The Journal of Molecular Diagnostics, Vol. 26, No. 5, May 2024

ELSEVIER

the Journal of Molecular Diagnostics
jmdjournal.org

Machine Learning–Supported Diagnosis of Small Blue Round Cell Sarcomas Using Targeted RNA Sequencing

Lea D. Schlieben,[*†‡] Maria Giulia Carta,[*†‡] Evgeny A. Moskalev,[*†‡] Robert Stöhr,[*†‡] Markus Metzler,[†‡§] Manuel Besendörfer,[¶] Norbert Meidenbauer,[†‡‖] Sabine Semrau,[†‡**] Rolf Janka,[††] Robert Grützmann,[†‡¶‡‡] Stefan Wiemann,[§§] Arndt Hartmann,[*†‡] Abbas Agaimy,[*†‡] Florian Haller,[*†‡] and Fulvia Ferrazzi[*†‡¶¶]

From the Institute of Pathology, the Departments of Pediatrics,[§] Pediatric Surgery,[¶] Internal Medicine 5–Hematology and Oncology,[‖] Radiation Oncology,** Radiology,[††] Surgery,[‡‡] and Nephropathology,[¶¶] Institute of Pathology, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen; the Comprehensive Cancer Center Erlangen-EMN,[†] Erlangen; the Bavarian Cancer Research Center,[‡] Erlangen; and the Division of Molecular Genome Analysis,[§§] German Cancer Research Center, Heidelberg, Germany*

Accepted for publication February 8, 2024.

Address correspondence to Fulvia Ferrazzi, Ph.D., Department of Nephropathology, Institute of Pathology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Krankenhausstr. 8-10, Erlangen 91054, Germany.
E-mail: fulvia.ferrazzi@uk-erlangen.de.

Small blue round cell sarcomas (SBRCSs) are a heterogeneous group of tumors with overlapping morphologic features but markedly varying prognosis. They are characterized by distinct chromosomal alterations, particularly rearrangements leading to gene fusions, whose detection currently represents the most reliable diagnostic marker. Ewing sarcomas are the most common SBRCSs, defined by gene fusions involving *EWSR1* and transcription factors of the ETS family, and the most frequent non–*EWSR1*-rearranged SBRCSs harbor a *CIC* rearrangement. Unfortunately, currently the identification of *CIC*::*DUX4* translocation events, the most common *CIC* rearrangement, is challenging. Here, we present a machine-learning approach to support SBRCS diagnosis that relies on gene expression profiles measured via targeted sequencing. The analyses on a curated cohort of 69 soft-tissue tumors showed markedly distinct expression patterns for SBRCS subgroups. A random forest classifier trained on Ewing sarcoma and *CIC*-rearranged cases predicted probabilities of being *CIC*-rearranged >0.9 for *CIC*-rearranged–like sarcomas and <0.6 for other SBRCSs. Testing on a retrospective cohort of 1335 routine diagnostic cases identified 15 candidate *CIC*-rearranged tumors with a probability >0.75, all of which were supported by expert histopathologic reassessment. Furthermore, the multigene random forest classifier appeared advantageous over using high *ETV4* expression alone, previously proposed as a surrogate to identify *CIC* rearrangement. Taken together, the expression-based classifier can offer valuable support for SBRCS pathologic diagnosis. *(J Mol Diagn 2024, 26: 387–398; https://doi.org/10.1016/j.jmoldx.2024.02.002)*

Small blue round cell sarcomas (SBRCSs) are a heterogeneous group of mesenchymal neoplasms predominantly affecting children and young adults. According to the current World Health Organization classification of soft-tissue tumors (STTs), SBRCSs are composed of relatively undifferentiated cells with small blue nuclei and a round to ovoid cytomorphology in hematoxylin and eosin staining.[1] Overlapping histologic and morphologic characteristics make the routine morphology-based classification of SBRCSs into specific subgroups a diagnostic challenge. These subgroups are characterized by different progression and sometimes treatment strategies, necessitating tumor stratification. Advances in both cytogenetics and molecular genetics have led to the detection of tumor-specific

Supported by Deutsche Forschungsgemeinschaft (German Research Foundation) grants TRR 305 [projects Z01 (F.F.) and Z02 (A.H.)] and TRR 374 (project INF) (F.F.); and confocal imaging was funded by Deutsche Forschungsgemeinschaft grant INST 410/91-1 FUGG.

L.D.S. and M.G.C. contributed equally to this work.

Current address of L.D.S., School of Medicine, Institute of Human Genetics, Klinikum Rechts der Isar, Technical University of Munich, Munich; and Institute of Neurogenomics, Computational Health Center, Helmholtz Zentrum München, Neuherberg, Germany.

https://doi.org/10.1016/j.jmoldx.2024.02.002

alterations in subsets of SBRCSs, which are characterized by particular recurrent chromosomal rearrangements and associated specific gene fusions.[2] The tumorigenic activity of gene fusions is exerted either by the formation of encoded chimeric fusion proteins or by the aberrant expression of a normal gene product through hijacking of regulatory elements of the translocation partner.[3,4] Molecular pathologic identification of characteristic gene fusion events has gained increasing importance in sarcoma diagnostics,[5] and RNA sequencing (RNA-seq)–based methods now are used extensively in the molecular diagnostic routine workup of STTs.

Ewing sarcomas (ESs) are the most common type of SBRCSs and harbor gene fusions involving *EWSR1* and transcription factors of the ETS family, predominantly *FLI1* (85%) and *ERG* (10%). In contrast, atypical ES or Ewing-like sarcomas are a group of undifferentiated round cell sarcomas that resemble classic ES morphologically to a variable extent, but lack the molecular hallmark for ES, the *EWSR1*::ETS fusion. Thus far, four main types of Ewing-like sarcomas have been described: *CIC*-rearranged sarcomas,[6] *BCOR*-rearranged sarcomas,[7] sarcomas with a fusion between *EWSR1* and a gene not belonging to the ETS family members, and unclassified round cell sarcomas.[8] SBRCSs with *CIC* rearrangement represent the most prevalent group within non–*EWSR1*-rearranged SBRCSs. The frequency of *CIC*-rearranged sarcomas has been estimated as 3% to 5% of SBRCSs in both adult and pediatric populations.[9,10] From a clinical perspective, these tumors present most commonly as deep STTs in the extremities.[9] Patients generally are older than those with classic ES, with a peak incidence in the fourth decade of life, in contrast to ES, in which an onset in the second to third decade of life is common.[9,11,12] *CIC*-rearranged sarcomas have a more aggressive behavior when compared with classic ES, although the available data are limited to retrospective case reports and case series.[9,11,13] The most common *CIC* rearrangement events are *CIC*::*DUX4* gene fusions. However, given the complex structure of the *DUX* gene locus, which harbors multiple gene copies with highly similar sequences, both fluorescence *in situ* hybridization (FISH)- and next-generation sequencing–based methods have difficulties in the correct identification of *CIC*::*DUX4* translocation events.[14,15] In particular, an inability of different RNA-seq approaches to identify the *CIC*::*DUX4* gene fusion has been observed.[15–17] Yet, gene fusions involving transcription factors are expected to have a significant effect on gene expression patterns. Indeed, the *CIC*::*DUX4* fusion protein has been shown to up-regulate the expression of the polyomavirus enhancer activator 3 (PEA3) subfamily of ETS transcription factors (eg, *ETV1*, *ETV4*, and *ETV5*).[6] Thus, the up-regulated expression of ETV1, ETV4, and WT1 have been introduced as diagnostic markers for *CIC*-rearranged sarcomas by immunohistochemistry,[10,18,19] RNA expression analysis,[10,15] and RNA *in situ* hybridization.[20] In particular, previous studies have shown that immunohistochemical stainings for ETV4 and WT1 are helpful diagnostic markers in the workup of SBRCSs, yet not sufficient to identify all *CIC*-rearranged cases in a specific way.[18,19] On the other hand, the molecular pathologic detection of *CIC* rearrangements, which still represents the only reliable diagnostic marker, remains challenging.

