Analysis of Gene Expression Profiles of Soft Tissue Sarcoma Using a Combination of Knowledge-Based Filtering with Integration of Multiple Statistics

Anna Takahashi1,9, Robert Nakayama2,3,*, Nanako Ishibashi4,*, Ayano Doi1,2,*, Risa Ichinohe2,5, Yoriko Ikuyo7, Teruyoshi Takahashi7, Shigetaka Marui7, Koji Yasuhara7, Tetsuro Nakamura7, Shintaro Sugita8, Hiromi Sakamoto2, Teruhiko Yoshida2, Tadashi Hasegawa8,9, Hiro Takahashi1,2,6,9

1 Plant Biology Research Center, Chubu University, Kasugai, Aichi, Japan, 2 Division of Genetics, National Cancer Center Research Institute, Tokyo, Japan, 3 Department of Orthopaedic Surgery, Keio University School of Medicine, Tokyo, Japan, 4 Division of Biological Science, Graduate School of Science, Nagoya University, Nagoya, Aichi, Japan, 5 Faculty of Horticulture, Chiba University, Matsudo, Chiba, Japan, 6 Division of Surgical Pathology, Sapporo Medical University School of Medicine, Sapporo, Hokkaido, Japan, 7 Faculty of Horticulture, Chiba University, Kasugai, Aichi, Japan, 8 Department of Surgical Pathology, Sapporo Medical University School of Medicine, Sapporo, Hokkaido, Japan, 9 Pathology Division, National Cancer Center Hospital, Tokyo, Japan

Abstract
The diagnosis and treatment of soft tissue sarcomas (STS) have been difficult. Of the diverse histological subtypes, undifferentiated pleomorphic sarcoma (UPS) is particularly difficult to diagnose accurately, and its classification per se is still controversial. Recent advances in genomic technologies provide an excellent way to address such problems. However, it is often difficult, if not impossible, to identify definitive disease-associated genes using genome-wide analysis alone, primarily because of multiple testing problems. In the present study, we analyzed microarray data from 88 STS patients using a combination method that used knowledge-based filtering and a simulation based on the integration of multiple statistics to reduce multiple testing problems. We identified 25 genes, including hypoxia-related genes (e.g., MIF, SOD1, P4HA1, ENO1, and STAT1) and cell cycle- and DNA repair-related genes (e.g., TACC3, PRDX1, PRKDC, and H2AFY). These genes showed significant differential expression among histological subtypes, including UPS, and showed associations with overall survival. STAT1 showed a strong association with overall survival in UPS patients (logrank $p = 1.84 \times 10^{-6}$ and adjusted $p$ value $2.99 \times 10^{-3}$ after the permutation test). According to the literature, the 25 genes selected are useful not only as markers of differential diagnosis but also as prognostic/predictive markers and/or therapeutic targets for STS. Our combination method can identify genes that are potential prognostic/predictive factors and/or therapeutic targets in STS and possibly in other cancers. These disease-associated genes deserve further preclinical and clinical validation.

Introduction
Recent advances in genomic technologies offer an excellent opportunity to determine the complete biological characteristics of neoplastic tissues, resulting in improved diagnosis, treatment selection, rational classification based on molecular carcinogenesis, and identification of therapeutic targets. The diagnosis and treatment of soft tissue sarcomas (STS) have been difficult because STSs comprise a group of highly heterogeneous tumors in terms of histopathology, molecular signature, histological grade, and primary site. These tumors have generally been classified into subtypes according to their histological resemblance to normal tissue. The Fédération Française des Centres de Lutte Contre le Cancer (FNCLCC) grading system was defined more than 20 years ago and is still the most commonly used grading system for STS [1,2]. Treatment of STS is based on both histological subtype and histological grade. The understanding gained regarding the molecular pathology of cancer in recent decades suggests that some tumor types exhibit stand-alone recurrent genetic aberrations, such as chromosomal translocations, that result in gene fusions, e.g., SYT-SSX in synovial sarcoma (SS) [3], TLS-CHOP in myxoid/round cell liposarcoma (MLS) [4], and KIF5B-RET in lung adenocarcinoma [5], or somatic mutations, e.g., KIT in gastrointestinal stromal tumors (GIST) [6] and 26 mutated genes (TP53, KRAS, EGFR, and 23 other genes) in lung adenocarcinoma [7].
 nomal [7]. The molecular markers specific to each tumor type are useful for tumor classification [6]. In contrast, several malignant tumors, such as malignant fibrous histiocytoma (MFH), are characterized by numerous nonrecurrent, complex chromosomal aberrations, and they frequently show overlapping histological features and immunophenotypes that are difficult for pathologists to interpret [9]. In particular, the diagnosis of MFH has been a controversial issue [10–13]. MFH is the most common soft tissue sarcoma in adults. It has a wide range of histological subtypes [13]. For this reason, discrimination between MFH and other STSs is difficult, but this discrimination is necessary because there are significant differences in the 5-year survival rates of the STS subtypes [14]: 100% for well-differentiated liposarcoma (WLS), 71% for synovial sarcoma (SS), 46% for pleomorphic MFH, and 92% for myxofibrosarcoma (MFS). MFH was renamed undifferentiated pleomorphic sarcoma (UPS) in 2002 by the World Health Organization (WHO) [15]. MFS was considered a subtype of MFH before this classification; WHO reclassified MFS as another subtype of STS [15]. Discrimination between UPS and MFS is particularly difficult [14] because of their histological similarities and because of the considerable heterogeneity of UPS [13]. UPS was previously characterized by global gene expression analysis using analysis of variance (ANOVA) and clustering analysis [13]. Although some possible prognostic factors were identified, the list of factors was not complete because the study was conducted without information on patient outcomes. In the present study, we hypothesized that some genes can serve both as diagnostic markers for histological subtyping and as prognostic markers of overall survival in STS. We used a combination of statistical and bioinformatic methods to identify those genes.

