TABLE 3. Genomic Breakpoints of Chimeric Transcripts

	Genomic breakpoints[a]						
	5′-partner gene			3′-partner gene			
Sample	Symbol	Chr	Genomic position	Symbol	Chr	Genomic position	Flanking sequences[b]
K6	*POLR2G*	11	62530558	*CYP1A2*	15	75045983	TAGTCTCTCGGAAGATCTGGGTTGGGTTCT|GAGAATTGCTTGAACTCTGGAGGTAGAGGC
K7	*AC010724.1*	15	83207075	*CPEB1*	15	83219352	GAGATTATTGAAGTAGATCCTGACACTAAG|GAAATTGGCTCCTCTCTTGTAACTTCTGCC
K9	*SEMA6A*	5	115796806	*CAMK4*	5	110823275	CGTAAGAAATTTGGTACATAAGCTGGTATT|TTAATCCAATTCATCCAAATTATTCTATCG
K13	*ASAP1*	8	131070249	*ADCY8*	8	131862252	GGCAGACAACGATGACGAGCTCACATTCAT|TGCAAAGTTTCTCAATAGAGAGAGTGCTCT
K15	*CPSF3*	2	9578689	*ASAP2*	2	9532071	ACCCTGTCACCCAGGCTGGAGTGTGGTGGC|ACAATCATGGCTCACTGCAGCCTCCAACTC

[a] National Center for Biotechnology Information Database (Genome Build 37).
[b] The genomic breakpoints are indicated by a vertical bar.

Human hematologic (Shima and Kitabayashi, 2011) and soft tissue malignancies (Cantile et al., 2013), prostatic adenocarcinoma (Tomlins et al., 2005), and distinct subtypes of lung adenocarcinoma (Soda et al., 2007; Kohno et al., 2012; Takeuchi et al., 2012) show "addiction" for gene fusion events. Although their incidence is low, fusion events involving the transcription factor *TFE3* gene have been reported in RCCs: RCC associated with Xp11.2 translocation, which harbors *TFE3* fusion, is considered to represent a distinct subtype according to the World Health Organization (WHO) classification (Eble et al., 2004). Moreover, fusion events including anaplastic lymphoma kinase (*ALK*), such as *TMP3-ALK*, *EML4-ALK*, and *VCL-ALK* fusion, have been reported in a distinct group of RCCs, including so-called "unclassified RCC" and papillary RCC in adults (Sugawara et al., 2012) and pediatric RCCs associated with the sickle cell trait (Debelenko et al., 2011; Mariño-Enríquez et al., 2011), based on fluorescence in situ hybridization (FISH) and immunohistochemistry. These findings have prompted us to perform comprehensive exploration of chimeric transcripts in the most common subtype, clear cell RCC, using next-generation sequencing technology. In the present study, to clarify the participation of expression of chimeric transcripts in renal carcinogenesis, whole transcriptome analysis was performed using tissue specimens of 68 clear cell RCCs in adults.

MATERIALS AND METHODS

Patients and Tissue Samples

The initial cohort subjected to whole transcriptome analysis comprised 68 samples of cancerous tissue (T) and 11 samples of non-cancerous renal cortex tissue (N) obtained from materials that had been surgically resected from 68 patients with primary clear cell RCCs. There were 49 men and 19 women with a mean (±standard deviation) age of 62.3 ± 11.0 years (range, 36 to 85 years). All patients underwent nephrectomy at the National Cancer Center Hospital, Tokyo, and had not received any preoperative treatment. Two expert pathologists specializing in genitourinary pathology, E.A. and Y.K., examined all histological slides and performed histological diagnosis in accordance with the WHO classification (Eble et al., 2004). All the tumors were graded on the basis of previously described criteria (Fuhrman et al., 1982) and classified according to the macroscopic configuration (Arai et al., 2006) and the pathological Tumor-Node-Metastasis (TNM) classification (Sobin et al., 2009). As a positive control for chimeric transcript detection, two T samples showing histological findings compatible with Xp11.2 translocation RCC based on the WHO criteria were also subjected to whole transcriptome analysis. For comparison, three T samples of papillary RCCs diagnosed in accordance with the WHO criteria were also subjected to whole transcriptome analysis.

The second cohort subjected to quantitative reverse transcription-polymerase chain reaction (RT-PCR) analysis comprised 26 paired T and N samples obtained from materials that had been surgically resected from 26 other patients with primary clear cell RCCs. These patients comprised 17 men and nine women with a mean (±standard