In this study it was hypothesized that the use of gene expression profiles extracted from a widely used routine diagnostic RNA hybrid capture–based next-generation sequencing panel might offer effective support in the identification of *CIC*-rearranged sarcomas and be advantageous with respect to relying only on the expression of the PEA3 subfamily of ETS transcription factors. Machine learning methods applied to gene expression profiles have allowed the successful development of diagnostic and prognostic tools for different cancer entities. For example, different breast cancer expression-based classifiers are used in clinical practice,[21,22] and several expression-based signatures have been proposed for lymphoid neoplasms, such as for peripheral T-cell lymphoma[23] or B-cell lymphomas.[24,25] Here, an expression-based classifier to predict the likelihood of any given SBRCS of harboring a *CIC* rearrangement was developed and tested on a large retrospective cohort of routine molecular pathology diagnostic cases.

Key Points

- Small blue round cell sarcomas (SBRCSs) are a heterogeneous group of tumors with overlapping morphologic features but divergent genotypes and markedly varying prognosis.
- The detection of recurrent gene fusions represents the most reliable diagnostic marker for SBRCSs. However, it can be challenging with currently available methods, as in the case of the *CIC*::*DUX4* translocation events, the most common *CIC* rearrangement in sarcomas.
- Bioinformatics analyses of the gene expression profiles measured via targeted sequencing show that SBRCS subgroups are characterized by markedly distinct expression patterns.
- The machine-learning classifier based on gene expression profiles that is presented here is a valuable tool to support the pathologic diagnosis of SBRCSs.

Materials and Methods

Patient Samples

A total number of 1404 tumors analyzed at the Institute of Pathology (Erlangen, Germany) were used for this study, including mainly STTs, salivary gland carcinomas, and kidney tumors. A cohort of 69 STTs with a confirmed diagnosis by an expert STT pathologist (A.A.) was used for the development of the machine learning classifier. These 69 STT samples hereafter are referred to as the *curated cohort*, whereas the remaining 1335 cases were used as the *test*

cohort. The curated cohort comprised 31 tumors within the morphologic spectrum of SBRCSs: 13 classic ESs with *EWSR1* gene fusions, 3 *BCOR*-rearranged sarcoma samples with *BCOR* gene fusions confirmed by immunostaining, and 15 SBRCSs with morphologic and immunophenotypic features of *CIC*-rearranged sarcomas that were negative for *EWSR1* gene rearrangements by FISH. Two of the *BCOR*-rearranged samples were two distinct sarcoma samples from the same patient, 2 years apart. The additional 38 STTs included 12 synovial sarcomas, 10 solitary fibrous tumors, 6 myxoid liposarcomas, and 10 cases of dermatofibrosarcoma protuberans. All samples were used in accordance with the ethical guidelines for the use of retrospective tissue samples provided by the local ethics committee of the Friedrich-Alexander-Universität Erlangen-Nuremberg (Ethics Committee statements 24.01.2005 and 18.01.2012).

Targeted RNA Sequencing

Tumor RNA was isolated from microdissected formalin-fixed, paraffin-embedded tissues (approximately five sections, 6- to 8-μm thick) using the RNeasy formalin-fixed, paraffin-embedded RNA kit (Qiagen, Hilden, Germany). The quality of RNA samples was assessed using RNA ScreenTape on a TapeStation 4200 (Agilent, Santa Clara, CA). Only samples with >30% of RNA fragments exceeding 200 nucleotides were included for library preparation. Libraries for sequencing were prepared using the Illumina TruSight RNA Fusion Panel (Illumina, Inc., San Diego, CA), which enriches for 507 genes frequently involved in gene fusions[26] and covers an approximately 1.44-megabase region. A total of 500 ng RNA was used as input for library preparation, prepared according to the manufacturer's instructions. Amplified libraries were sequenced either on a MiSeq instrument or on a NextSeq 550 instrument (Illumina) with 151-bp paired-end reads with generally >3 million reads per sample. Illumina's sequenced libraries were preprocessed via the Illumina RNA-seq Alignment BaseSpace App. Within the Illumina App, sequencing reads were aligned to the human reference genome (University of California Santa Cruz, hg19) relying on the STAR aligner,[27] and the Manta gene fusion caller[28] was used for gene fusion detection.

Additional libraries for a subset of samples were prepared with the Archer FusionPlex Sarcoma Kit (Archer, Boulder, CO) according to the manufacturer's instructions. Library was prepared using 250 ng total RNA. Libraries were sequenced with 151-bp paired-end reads on a MiSeq instrument, leading to a minimum of 1.5 million reads per sample. Sequencing data were analyzed using Archer Analysis Suite v6.2.7.

FISH Analysis

Freshly cut tissue sections were used for the detection of *CIC* gene locus rearrangements using a commercial probe according to the manufacturer's instructions (Zytolight SPEC CIC Dual Color Break Apart Probe, Z-2285-50; Zytovision GmbH, Bremerhaven, Germany). A *CIC*-rearranged sarcoma (CIC_R_11) from the curated cohort was used as a positive control, whereas a *BCOR*-rearranged sarcoma (BCOR_02) from the curated cohort was used as a negative control. Fifty nuclei were counted per case. A signal was considered to be rearranged if the distance between the green and orange signals was at least the size of a single signal. A case was considered rearranged if >20% of the signals were rearranged. Image acquisition and high-resolution microscopy were performed using a LSM800 confocal laser scanning microscope equipped with an Airyscan detector (Zeiss, Oberkochen, Germany) and the ZEN 2.3 (blue edition) software with Airyscan image processing.

Expression Data Analysis

Illumina targeted RNA-seq data were analyzed within the R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria, *https://www.R-project.org*)/Bioconductor version 3.13 environment,[29] starting from the raw counts obtained via the Illumina RNA-seq Alignment BaseSpace App. First, only the curated cohort samples were considered. Relying on the DESeq2 package version 1.34.0,[30] size factor normalization was performed for the raw counts and variance-stabilized transformed (VST) counts were obtained. A variance-stabilizing transformation is a data transformation such that the variance of transformed values is approximately independent of the mean.[31] The use of VST counts facilitates visualization and analyses of gene expression such as clustering or classification, which generally work better with variables with similar distributions. In this study, VST counts were used as the expression measure. Principal component analysis was performed on the VST counts of the 100 genes with the highest variance across all samples and the plot of the first two principal components was generated with ggplot2 v.3.3.6.[32] Expression heatmaps were generated with the pheatmap package 1.0.12 using mean-centered VST counts (ie, VST counts from which, for each gene, the mean VST count across samples was subtracted). Dot plots of the VST counts of single genes were generated using ggplot2. Differential expression analysis of SBRCS samples was performed, relying on the DESeq2 package, starting from the raw count data of only the SBRCS samples. Genes with a Benjamini-Hochberg–adjusted P value smaller than 0.01 were considered differentially expressed.