Many statistical and bioinformatic methods have been proposed for global biological information analysis in the past 3 decades. For example, basic local alignment search tool (BLAST) [16], ClustalW [17], BLAST-based algorithm for the identification of upstream ORFs with conserved amino acid sequences (BAIUCAS) [18], and G4 DNA motif region finder by R (G4MR-FindeR) [19] have been used for sequence analysis; hierarchical clustering [20], fuzzy k-means [21], and fuzzy adaptive resonance theory (FuzzyART) [22,23] have been used for gene cluster analysis; gene set enrichment analysis (GSEA) [24], modified signal-to-noise (S2N) [25], and projective adaptive resonance theory (PART) [26,27] have been used for gene selection; fuzzy neural network (FNN) [28,29] and boosted fuzzy classifier with a SWEEP operator (BFCS) [30–32] have been used for the construction of prediction models; and IntPath [33] and Stringent DDI-based Prediction (BFCS) [34] were used for analysis of pathways and protein–protein interactions. The use of statistical or bioinformatic analysis is practical and useful for clinical diagnosis [35–37] and the identification of marker genes [38–43]. In the present study, we focused on microarray data analysis; however, the analysis of data obtained using next-generation sequencing technologies [44] is a subject of an upcoming project.

Global analysis of gene expression is a powerful method for the identification of prognostic/predictive factors and/or therapeutic targets. However, it is often difficult, if not impossible, to identify definitive disease-associated genes using genome-wide analysis alone, primarily because of multiple testing problems. In this situation, knowledge-based approaches, such as knowledge-based fuzzy adaptive resonance theory (KB-FuzzyART) [45] and knowledge-based single nucleotide polymorphism (KB-SNP) [46,47], are effective and interpretable [48–50]. Online Mendelian Inheritance in Man (OMIM) is a continuously updated catalog of human genes and genetic disorders and traits. In the present study, we used OMIM as a knowledge source for narrowing the list of candidate genes and applied the OMIM-based method to gene expression data from STS patients. Thus, we identified 25 genes that showed significant differential expression among histological subtypes, including UPS, and showed associations with overall survival. According to the literature, these genes are useful not only as diagnostic markers for the discrimination of molecular pathway-based subtypes but also as prognostic/predictive markers and/or therapeutic targets for STS. Moreover, these genes are useful for understanding the mechanisms underlying tumor progression or metastasis and for the rational design of anticancer therapies.

| Table 1. Characteristics of the 88 patients with soft tissue sarcoma. |

| Characteristics | STS patients (n=88) |
|-----------------|---------------------|
| Gender          | Male 46             |
|                 | Female 42           |
| Age             | Median 54           |
|                 | MAD 19              |
| Histological type| UPS 20              |
|                 | MLS 20              |
|                 | SS 17               |
|                 | MFS 15              |
|                 | LMS 6               |
|                 | FS 5                |
| Histological grade| MPNST 5            |
| Relapse events  | Metastasis 43       |

STS: soft tissue sarcoma, MAD: Median absolute deviation, UPS: undifferentiated pleomorphic sarcoma, MLS: myxoid liposarcoma, SS: synovial sarcoma, MFS: myxofibrosarcoma, LMS: leiomyosarcoma, FS: fibrosarcoma, MPNST: malignant peripheral nerve sheath tumor.

doi:10.1371/journal.pone.0106801.t001
therapeutics. Therefore, our combination method of knowledge-based filtering and simulation based on the integration of multiple statistics can identify potential prognostic/predictive factors and/or therapeutic targets in STS and possibly in other cancers.

**Materials and Methods**

**Ethics statement**

The study was conducted according to the principles expressed in the Declaration of Helsinki. The ethics committee of the National Cancer Center approved the study protocol. All patients provided written informed consent.

**Patients and tumor samples**

The characteristics of the 88 STS patients (20 with UPS, 15 with MFS, 17 with SS, 20 with myxoid liposarcoma [MLS], 6 with leiomyosarcoma [LMS], 5 with fibrosarcoma [FS], and 5 with a malignant peripheral nerve sheath tumor [MPNST]) enrolled in this study are shown in Table 1. All patients had received a histological diagnosis of primary soft tissue tumor at the National Cancer Center Hospital, Tokyo, between 1996 and 2002 [51], as shown in Table S1. Tumor samples were obtained at the time of excision and were cryopreserved in liquid nitrogen.

**Microarray analysis**

For RNA extraction, trained pathologists carefully excised the tissue samples from the main tumor, leaving a margin free from the surrounding nontumorous tissue. The elimination of nontumorous stromal cells is necessary for gene expression analysis of carcinomas because tumor tissues contain a significant number of nontumorous stromal cells, including fibroblasts, endothelial cells, and inflammation-associated cells. STS contains non-tumorous...
stromal cells that are difficult to exclude because STS originates from mesenchymal cells. However, in STS, the tumor tissue contains very few non-tumorous stromal cells and therefore unlikely to confound the analysis. Hence, laser microdissection was not performed in this study. Total RNA samples extracted from the bulk tissue specimens were labeled with biotin and hybridized to high-density oligonucleotide microarrays (Human Genome U133A 2.0 Array; Affymetrix, Santa Clara, CA, USA) comprising 22,283 probe sets representing 18,400 transcripts, according to the manufacturer’s instructions. The scanned array data were processed using the Affymetrix Microarray Suite v.5.1 software (MAS5), which scaled the average intensity of all the genes on each array to the target signal of 1000. The microarray data from the present study are available in the Genome Medicine Database of Japan (GeMDBJ) [52] (https://gemdbj.nibio.go.jp/dgdb/) under the accession number EXPR058P.

Data preprocessing
We excluded 68 control probe sets and 2343 genes that were subject to cross-hybridization according to NetAffx Annotation.
Furthermore, we excluded those genes for which more than 50% (44/88) of the samples showed an absent call (i.e., the detection call determined by MAS5 based on the p value of the one-sided Wilcoxon signed-rank test; an absent call corresponds to $p \geq 0.065$, which is the default threshold in MAS5). An absent call indicates that the expression signal was undetectable. Genes showing low variance, i.e., a signal range value (95th percentile to 5th percentile) of less than 2000, were excluded [40]. Furthermore, we conducted an OMIM-based reduction of the number of candidate genes. In total, 1412 genes were selected, to which we applied log-transformation or binarization using the median value as a threshold for each gene, as shown in Fig. 1. The 2 types of datasets, log-transformed and binarized, were used for ANOVA and the logrank test, respectively.

