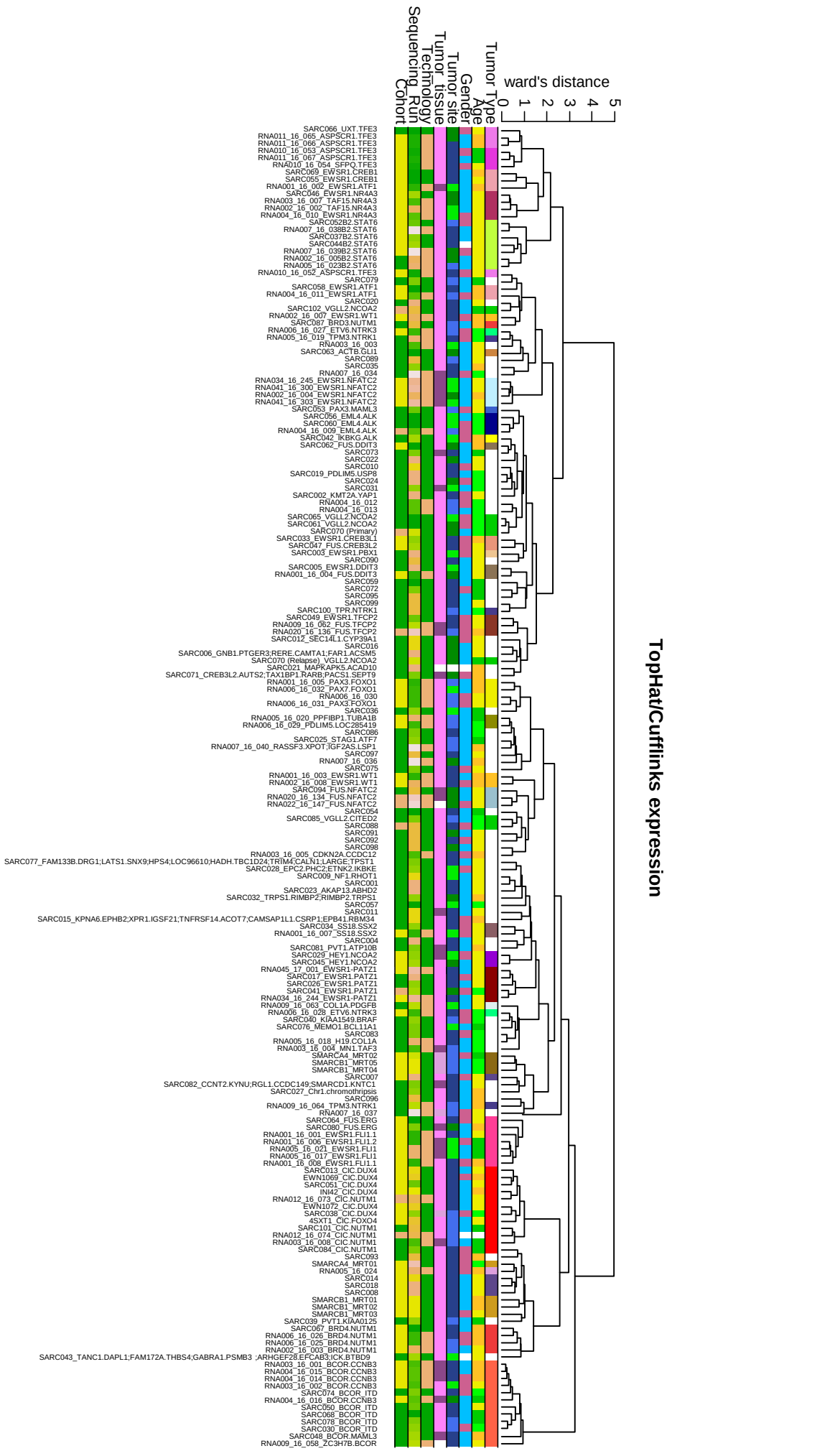
Watson et al., Supplementary Figure S2

B (continued)