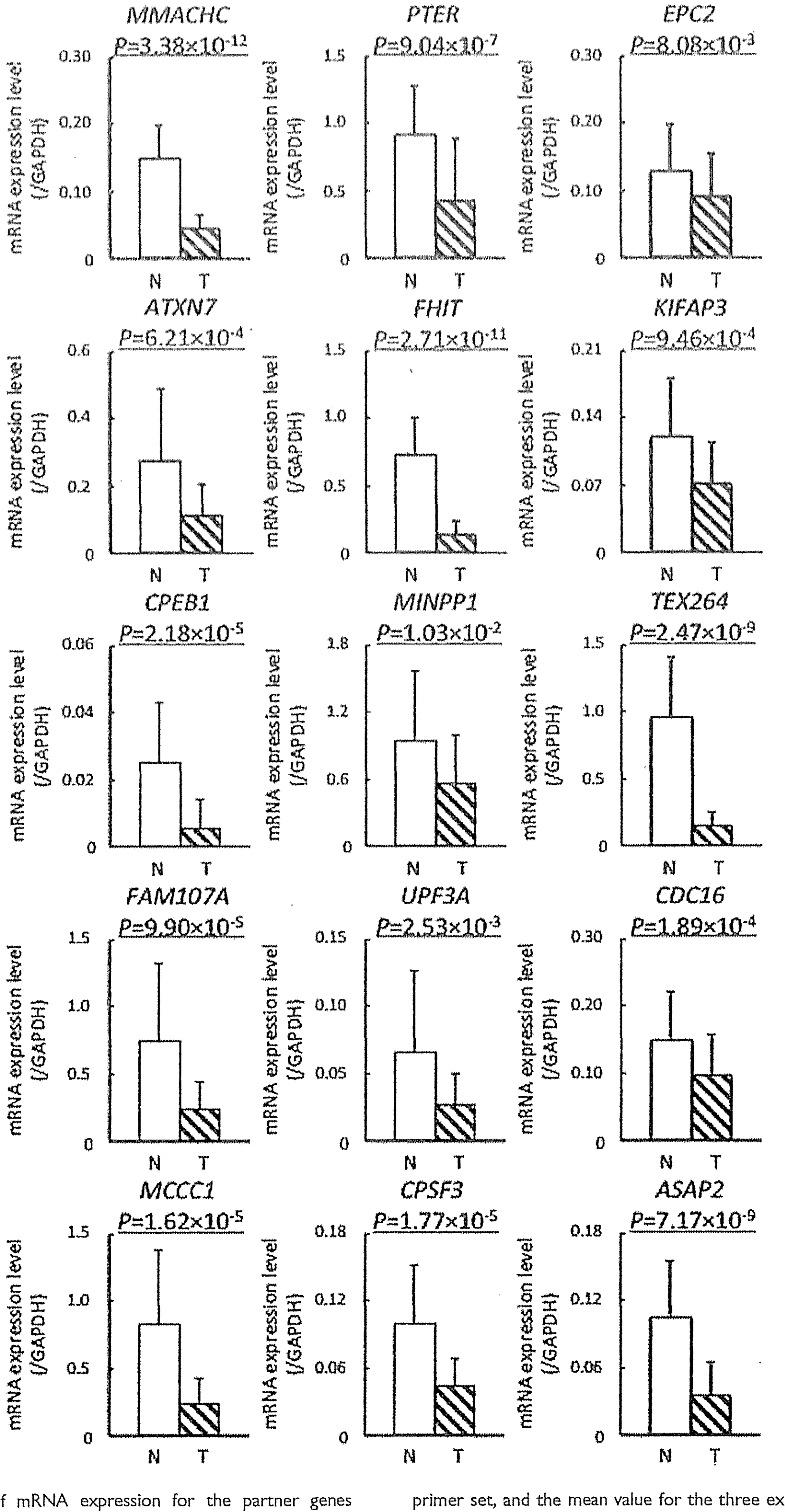

Figure 1. Levels of mRNA expression for the partner genes involved in chimeric transcripts in 26 paired samples of tumorous tissue (T) and non-cancerous renal cortex tissue (N) in the second cohort. mRNA expression was analyzed using custom TaqMan Gene Expression Assays on the 7500 Fast Real-Time PCR System (Life Technologies) employing the relative standard curve method. The probes and PCR primer sets used are summarized in Supporting Information Table S6. Experiments were performed in triplicate for each sample-primer set, and the mean value for the three experiments was used as the CT value. All CT values were normalized to that of *GAPDH* in the same sample. Levels of mRNA expression for the *MMACHC, PTER, EPC2, ATXN7, FHIT, KIFAP3, CPEB1, MINPP1, TEX264, FAM107A, UPF3A, CDC16, MCCC1, CPSF3*, and *ASAP2* genes were significantly reduced in T samples (shaded column) relative to N samples (white column). Bar, standard deviation.

TABLE 4. Correlations Between Levels of mRNA Expression for Each of the Partner Genes Involved in Chimeric Transcripts in Tumorous Tissue Samples and Clinicopathological Parameters Reflecting Tumor Aggressiveness in the Second Cohort

		MMACHC		*PTER*		*EPC2*		*ATXN7*	
Clinicopathological parameters	Number of tumors	Expression[a]	*P*	Expression[a]	*P*	Expression[a]	*P*	Expression[a]	*P*
Macroscopic configuration[b]									
Type 1	13	0.0528 ± 0.0226	1.61×10^{-2}[c]	0.558 ± 0.557	1.88×10^{-1}[c]	0.114 ± 0.065	1.23×10^{-1}[c]	0.125 ± 0.099	2.12×10^{-1}[c]
Type 2	5	0.0250 ± 0.0116		0.324 ± 0.400		0.0546 ± 0.0480		0.0598 ± 0.0374	
Type 3	8	0.0384 ± 0.0202		0.277 ± 0.248		0.0784 ± 0.0596		0.128 ± 0.099	
Histological grades[d,e]									
G1	1	0.086	5.05×10^{-2}[c]	0.574	1.15×10^{-1}[c]	0.246	4.65×10^{-3}[c]	0.228	1.64×10^{-1}[c]
G2	6	0.0443 ± 0.0222		0.530 ± 0.608		0.112 ± 0.083		0.134 ± 0.144	
G3	11	0.0489 ± 0.0233		0.498 ± 0.479		0.106 ± 0.033		0.124 ± 0.082	
G4	8	0.0285 ± 0.0105		0.232 ± 0.309		0.0373 ± 0.0185		0.0696 ± 0.0311	
Vascular involvement[f]									
Negative	13	0.0533 ± 0.0254	2.56×10^{-2}[g]	0.571 ± 0.433	8.60×10^{-3}[g]	0.125 ± 0.067	7.24×10^{-3}[g]	0.139 ± 0.100	5.01×10^{-2}[g]
Positive	13	0.0327 ± 0.0132		0.282 ± 0.452		0.0586 ± 0.0383		0.0876 ± 0.078	
Growth pattern[e]									
Expansive	21	0.046 ± 0.0239	2.00×10^{-1}[g]	0.482 ± 0.492	1.57×10^{-1}[g]	0.0954 ± 0.0659	7.53×10^{-1}[g]	0.121 ± 0.100	7.05×10^{-1}[g]
Infiltrative	5	0.0306 ± 0.0077		0.194 ± 0.129		0.0762 ± 0.0541		0.0836 ± 0.035	
Tumor necrosis									
Negative	16	0.0503 ± 0.0242	1.69×10^{-2}[g]	0.548 ± 0.513	1.44×10^{-2}[g]	0.120 ± 0.064	7.13×10^{-4}[g]	0.142 ± 0.106	2.68×10^{-2}[g]
Positive	10	0.0314 ± 0.0134		0.232 ± 0.277		0.0472 ± 0.0274		0.0672 ± 0.0281	
Renal pelvic invasion									
Negative	23	0.0451 ± 0.0230	1.34×10^{-1}[g]	0.469 ± 0.471	8.46×10^{-3}[g]	0.0973 ± 0.0649	1.57×10^{-1}[g]	0.120 ± 0.095	2.11×10^{-1}[g]
Positive	3	0.0270 ± 0.0030		0.101 ± 0.018		0.0487 ± 0.0218		0.0620 ± 0.0265	
Distant metastasis									
Negative	24	0.0450 ± 0.0221	2.46×10^{-2}[g]	0.413 ± 0.460	4.98×10^{-1}[g]	0.0974 ± 0.0625	5.54×10^{-2}[g]	0.120 ± 0.093	3.69×10^{-2}[g]
Positive	2	0.0185 ± 0.0007		0.589 ± 0.572		0.0235 ± 0.0021		0.0365 ± 0.0106	
Pathological TNM stage[h]									
Stage I	13	0.0430 ± 0.0166	5.54×10^{-2}[c]	0.561 ± 0.573	4.69×10^{-1}[c]	0.104 ± 0.056	1.78×10^{-1}[c]	0.112 ± 0.095	1.87×10^{-1}[c]
Stage II	3	0.0277 ± 0.0155		0.204 ± 0.232		0.0607 ± 0.0503		0.112 ± 0.099	
Stage III	8	0.0549 ± 0.0287		0.251 ± 0.163		0.100 ± 0.078		0.135 ± 0.097	
Stage IV	2	0.0185 ± 0.0007		0.589 ± 0.572		0.0235 ± 0.0021		0.0365 ± 0.0106	