Simulation based on the combination of a permutation test and the integration of multiple statistics

We previously proposed a statistical simulation based on a permutation test and the integration of multiple statistics [51]. This method was used in the present study. We first calculated $p$ values using ANOVA to discriminate among histological subtypes, including UPS, MFS, SS, and MLS. We also calculated $p$ values by means of the logrank test in the survival analysis of all STS patients in relation to the 1412 filtered genes. We defined the integrated statistic $p^*$ as $p_1 \cdot p_2$, where $p_1$ is the $p$ value from ANOVA and $p_2$ is the $p$ value from the logrank test. The same STS patients ($n = 72$; 20 UPS, 15 MFS, 17 SS, and 20 MLS patients) were used in both of these tests. The integrated statistic $p^*$ could be underestimated by the use of 72 common samples. Therefore, to cancel this influence, we conducted a simulation based on the permutation test, as shown in Fig. 1, to estimate the adjusted $p^*$ values as well as the multiple testing problems.

Statistical analysis

The median value of the gene expression signals for each gene was calculated, and the patients were distributed into 2 groups using the median value as a threshold for each gene. Logrank tests [53] were performed for overall survival of STS patients for each

### Table 3. Correlation analysis based on Spearman’s rank correlation coefficient between gene expression data and the histological grade (or metastasis status).

| Affymetrix probe ID | Accession no. | Gene symbol | With histological grade | With metastasis |
|---------------------|---------------|-------------|------------------------|----------------|
|                     |               |             | $p$                    | $p$ value      |
|                     |               |             | $r$                    | $p$ value      |
| 200832_s_at         | AB032261      | SCD1        | 0.0191                 | 8.60E-01       |
| 200887_s_at         | NM_007315     | STAT1       | 0.146                  | 1.73E-01       |
| 201231_s_at         | NM_001428     | ENO1/MBP1   | 0.356                  | 6.66E-04       |
| 201508_at           | NM_001552     | IGFBP4      | 0.247                  | 2.04E-02       |
| 202236_s_at         | NM_003051     | SLC16A1/MCT1| 0.400                  | 1.12E-04       |
| 202870_s_at         | NM_001255     | CDC20       | 0.413                  | 6.27E-05       |
| 203065_s_at         | NM_001753     | CAV1        | 0.250                  | 1.87E-02       |
| 203323_at           | BF197655      | CAV2        | 0.363                  | 5.11E-04       |
| 203554_x_at         | NM_004219     | PTG1        | 0.402                  | 1.05E-04       |
| 207011_s_at         | NM_002621     | PTK7        | 0.265                  | 1.26E-02       |
| 207168_s_at         | NM_004893     | H2AFY/H2AX  | 0.411                  | 7.03E-05       |
| 207543_s_at         | NM_000917     | P4HA1       | 0.449                  | 1.12E-05       |
| 208680_at           | L19184        | PRDX1       | 0.258                  | 1.51E-02       |
| 208694_at           | U47077        | PRKDC/DNA-PKcs| 0.409               | 7.64E-05       |
| 208767_s_at         | AW149681      | LAPT4B      | 0.329                  | 1.75E-03       |
| 209030_s_at         | NM_014333     | CADM1/TSLC1| 0.196                  | 6.70E-02       |
| 209031_at           | AL519710      | CADM1/TSLC1| 0.231                  | 3.03E-02       |
| 209543_s_at         | MB1104        | CD34        | 0.363                  | 5.11E-04       |
| 210495_x_at         | AF130095      | FN1         | 0.286                  | 6.99E-03       |
| 210559_s_at         | D88357        | CDK1/CD2    | 0.435                  | 2.34E-05       |
| 212097_at           | AU147399      | CAV1        | 0.237                  | 2.64E-02       |
| 212464_s_at         | X02761        | FN1         | 0.286                  | 6.99E-03       |
| 212794_s_at         | UB8968        | ENO1/MBP1   | 0.387                  | 1.97E-04       |
| 217871_s_at         | NM_002415     | MIF         | 0.421                  | 4.41E-05       |
| 218308_at           | NM_006342     | TACC3       | 0.333                  | 1.52E-03       |
| 218502_s_at         | NM_014112     | TRPS1       | 0.276                  | 9.23E-03       |
| 218755_at           | NM_005733     | KIF20A/MKlp2| 0.407               | 8.35E-05       |
| 219918_s_at         | NM_018123     | ASPM        | 0.399                  | 1.16E-04       |
| 220942_x_at         | NM_014367     | FAM162A/HGTD-P| 0.151               | 1.60E-01       |

doi:10.1371/journal.pone.0106801.t003

(www.affymetrix.com). Furthermore, we excluded those genes for which more than 50% (44/88) of the samples showed an absent call (i.e., the detection call determined by MAS5 based on the $p$ value of the one-sided Wilcoxon signed-rank test; an absent call corresponds to $p \geq 0.065$, which is the default threshold in MAS5). An absent call indicates that the expression signal was undetectable. Genes showing low variance, i.e., a signal range value (95th percentile to 5th percentile) of less than 2000, were excluded [40]. Furthermore, we conducted an OMIM-based reduction of the number of candidate genes. In total, 1412 genes were selected, to which we applied log-transformation or binarization using the median value as a threshold for each gene, as shown in Fig. 1. The 2 types of datasets, log-transformed and binarized, were used for ANOVA and the logrank test, respectively.
gene. We also calculated Spearman’s rank correlation coefficients to assess the relationships between gene expression signals and histological grades [54] or incidence of tumor metastases. We considered data obtained after 50 months of follow-up as censored data in the analysis of the logrank test, similar to the procedure followed in our previous study [51]. Kaplan-Meier curves [55] based on histological subtype were constructed for all STS patients.