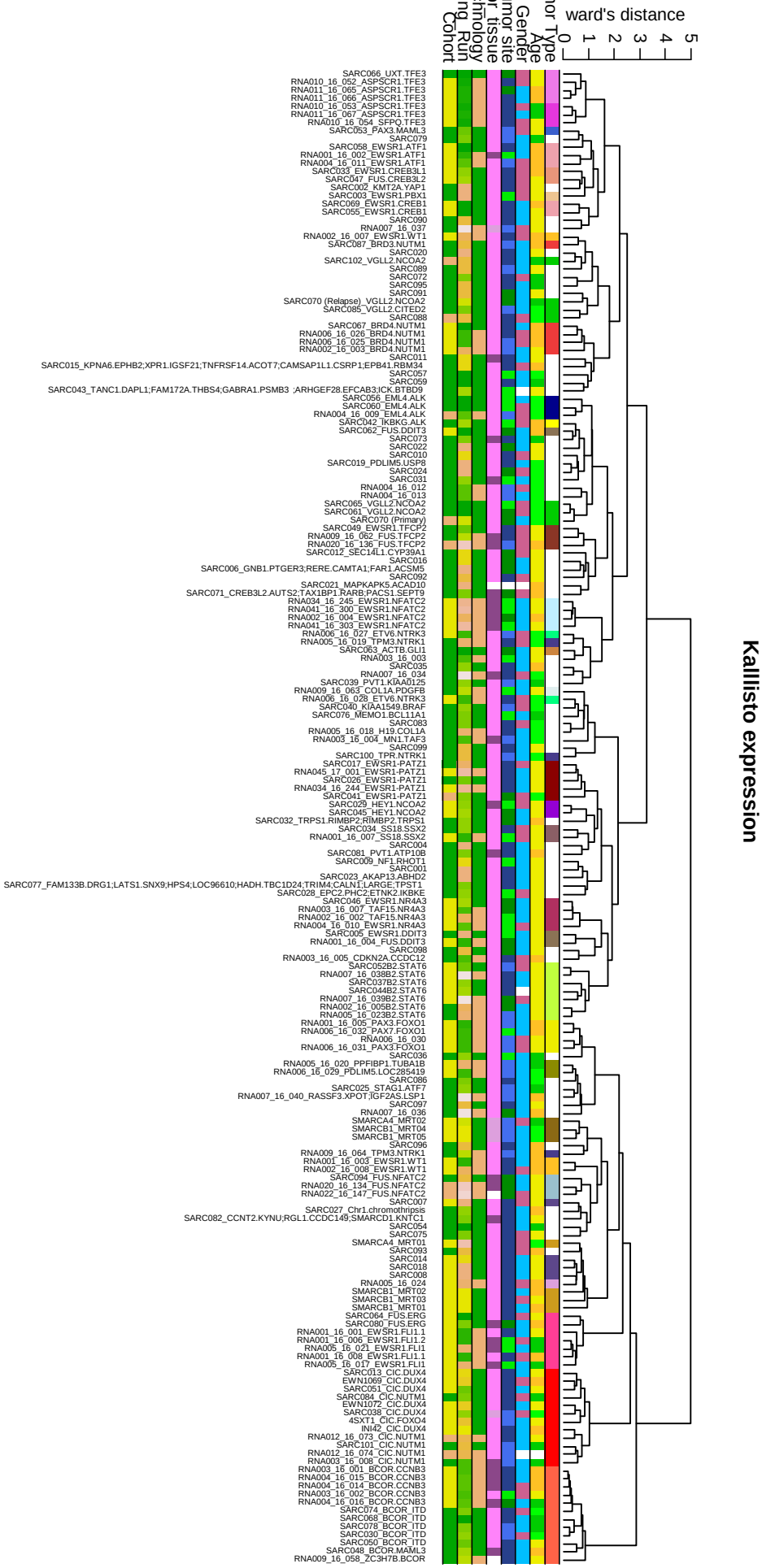
Watson et al., Supplementary Figure S2

C

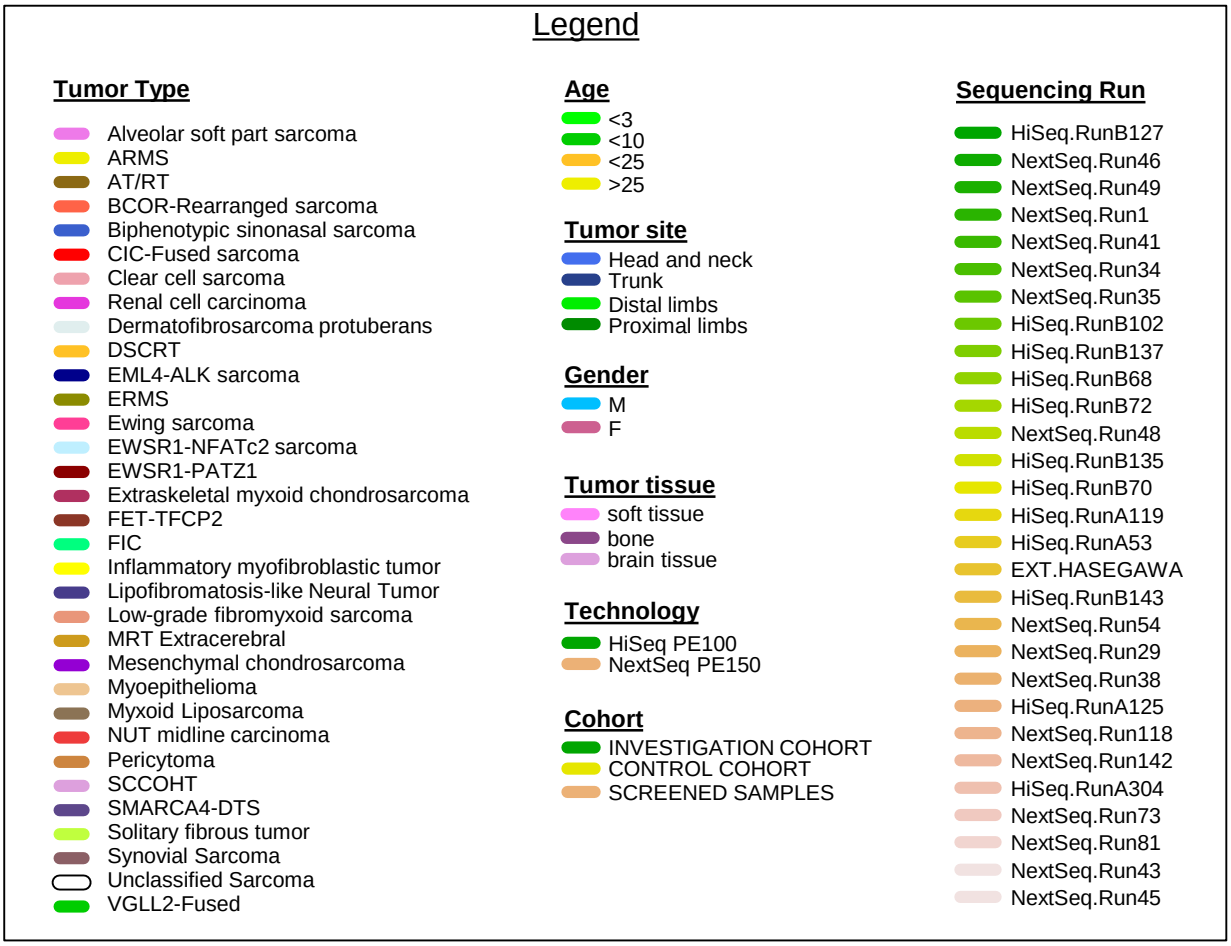


Watson et al., Supplementary Figure S2

D

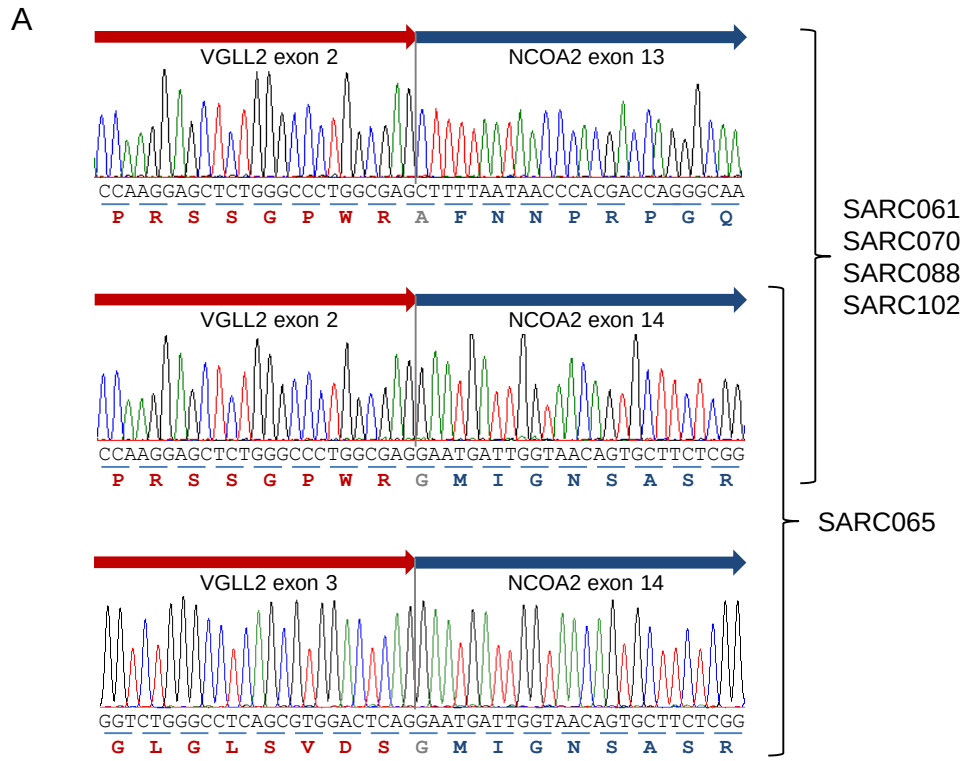


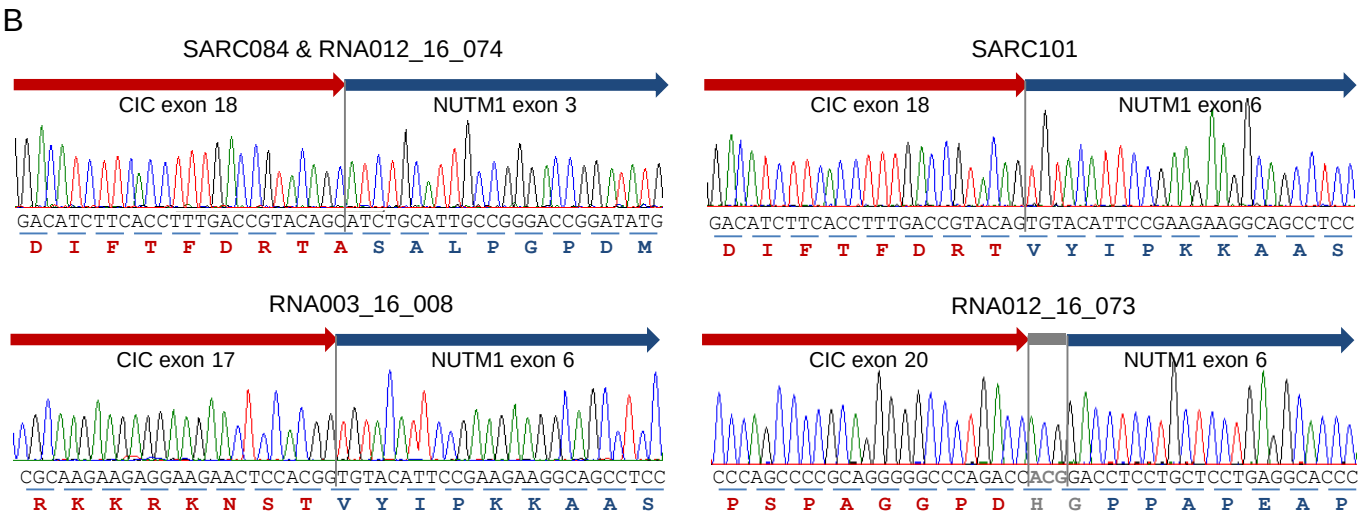
Watson et al., Supplementary Figure S2



Supplementary Figure S2: Detailed unsupervised analyses.

A) Zoom on the center region of the TSNE analysis (Figure 1A) with another angle, demonstrating dispersion of EWSR1-CREB3L1/2, NTRK1-, NTRK3- and VGLL2-fused tumors, and at a lesser extend EWSR1-CREB1/ATF1 and FET-DDIT3-fused tumors, as compared to the more defined EWSR1-NR4A3, EWSR1-WT1 or EWSR1-PATZ1 groups. B). Details of the clustering analysis (identical to Figure 1B) showing all the fusion genes identified. C&D) Expression profiles were generated using two different methods: in C) sequencing reads were aligned with TopHat2 and expression profiles were extracted using Cufflinks2 tool. In D) expression profiles were extracted with Kallisto tools that does not require prior alignment. In both cases, hierarchical clustering using the ten percent most variant genes based on interquartile range, proved to be quite similar. The main difference with Figure 1 concerned the VGLL2-fused samples that split in two or more groups. No clustering bias due to either patient ‘s clinical data (age, tumor site, gender or tissue type) or experimental condition (technology used or sequencing run) could be observed.