		FHIT		*KIFAP3*		*CPEB1*		*TEX264*	
Clinicopathological parameters	Number of Tumors	Expression[a]	*P*	Expression[a]	*P*	Expression[a]	*P*	Expression[a]	*P*
Macroscopic configuration[b]									
Type 1	13	0.177 ± 0.125	4.73×10^{-2}[c]	0.0884 ± 0.0432	3.82×10^{-2}[c]	0.00369 ± 0.00572	3.31×10^{-1}[c]	0.155 ± 0.111	9.22×10^{-1}[c]
Type 2	5	0.0782 ± 0.0187		0.0362 ± 0.0222		0.0058 ± 0.0102		0.142 ± 0.104	
Type 3	8	0.100 ± 0.069		0.0675 ± 0.0345		0.0084 ± 0.0113		0.143 ± 0.086	
Histological grades[d,e]									
G1	1	0.280	6.62×10^{-3}[c]	0.118	1.18×10^{-1}[c]	0.004	4.23×10^{-2}[c]	0.148	9.75×10^{-2}[c]
G2	6	0.211 ± 0.172		0.0862 ± 0.0535		0.00383 ± 0.00722		0.233 ± 0.154	
G3	11	0.130 ± 0.044		0.0718 ± 0.0259		0.00173 ± 0.00168		0.142 ± 0.068	
G4	8	0.0631 ± 0.0308		0.0556 ± 0.0481		0.0123 ± 0.0119		0.0955 ± 0.0427	
Vascular involvement[f]									
Negative	13	0.178 ± 0.126	1.20×10^{-2}[g]	0.0839 ± 0.0384	5.68×10^{-2}[g]	0.00492 ± 0.00742	1.00[g]	0.155 ± 0.112	6.50×10^{-1}[g]
Positive	13	0.0898 ± 0.0503		0.0599 ± 0.0421		0.00615 ± 0.00978		0.143 ± 0.088	
Growth pattern[e]									
Expansive	21	0.150 ± 0.109	4.09×10^{-2}[g]	0.0752 ± 0.0446	5.69×10^{-1}[g]	0.00433 ± 0.00664	1.05×10^{-1}[g]	0.158 ± 0.104	3.40×10^{-1}[g]
Infiltrative	5	0.0660 ± 0.0410		0.0580 ± 0.0207		0.0106 ± 0.0139		0.112 ± 0.066	
Tumor necrosis									
Negative	16	0.174 ± 0.114	5.55×10^{-4}[g]	0.0820 ± 0.0381	3.09×10^{-2}[g]	0.00275 ± 0.00449	4.08×10^{-2}[g]	0.183 ± 0.110	6.05×10^{-3}[g]
Positive	10	0.0699 ± 0.0316		0.0558 ± 0.0431		0.0100 ± 0.0115		0.0943 ± 0.0395	
Renal pelvic invasion									
Negative	23	0.143 ± 0.107	1.34×10^{-1}[g]	0.0763 ± 0.0420	6.38×10^{-2}[g]	0.00461 ± 0.00639	5.94×10^{-1}[g]	0.156 ± 0.103	3.12×10^{-1}[g]
Positive	3	0.0643 ± 0.0397		0.0380 ± 0.0044		0.0127 ± 0.0193		0.0922 ± 0.0222	
Distant metastasis									
Negative	24	0.141 ± 0.106	5.54×10^{-2}[g]	0.0755 ± 0.0409	3.69×10^{-2}[g]	0.00458 ± 0.00790	5.54×10^{-2}[g]	0.156 ± 0.100	5.54×10^{-2}[g]
Positive	2	0.0520 ± 0.0014		0.0285 ± 0.0078		0.0170 ± 0.0099		0.0638 ± 0.0021	
Pathological TNM stage[h]									
Stage I	13	0.161 ± 0.124	1.40×10^{-1}[c]	0.0730 ± 0.0375	7.75×10^{-2}[c]	0.00300 ± 0.00478	1.15×10^{-1}[c]	0.153 ± 0.111	2.26×10^{-1}[c]
Stage II	3	0.0893 ± 0.0302		0.0400 ± 0.0271		0.00100 ± 0.00100		0.195 ± 0.116	
Stage III	8	0.127 ± 0.090		0.0930 ± 0.0446		0.0085 ± 0.0117		0.147 ± 0.083	
Stage IV	2	0.0520 ± 0.0014		0.0285 ± 0.0078		0.0170 ± 0.0099		0.0638 ± 0.0021	

		FAM107A		*CDC16*		*CPSF3*		*ASAP2*	
Clinicopathological parameters	Number of Tumors	Expression[a]	*P*	Expression[a]	*P*	Expression[a]	*P*	Expression[a]	*P*
Macroscopic configuration[b]									
Type 1	13	0.312 ± 0.184	5.51×10^{-2}[c]	0.113 ± 0.054	1.35×10^{-1}[c]	0.0476 ± 0.0255	2.78×10^{-1}[c]	0.0455 ± 0.0346	1.85×10^{-1}[c]

TABLE 4. (Continued)

Clinicopathological parameters	Number of Tumors	*FAM107A* Expression[a]	*FAM107A* *P*	*CDC16* Expression[a]	*CDC16* *P*	*CPSF3* Expression[a]	*CPSF3* *P*	*ASAP2* Expression[a]	*ASAP2* *P*
Type 2	5	0.0986 ± 0.0779		0.0584 ± 0.0377		0.0260 ± 0.0151		0.0218 ± 0.0191	
Type 3	8	0.203 ± 0.242		0.0926 ± 0.0753		0.0465 ± 0.0286		0.0250 ± 0.0227	
Histological grades[d,e]									
G1	1	0.685	$\underline{2.14\times10^{-2\mathrm{c}}}$	0.172	$\underline{1.28\times10^{-2\mathrm{c}}}$	0.0613	$1.30\times10^{-1\mathrm{c}}$	0.112	$\underline{1.93\times10^{-2\mathrm{c}}}$
G2	6	0.209 ± 0.140		0.113 ± 0.083		0.0575 ± 0.0335		0.0399 ± 0.0320	
G3	11	0.313 ± 0.197		0.116 ± 0.049		0.0464 ± 0.0236		0.0411 ± 0.0252	
G4	8	0.100 ± 0.129		0.0475 ± 0.0206		0.0256 ± 0.0118		0.0122 ± 0.0068	
Vascular involvement[f]									
Negative	13	0.258 ± 0.182	$3.11\times10^{-1\mathrm{g}}$	0.123 ± 0.063	$\underline{2.56\times10^{-2\mathrm{g}}}$	0.0518 ± 0.0271	$1.13\times10^{-1\mathrm{g}}$	0.0508 ± 0.030	$\underline{7.95\times10^{-4\mathrm{g}}}$
Positive	13	0.217 ± 0.226		0.0696 ± 0.0458		0.0344 ± 0.0214		0.0185 ± 0.0203	
Growth pattern[e]									
Expansive	21	0.255 ± 0.211	$4.47\times10^{-1\mathrm{g}}$	0.0975 ± 0.0608	$8.01\times10^{-1\mathrm{g}}$	0.0437 ± 0.0263	$9.00\times10^{-1\mathrm{g}}$	0.0372 ± 0.0316	$4.09\times10^{-1\mathrm{g}}$
Infiltrative	5	0.166 ± 0.159		0.0914 ± 0.0662		0.0409 ± 0.0246		0.0240 ± 0.0211	
Tumor necrosis									
Negative	16	0.317 ± 0.205	$\underline{2.24\times10^{-3\mathrm{g}}}$	0.121 ± 0.063	$\underline{5.02\times10^{-3\mathrm{g}}}$	0.0535 ± 0.0266	$\underline{1.44\times10^{-2\mathrm{g}}}$	0.0475 ± 0.0317	$\underline{2.77\times10^{-3\mathrm{g}}}$
Positive	10	0.110 ± 0.117		0.0567 ± 0.0278		0.0266 ± 0.0115		0.0142 ± 0.0086	
Renal pelvic invasion									
Negative	23	0.259 ± 0.205	$7.85\times10^{-2\mathrm{g}}$	0.100 ± 0.063	$3.52\times10^{-1\mathrm{g}}$	0.0453 ± 0.0263	$3.95\times10^{-1\mathrm{g}}$	0.0368 ± 0.0313	$4.42\times10^{-1\mathrm{g}}$
Positive	3	0.0726 ± 0.0602		0.0650 ± 0.0128		0.0263 ± 0.0073		0.0179 ± 0.0058	
Distant metastasis									
Negative	24	0.256 ± 0.199	$5.54\times10^{-2\mathrm{g}}$	0.102 ± 0.059	$\underline{2.46\times10^{-2\mathrm{g}}}$	0.0457 ± 0.0249	$\underline{1.23\times10^{-2\mathrm{g}}}$	0.0363 ± 0.0307	$3.94\times10^{-1\mathrm{g}}$
Positive	2	0.0165 ± 0.0030		0.0245 ± 0.0007		0.0125 ± 0.0019		0.0151 ± 0.0005	
Pathological TNM stage[h]									
Stage I	13	0.246 ± 0.152	$2.72\times10^{-1\mathrm{c}}$	0.109 ± 0.053	$\underline{4.59\times10^{-2\mathrm{c}}}$	0.0445 ± 0.0279	$9.67\times10^{-2\mathrm{c}}$	0.0404 ± 0.0279	$1.86\times10^{-1\mathrm{c}}$
Stage II	3	0.209 ± 0.167		0.0513 ± 0.0307		0.0324 ± 0.0170		0.0103 ± 0.0082	
Stage III	8	0.290 ± 0.284		0.110 ± 0.071		0.0526 ± 0.0220		0.0394 ± 0.0376	
Stage IV	2	*0.0165 ± 0.0030*		*0.0245 ± 0.0007*		*0.0125 ± 0.0019*		*0.0151 ± 0.0005*	