Random Forest Classifier

A random forest (RF) classifier[33] was trained using the randomForest version 4.7-1.1 package.[34] To build the training data set, only the *CIC*-rearranged and ES samples

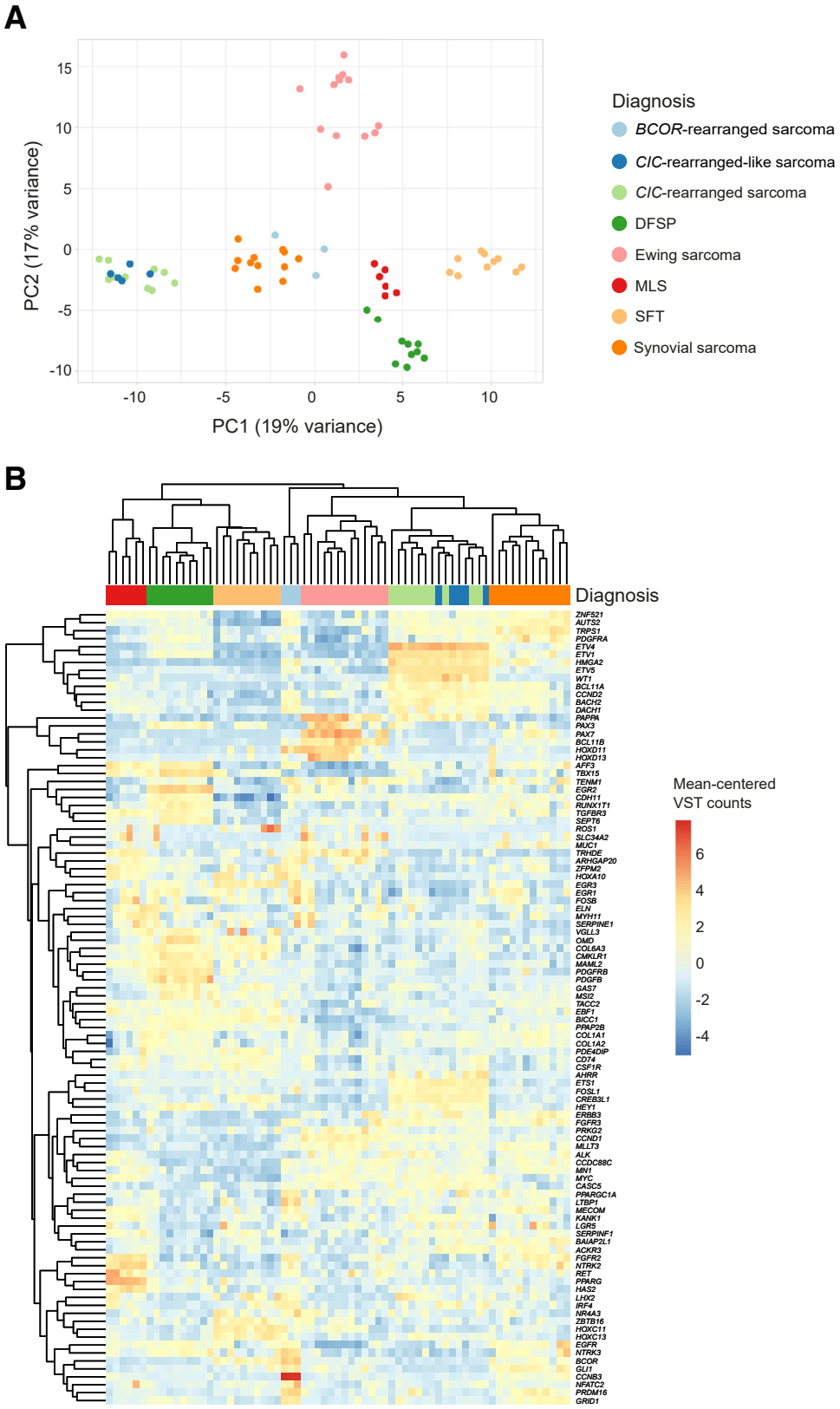

Figure 1 Unsupervised expression data analysis. **A:** Principal component (PC) analysis plot based on the expression [variance-stabilized transformed (VST) counts] of the 100 genes with the highest variability across the 69 soft-tissue tumors of the curated cohort. **B:** Expression heatmap of the same top 100 variable genes. Gene expression values in the heatmap are mean-centered VST counts and are color-coded so that red represents up-regulation and blue represents down-regulation. DFSP, dermatofibrosarcoma protuberans; MLS, myxoid liposarcoma; SFT, solitary fibrous tumor.

contained in the curated cohort were considered. The associated raw gene counts were transformed into VST counts and the top 100 variable genes (ie, the 100 genes with the highest variance based on their VST counts) were identified. The VST counts of these 100 genes were used as training data for the RF classifier, with the number of trees set to 100 and the number of genes randomly sampled at each tree split set to 10. Gene importance was assessed by the mean decrease in the Gini index[35] and used to select a subset of top important genes to be explored further. The Gini index is a measure of the impurity of a node in a tree, which equals 0 when all samples associated to a node belong to one class, and reaches the maximum (corresponding to 0.5 in a binary setting) when the classes are represented equally. During training of a tree, the ability of the expression value of a gene to separate samples belonging to the two classes is assessed via the decrease in the Gini index. Thus, in a RF, the overall mean decrease in the Gini index of a gene offers a measure of the gene's discriminative power. Given the small size of the training set and the presence of multiple genes likely to be able to discriminate well between the two classes, to assess the stability of the top important genes, another nine RFs were trained, each time using a different random seed to ensure different random initializations as well as reproducibility. Then, the median, as well as the minimum and maximum,

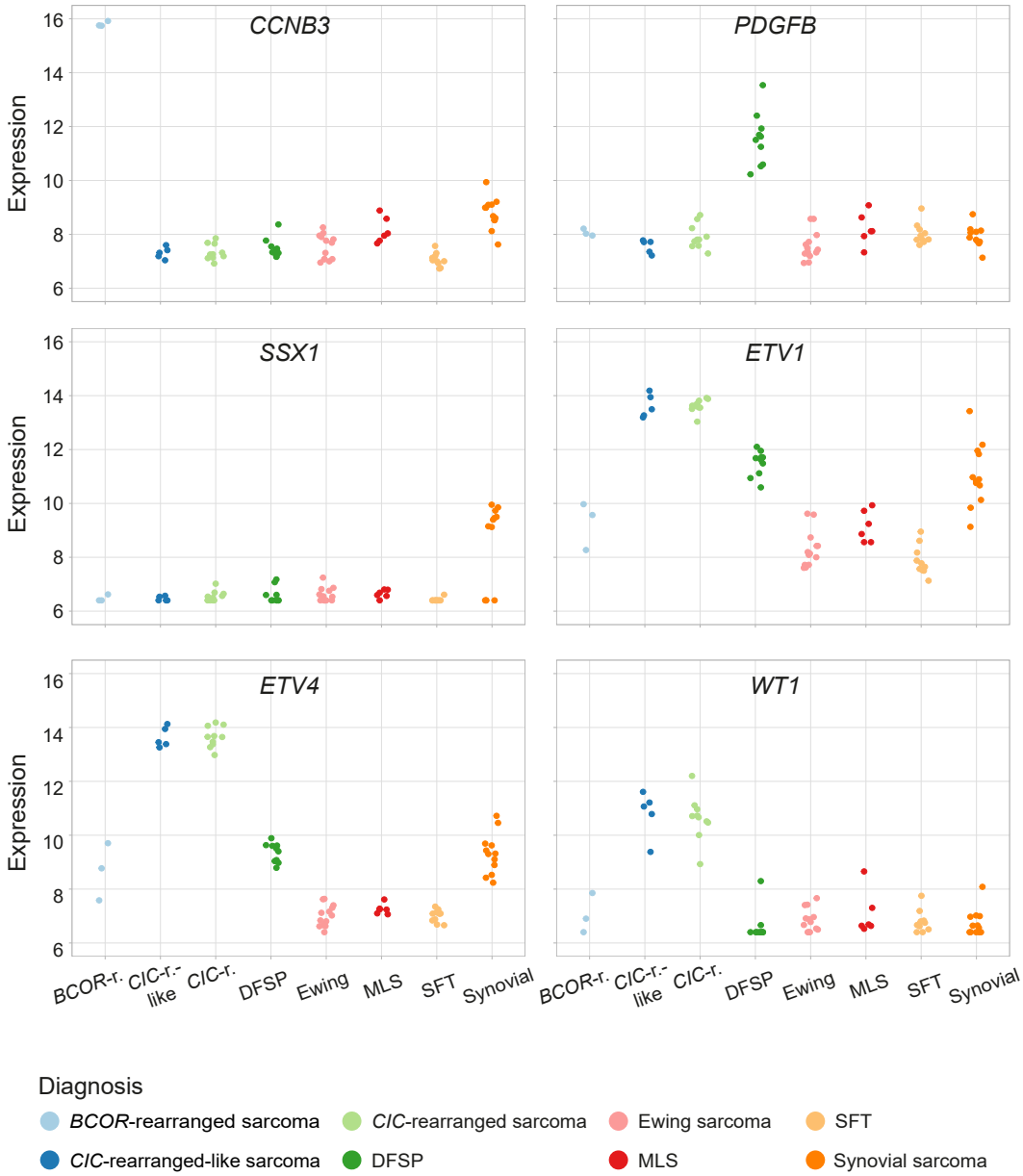

Figure 2 Expression profiles of selected genes. Expression (variance-stabilized transformed counts) dot plots of selected genes showing up-regulation in distinct soft-tissue tumor entities: *CCNB3*, *PDGFB*, *SSX1*, *ETV1*, *ETV4*, and *WT1*. *BCOR*-r., *BCOR*-rearranged sarcoma; *CIC*-r., *CIC*-rearranged sarcoma; *CIC*-r.-like, *CIC*-rearranged-like sarcoma; DFSP, dermatofibrosarcoma protuberans; MLS, myxoid liposarcoma; SFT, solitary fibrous tumor.