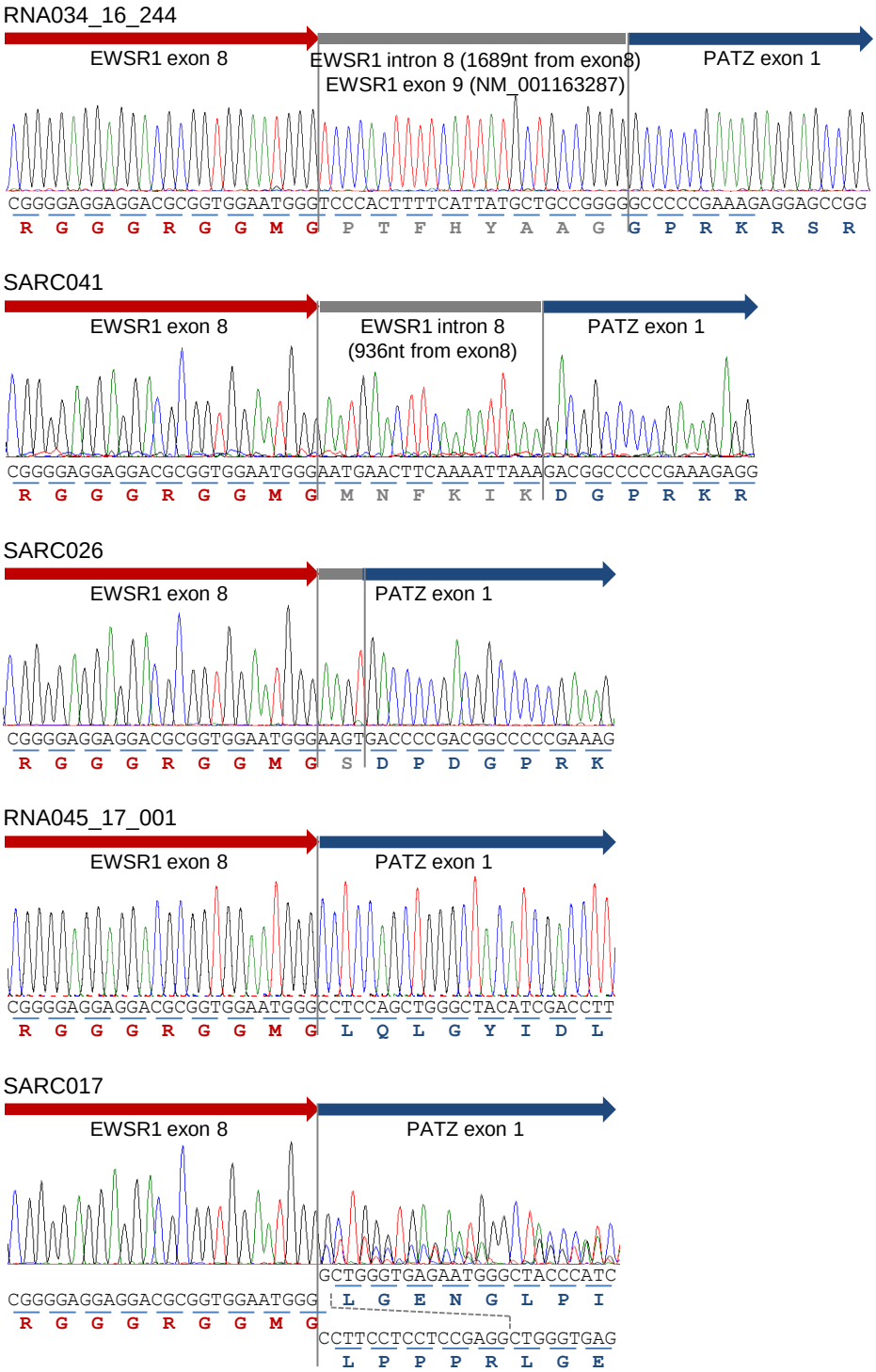




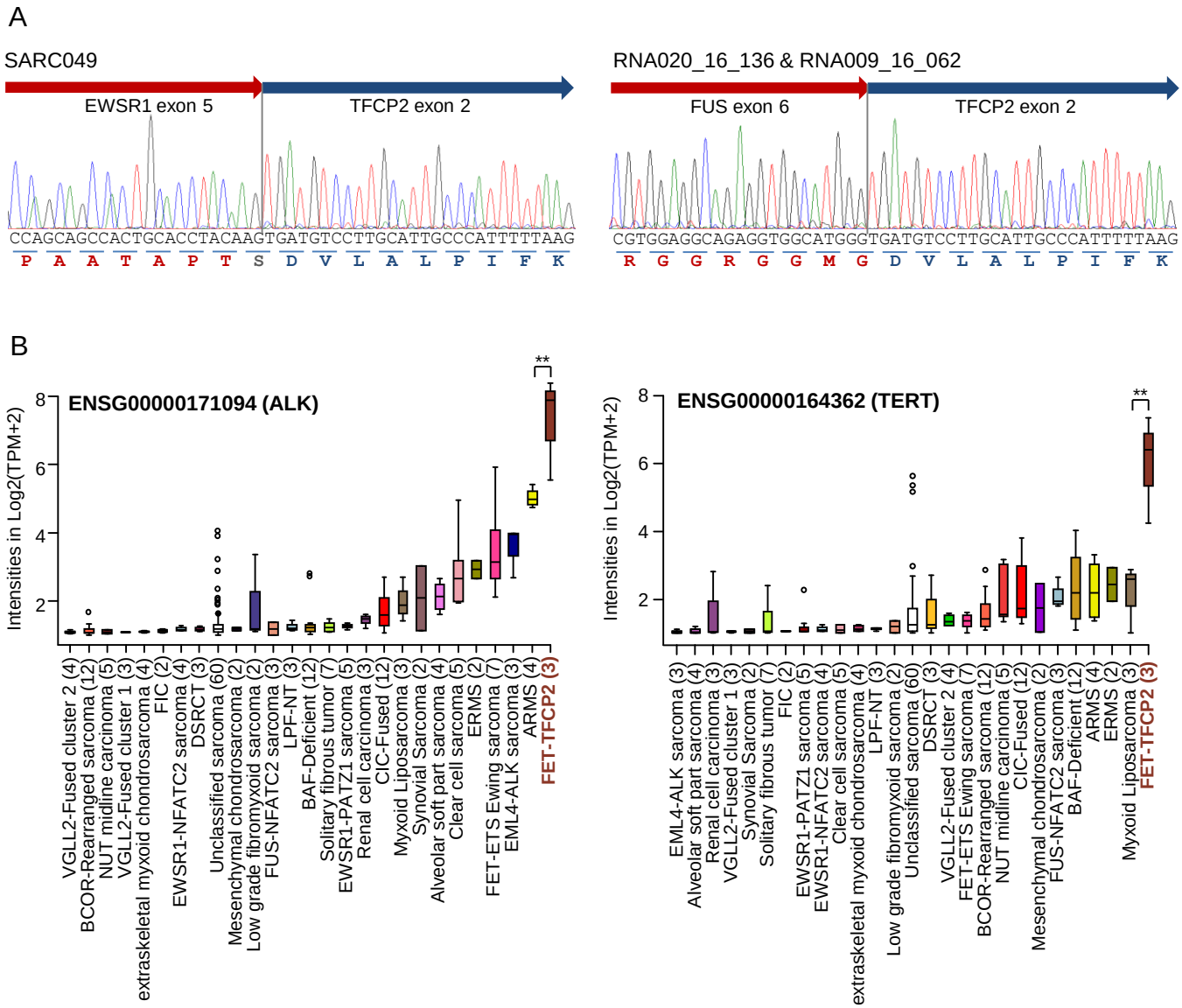
C

```
    WT CSKDLEAFNPESKELLDLVEFTNEIQTLLGSSVEWLHPSDLASDNYW*
SARC068 CSKDLEAFNPESKELLDLVEFTNEIQTLLGSSVSLEAFNPESKELLDLVEFTNEIQTLLGSSVEWLHPSDLASDNYW*
SARC050 CSKDLEAFNPESKELLDLVEFTNEIQTLLGSSVFMEAFNPESKELLDLVEFTNEIQTLLGSSVEWLHPSDLASDNYW*
SARC078 CSKDLEAFNPESKELLDLVEFTNEIQTLLGSSVEWLHPSDLASDKELLDLVEFTNEIQTLLGSSVEWLHPSDLASDNYW*
SARC074 CSKDLEAFNPESKELLDLVEFTNEIQTLLGSSVEWLHPSDLASDELLDLVEFTNEIQTLLGSSVEWLHPSDLASDNYW*
SARC030 CSKDLEAFNPESKELLDLVEFTNEIQTLLGSSVEWLHPSDLASDNYWLDLVEFTNEIQTLLGSSVEWLHPSDLASDNYW*
```