[a] Average mRNA levels/GAPDH ± standard deviation.
[b] Macroscopic configuration was evaluated on the basis of previously described criteria (Arai et al., 2006).
[c] Kruskal–Wallis test. *P* values of < 0.05 are underlined.
[d] All the tumors were graded on the basis of previously described criteria (Fuhrman et al., 1982).
[e] If the tumor showed heterogeneity, the most aggressive features of the tumor were described.
[f] The presence or absence of vascular involvement was examined microscopically on slides stained with hematoxylin-eosin and elastica van Gieson.
[g] Mann–Whitney *U* test. *P* values of <0.05 are underlined.
[h] All the tumors were classified according to the pathological Tumor-Node-Metastasis classification (Sobin et al., 2009). Although no significant correlation between expression of any of the 26 chimeric transcripts and clinicopathological parameters was observed in the initial cohort (Supporting Information, Table S4), downregulation of mRNA levels for each of the partner genes did show significant correlations with the above clinicopathological parameters in the second cohort.

deviation) age of 57.1 ± 10.8 years (range, 33–81 years). Copy number analysis using the HumanOmni1-Quad BeadChip (Illumina, San Diego, CA) and Global Parameter Hidden Markov Model (http://bioinformatics.ustc.edu.cn/gphmm/; Li et al., 2011) revealed copy number alterations in chromosome 3 in all 91 clear cell RCCs in the initial and second cohorts (with three exceptions, Supporting Information Tables S1 and S2 and Supporting Information Fig. S1). These findings were considered to be the hallmark of clear cell RCCs in the initial and second cohorts. The clinicopathological parameters of RCCs belonging to the initial and second cohorts are summarized in Table 1.

Tissue specimens were taken and frozen immediately after surgical removal, and thereafter stored in liquid nitrogen until use. These tissue specimens were provided by the National Cancer Center Biobank, Tokyo. This study was approved by the Ethics Committees of the National Cancer Center and National Center for Global Health and Medicine, Tokyo, and was performed in accordance with the Declaration of Helsinki. All the patients provided written informed consent prior to inclusion in the study.

Whole Transcriptome Analysis

Total RNA was isolated using TRIzol reagent (Life Technologies, Carlsbad, CA). A total of 84 (73 T and 11 N) samples in the initial cohort were subjected to whole transcriptome analysis. Sequencing libraries were prepared from 1.0 to 2.5 μg of total RNA using an mRNA-Seq Sample Prep Kit or a TruSeq RNA Sample Prep Kit (Illumina), according to the manufacturer's standard protocols. An mRNA-Seq Sample Prep Kit was used for libraries of 35 (30 T and 5 N) samples, and these libraries were prepared using a procedure including a gel purification step, in which a

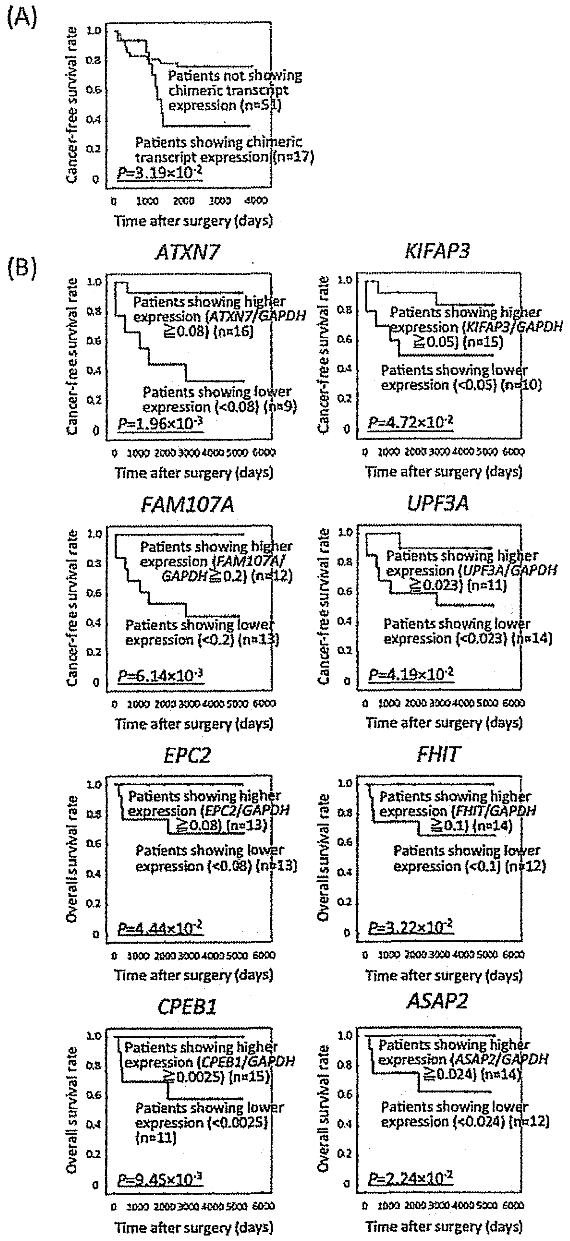

Figure 2. Kaplan–Meier survival curves of patients with clear cell RCCs in the initial (A) and second (B) cohorts. (A) Expression of any of 26 chimeric transcripts was inversely correlated with the cancer-free survival rate of patients in the initial cohort (the log-rank test, $P = 3.19 \times 10^{-2}$). (B) ROC curves were generated for levels of mRNA expression of each partner gene of chimeric transcripts, and the thresholds were set to the top left corner of the graph (data not shown). Using these thresholds, Kaplan–Meier curves were generated. mRNA levels for the *ATXN7* ($P = 1.96 \times 10^{-3}$), *KIFAP3* ($P = 4.72 \times 10^{-2}$), *FAM107A* ($P = 6.14 \times 10^{-3}$), and *UPF3A* ($P = 4.19 \times 10^{-2}$) genes in T samples were inversely correlated with the cancer-free survival rate of patients who underwent complete resection ($n = 25$), whereas those for the *EPC2* ($P = 4.44 \times 10^{-2}$), *FHIT* ($P = 3.22 \times 10^{-2}$), *CPEB1* ($P = 9.45 \times 10^{-3}$), and *ASAP2* ($P = 2.24 \times 10^{-2}$) genes were inversely correlated with the overall survival rate of all patients ($n = 26$) in the second cohort.

fraction of 250–300 bp (insert size: 150–200 bp) was collected. A TruSeq RNA Sample Prep Kit was used for libraries of the other 49 (43 T and 6 N) samples, and these libraries were prepared without gel purification. The resulting libraries were subjected to paired-end sequencing of 50-base reads on a GAIIx or HiSeq2000 sequencer (Illumina).

Detection of Chimeric Transcripts

To avoid multiple counting of each chimeric transcript, RNA sequencing data were used after removal of paired-end reads with the identical nucleotide sequence, which had probably been derived from PCR duplicates during library preparation. For prediction of chimeric transcripts, the deFuse program version 0.4.3 was used (McPherson et al., 2011). After applying default filtering of this program, potential alternative splicing and read-through products that the program predicted were eliminated, and candidates that had exon boundary junctions were selected. Finally, we discarded candidates that were also predicted from the data of 11 N samples.

RT-PCR and Sanger Sequencing

cDNA was reverse-transcribed from the total RNA (500 ng) of the initial cohort samples, in which candidate chimeric transcripts were detected by whole transcriptome analysis, using random primers and Superscript III RNase H^{-}Reverse Transcriptase (Life Technologies). cDNA (corresponding to 10 ng total RNA) was subjected to PCR amplification using an optimal DNA polymerase among AmpliTaq Gold DNA Polymerase (Life Technologies), HotStar Taq DNA polymerase (Qiagen, Hilden, Germany) or KAPA Taq DNA Polymerase (KAPA Biosystems, Woburn, MA). The PCR products were separated electrophoretically on 2% agarose gel and stained with ethidium bromide to confirm that specific products of the size estimated on the basis of whole transcriptome analysis were obtained, and that no nonspecific products appeared on amplification. The PCR products were then directly sequenced in both directions using the same primers with the BigDye Terminator v3.1 Cycle Sequencing kit and an ABI 3130xl DNA Sequencer (Life Technologies).