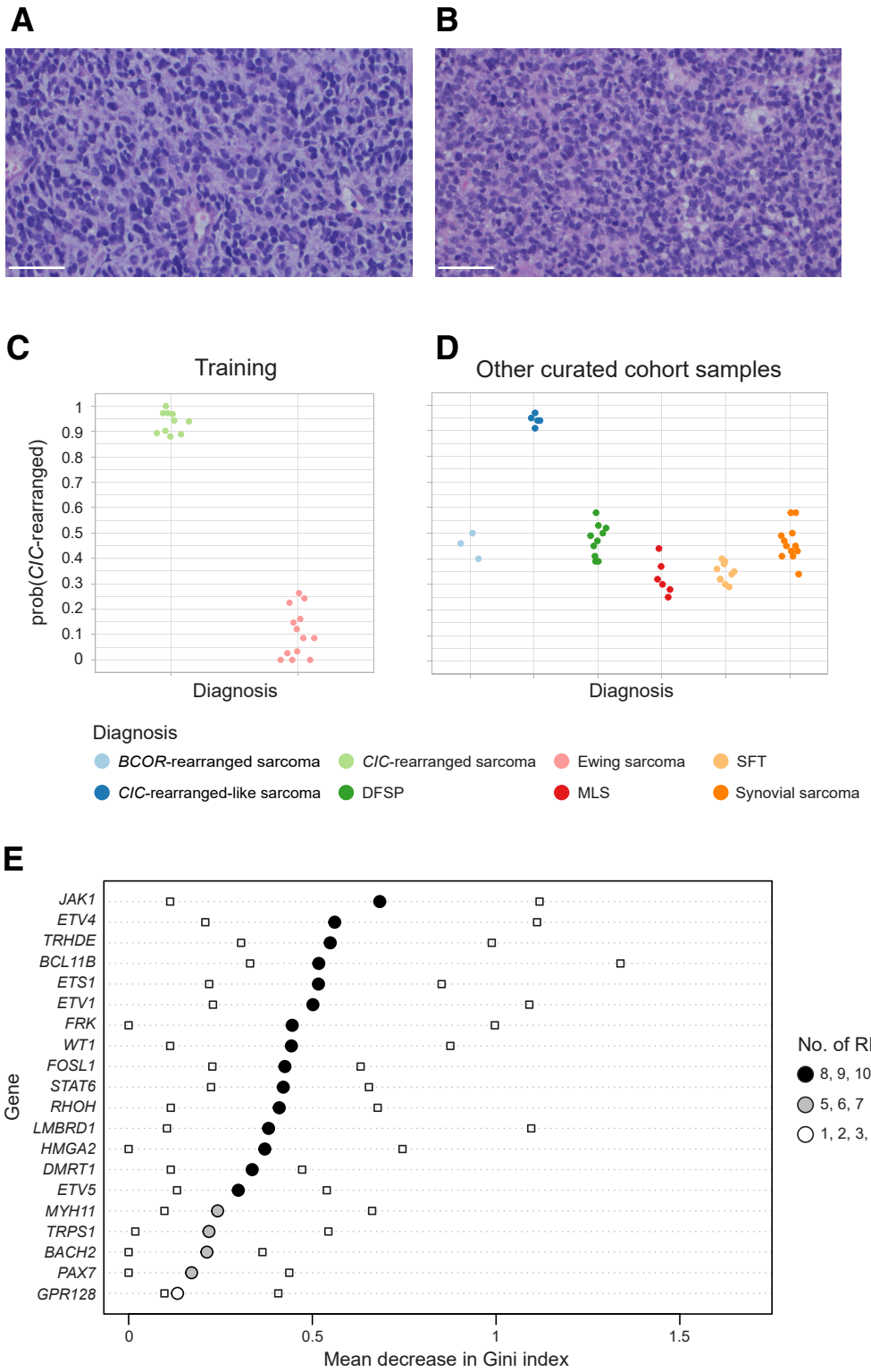

Figure 3 Expression-based classifier to predict the probability of being *CIC*-rearranged. Histopathologic hematoxylin and eosin stain of an exemplary *CIC*-rearranged (**A**) and Ewing sarcoma (**B**) sample. Dot plots of out-of-bag probabilities of being *CIC*-rearranged in the training set (**C**) and predicted probabilities of being *CIC*-rearranged for curated cohort samples not included in the training (**D**). **E:** Variable importance plot of the top 20 genes, sorted according to their median importance (assessed by mean decrease in the Gini index) across the 10 random forests (RFs). For each gene, the **dot** shows the median variable importance across the 10 RFs, and the **squares** show the minimum and maximum variable importance. Dot color is associated with the number of RFs in which the gene was included in the top 20 predictors. Scale bars = 100 μm. DFSP, dermatofibrosarcoma protuberans; MLS, myxoid liposarcoma; SFT, solitary fibrous tumor.

Table 1 Candidate *CIC*-Rearranged Cases Identified by the Classifier

Sample identifier	Predicted probability of being *CIC*-rearranged	Identified candidate fusion	Age, years/sex/ localization	Initial diagnosis in pathology report
C1	0.99	None	64/F/lower jaw	Unclassified highly malignant epithelioid neoplasm
C2	0.98	None	38/M/thoracic wall	Ewing-like sarcoma, most likely *CIC*-rearranged
C3	0.97	None	54/M/tongue	Undifferentiated round cell sarcoma, NOS
C4	0.94	None	19/F/upper arm	Unclassified highly malignant round and spindled cell sarcoma
C5	0.94	None	32/F/neck	Ewing-like sarcoma, most likely *CIC*-rearranged
C6	0.93	None	25/F/thoracic wall	Ewing-like sarcoma, most likely *CIC*-rearranged
C7	0.92	None	48/F/lower arm	Ewing-like sarcoma, most likely *CIC*-rearranged
C8	0.89	None	21/F/foot	Ewing-like sarcoma, most likely *CIC*-rearranged
C9	0.87	None	42/M/arm	Unclassified epithelioid soft tissue neoplasm of unknown histogenesis
C10	0.86	*CIC*::*NUTM1*	44/F/neck	*CIC*::*NUTM1* rearranged neoplasm
C11	0.85	None	34/F/knee	Ewing-like sarcoma, most likely *CIC*-rearranged
C12	0.84	None	17/F/retroperitoneum	Unclassified small round cell sarcoma
C13	0.82	None	25/F/leg	Ewing-like sarcoma, most likely *CIC*-rearranged
C14	0.82	None	65/F/arm	Ewing-like sarcoma, most likely *CIC*-rearranged
C15	0.8	None	14/F/axilla	Unclassified highly malignant epithelioid neoplasm

The table columns show sample identifier, predicted probability of being *CIC*-rearranged, identified gene fusion during primary routine pathology workup, age, sex, anatomic localization, and initial diagnosis in the pathology report.