D



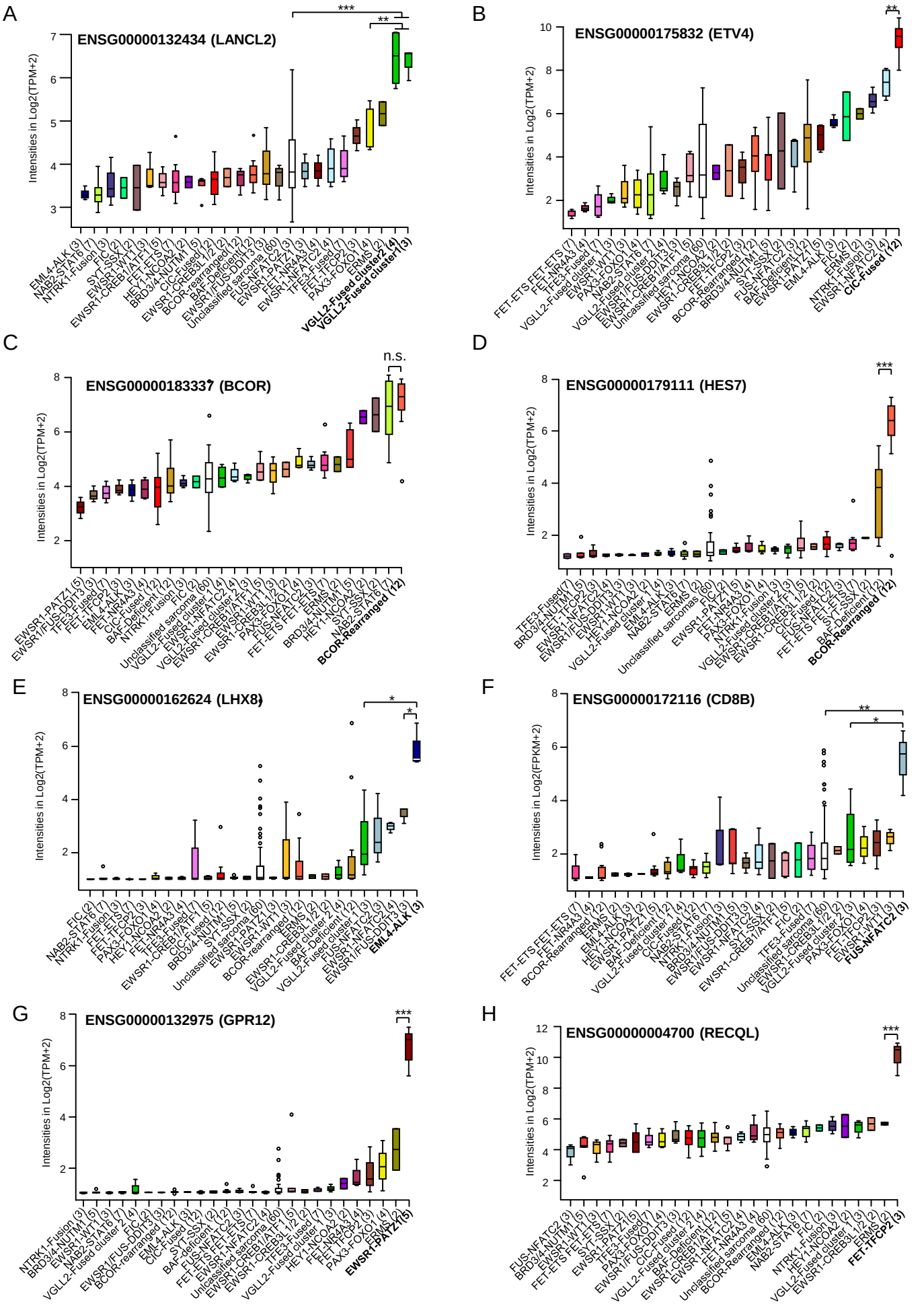
Supplementary Figure S3: Details of *VGLL2-NCOA2/CITED2* (A), *CIC-NUTM1* (B), *BCOR*-ITD (C) and *EWSR1-PATZ1* (D) fusion points.
Sanger sequences of the fusion points are given in A, B and D together with amino acid sequences of the translated fusion genes. Amino acid sequences of the internal tandem duplications in exon 15 of *BCOR* are shown for *BCOR*-ITD-positive tumors (C).