Genomic PCR and Sanger Sequencing

High-molecular-weight genomic DNA was extracted from the initial cohort samples, in which candidate chimeric transcripts were verified by the above RT-PCR and Sanger sequencing, using phenol–chloroform followed by dialysis. Genomic

DNA (10 ng) was subjected to PCR amplification using an optimal DNA polymerase among AmpliTaq Gold DNA Polymerase, Platinum Taq DNA-polymerase high fidelity (Life Technologies), HotStar Taq DNA polymerase (Qiagen) or KAPA Taq DNA Polymerase (KAPA Biosystems). The PCR products were separated electrophoretically on 1% agarose gel and stained with ethidium bromide to confirm that no nonspecific products appeared on amplification. The PCR products were then directly sequenced in both directions using the same primers with the BigDye Terminator v3.1 Cycle Sequencing kit and an ABI 3130xl DNA Sequencer (Life Technologies).

Quantitative RT-PCR Analysis

cDNA was reverse transcribed from total RNA (500 ng) of the 26 paired T and N samples of the second cohort using random primers and Superscript III RNase H$^-$Reverse Transcriptase (Life Technologies). mRNA expression was analyzed using custom TaqMan Gene Expression Assays and TaqMan Fast Advanced Master Mix (Life Technologies) on a 7500 Fast Real-Time PCR System (Life Technologies) employing the relative standard curve method. Experiments were performed in triplicate for each sample-primer set, and the mean value for the three experiments was used as the CT value. All CT values were normalized to that of *GAPDH* in the same sample.

Statistics

Differences in clinicopathological parameters between the initial and second cohorts were assayed by Mann–Whitney *U* test and Fisher's exact test. Differences in the levels of mRNA expression between N and T samples were examined by Mann–Whitney *U* test. Correlations between levels of mRNA expression and clinicopathological parameters were assayed by Kruskal–Wallis test and Mann–Whitney *U* test. Receiver operating characteristic (ROC) curves were generated for the levels of mRNA expression of each partner gene involved in the chimeric transcripts, and the thresholds were set at the top left corner of the graph. Subsequently, the impact of chimeric transcript expression and downregulation of mRNA levels for each partner gene on patient outcome was analyzed by the Kaplan–Meier method using the set thresholds and the log-rank test. Differences at $P < 0.05$ were considered to be significant.

RESULTS

Identification of Novel Chimeric Transcripts in RCCs of the Initial Cohort

We performed RNA sequencing of 68 T samples (K1 to K68) and 11 N samples in the initial cohort, and a T sample (K69) showing histological findings compatible with Xp11.2 translocation RCC. At least 30,000,000 reads (average read count 50,000,000) were obtained for each sample. The deFuse program version 0.4.3 (McPherson et al., 2011) provided 3,746 fusion gene candidates from the data obtained using the 69 T samples by applying default filtering. From those candidates, 95 were extracted by eliminating potential alternative splicing and read-through products that the program predicted, and by selecting candidates that had exon boundary junctions. Next, candidates that were predicted even from the data obtained using the 11 N samples were discarded, and finally 35 candidates were obtained. Three candidates were abandoned because of difficulty with the primer design and shortage of samples, and then RT-PCR analysis was performed for the 32 candidates in the same T sample. The PCR and sequencing primers used are shown in Supporting Information Table S3.

After a T sample (K96) of Xp11.2 translocation RCC and three T samples of papillary RCCs (K97 to K99) had been additionally analyzed for comparison, expression of 33 fusion transcripts (including two transcripts [*MMACHC-BX004987.7* and *TFE3-RBM10*] consisting of the same partner gene sets with a different exon boundary or a different transcriptional direction and three transcripts sharing a partner gene, *TFE3*) from the 61 genes was finally verified by RT-PCR, and the exon boundaries and flanking sequences were determined by Sanger sequencing analysis (Table 2 and Supporting Information Fig. S2).

Previously reported in-frame fusion transcripts including *TFE3* (Table 2B; Clark et al., 1997), which are attributable to translocation of the X chromosome, were detected in samples K69 and K96 showing histological findings compatible with Xp11.2 translocation RCC, indicating the reliability of our study. Other than *TFE3* fusion transcripts, three additional transcripts (*EEF2-ENHO*, *PARG-BMS1*, and *RAGE-EML1*, Table 2B) and one additional transcript (*DPP6-ACTR3B*, Table 2B), which have never been reported in RCCs, were also detected in the K69 and K96, respectively. *NONO-TFE3*, *PARG-BMS1*, *RAGE-EML1*, *RBM10-TFE3*, and *DPP6-ACTR3B* transcripts were predicted to generate in-frame chimeric proteins. These observations of additional chimeric transcripts in K69 and K96 were different from the previously reported characteristics of RCCs associated with Xp11.2 translocation [Pflueger et al. (2013) reported that expression of the *TMED6-COG8* chimeric transcript and higher expression levels of the *EEF1A2* and *CNTN3* genes characterize RCCs associated with Xp11.2 translocation]. K69 showed grade 3 histology, vascular involvement, and tumor necrosis in surgically resected materials, and the patient developed lymph node metastasis 6 months after surgery, whereas K96 showed grade 3 histology. Such phenotypes, especially those of K69, which are more aggressive than those generally described for RCCs with Xp11.2 translocation (Eble et al., 2004), may be attributable to expression of multiple additional chimeric transcripts. Conversely, in three papillary RCCs (K97 to K99) analyzed for comparison, no chimeric transcript was detected.

All 26 chimeric transcripts detected in the initial cohort of clear cell RCCs (Table 2A) have never been reported previously. Even though chimeric transcripts involving the *FHIT* and *TERT* genes have recently been sequenced by The Cancer Genome Atlas (TCGA, The Cancer Genome Atlas Research Network, 2013), the partner gene of *FHIT*, *FAM172A*, and that of *TERT*, *PDCD6*, listed in TCGA each differed from those (*ATXN7* and *TPPP*) in the present study. Each of the detected chimeric transcripts was expressed in a single clear cell RCC. *ANTXR1-GKN1*, *ERBB2-LTBP4*, *POLR2G-CYP1A2*, *AC010724.1-CPEB1*, and *CPSF3-ASAP2* chimeric transcripts were predicted to generate in-frame chimeric proteins, whereas other chimeric transcripts resulted in a premature stop codon in the 3′-partner gene or were generated in the untranslated regions.

The chimeric transcripts were expressed in 17 clear cell RCCs in the initial cohort [17/68 (Table 2), 25%]. Samples K1 and K5 had multiple chimeric transcripts (Table 2). No significant correlation between expression of any of 26 chimeric transcripts and clinicopathological parameters was observed in the initial cohort (Supporting Information Table S4). However, when examined individually, each clear cell RCC with chimeric transcripts showed tumor aggressiveness: e.g., K11 carrying a *TEX264-FAM107A* chimeric transcript showed a type 3 macroscopic configuration and K15 carrying a *CPSF3-ASAP2* chimeric transcript showed a type 3 macroscopic configuration, grade 4 histology, vascular involvement, an invasive growth pattern, and tumor necrosis. Moreover, expression of any of the 26 chimeric transcripts was inversely correlated with the cancer-free survival rate of patients in the initial cohort (the period covered ranged from 42 to 4,783 days [mean, 2,015 days]; log-rank test, $P = 3.19 \times 10^{-2}$; Fig. 2).

Identification of Genomic Breakpoints in RCCs of the Initial Cohort

Long-range genomic PCR and Sanger sequencing were performed for 17 clear cell RCCs (K1 to K17) harboring chimeric transcripts using the primers shown in Supporting Information Table S5. Genomic breakpoints for five chimeric transcripts, *POLR2G-CYP1A2*, *AC010724.1-CPEB1*, *SEMA6A-CAMK4*, *ASAP1-ADCY8*, and *CPSF3-ASAP2*, were successfully revealed, but the genomic PCR failed for the other transcripts. The genomic breakpoints for these five chimeric transcripts are summarized in Table 3.