OMIM
OMIM is a continuously updated catalog of human genes and genetic disorders and traits, with a focus on the molecular relationship between genetic variation and phenotypic expression. The list of MIM gene accession numbers associated with keywords related to cancer was obtained from the OMIM website (http://www.omim.org/). We used several keywords related to cancer, including “cancer,” “carcinoma,” “sarcoma,” “tumor,” and “neoplasm,” to create the MIM gene accession number list. There were 4394 MIM gene accession numbers, as shown in Table S2. The final MIM gene accession number list was obtained on January 10, 2014.

Ensembl
Ensembl is a joint project between EMBL-EBI and the Sanger Centre to develop software that produces and maintains automatic annotation of eukaryotic genomes [56]. We converted MIM numbers to the Affymetrix probe set IDs of the Human Genome

Figure 3. Heatmap and hierarchical clustering analyses. Twenty-nine probe sets were extracted using a simulation based on the permutation test (with adjusted \( p < 0.05 \)). The 29 probe sets were roughly divided into 4 clusters (clusters A–D). Columns represent probe sets, and rows represent samples. Red and green indicate high and low expression, respectively. UPS: undifferentiated pleomorphic sarcoma, MLS: myxoid liposarcoma, SS: synovial sarcoma, MFS: myxofibrosarcoma.

doi:10.1371/journal.pone.0106801.g003
Gene Expression Analysis of STS

Multiple testing correction

The Bonferroni correction is a method used to address the problem of multiple comparisons (also known as the multiple testing problem). It is considered the simplest and most conservative method for control of the family-wise error rate (FWER). False discovery rate (FDR) controlling procedures, such as the Benjamini-Hochberg (BH) method, are more powerful (i.e., less conservative) than the FWER procedures, but their use increases the likelihood of false positives within the rejected hypothesis. In the present study, the BH method was used to calculate the \( q \) value. The \( p \) value is defined as an FDR analog of the \( p \) value.

Heatmap and hierarchical clustering analyses

A heatmap was created using the R program (function heatmap.2 in Package gplots) for the log-transformed and scaled gene expression data of selected genes. Hierarchical clustering was also conducted using the Euclidean distance and complete linkage (default parameters of function heatmap.2).

Results

Kaplan-Meier curves for 4 histological subtypes

Kaplan-Meier curves based on a histological subtype were constructed for all STS patients, as shown in Fig. 2. This figure shows that MFS had a good prognosis, MLS and SS had intermediate prognoses, and UPS had a poor prognosis. Although the logrank test yielded statistically significant results \((p<0.05)\) in histological types, we conducted gene expression analysis to select molecular markers for more accurate diagnosis in accordance with the analysis.

Extraction of genes that are both diagnostic and prognostic markers, by means of a simulation using the permutation test

To extract genes that are both diagnostic markers (for discrimination of histological subtypes) and prognostic markers (of overall survival in STS), we applied a simulation based on the combination of a permutation test and the integration of multiple statistics into 1412 prefiltered probe sets of microarray data obtained from STS patients. As shown in Table 2, 29 probe sets, representing 25 genes, were extracted (adjusted \( p \) value <0.05).

Association analysis of the histological grade (or metastasis status) and gene expression data for the 25 selected genes

We next used Spearman’s rank correlation analysis to analyze the association between the gene expression level in STS patients and the histological grade (or metastasis status), as shown in Table 3. Table 3 shows that genes with positive \( \rho \) were upregulated in highly malignant tumors, whereas genes with negative \( \rho \) were downregulated in highly malignant tumors. The expression levels of almost all of the 25 genes were associated with either the histological grade or metastasis. However, stearoyl-CoA desaturase 1 (SCD1) and signal transducer and activator of transcription 1 (STAT1) were not associated with either the histological grade or metastasis. However, stearoyl-CoA desaturase 1 (SCD1) and signal transducer and activator of transcription 1 (STAT1) were not associated with either the histological grade (SCD1: \( \rho = -0.0191, \ p = 0.860; \) STAT1: \( \rho = -0.146, \ p = 0.173 \) or metastasis (SCD1: \( \rho = 0.0237, \ p = 0.626; \) STAT1: \( \rho = -0.177, \ p = 0.0995 \)). This result indicates that SCD1 and STAT1 expression levels can be related to the overall survival of STS patients but not to metastasis. Therefore, these data suggest that SCD1 and STAT1 expression levels can...
Table 4. Pairwise comparison between histological types using Welch’s t test for 29 probe sets.