F, female; M, male.

variable importance (mean decrease in the Gini index) of the top identified genes was calculated across the 10 trained RFs. In addition, for each of the top genes, the number of RFs (minimum number of RFs = 1, maximum number of RFs = 10) in which the gene was found within the top important genes corresponding to each RF was assessed. Internal assessment of the RF classifier performance occurred via out-of-bag evaluation.[33] Indeed, because each decision tree in a RF typically is trained on a bootstrap sample of the training set, out-of-bag evaluation utilizes, for each sample, the predicted classes by the decision trees that did not use the sample during training. For each sample, a classification (*CIC*-rearranged or Ewing) is provided by each tree in the RF. The predicted probability of being *CIC*-rearranged [hereafter indicated as prob(*CIC*-rearranged)] is taken equal to the proportion of trees in the RF classifying the case as *CIC* rearranged.

Afterward, the RF model was used to calculate the predicted prob(*CIC*-rearranged) for all samples of the curated cohort not included in the training. Finally, the RF model was used to obtain predictions on the 1335 test cases. In both cases, the transformed counts to be used for the RF prediction were obtained for each sample as follows: first, raw counts were normalized using the size factor estimated using the gene-wise geometric means calculated on the training data as pseudo-reference sample; then, VST counts were obtained using the dispersion function pre-estimated on the training data. For each case of the test cohort, prob(*CIC*-rearranged) was assessed, and a probability cut-off value chosen on the basis of the distribution of all probability values was used to select candidate *CIC*-rearranged cases. The stability of the predicted probabilities for the candidate cases was explored by assessing the range of their predicted probabilities across the 10 trained RFs, as well as by assessing whether the median probability was higher than the selected cut-off value.

Results

Detection of Gene Fusions via Targeted RNA Sequencing

Gene fusions were analyzed in a curated cohort of 69 STT samples, including 31 SBRCSs and 38 other types of STTs (Supplemental Table S1), via targeted RNA-seq (Illumina). In all 38 STTs, RNA-seq revealed the expected gene fusion given their histopathologic diagnosis. Specifically, 8 and 4 of the 12 synovial sarcomas, respectively, were identified to harbor a *SS18*::*SSX1* and *SS18*::*SSX2* gene fusion, all 10 solitary fibrous tumors had a *NAB2*::*STAT6* gene fusion, all 6 myxoid liposarcomas had a *FUS*::*DDIT3* gene fusion, and all 10 dermatofibrosarcoma protuberans showed a *COL1A1*::*PDGFB* gene fusion.

The 31 SBRCS samples included 13 classic ESs, 3 *BCOR*-rearranged sarcoma samples with positive BCOR immunostaining, and 15 SBRCSs with morphologic features of *CIC*-rearranged sarcomas. Targeted RNA-seq allowed the identification of gene fusions in all 13 ESs, with 10 samples harboring a *EWSR1*::*FLI1* gene fusion, and the other 3 samples harboring a *EWSR1*::*ERG*, *EWSR1*::*FEV*, and *FUS*::*ERG* gene fusion, respectively. All three *BCOR*-rearranged sarcoma samples harbored a *BCOR*::*CCNB3* gene fusion. However, no fusion was detected for the 15 SBRCSs with morphologic features of

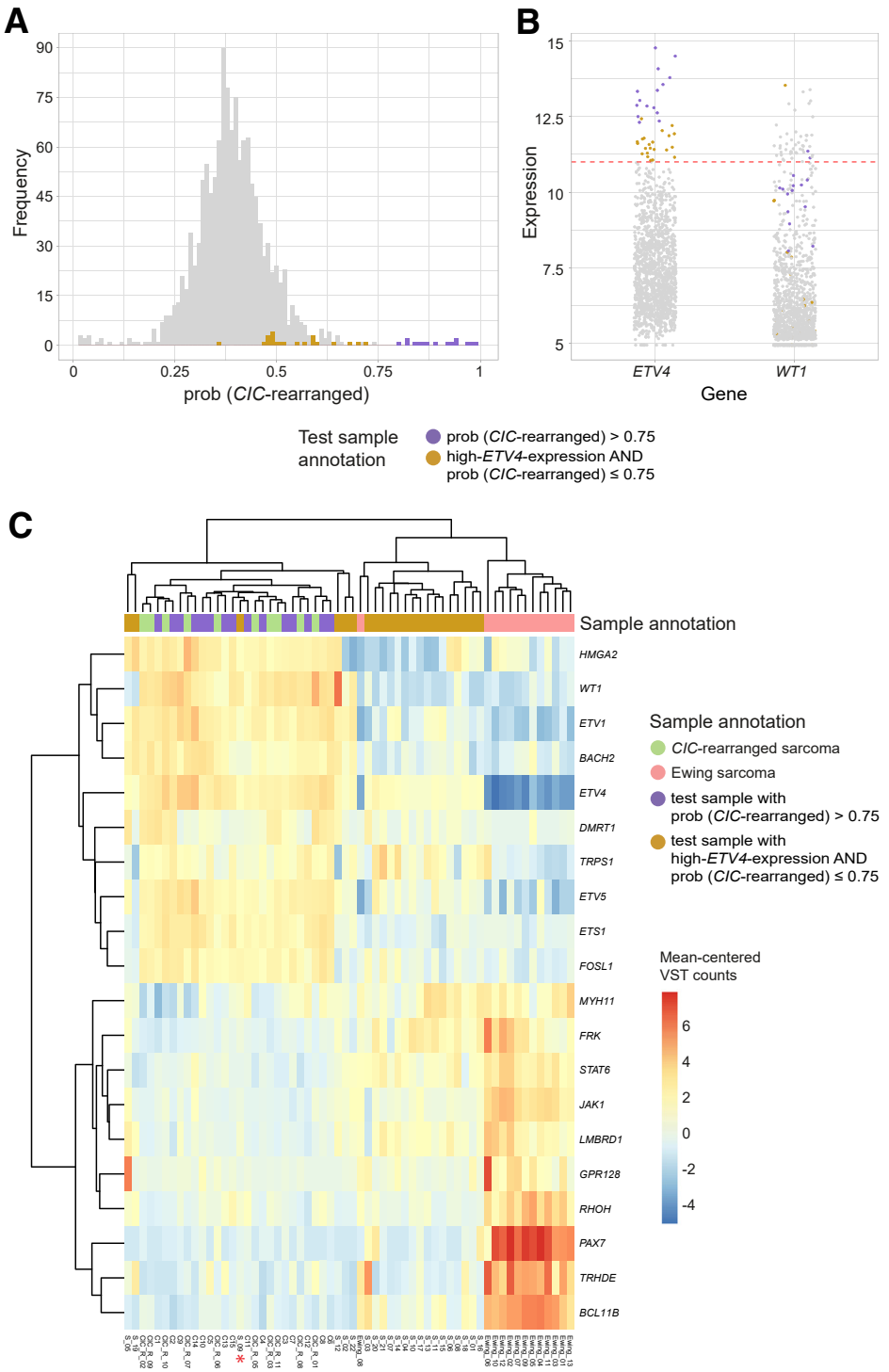

Figure 4 Test of the expression-based classifier on a cohort of 1335 routine diagnostics cases. **A:** Histogram of the predicted probabilities of being *CIC*-rearranged in the test cohort samples ($N = 1335$). **B:** Expression values [variance-stabilized transformed (VST) counts] of *ETV4* and *WT1* in the test cohort samples. The **red dotted line** indicates the threshold used to identify high-*ETV4*-expression samples (*ETV4* > 11 VST counts, corresponding to the 97th percentile). **A** and **B:** Purple color indicates the candidate *CIC*-rearranged samples predicted by the random forest classifier, that is, the test samples with probability of being *CIC*-rearranged [prob(*CIC*-rearranged)] > 0.75; dark yellow indicates test samples with high-*ETV4*-expression and prob(*CIC*-rearranged) ≤ 0.75. **C:** Expression heatmap based on the top 20 predictors identified by the random forest. Green indicates the *CIC*-rearranged samples of the training set; pink indicates the Ewing sarcoma samples of the training set; purple indicates the candidate *CIC*-rearranged samples; and dark yellow indicates the high-*ETV4* expression samples with prob(*CIC*-rearranged) ≤0.75. The **red asterisk** indicates the false-negative *CIC*-rearranged case [ie, the case (S_09) compatible with *CIC*-rearranged sarcoma diagnosis that was predicted with a probability of 0.72]. Gene expression values in the heatmap are mean-centered VST counts and are color-coded so that red represents up-regulation and blue represents down-regulation.