Levels of mRNA Expression for the Genes Involved in Chimeric Transcripts

The levels of mRNA expression for 20 representative partner genes involved in chimeric transcripts in the initial cohort were quantitatively examined in 26 paired T and N samples in the second cohort. The probes and PCR primer sets used are shown in Supporting Information Table S6.

The levels of mRNA expression for the *MMACHC*, *PTER*, *EPC2*, *ATXN7*, *FHIT*, *KIFAP3*, *CPEB1*, *MINPP1*, *TEX264*, *FAM107A*, *UPF3A*, *CDC16*, *MCCC1*, *CPSF3*, and *ASAP2* genes were significantly reduced in T samples relative to the corresponding N samples (Fig. 1, Mann–Whitney *U* test, $P = 3.38 \times 10^{-12}$, $P = 9.04 \times 10^{-7}$, $P = 8.08 \times 10^{-3}$, $P = 6.21 \times 10^{-4}$, $P = 2.71 \times 10^{-11}$, $P = 9.46 \times 10^{-4}$, $P = 2.18 \times 10^{-5}$, $P = 1.03 \times 10^{-2}$, $P = 2.47 \times 10^{-9}$, $P = 9.90 \times 10^{-5}$, $P = 2.53 \times 10^{-3}$, $P = 1.89 \times 10^{-4}$, $P = 1.62 \times 10^{-5}$, $P = 1.77 \times 10^{-5}$, and $P = 7.17 \times 10^{-9}$, respectively). The levels of mRNA expression for the *MMACHC*, *PTER*, *EPC2*, *ATXN7*, *FHIT*, *KIFAP3*, *CPEB1*, *TEX264*, *FAM107A*, *CDC16*, *CPSF3*, and *ASAP2* genes in T samples in the second cohort were significantly correlated with clinicopathological parameters reflecting tumor aggressiveness, such as invasive macroscopic configuration, higher histological grades, vascular involvement, invasive growth pattern, tumor necrosis, renal pelvic invasion, distant metastasis, and higher TNM stages (Table 4). Moreover, mRNA levels for the *ATXN7* ($P = 1.96 \times 10^{-3}$), *KIFAP3* ($P = 4.72 \times 10^{-2}$), *FAM107A* ($P = 6.14 \times 10^{-3}$), and *UPF3A* ($P = 4.19 \times 10^{-2}$) genes in T samples were inversely correlated with the cancer-free survival rate, whereas those for the *EPC2* ($P = 4.44 \times 10^{-2}$), *FHIT* ($P = 3.22 \times 10^{-2}$), *CPEB1* ($P = 9.45 \times 10^{-3}$), and *ASAP2* ($P = 2.24 \times 10^{-2}$) genes were inversely correlated with the overall survival rate in the second cohort (the period covered ranged from 88 to 5,207 days [mean, 3,038 days], the log-rank test; Fig. 2).

DISCUSSION

To comprehensively explore chimeric transcripts in clear cell RCCs, whole transcriptome analysis was performed using tissue specimens. The significance of generation of chimeric transcripts has not been fully elucidated in adult solid tumors other than well-studied exceptions, such as sarcomas and adenocarcinomas of the prostate and the lung. Although previous reports of fusion events involving the *ALK* gene based on FISH and immunohistochemistry have been restricted to nonclear cell RCCs (Sugawara et al., 2012), when comprehensively explored using next-generation sequencing technology, chimeric transcripts were detected in 25% (17/68) of the clear cell RCCs. In some RCCs (K1 and K5), multiple chimeric transcripts were observed. Moreover, the genomic breakpoints revealed for five chimeric transcripts in clear cell RCCs indicate that such transcripts have actually arisen through genomic rearrangement. Gene fusion events may thus play a greater role in renal carcinogenesis than previously anticipated. Conversely, mechanisms other than genomic rearrangements (Yuan et al., 2013), e.g., trans-splicing (Li et al., 2008), may generate chimeric transcripts in which genomic breakpoints are not revealed.

The WHO classification defines RCC associated with Xp11.2 translocation, which involves *TFE3* fusion, as a distinct subtype (Eble et al., 2004). Diagnosis of RCC associated with Xp11.2 translocation depends on detection of *TFE3* protein overexpression using immunohistochemistry or detection of gene fusion using FISH and/or RT-PCR analysis (Green et al., 2013; Rao et al., 2013). The procedure for final diagnosis of RCC associated with Xp11.2 translocation differs from that for other RCC subtypes, such as clear cell RCC, papillary RCC, and chromophobe RCC, which generally can be diagnosed on the basis of histological observation. As RCC associated with Xp11.2 translocation and other RCCs were lumped into the same category as the RCC subtypes, the final diagnosis of RCC subtypes could not be made based solely on conventional histological examination of surgically resected or biopsy specimens. As the present comprehensive study demonstrated multiple chimeric transcripts in various RCCs, it seems that the use of Xp11.2 translocation as the only criterion for defining a distinct subtype of RCC may not be a rational approach. The classification of RCC subtypes should therefore be revised after a more comprehensive appraisal of correlations between histological features and genetic background.

All 26 chimeric transcripts detected in the initial cohort were novel chimeric transcripts that have never been reported previously in RCCs. However, only five of them were predicted to generate in-frame chimeric proteins in clear cell RCCs. Expression microarray analysis did not necessarily suggest prominent overexpression of in-frame chimeric transcripts in the initial cohort (data not shown). Moreover, in-frame chimeric transcripts observed in clear cell RCCs do not necessarily result in constitutive activation of protein kinases, which frequently cause addiction for gene fusion events. Conversely, many genes for which reduced expression and/or tumor suppressive function have been reported in human cancers were included in chimeric transcripts observed in the initial cohort. Therefore, we examined the levels of mRNA expression for 20 representative genes involved in chimeric transcripts in the second cohort (Supporting Information Table S6) and revealed significantly reduced mRNA expression of the *MMACHC, PTER, EPC2, ATXN7, FHIT, KIFAP3, CPEB1, MINPP1, TEX264, FAM107A, UPF3A, CDC16, MCCC1, CPSF3*, and *ASAP2* genes in T samples in the second cohort (Fig. 1).

It has been reported that reduced expression of the *MMACHC* gene, which participates in intracellular trafficking of cobalamin, can result in increased tumorigenicity and methionine dependence of cancer cells (Loewy et al., 2009). Although its implication in human cancers has been unclear, the *PTER* gene was first cloned as a rat homolog of bacterial phosphotriesterase, and its expression in the normal proximal tubules of the kidney has been reported (Davies et al., 1997). Single nucleotide polymorphism (SNP) of the *EPC2* gene has been reported to be associated with response to gemcitabine in human cancer cell lines (Jarjanazi et al., 2008). SNP of the *ATXN7* gene, which encodes a subunit of the GCN5 histone acetyltransferase-containing coactivator complex (Helmlinger et al., 2006), is reportedly associated with susceptibility to breast cancer (Milne et al., in press). The fragile *FHIT* gene, encompassing the chromosomal fragile site *FRA3B*, is a target of DNA damage-induced cancer initiation and progression through modulation of genomic stability (Karras et al., 2014). KIFAP3 is colocalized with KIF3, which participates in subcellular transport of several cancer-related proteins including beta-catenin and cadherins (Tanuma et al., 2009). Down regulation of *CPEB1*, which participates in the regulation of mRNA translation and processing of the 3′ untranslated region (Bava et al., 2013), has been reported in human cancers (Caldeira et al., 2012). As has been reported for the *PTEN* gene, somatic mutation and germline variants of the *MINPP1* gene, located in proximity to *PTEN* in 10q23.3, have been reported in patients with follicular thyroid tumors (Gimm et al., 2001). *FAM107A* was first identified in a commonly deleted region in 3p21 in RCCs (Wang et al., 2000), and transfection of this gene induces growth suppression and apoptosis of FAM107A-negative cancer cell lines (Wang et al., 2000; Liu et al., 2009). UPF3A is a crucial factor of nonsense-mediated decay, an RNA decay pathway that downregulates aberrant mRNAs (Chan et al., 2009). CDC16 is a component of the Anaphase Promoting Complex/Cyclosome (APC/C), which governs cell cycle progression and has crucial functions in maintaining genomic integrity and tumorigenesis (Zhang et al., 2014). Genetic imbalance at the *MCCC1* gene locus has recently been reported in clinical specimens of oral squamous cell carcinoma (Ribeiro et al., 2014). *CPSF3* is required for site-specific endonucleolytic cleavage and poly (A) addition (Keller and Minvielle-Sebastia, 1997) and directly interacts with (Zhu et al., 2009) tumor suppressor gene product CSR1 (Yu et al., 2006). The src homology 3 domain of the paxillin-binding protein (Kondo et al., 2000; Coutinho-Camillo et al., 2006), *ASAP2*, directly interacts with the SAMP repeat region of the *APC* tumor suppressor gene (Matsui et al., 2008).