| Affymetrix probe ID | Accession no. | Gene symbol | UPS vs. MFS | UPS vs. SS | UPS vs. MLS |
|---------------------|---------------|-------------|-------------|------------|-------------|
|                     |               |             | p value | q value | p value | q value | p value | q value |
| 200832_s_at         | AB032261      | SCD1        | 7.36E-05 | *        | 8.87E-04 | 1.06E-03 | *        | 2.56E-03 | 3.52E-01 | 4.26E-01 |
| 200887_s_at         | NM_007315     | STAT1       | 2.81E-01 | *        | 4.07E-01 | 1.54E-03 | *        | 3.19E-03 | 2.04E-01 | 2.69E-01 |
| 201231_s_at         | NM_001428     | ENO1/MBP1   | 1.06E-04 | *        | 8.87E-04 | 4.73E-08 | *        | 6.85E-07 | 4.27E-06 | 1.42E-05 |
| 201508_at           | NM_001552     | IGFBP4      | 4.21E-02 | 1.15E-01 | 7.39E-03 | 1.37E-02 | *        | 1.13E-02 | 7.25E-02 | 1.00E-01 |
| 202236_s_at         | NM_003051     | SLC16A1/MCT1| 1.54E-01 | 2.80E-01 | 3.92E-01 | 4.06E-01 | *        | 1.25E-03 | 6.49E-04 | 1.78E-05 |
| 202870_s_at         | NM_001255     | CDC20       | 2.10E-01 | 3.58E-01 | 1.23E-03 | 2.74E-03 | *        | 6.26E-06 | 6.66E-01 | 1.78E-05 |
| 203065_s_at         | NM_001753     | CAV1        | 8.76E-01 | 8.76E-01 | 5.56E-07 | 2.69E-06 | *        | 5.31E-01 | 5.93E-01 | 9.31E-01 |
| 203323_at           | BF197655      | CAV2        | 3.76E-01 | 4.96E-01 | 8.95E-05 | 2.60E-04 | *        | 1.59E-08 | 2.37E-07 | 2.31E-07 |
| 20354_x_at          | NM_004219     | PTG1        | 6.14E-03 | 2.23E-02 | 4.21E-03 | 6.78E-03 | *        | 9.19E-01 | 9.19E-01 | 9.19E-01 |
| 207011_s_at         | NM_002821     | PTK7        | 4.37E-02 | 1.15E-01 | 1.18E-01 | 1.37E-01 | *        | 1.87E-05 | 1.87E-05 | 1.87E-05 |
| 207542_s_at         | NM_009917     | P4HA1       | 1.22E-04 | *        | 8.87E-04 | 2.64E-02 | *        | 3.48E-02 | 2.51E-03 | 4.05E-03 |
| 208650_at           | L19184        | PRED1       | 1.84E-03 | *        | 7.61E-03 | 5.31E-05 | *        | 1.93E-04 | 1.36E-08 | 2.31E-07 |
| 208694_at           | U47077        | PRKDC/DNA-Pkcs| 5.49E-02 | 1.33E-01 | 9.76E-01 | 9.76E-01 | *        | 1.13E-03 | 2.08E-03 | 2.08E-03 |
| 208767_s_at         | AW149681      | LAPTMB4     | 1.12E-04 | *        | 8.87E-04 | 2.64E-02 | *        | 3.48E-02 | 2.51E-03 | 4.05E-03 |
| 209030_s_at         | NM_014333     | CADM1/TSCL1 | 6.04E-02 | 1.35E-01 | 2.67E-07 | 1.82E-06 | *        | 6.66E-01 | 6.66E-01 | 6.66E-01 |
| 209031_at           | ALS19710      | CADM1/TSCL1| 6.04E-02 | 1.35E-01 | 2.67E-07 | 1.82E-06 | *        | 6.66E-01 | 6.66E-01 | 6.66E-01 |
| 209543_s_at         | M81104        | CD34        | 4.37E-02 | 1.15E-01 | 1.18E-01 | 1.37E-01 | *        | 1.87E-05 | 1.87E-05 | 1.87E-05 |
| 210495_x_at         | AF130095      | FN1         | 4.83E-01 | 5.61E-01 | 2.30E-03 | 4.27E-03 | 3.58E-06 | *        | 1.42E-05 | 1.42E-05 |
| 210559_s_at         | D88357        | CDX1/CDX2   | 7.05E-02 | 1.46E-01 | 2.35E-02 | 3.24E-02 | 3.57E-06 | *        | 1.42E-05 | 1.42E-05 |
| 212097_at           | AU147399      | CAV1        | 6.34E-01 | 6.91E-01 | 3.14E-07 | 1.82E-06 | 4.16E-01 | 4.16E-01 | 4.16E-01 | 4.16E-01 |
| 212464_s_at         | X02761        | FN1         | 5.22E-01 | 5.38E-01 | 2.33E-03 | 4.22E-03 | 2.07E-06 | *        | 1.20E-05 | 1.20E-05 |
| 217294_s_at         | U88968        | ENO1/MBP1   | 4.24E-04 | *        | 2.46E-03 | 4.07E-05 | *        | 1.69E-04 | 1.58E-07 | 1.50E-06 |
| 217871_s_at         | NM_002415     | MIF         | 5.31E-06 | *        | 1.54E-04 | 1.38E-01 | 1.54E-01 | 1.35E-05 | 3.27E-05 | 3.27E-05 |
| 218308t_at          | NM_006342     | TACC3       | 2.33E-01 | 3.80E-01 | 7.67E-04 | 2.02E-03 | 2.91E-05 | *        | 6.49E-05 | 6.49E-05 |
| 218502_s_at         | NM_014112     | TRPS1       | 3.64E-01 | 4.95E-01 | 5.21E-11 | 1.51E-09 | 1.85E-02 | *        | 2.68E-02 | 2.68E-02 |
| 218735_s_at         | NM_005733     | KIF20A/MKlp2| 4.44E-01 | 5.37E-01 | 9.97E-03 | 1.45E-02 | 4.41E-06 | *        | 1.42E-05 | 1.42E-05 |
| 219918_s_at         | NM_018123     | ASPM        | 1.11E-01 | 2.13E-01 | 2.25E-03 | 4.22E-03 | 7.89E-07 | *        | 5.72E-06 | 5.72E-06 |
| 220422_s_at         | NM_014367     | FAM162A/HGTD-P | 1.39E-03 | *        | 6.70E-03 | 3.81E-02 | *        | 4.60E-02 | 6.23E-01 | 6.66E-01 |

*q < 0.05. The p value was calculated using Welch’s t test, and the q value was calculated from the p value by means of the Benjamini-Hochberg method for the correction of multiple testing problems.

doi:10.1371/journal.pone.0106801.t004
be used in combination with the histological grade to predict the survival of STS patients.

Hierarchical clustering based on the gene expression pattern of the 25 selected genes

We performed hierarchical clustering for the 29 selected probe sets, representing 25 genes and 4 histological subtypes (UPS, MFS, MLS, and SS), as shown in Fig. 3. The genes were roughly classified into 4 clusters (clusters A, B, C, and D). Almost all genes were upregulated in both UPS and MFS. In addition, genes in cluster A were upregulated in SS, and genes in cluster D were upregulated in MLS.