CIC-rearranged sarcomas. These SBRCSs additionally were analyzed by a second RNA-seq panel (Archer). The use of this panel allowed identification of a *CIC*::*DUX4* gene fusion in 10 samples, which thus were classified as *CIC*-rearranged sarcomas. In addition, a *CIC* rearrangement was confirmed by FISH in one case (CIC_R_11) (Supplemental Table S1) with available paraffin material. No fusion could be identified for the five remaining SBRCSs with the second RNA-seq panel. Because these samples had morphologic features of *CIC* rearrangement, they are hereafter referred to as *CIC*-rearranged-like sarcomas. Furthermore, a *CIC* rearrangement was found by FISH in two of these *CIC*-rearranged cases (CIC_R_like_01 and CIC_R_like_05) (Supplemental Table S1) with available paraffin material, while there was no sufficient paraffin material for additional FISH analysis in the three other cases.

STT Subtypes Show Distinct Gene Expression Profiles

To explore whether the expression profiles measured with the Illumina TruSight targeted RNA-seq panel could be used to support the identification of *CIC*-rearranged sarcomas, unsupervised analyses of the SBRCS expression profiles were performed. A plot of the first two components obtained by principal component analysis of the expression of the top 100 variable genes revealed that tumor entities clustered according to diagnosis (Figure 1A). Notably, the five SBRCSs with morphologic features of *CIC* rearrangement clustered together with the 10 *CIC*-rearranged tumor samples. In addition, hierarchical clustering analysis of the expression profiles of the same 100 genes confirmed that tumor samples with the same diagnosis clustered together (Figure 1B). Furthermore, *CIC*-rearranged sarcomas and *CIC*-rearranged-like sarcomas displayed remarkably similar expression profiles characterized by up-regulation of the expression of PEA3 transcription factor genes (*ETV1*, *ETV4*, *ETV5*) and *WT1*. At the single-gene level, several STT entities were characterized by up-regulated genes involved in the respective gene fusion (eg, there was a high expression of *CCNB3* in *BCOR*-rearranged sarcomas, of *PDGFB* in dermatofibrosarcoma protuberans, and of *SSX1* in synovial sarcomas) (Figure 2). Similarly, *CIC*-rearranged sarcomas and *CIC*-rearranged-like sarcomas showed a distinct and similar up-regulated expression of the known *CIC*::*DUX4* target genes *ETV1*, *ETV4*, and *WT1* (Figure 2). Yet, up-regulation was not always seen in all cases of one STT entity or was not entirely specific to one entity. Thus, the expression pattern of a larger number of genes might be more predictive.

Differential gene expression analysis between the different SBRCS groups confirmed that *CIC*-rearranged and *CIC*-rearranged-like sarcomas showed similar expression profiles that were markedly different from the other SBRCSs. Indeed, 68 genes were differentially expressed between *CIC*-rearranged and *BCOR*-rearranged sarcomas (adjusted $P < 0.01$) (Supplemental Table S2), and 137 between *CIC*-rearranged and ESs (adjusted $P < 0.01$) (Supplemental Table S3). In both comparisons, *ETV1*, *ETV4*, and *WT1* were among the differentially expressed genes, showing up-regulation in *CIC*-rearranged sarcomas. Instead, only one gene (*LINC00982*) was differentially expressed (adjusted $P < 0.01$) (Supplemental Table S4) when comparing the *CIC*-rearranged sarcomas with the *CIC*-rearranged-like sarcoma samples.

Machine Learning—Based Prediction of *CIC*-Rearranged Cases

To explore the possibility of using gene expression profiles to support the diagnosis of SBRCSs, a gene expression-based classifier was trained. Because in the pathologic diagnostics the most critical differential diagnosis is between *CIC*-rearranged sarcoma (Figure 3A) and ES (Figure 3B), and these are the two most frequent SBRCSs, the classifier focused on these two SBRCSs. An RF classifier was trained on the expression profiles of the top 100 variable genes identified considering the *CIC*-rearranged and ES samples of the curated cohort. The predicted out-of-bag probabilities of being *CIC*-rearranged were greater than 0.85 for all *CIC*-rearranged samples and less than 0.3 for all Ewing samples (Figure 3C). Thus, the classifier could effectively support the differential diagnosis between *CIC*-rearranged and ESs. Results showed that only 39 genes had an importance (assessed via the mean decrease in the Gini index) greater than zero. Of these 39 genes used by the RF classifier, the top 20 important ones (ie, those with the highest mean decrease in the Gini index) included several of the genes identified as differentially expressed when comparing *CIC*-rearranged versus ESs, including *ETV4*, *ETV1*, and *WT1* (Figure 3E). The stability of these 20 top predictors was assessed by considering their importance measure obtained in nine additional RFs. Despite fluctuations in the variable importance measure, 15 of 20 genes were among the top 20 predictors in at least eight RFs, 4 of 20 genes were among the top 20 predictors in five to seven RFs, and only 1 gene was not found in the top 20 predictors in other RFs. To explore the possibility of using the classifier to identify candidate *CIC*-rearranged cases in a more varied SBRCS cohort, the RF was used to predict the probability of being *CIC*-rearranged for all other samples in the curated cohort. The predicted probabilities were greater than 0.9 for all five *CIC*-rearranged-like sarcoma samples and less than 0.6 for all other samples (Figure 3D). This validation suggested that the RF classifier effectively supports the identification of candidate *CIC*-rearranged sarcomas.

To test the classifier, a large collection of 1335 tumor samples was used, mainly including STTs, salivary gland carcinomas, and kidney tumors. These samples were analyzed during routine diagnostics at the Institute of Pathology (Erlangen, Germany) using the same Illumina TruSight Panel. On the basis of visual inspection of the distribution of predicted probabilities of being *CIC*-rearranged, a probability threshold of 0.75 was adopted to

identify 15 samples as candidate *CIC*-rearranged cases (Figure 4A and Table 1). The predicted probabilities appeared stable because the 15 samples had a median predicted probability of being *CIC*-rearranged higher than 0.75 over the 10 trained RFs (Supplemental Figure S1A), and a predicted probability higher than 0.75 in most of the RFs (Supplemental Figure S1B). For 14 of these cases no candidate gene fusion was identified during routine diagnostics, whereas 1 case harbored a *CIC*::*NUTM1* gene fusion. Careful review of these cases by STT expert pathologists (A.A. and F.H.) showed that all cases were compatible with a diagnosis of *CIC*-rearranged sarcoma. In addition, in eight of the cases a diagnosis of most likely *CIC*-rearranged sarcoma already had been made based on histomorphology and immunohistochemistry in the first diagnostic setting without RNA fusion analysis (Table 1). Furthermore, 1 of these 15 cases (C4) (Table 1) was available for FISH analysis and showed a *CIC* rearrangement (Supplemental Figure S2).