Although the *TEX264* gene has been simply identified as one of the protein-encoding open reading frames deposited in a database (Lamesch et al., 2007), the above characteristics of each of the partner genes suggest that down-regulation of the *MMACHC, PTER, EPC2, ATXN7, FHIT, KIFAP3, CPEB1, MINPP1, FAM107A, UPF3A, CDC16, MCCC1, CPSF3*, and *ASAP2* genes may participate in renal carcinogenesis. Moreover, the levels of mRNA expression for many of the partner genes in T samples were significantly correlated with the clinicopathological aggressiveness of RCCs (Table 4) and were inversely correlated with the cancer-free and/or overall survival rates of patients with clear cell RCCs (Fig. 2), indicating that such reduced expression may continue to play a role in multistage malignant progression during renal carcinogenesis.

Even if the same chimeric transcripts detected in the initial cohorts had been expressed in the second cohort, quantitative RT-PCR analysis for each partner gene would not have evaluated chimeric transcripts lacking target exons (Supporting Information Table S6). Therefore, to reveal the presence or absence of the same chimeric transcripts detected in the initial cohort, RT-PCR analysis using total RNA samples and the primer sets indicated in Supporting Information Table S3 and long-range PCR analysis using genomic DNA samples and the primer sets described in Supporting Information Table S5 were performed in the second cohort. These analyses did not detect the same chimeric transcripts in the second cohort (data not shown). As all detected chimeric transcripts were expressed only in a single clear cell RCC in the initial cohort, it is possible that the same chimeric transcripts may have been absent in the second cohort. Downregulation of mRNA levels for each of the genes described in Figure 1 in the second cohort would have been attributable to mechanisms other than expression of chimeric transcripts, such as gene deletion, DNA methylation status around the promoter regions and/or alterations in the expression levels, and accessibility of transcription factors. In fact, silencing of the *MMACHC* (Loewy et al., 2009) and *CPEB1* (Caldeira et al., 2012) genes due to DNA methylation, and gene deletion and DNA methylation of *FHIT* (Karras et al., 2014), have been reported in human cancers. However, further studies are needed to reveal the mechanisms responsible for downregulation of each of the partner genes in the second cohort.

Conversely, it is feasible that dysfunction of each partner gene is induced by generation of chimeric transcripts in the initial cohort of clear cell RCCs, as such mechanisms of tumor suppressor gene functional impairment have been reported in adult malignancies such as acute myeloid leukemia (McNerney et al., 2013). Even though prominent overexpression and/or constitutive activation of growth factors and/or protein kinases due to gene fusion events is rare, generation of chimeric transcripts may participate in renal carcinogenesis through dysfunction of tumor-related genes.

REFERENCES

Arai E, Kanai Y. 2010. Genetic and epigenetic alterations during renal carcinogenesis. Int J Clin Exp Pathol 4:58–73.

Arai E, Kanai Y, Ushijima S, Fujimoto H, Mukai K, Hirohashi S. 2006. Regional DNA hypermethylation and DNA methyltransferase (DNMT) 1 protein overexpression in both renal tumors and corresponding nontumorous renal tissues. Int J Cancer 119: 288–296.

Baldewijns MM, van Vlodrop IJ, Vermeulen PB, Soetekouw PM, van Engeland M, de Bruïne AP. 2010. VHL and HIF signalling in renal cell carcinogenesis. J Pathol 221:125–138.

Bava FA, Eliscovich C, Ferreira PG, Miñana B, Ben-Dov C, Guigó R, Valcárcel J, Méndez R. 2013. CPEB1 coordinates alternative 3′-UTR formation with translational regulation. Nature 495:121–125.

Caldeira J, Simões-Correia J, Paredes J, Pinto MT, Sousa S, Corso G, Marrelli D, Roviello F, Pereira PS, Weil D, Oliveira C, Casares F, Seruca R. 2012. CPEB1, a novel gene silenced in gastric cancer: A Drosophila approach. Gut 61:1115–1123.

Cantile M, Marra L, Franco R, Ascierto P, Liguori G, De Chiara A, Botti G. 2013. Molecular detection and targeting of EWSR1 fusion transcripts in soft tissue tumors. Med Oncol 30:412.

Chan WK, Bhalla AD, Le Hir H, Nguyen LS, Huang L, Gécz J, Wilkinson MF. 2009. A UPF3-mediated regulatory switch that maintains RNA surveillance. Nat Struct Mol Biol 16:747–753.

Clark J, Lu YJ, Sidhar SK, Parker C, Gill S, Smedley D, Hamoudi R, Linehan WM, Shipley J, Cooper CS. 1997. Fusion of splicing factor genes PSF and NonO (p54nrb) to the TFE3 gene in papillary renal cell carcinoma. Oncogene 15:2233–2239.

Coutinho-Camillo CM, Salaorni S, Sarkis AS, Nagai MA. 2006. Differentially expressed genes in the prostate cancer cell line LNCaP after exposure to androgen and anti-androgen. Cancer Genet Cytogenet 166:130–138.

Dalgliesh GL, Furge K, Greenman C, Chen L, Bignell G, Butler A, Davies H, Edkins S, Hardy C, Latimer C, Teague J, Andrews J, Barthorpe S, Beare D, Buck G, Campbell PJ, Forbes S, Jia M, Jones D, Knott H, Kok CY, Lau KW, Leroy C, Lin ML, McBride DJ, Maddison M, Maguire S, McLay K, Menzies A, Mironenko T, Mulderrig L, Mudie L, O'Meara S, Pleasance E, Rajasingham A, Shepherd R, Smith R, Stebbings L, Stephens P, Tang G, Tarpey PS, Turrell K, Dykema KJ, Khoo SK, Petillo D, Wondergem B, Anema J, Kahnoski RJ, Teh BT, Stratton MR, Futreal PA. 2010. Systematic sequencing of renal carcinoma reveals inactivation of histone-modifying genes. Nature 463:360–363.

Davies JA, Buchman VL, Krylova O, Ninkina NN. 1997. Molecular cloning and expression pattern of rpr-1, a resiniferatoxin-binding, phosphotriesterase-related protein, expressed in rat kidney tubules. FEBS Lett 410:378–382.

Debelenko LV, Raimondi SC, Daw N, Shivakumar BR, Huang D, Nelson M, Bridge JA. 2011. Renal cell carcinoma with novel VCL-ALK fusion: New representative of ALK-associated tumor spectrum. Mod Pathol 24:430–442.

Eble JN, Togashi K, Pisani P. 2004. Renal cell carcinoma. In: Eble JN, Sauter G, Epstein JI, Sesterhenn IA, editors. World Health Organization classification of tumours. Pathology and genetics. Tumours of the urinary system and male genital organs. Lyon: IARC, pp. 10–43.