Analysis of the distribution of histological subtypes based on gene expression levels

We performed PCA to calculate the first and second PCs using the 29 probe sets. Detailed information on PCA, including eigenvector, standard deviation, proportion of variance, and cumulative proportion, is provided in Tables S4 and S5. The distribution of the 4 histological subtypes of STS on the 2 axes is shown in Fig. 4. The 4 histological subtypes were clearly classified into 3 clusters (SS, MLS, and UPS+MFS). This result indicated that UPS and MFS had histological similarities and similar gene expression patterns. Therefore, to discriminate between UPS and MFS, we applied Welch’s t test and the BH method to the gene expression data from the 29 probe sets to discriminate UPS from SS and UPS from MLS, as shown in Table 4.

Classification of the 25 genes based on pairwise comparison of histological subtypes

We classified the 25 genes into 7 groups on the basis of 3 comparisons (UPS vs. MFS, UPS vs. SS, and UPS vs. MLS), as shown in Fig. 6. Only 3 genes, ENO1/MBP1, P4HA1, and PRDX1, were commonly selected (genes that were selected in the UPS vs. MFS comparison were also selected in the UPS vs. SS or UPS vs. MLS comparison). Furthermore, we compared the 25 genes selected in our study with the genes involved in the complexity index in sarcomas (CINSARC) [59] because the use of CINSARC (composed of 67 genes) instead of the FNCLCC grading system [1,2] was recently proposed for predicting metastasis in STS [59]. In this comparison, only 4 common genes, that is, pituitary tumor-transforming 1 (PTTG1), abnormal spindle-like microcephaly-associated protein (ASPM), cell division cycle protein 20 (CDC20), and kinesin family member 20A (KIF20A)/mitotic kinesin-like protein 2 (MKlp2), were extracted. The differential expression of these 4 genes was statistically significant (q < 0.05) for UPS vs. SS and for UPS vs. MLS, but not for UPS vs. MFS. These 4 genes belonged to cluster B, as shown in Fig. 3. Consequently, the 25 genes were classified into 7 groups on
Figure 8. A hypothetical regulation model of metabolic and signaling control in highly malignant STS. (A) Signaling pathways, excluding cell cycle and DNA repair. (B) Cell cycle and DNA repair pathways. The pink oval indicates the genes selected in the present study. MUFA, monounsaturated fatty acid; SFA, saturated fatty acid; SCD1, stearoyl-CoA desaturase 1; MIF, macrophage migration inhibitory factor; CXCR, CXC chemokine receptor; PI3K, phosphoinositide 3-kinase; MAPK, extracellular signal-regulated kinase; ERK, mitogen-activated protein kinase; PTTG1, pituitary tumor-transforming 1; ASPM, abnormal spindle-like microcephaly-associated protein; CDC20, cell division cycle protein 20; KIF20A, kinesin family member 20A; ENO1, enolase 1; P4HA, prolyl 4-hydroxylase subunit α; PRDX1, peroxiredoxin 1; FAM162A, family with sequence similarity 162, member A; STAT1, signal transducer and activator of transcription 1; CDK1, cyclin-dependent kinase 1; TACC3, transforming, acidic coiled-coil containing protein 3; PRKDC, protein kinase, DNA-activated, catalytic polypeptide; H2AFY, H2A histone family, member Y; SLC16A1, solute carrier family 16, member 1; VEGF, vascular endothelial growth factor; HIF, hypoxia inducible factor; PLOD2, procollagen-lysine,2-oxoglutarate 5-dioxygenase 2; NF-κB, nuclear factor-kappa B. 
doi:10.1371/journal.pone.0106801.g008
the basis of pairwise comparisons of histological subtypes, as shown in Fig. 4.

Survival analysis in UPS patients

We used the logrank test to analyze the survival of UPS patients. We selected the best \( p \) value for various thresholds (30th, 40th, 50th, 60th, 70th, and 80th percentiles) of gene expression signals in UPS patients for each probe set when the gene expression signals were binarized. Adjusted \( p \) values were obtained by adjusting the data for the multiple testing problem (6 thresholds \( \times \) 29 probe sets) based on the permutation test, as shown in Table S6. Only \( \text{STAT1} \) showed a statistically significant association with survival in UPS (logrank \( p \) value \( 1.84 \times 10^{-6} \) and adjusted \( p \) value \( 2.99 \times 10^{-3} \) after the permutation test). Fig. 7 shows that \( \text{STAT1} \)-positive and \( \text{STAT1} \)-negative groups had clearly different survival curves based on the Kaplan-Meier method.

Discussion

In the present study, we conducted a simulation based on a permutation test to extract genes that are both diagnostic markers (for discrimination of histological subtypes) and prognostic markers (for overall survival in STS). As shown in Table 2, 25 genes were extracted, and their adjusted \( p \) values were statistically significant (adjusted \( p < 0.05 \)). We analyzed studies related to these 25 genes and found many reports suggesting that these 25 genes are effective prognostic/predictive factors or therapeutic targets, as shown in Table S7, according to the literature (See Supplementary Discussion).

Although we did not try to identify the molecular mechanisms behind the 25 selected genes, several published studies have examined pathways related to these 25 genes, as shown in Table S7 and Fig. 8. These 25 genes are roughly classified into 4 types, namely, hypoxia-related genes (\( \text{MIF, SC1D, P4HA1, ENO1/MBP1, FAM162A/HGTD-P, SLC16A1/MCT1, FN1, and STAT1} \)), cell cycle- and DNA repair-related genes (\( \text{ASPM, CDK1/CDC2, CDC20, KIF20A/MKlp2, PTTG1, TACC3, PRDX1, PRKDC/DNA-PKcs, and H2AFY/H2AX} \)), growth factor signal transduction-related genes, and other genes. Cell cycle- and DNA repair-related genes, hypoxia-induced genes, and growth factor signal transduction-related genes are key players in tumor growth, angiogenesis, metabolism, invasion, and metastasis in various types of cancer. In fact, these processes are attenuated by the inhibition or killing of many of these 25 genes, as shown in Table S7. These genes are therefore possible prognostic/predictive markers and/or therapeutic targets.