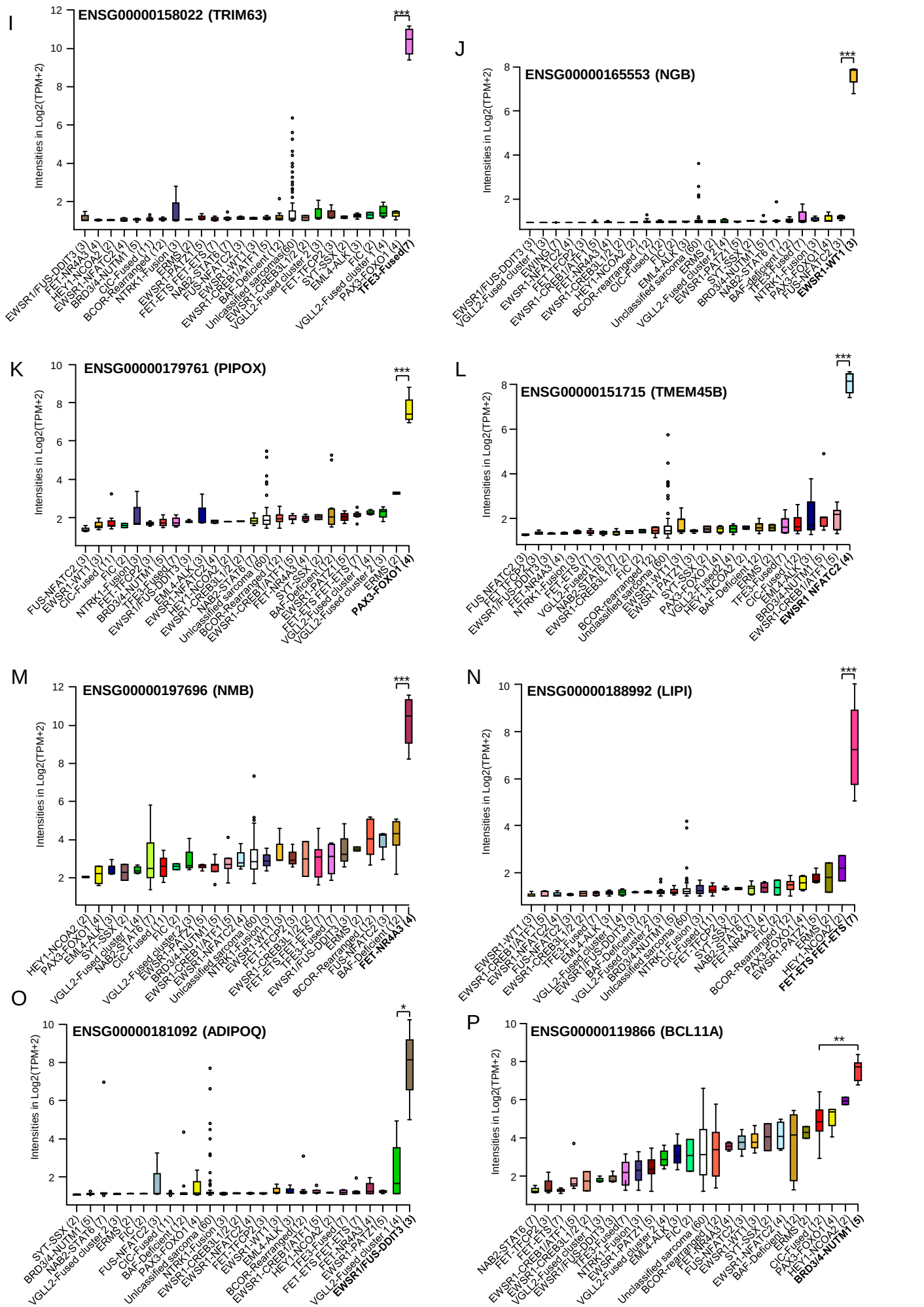
Supplementary Figure S4: Sequence of the fusion points and genes expressed in FET-*TFCP2*-positive tumors

A. Sanger sequencing of *EWSR1-TFCP2*-positive sample (SARC049) and *FUS-TFCP2*-positive samples (RNA020_16_136, similar to RNA009_16_062)