Unsupervised analyses based on the expression of the 20 top predictors identified via the RF classifier showed that the 15 candidate cases clustered together with the *CIC*-rearranged samples of the training set (Figure 4C). *ETV4* expression was very high in all 15 candidate cases, validating the approach. However, these cases were not those with the highest *WT1* expression (Figure 4B). This suggested that *ETV4* expression was a key element in the classifier predictions, whereas the use of *WT1* expression alone might have missed these cases. To explore further whether the multigene RF classifier is superior to using *ETV4* expression alone, all 37 test samples with high-*ETV4*-expression (ie, *ETV4* expression higher than the 97th percentile of *ETV4* gene expression across all test samples) were investigated. This *ETV4* expression threshold was chosen on the basis of the curated cohort as one that could well separate *CIC*-rearranged and *CIC*-rearranged-like cases from the others (Figure 2). These 37 high-*ETV4*-expression cases included the 15 candidate cases identified by the RF and 22 additional cases. Of these 22 cases, 1 case clustered with the *CIC*-rearranged samples of the training set and 21 cases had discordant expression profiles, which differed from the *CIC*-rearranged cases (Figure 4C). The case clustering with the *CIC*-rearranged samples (S_09) (Table 2) was an additional candidate *CIC*-rearranged case on the basis of its expression profile. This high-*ETV4*-expression case had a predicted probability of 0.72 and indeed was compatible with a *CIC*-rearranged sarcoma on the basis of the STT expert pathologists' review. Because this case was an external case that had been sent for consultation, no paraffin material was available for FISH analysis. However, all of the other 21 tumors with a probability less than 0.75 but high *ETV4* expression were not compatible with *CIC*-rearranged sarcomas, and included carcinomas of the salivary glands, sinonasal undifferentiated carcinoma, and epithelioid sarcoma of proximal type. A reassessment of all 1335 cases on the basis of their clinicopathologic data and morphology suggested that the 15 candidate *CIC*-rearranged samples identified by the classifier with prob(*CIC*-rearranged) > 0.75 and the additional sample S_09 were the only 16 cases with a diagnosis compatible with *CIC*-rearranged sarcoma. Taken together, these results show that the RF classifier performs better than *ETV4* expression alone.

Discussion

Here, the usefulness of gene expression profiles measured with the Illumina TruSight RNA Fusion panel to support the diagnosis of STTs was evaluated. In the curated cohort of 69 STTs, unsupervised analysis of expression data revealed that tumor entities characterized by distinct gene fusions clustered together according to their gene expression profiles. Notably, the 10 *CIC*-rearranged sarcomas and the five SBRCSs with a *CIC*-rearranged-like histomorphology showed very similar expression patterns. In addition, all 15 samples showed an up-regulation of PEA3 transcription factor genes (*ETV1*, *ETV4*, *ETV5*) and *WT1*, in agreement with previous findings that PEA3 genes are transcriptional targets of the chimeric *CIC*::*DUX4* fusion protein.[6,10,15,18–20] Given the observed subtype-specific transcriptional profiles revealed by the targeted RNA-seq panel, the possibility of building an expression-based RF classifier focusing on *CIC*-rearranged and ES cases, the two most common SBRCS subgroups and those with the most critical differential diagnosis, was explored further. For the other samples in the curated cohort, the RF predicted probabilities greater than 0.9 for the *CIC*-rearranged-like cases and less than 0.6 for the other SBRCSs. When tested on a large test cohort of 1335 routine diagnostic cases, the classifier identified 15 cases with high probability of being *CIC*-rearranged, including 1 case with a *CIC*::*NUTM1* gene fusion.[36] The compatibility with a *CIC*-rearranged sarcoma

Table 2 Additional Candidate *CIC*-Rearranged Case Identified on the Basis of Its Expression Profile

Sample identifier	Predicted probability of being *CIC*-rearranged	Identified candidate fusion	Age, years/sex/ localization	Initial diagnosis in pathology report
S_09	0.72	None	47/F/hand	Unclassified small round cell sarcoma

The table columns show sample identifier, predicted probability of being *CIC*-rearranged, identified gene fusion during primary routine pathology workup, age, sex, anatomic localization, and initial diagnosis in the pathology report.

F, female.

diagnosis was confirmed for all of them by their re-evaluation by expert soft-tissue pathologists. In addition, one case with available paraffin material was confirmed to harbor *CIC* rearrangement by FISH.

Results confirmed the importance of *ETV4* as a marker for *CIC*-rearranged sarcomas, however, they also highlighted advantages of using a multigene assessment. Indeed, all 15 candidate cases had very high *ETV4* expression value. However, when considering the additional 22 samples with *ETV4* expression higher than approximately the 97th percentile, the predicted probabilities by the RF in most cases were markedly lower, correctly indicating that these samples were not *CIC*-rearranged. The expression profiles of 21 of these 22 cases differed from those of the training *CIC*-rearranged cases and re-evaluation by the expert STT pathologists indeed ruled out a *CIC*-rearranged sarcoma diagnosis for all 21. On the other hand, only one case of potential *CIC*-rearranged sarcoma with a probability of 0.72 would have been interpreted as a false negative. A reassessment of this case revealed that this tumor had been classified as likely *CIC*-rearranged sarcoma in the initial pathology report based on morphology and immunohistochemistry. However, the sample harbored very poor RNA quality, which might have affected the measured expression profiles and in turn the predicted probability by the RF. In summary, a prediction based on *ETV4* expression alone, with an expression threshold of approximately the 97th *ETV4* expression percentile, would have identified one more *CIC*-rearranged case correctly, but with a precision (ie, percentage of correctly identified *CIC*-rearranged cases out of all cases predicted as *CIC*-rearranged) of 43% (16 of 37). Instead, the RF classifier achieved a precision of 100% (15 of 15). This suggests that more specific diagnostic support might be achievable by using the RF multigene classifier, while high *ETV4* expression is a sensitive but less-specific screening approach. Furthermore, the RF provides a quantitative and observer-independent prediction score. To assess the sensitivity and specificity of the RF properly, it would be necessary to have an alternative, ground-truth criterion able to identify all positive cases (ie, all *CIC*-rearranged sarcomas), in the test set of 1335 cases. The reassessment of all 1335 cases on the basis of their clinicopathologic data and morphology suggested that these 16 samples were the only cases compatible with *CIC*-rearranged sarcoma. Thus, a threshold of prob(*CIC*-rearranged) $>$ 0.75 identified 15 of 16 cases with no false-positive results, whereas a threshold $\geq$0.72 would have identified all 16 cases with one false-positive case. Accordingly, the identification of *CIC*-rearranged cases with the RF classifier appears highly specific. Yet, even a case with high prob(*CIC*-rearranged) by the RF classifier needs critical interpretation on the basis of morphology, immunoprofiling, and clinicopathologic data by the pathologist.

Taken together, this study shows that the bioinformatical analysis of targeted RNA-seq expression patterns can effectively support a SBRCS diagnosis. The failure to detect *CIC*::*DUX4* rearrangements by next-generation sequencing–based RNA fusion analysis emphasizes the need for careful evaluation of novel diagnostic assays using known rearranged cases and expertise of a laboratory specialized in STTs. In the future, the availability of a larger and well-curated training cohort of STTs could mitigate the well-known curse of dimensionality issue, associated with having a number of features (genes) significantly higher than the number of samples. A larger training set also could allow extension of the approach to a multiclass prediction tool, able to predict probabilities associated with different STT subtypes. In addition, this approach may be extended to separate STT entities without specific gene fusion events, such as dedifferentiated liposarcoma.

Acknowledgments

We thank Dr. Silvia Vergarajauregui and Prof. Felix Engel for support with confocal imaging, and Simone Hebele and Nadine Pohl for excellent technical assistance.

Author Contributions

L.D.S. acquired and analyzed the data; M.G.C. performed bioinformatical analyses; L.D.S., M.G.C., F.H., and F.F. interpreted the data and wrote the manuscript; E.A.M. acquired and analyzed the data; R.S., M.M., M.B., N.M., S.S., R.J., R.G., S.W., and A.H. participated in patient collection and/or provided resources; A.A. participated in patient collection, analyzed the data, and wrote the manuscript; F.H. and F.F. conceived and designed the study; F.H. supervised data acquisition and analysis; and F.F. supervised bioinformatical analyses. All authors read and approved the final version of the manuscript.

Disclosure Statement

None declared.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2024.02.002*.