\( \text{STAT1} \) expression was found to be strongly associated with survival in UPS patients. \( \text{STAT1} \) interacts directly with \( \text{p53} \) and induces cell growth arrest and apoptosis, as shown in Fig. 8. Although \( \text{STAT1} \) is repressed by HIF-1, the \( \text{STAT1} \)-positive group among the UPS patients had a better prognosis, even when hypoxia-related genes were upregulated. Therefore, \( \text{STAT1} \) is a possible novel, independent prognostic/predictive factor of STS, particularly UPS.

In the diagnosis of STS, classification of UPS is the most controversial topic. Among the 25 selected genes, hypoxia-related genes (\( \text{MIF, SC1D, P4HA1, ENO1/MBP1, FAM162A/HGTD-P, SLC16A1/MCT1, FN1, and STAT1} \)) are present in this study. In particular, the genes \( \text{MIF, SC1D, P4HA1, ENO1/MBP1, and FAM162A/HGTD-P} \) are differentially expressed between UPS and MFS, as shown in Fig. 6 and Table 4. Furthermore, \( \text{STAT1} \) is a prognostic marker in UPS patients, as shown in Fig. 7. Therefore, these hypoxia-related genes are promising prognostic and therapeutic targets and, if validated, may improve the treatment/diagnosis of this type of cancer. Further research is needed regarding the hypoxia-related pathways in highly malignant STS.

We manually constructed a hypothetical regulation model (Figure 8) of metabolic and signaling control in highly malignant STS. Nevertheless, according to the literature, a part of these networks could be automatically predicted by pathway and interaction analyses. For example, pathways of the cell cycle and the DNA damage response were identified by ImPath [33, 60, 61] with statistical significance (\( q \) value \( < 0.05 \)), as shown in Table S8. Interaction networks of the cell cycle (\( \text{ASPM, CDK1, CDC20, KIF20A, PTTG1, PRKDC, and TACC3} \) and \( \text{HIF-1} \) (\( \text{MIF, ENO1, and PRDX1} \)) were identified by means of STRING [62], as shown in Fig. S1. Nonetheless, these tools should be used with appropriate parameters [34, 60, 61]. Such tools are more effective methods when large numbers of candidate genes are extracted.

In summary, we analyzed microarray gene expression data from 88 STS patients using a combination method involving knowledge-based filtering and a simulation based on the integration of multiple statistics to reduce multiple testing problems. Our combination method automatically identified 25 genes in the gene expression data from STS. These genes showed significant differential expression between different histological subtypes, including UPS, and showed associations with survival in STS. Furthermore, we conducted a bibliographic survey in terms of cancer progression for the 25 identified genes, and substantial evidence was uncovered in the literature. These genes were roughly classified into 4 types, namely, hypoxia-related genes, cell cycle- and DNA repair-related genes, growth factor signal transduction-related genes, and other genes. \( \text{STAT1} \) showed a statistically significant association with the survival of UPS patients (logrank adjusted \( p = 0.00299 \)). Although only a few studies have investigated the association of these genes with survival in STS, many recent studies have reported that these genes are prognostic factors and/or therapeutic targets in other types of cancers. Therefore, these results suggest that our combination method is capable of identifying genes that are potential prognostic/predictive factors and/or therapeutic targets in STS and possibly in other cancers. These disease-associated genes deserve further preclinical and clinical validation.

Supporting Information

Figure S1 The pathways predicted by STRING from the 25 selected genes.

Table S1 Clinical data of the 88 patients with soft tissue sarcoma. UPS: undifferentiated pleomorphic sarcoma, MFS: myxofibrosarcoma, LMS: leiomyosarcoma, SF: fibrosarcoma, MPNST: malignant peripheral nerve sheath tumor. Tumor metastasis indicates the incidence of tumor metastasis in UPS patients.

Table S2 The MIM number list.

Table S3 Selected Affymetrix probe IDs.

Table S4 Information on PCA, including the eigenvector, standard deviation, proportion of variance, and cumulative proportion for 29 probe sets. PCA: principal component analysis, PC: principal components.
Table S5  Information on PCA, including the eigenvector, standard deviation, proportion of variance, and cumulative proportion for 9 probe sets. PCA: principal component analysis, PC: principal components.

Table S6  Survival analysis in UPS using the logrank test. Adjusted \( p \) values were calculated using the permutation test (100,000 repeats) from logrank \( p \) values.

Table S7  Gene or pathway annotations and likelihood as prognostic/predictive factors and/or therapeutic targets. Adjusted \( p \) values were calculated using the permutation test (100,000 repeats) from logrank \( p \) values.

Table S8  Pathway analysis in IntPath. \( k \): genes from the overlap between genes in the list and genes in the pathway; \( n \): the number of genes in the input gene list; \( m \): the number of genes in the identified pathways; \( N \): the total number of genes. The \( p \) values were calculated using the hypergeometric test; the \( q \) values were calculated from the \( p \) values using the Benjamini-Hochberg (BH) method.