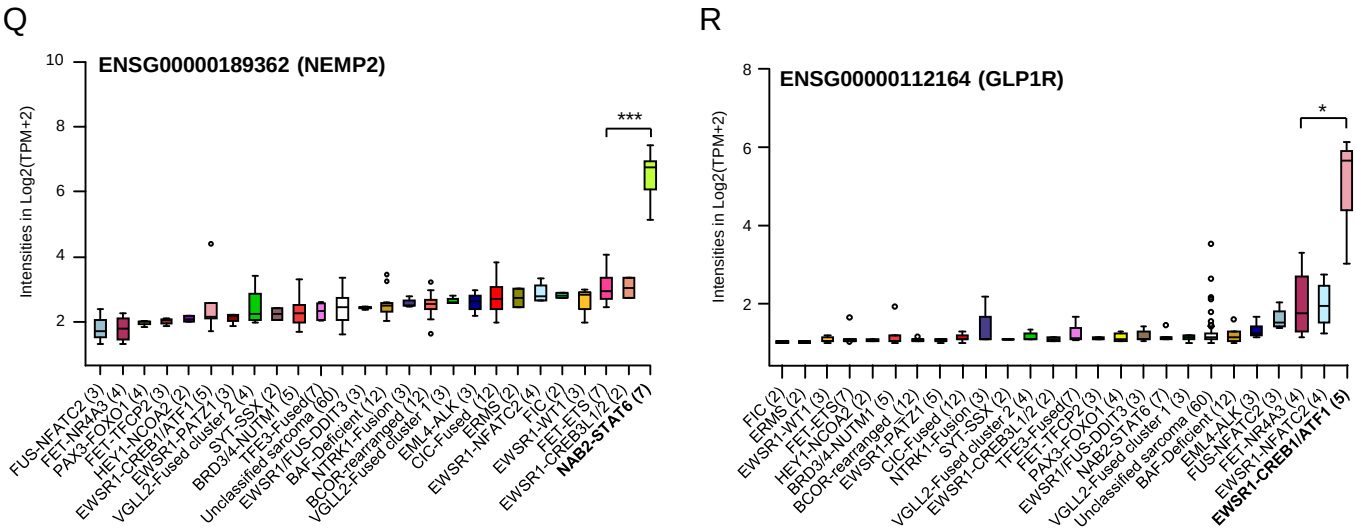
B. Boxplots for *ALK* and *TERT* gene expression level as log2(transcript per million + 2) across all tumor samples demonstrate strong signals in *FET-TFCP2* tumor samples. Number of samples for each boxplot is indicated under brackets. ** : Welsh t-test p-value < 0.01.



Watson et al., Supplementary Figure S5



Watson et al., Supplementary Figure S5



Supplementary Figure S5: Expression of the most specific gene for each tumor entity (with more than 2 samples)

A. Expression level of *LANCL2* gene throughout all samples demonstrating its specificity for *VGLL2*-Fused tumors
B. Expression level of *ETV4* gene throughout all samples demonstrating its specificity for *CIC*-Fused tumors
C. Expression level of *BCOR* gene throughout all samples demonstrating its **lack** of specificity for *BCOR*-rearranged tumors
D. Expression level of *HES7* gene throughout all samples demonstrating its specificity for *BCOR*-rearranged tumors
E. Expression level of *LHX8* gene throughout all samples demonstrating its specificity for *EML4-ALK*-positive tumors
F. Expression level of *CD8B* gene throughout all samples demonstrating its specificity for *FUS-NFATC2*-positive tumors
G. Expression level of *GPR12* gene throughout all samples demonstrating its specificity for *EWSR1-PATZ1*-positive tumors
H. Expression level of *RECQL* gene throughout all samples demonstrating its specificity for *FET-TFCP2*-positive tumors
I. Expression level of *TRIM63* gene throughout all samples demonstrating its specificity for *TFE3*-fused sarcoma
J. Expression level of *NGB* gene throughout all samples demonstrating its specificity for *EWSR1-WT1* desmoplastic small round cell tumors
K. Expression level of *PIPOX* gene throughout all samples demonstrating its specificity for *PAX-FOXO* alveolar rhabdomyosarcoma
L. Expression level of *TMEM45* gene throughout all samples demonstrating its specificity for *EWSR1-NFATC2*-positive tumors
M. Expression level of *NMB* gene throughout all samples demonstrating its specificity for *FET-NR4A3* extraskeletal myxoid chondrosarcoma
N. Expression level of *LIPI* gene throughout all samples demonstrating its specificity for *FET-ETS* Ewing sarcoma
O. Expression level of *ADIPOQ* gene throughout all samples demonstrating its specificity *FET-DDIT3* myxoid liposarcoma
P. Expression level of *BCL11A* gene throughout all samples demonstrating its specificity for *NUTM1-BRD3/4* NUT-midline carcinoma
Q. Expression level of *NEMP2* gene throughout all samples demonstrating its specificity for *NAB2-STAT6* Solitary Fibrous Tumors
R. Expression level of *GLP1R* gene throughout all samples demonstrating its specificity for *FET-CREB1/ATF1* clear cell sarcoma

Number of tumors by groups are indicated under brackets. *, ** and ***: Welsh t-test p-value < 0.05, 0.01 and 10^{-4}, respectively, n.s.: not significant