References

1. WHO Classification of Tumours Editorial Board: Soft Tissue and Bone Tumours. Vol 3, ed 5. Lyon, France: IARC Press, 2020
2. Antonescu C: Round cell sarcomas beyond Ewing: emerging entities. Histopathology 2014, 64:26–37
3. Antonescu CR: The role of genetic testing in soft tissue sarcoma. Histopathology 2006, 48:13–21
4. Mitelman F, Johansson B, Mertens F: The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 2007, 7:233–245
5. Machado I, Yoshida A, Morales MGN, Abrahao-Machado LF, Navarro S, Cruz J, Lavernia J, Parafioriti A, Picci P, Llombart-

Bosch A: Review with novel markers facilitates precise categorization of 41 cases of diagnostically challenging, "undifferentiated small round cell tumors". A clinicopathologic, immunophenotypic and molecular analysis. Ann Diagn Pathol 2018, 34:1−12
6. Kawamura-Saito M, Yamazaki Y, Kaneko K, Kawaguchi N, Kanda H, Mukai H, Gotoh T, Motoi T, Fukayama M, Aburatani H, Takizawa T, Nakamura T: 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
7. Pierron G, Tirode F, Lucchesi C, Reynaud S, Ballet S, Cohen-Gogo S, Perrin V, Coindre JM, Delattre O: A new subtype of bone sarcoma defined by BCOR-CCNB3 gene fusion. Nat Genet 2012, 44:461−466
8. Sbaraglia M, Righi A, Gambarotti M, Dei Tos AP: Ewing sarcoma and Ewing-like tumors. Virchows Arch 2020, 476:109−119
9. Antonescu CR, Owosho AA, Zhang L, Chen S, Deniz K, Huryn JM, Kao YC, Huang SC, Singer S, Tap W, Schaefer IM, Fletcher CD: Sarcomas with CIC-rearrangements are a distinct pathologic entity with aggressive outcome: a clinicopathologic and molecular study of 115 cases. Am J Surg Pathol 2017, 41:941−949
10. Specht K, Sung YS, Zhang L, Richter GH, Fletcher CD, Antonescu CR: Distinct transcriptional signature and immunoprofile of CIC-DUX4 fusion-positive round cell tumors compared to EWSR1-rearranged Ewing sarcomas: further evidence toward distinct pathologic entities. Genes Chromosomes Cancer 2014, 53:622−633
11. Italiano A, Sung YS, Zhang L, Singer S, Maki RG, Coindre JM, Antonescu CR: High prevalence of CIC fusion with double-homeobox (DUX4) transcription factors in EWSR1-negative undifferentiated small blue round cell sarcomas. Genes Chromosomes Cancer 2012, 51:207−218
12. Gaspar N, Hawkins DS, Dirksen U, Lewis IJ, Ferrari S, Le Deley MC, Kovar H, Grimer R, Whelan J, Claude L, Delattre O, Paulussen M, Picci P, Sundby Hall K, van den Berg H, Ladenstein R, Michon J, Hjorth L, Judson I, Luksch R, Bernstein ML, Marec-Berard P, Brennan B, Craft AW, Womer RB, Juergens H, Oberlin O: Ewing sarcoma: current management and future approaches through collaboration. J Clin Oncol 2015, 33:3036−3046
13. Yoshimoto T, Tanaka M, Homme M, Yamazaki Y, Takazawa Y, Antonescu CR, Nakamura T: CIC-DUX4 induces small round cell sarcomas distinct from Ewing sarcoma. Cancer Res 2017, 77:2927−2937
14. Yoshida A, Arai Y, Kobayashi E, Yonemori K, Ogura K, Hama N, Mukai W, Motoi T, Kawai A, Shibata T, Hiraoka N: CIC break-apart fluorescence in-situ hybridization misses a subset of CIC-DUX4 sarcomas: a clinicopathological and molecular study. Histopathology 2017, 71:461−469
15. Kao YC, Sung YS, Chen CL, Zhang L, Dickson BC, Swanson D, Vaiyapuri S, Latif F, Alholle A, Huang SC, Hornick JL, Antonescu CR: ETV transcriptional upregulation is more reliable than RNA sequencing algorithms and FISH in diagnosing round cell sarcomas with CIC gene rearrangements. Genes Chromosomes Cancer 2017, 56:501−510
16. Machado I, Navarro L, Pellin A, Navarro S, Agaimy A, Tardio JC, Karseladze A, Petrov S, Scotlandi K, Picci P, Llombart-Bosch A: Defining Ewing and Ewing-like small round cell tumors (SRCT): the need for molecular techniques in their categorization and differential diagnosis. A study of 200 cases. Ann Diagn Pathol 2016, 22:25−32
17. Panagopoulos I, Bjerkehagen B, Gorunova L, Berner JM, Boye K, Heim S: Several fusion genes identified by whole transcriptome sequencing in a spindle cell sarcoma with rearrangements of chromosome arm 12q and MDM2 amplification. Int J Oncol 2014, 45:1829−1836
18. Hung YP, Fletcher CD, Hornick JL: Evaluation of ETV4 and WT1 expression in CIC-rearranged sarcomas and histologic mimics. Mod Pathol 2016, 29:1324−1334
19. Le Guellec S, Velasco V, Perot G, Watson S, Tirode F, Coindre JM: 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
20. Smith SC, Palanisamy N, Martin E, Almenara J, McHugh JB, Choi EK, Lucas DR, Betz BL, Thomas D, Patel RM: The utility of ETV1, ETV4 and ETV5 RNA in-situ hybridization in the diagnosis of CIC-DUX sarcomas. Histopathology 2017, 70:657−663
21. Prat A, Ellis MJ, Perou CM: Practical implications of gene-expression-based assays for breast oncologists. Nat Rev Clin Oncol 2011, 9:48−57
22. Taherian-Fard A, Srihari S, Ragan MA: Breast cancer classification: linking molecular mechanisms to disease prognosis. Brief Bioinform 2015, 16:461−474
23. Iqbal J, Wright G, Wang C, Rosenwald A, Gascoyne RD, Weisenburger DD, et al: Gene expression signatures delineate biological and prognostic subgroups in peripheral T-cell lymphoma. Blood 2014, 123:2915−2923
24. Bobee V, Drieux F, Marchand V, Sater V, Veresezan L, Picquenot JM, Viailly PJ, Lanic MD, Viennot M, Bohers E, Oberic L, Copie-Bergman C, Molina TJ, Gaulard P, Haioun C, Salles G, Tilly H, Jardin F, Ruminy P: Combining gene expression profiling and machine learning to diagnose B-cell non-Hodgkin lymphoma. Blood Cancer J 2020, 10:59
25. Zhang W, Ao Q, Guan Y, Zhu Z, Kuang D, Li MMQ, Shen K, Zhang M, Wang J, Yang L, Cai H, Wang Y, Young KH, Zhou J, Xiao M: A novel diagnostic approach for the classification of small B-cell lymphoid neoplasms based on the NanoString platform. Mod Pathol 2022, 35:632−639
26. Sun L, McNulty SN, Evenson MJ, Zhu X, Robinson JA, Mann PR, Duncavage EJ, Pfeifer JD: Clinical implications of a targeted RNA-sequencing panel in the detection of gene fusions in solid tumors. J Mol Diagn 2021, 23:1749−1760
27. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013, 29:15−21
28. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, Cox AJ, Kruglyak S, Saunders CT: Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 2016, 32:1220−1222
29. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004, 5:R80
30. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014, 15:550
31. Anders S, Huber W: Differential expression analysis for sequence count data. Genome Biol 2010, 11:R106
32. Wickham H: ggplot2: Elegant Graphics for Data Analysis. New York, Springer-Verlag, 2016
33. Breiman L: Random forests. Mach Learn 2001, 45:5−32
34. Liaw A, Wiener M: Classification and regression by randomForest. R News 2002, 2:18−22
35. Krzywinski M, Altman N: Classification and regression trees. Nat Methods 2017, 14:755−756
36. Schaefer IM, Dal Cin P, Landry LM, Fletcher CDM, Hanna GJ, French CA: CIC-NUTM1 fusion: a case which expands the spectrum of NUT-rearranged epithelioid malignancies. Genes Chromosomes Cancer 2018, 57:446−451