References

1. Trojani M, Contesso G, Cointe JM, Rousse J, Bui NB, et al. (1984) Soft-tissue sarcomas of adults; study of pathologic prognostic features and definition of a histopathogenic grading system. J Cancer 33: 37–42.
2. Guillou L, Cointe JM, Bonchon F, Nguyen BB, Terrier P, et al. (1997) Comparative study of the National Cancer Institute and French Federation of Cancer Centers Sarcoma Group grading systems in a population of 410 adult patients with soft tissue sarcoma. J Clin Oncol 15: 350-362.
3. Clark J, Roesquez PJ, Crew AJ, Gill S, Shipton J, et al. (1994) Identification of novel genes, SVT and SNX, involved in the \( k \), \( l \), \( p \), \( q \), \( r \), \( s \), \( t \), \( u \), \( v \), \( w \) translocation found in human synovial sarcoma. Nat Genet 7: 502–508.
4. Antoncuse CR, Elahi A, Humphrey M, Liu MY, Healey JH, et al. (2000) Specificity of TLS-CHOP rearrangement for classic myoid/rond cell liposarcoma: absence in predominantly myoid well-differentiated liposarcomas. J Mol Diagn 2: 132–138.
5. Kohno T, Ichikawa H, Totsuki Y, Yasuda K, Hiramoto M, et al. (2012) KIF5B-RET fusions in lung adenocarcinoma. Nat Med 18: 375–377.
6. Lux ML, Rubin BP, Baile TL, Chen CJ, Macure T, et al. (2000) KIT extracellular and kinase domain mutations in gastrointestinal stromal tumors. Am J Pathol 156: 791–795.
7. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, et al. (2008) Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455: 1069–1075.
8. Ando T, Suguro M, Hanai T, Kobayashi T, Honda H, et al. (2002) Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma. Jpn J Cancer Res 93: 1207–1212.
9. Takahashi H, Manabe K, Ando T, Kobayashi T, Honda H (2004) Prognostic predictor with multiple fuzzy neural models using expression profiles from DNA microarray for metastases of breast cancer. J Biosci Bioeng 98: 193–199.
10. Takahashi H, Honda H (2006) Modified signal-to-noise: a new simple and practical gene filtering approach based on the concept of projective adaptive resonance theory (PART) filtering method. Bioinformatics 22: 1662–1664.
11. Takahashi H, Kobayashi T, Honda H (2005) Construction of robust prognostic predictors by using projective adaptive resonance theory as a gene filtering method. Bioinformatics 21: 179–186.
12. Kawaiura T, Takahashi H, Honda H (2008) Proposal of new gene filtering method, BagPART, for gene expression analysis with small sample. J Biosci Bioeng 105: 81–84.
cancer from gene expression based on boosting and projective adaptive resonance theory. J Biosci Bioeng 102: 46–52.
38. Matsuo N, Hase M, Makino M, Takahashi H, Banno H (2009) Identification of ENHANCER OF SHOOT REGENERATION 1-upregulated genes during in vitro shoot regeneration. Plant Biotechnol 36: 385–393.
39. Yajima J, Kumazaka MY, Naito Y, Yoshikawa T, Takahashi H, et al. (2012) Reduced GNG2 expression levels in mouse malignant melanomas and human melanoma cell lines. Am J Cancer Res 2: 322–329.
40. Sano M, Aoyagi K, Takahashi H, Kawamura T, Mahuchi T, et al. (2010) Forkhead box A1 transcriptional pathway in KRT7-expressing esophageal squamous cell carcinomas with extensive lymph node metastasis. Int J Oncol 36: 321–330.
41. Chiba Y, Mineta K, Hirai MY, Suzuki Y, Kanaya S, et al. (2013) Changes in mRNA stability associated with cold stress in Arabidopsis cells. Plant Cell Physiol 54: 180–194.
42. Nakagawa A, Takahashi H, Kojima S, Sato N, Ohga K, et al. (2012) Berberine enhances defects in the establishment of leaf polarity in asymmetric leaves1 and asymmetric leaves2 of Arabidopsis thaliana. Plant Mol Biol 79: 569–581.
43. Yoshimura K, Morii T, Yokoyama K, Koike Y, Tanabe N, et al. (2011) Identification of alternative splicing events regulated by an Arabidopsis serine/arginine-like protein, atSRF-A, in response to high-light stress using a tiling array. Plant Cell Physiol 52: 1786–1805.
44. Portal D, Zhou H, Zhao B, Kharchenko PV, Lowry E, et al. (2013) Epstein-Barr virus nuclear antigen leader protein localizes to promoters and enhancers with increased coverage. BMC Syst Biol 7 Suppl 6: S6.
45. Takahashi H, Iwakawa H, Nakao S, Ojio T, Morishita R, et al. (2008) Dual regulation of ETTIN (ARF3) gene expression by AS1–AS2, which maintains the DNA methylation level, is involved in stabilization of leaf adaxial-abaxial patterning in Arabidopsis. Development 140: 1958–1969.
46. Takahashi H, Nakayama R, Hayashi S, Nemoto T, Murase Y, et al. (2013) Meta-Phage migration inhibitory factor and stearoyl-CoA desaturase 1: potential prognostic markers for soft tissue sarcomas based on bioinformatics analyses. PLoS One 8: e78250.
47. Yoshida T, Otto H, Kuchiba A, Saeke N, Sakamoto H (2010) Genome-wide germ-line analyses on cancer susceptibility and GaMDiBj database: Gastric cancer as an example. Cancer Sci 101: 1582–1589.
48. Mantel N (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 50: 163–170.
49. Hasegawa T, Yokoyama R, Lee YH, Shimoda T, Bepu Y, et al. (2000) Prognostic relevance of a histological grading system using MIB-1 for adult soft-tissue sarcoma. Oncology 58: 66–74.
50. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Amer Statist Assn 53: 457–481.
51. Flick P, Amode MR, Barrell D, Beal K, Bills K, et al. (2014) Ensembl 2014. Nucleic Acids Res 42: D749–D755.
52. Pearson K (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. Philos Mag 2 559–572.
53. Mantel N (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 50: 163–170.
54. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc serB 57: 289–300.
55. Chibon F, Lagarde P, Salas S, Perot G, Brouste V, et al. (2010) Validated prediction of H. sapiens-M. tuberculosis H37Rv protein-protein interactions. BMC Bioinformatics 11: 321–330.
56. Fleck P, Amode MR, Barrell D, Beal K, Bills K, et al. (2014) Ensembl 2014. Nucleic Acids Res 42: D749–D755.
57. Pearson K (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. Philos Mag 2 559–572.
58. Zhou H, Gao S, Nguyen NN, Fan M, Jin J, et al. (2014) Stringent homology-based prediction of H. sapiens- M. tuberculosis H37Rv protein-protein interactions. Biol Direct 9: 5.
59. Zhou H, Rezaei J, Hugo W, Gao S, Jin J, et al. (2013) Stringent DDE-based prediction of H. sapiens-M. tuberculosis H37Rv protein-protein interactions. BMC Syst Biol 7 Suppl 6: S6.
60. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41: D808–D815.