Fuhrman SA, Lasky LC, Limas, C. 1982. Prognostic significance of morphologic parameters in renal cell carcinoma. Am J Surg Pathol 6:655–663.

Gimm O, Chi H, Dahia PL, Perren A, Hinze R, Komminoth P, Dralle H, Reynolds PR, Eng C. 2001. Somatic mutation and germline variants of MINPP1, a phosphatase gene located in proximity to PTEN on 10q23.3, in follicular thyroid carcinomas. J Clin Endocrinol Metab 86:1801–1805.

Green WM, Yonescu R, Morsberger L, Morris K, Netto GJ, Epstein JI, Illei PB, Allaf M, Ladanyi M, Griffin CA, Argani P. 2013. Utilization of a TFE3 break-apart FISH assay in a renal tumor consultation service. Am J Surg Pathol 37:1150–1163.

Guo G, Gui Y, Gao S, Tang A, Hu X, Huang Y, Jia W, Li Z, He M, Sun L, Song P, Sun X, Zhao X, Yang S, Liang C, Wan S, Zhou F, Chen C, Zhu J, Li X, Jian M, Zhou L, Ye R, Huang P, Chen J, Jiang T, Liu X, Wang Y, Zou J, Jiang Z, Wu R, Wu S, Fan F, Zhang Z, Liu L, Yang R, Liu X, Wu H, Yin W, Zhao X, Liu Y, Peng H, Jiang B, Feng Q, Li C, Xie J, Lu J, Kristiansen K, Li Y, Zhang X, Li S, Wang J, Yang H, Cai Z, Wang J. 2012. Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. Nat Genet 44:17–19.

Helmlinger D, Hardy S, Abou-Sleymane G, Eberlin A, Bowman AB, Gansmüller A, Picaud S, Zoghbi HY, Trottier Y, Tora L, Devys D. 2006. Glutamine-expanded ataxin-7 alters TFTC/STAGA recruitment and chromatin structure leading to photoreceptor dysfunction. PLoS Biol 4:e67.

Jarjanazi H, Kiefer J, Savas S, Briollais L, Tuzmen S, Pabalan N, Ibrahim-Zada I, Mousses S, Ozcelik H. 2008. Discovery of genetic profiles impacting response to chemotherapy: Application to gemcitabine. Hum Mutat 29:461–467.

Karras JR, Paisie CA, Huebner K. 2014. Replicative Stress and the FHIT Gene: Roles in tumor suppression, genome stability and prevention of carcinogenesis. Cancers (Basel) 6:1208–1219.

Keller W, Minvielle-Sebastia L. 1997. A comparison of mammalian and yeast pre-mRNA 3′-end processing. Curr Opin Cell Biol 9:329–336.

Kohno T, Ichikawa H, Totoki Y, Yasuda K, Hiramoto M, Nammo T, Sakamoto H, Tsuta K, Furuta K, Shimada Y, Iwakawa R, Ogiwara H, Oike T, Enari M, Schetter AJ, Okayama H, Haugen A, Skaug V, Chiku S, Yamanaka I, Arai Y, Watanabe S, Sekine I, Ogawa S, Harris CC, Tsuda H, Yoshida T, Yokota J, Shibata T. 2012. KIF5B-RET fusions in lung adenocarcinoma. Nat Med 18:375–377.

Kondo A, Hashimoto S, Yano H, Nagayama K, Mazaki Y, Sabe H. 2000. A new paxillin-binding protein, PAG3/Papalpha/KIAA0400, bearing an ADP-ribosylation factor GTPase-activating protein activity, is involved in paxillin recruitment to focal adhesions and cell migration. Mol Biol Cell 11:1315–1327.

Lamesch P, Li N, Milstein S, Fan C, Hao T, Szabo G, Hu Z, Venkatesan K, Bethel G, Martin P, Rogers J, Lawlor S, McLaren S, Dricot A, Borick H, Cusick ME, Vandenhaute J, Dunham I, Hill DE, Vidal M. 2007. hORFeome v3.1: A resource of human open reading frames representing over 10,000 human genes. Genomics 89:307–315.

Li A, Liu Z, Lezon-Geyda K, Sarkar S, Lannin D, Schulz V, Krop I, Winer E, Harris L, Tuck D. 2011. GPHMM: An integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays. Nucleic Acids Res 39:4928–4941.

Li H, Wang J, Mor G, Sklar J. 2008. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. Science 321:1357–1361.

Liu Q, Zhao XY, Bai RZ, Liang SF, Nie CL, Yuan Z, Wang CT, Wu Y, Chen LJ, Wei YQ. 2009. Induction of tumor inhibition and apoptosis by a candidate tumor suppressor gene DRR1 on 3p21.1. Oncol Rep 22:1069–1075.

Ljungberg B, Campbell SC, Choi HY, Jacqmin D, Lee JE, Weikert S, Kiemeney LA. 2011. The epidemiology of renal cell carcinoma. Eur Urol 60:615–621.

Loewy AD, Niles KM, Anastasio N, Watkins D, Lavoie J, Lerner-Ellis JP, Pastinen T, Trasler JM, Rosenblatt DS. 2009. Epigenetic modification of the gene for the vitamin B(12) chaperone MMACHC can result in increased tumorigenicity and methionine dependence. Mol Genet Metab 96:261–267.

Mariño-Enríquez A, Ou WB, Weldon CB, Fletcher JA, Pérez-Atayde AR. 2011. ALK rearrangement in sickle cell trait-associated renal medullary carcinoma. Genes Chromosomes Cancer 50:146–153.

Matsui C, Kaieda S, Ikegami T, Mimori-Kiyosue Y. 2008. Identification of a link between the SAMP repeats of adenomatous polyposis coli tumor suppressor and the Src homology 3 domain of DDEF. J Biol Chem 283:33006–33020.

McNerney ME, Brown CD, Wang X, Bartom ET, Karmakar S, Bandlamudi C, Yu S, Ko J, Sandall BP, Stricker T, Anastasi J, Grossman RL, Cunningham JM, Le Beau MM, White KP. 2013. CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia. Blood 121:975–983.

McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, Pacheco M, Marra MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP. 2011. deFuse: An algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol 7: e1001138.

Milne RL, Burwinkel B, Michailidou K, Arias-Perez JI, Zamora MP, Menéndez-Rodríguez P, Hardisson D, Mendiola M, González-Neira A, Pita G, Alonso MR, Dennis J, Wang Q, Bolla MK, Swerdlow A, Ashworth A, Orr N, Schoemaker M, Ko YD, Brauch H, Hamann U; The GENICA Network, Andrulis IL, Knight JA, Glendon G, Tchatchou S, Investigators K; Australian Ovarian Cancer Study Group, Matsuo K, Ito H, Iwata H, Tajima K, Li J, Brand JS, Brenner H, Dieffenbach AK, Arndt V, Stegmaier C, Lambrechts D, Peuteman G, Christiaens MR, Smeets A, Jakubowska A, Lubinski J, Jaworska-Bieniek K, Durda K, Hartman M, Hui M, Lim WY, Chan CW, Marme F, Yang R, Bugert P, Lindblom A, Margolin S, García-Closas M, Chanock SJ, Lissowska J, Figueroa JD, Bojesen SE, Nordestgaard BG, Flyger H, Hooning MJ, Kriege M, van den Ouweland AM, Koppert LB, Fletcher O, Johnson N, Dos-Santos-Silva I, Peto J, Zheng W, Deming-Halverson S, Shrubsole MJ, Long J, Chang-Claude J, Rudolph A, Seibold P, Flesch-Janys D, Winqvist R, Pylkäs K, Jukkola-Vuorinen A, Grip M, Cox A, Cross SS, Reed MW, Schmidt MK, Broeks A, Cornelissen S, Braaf L, Kang D, Choi JY, Park SK, Noh DY, Simard J, Dumont M, Goldberg MS, Labrèche F, Fasching PA, Hein A, Ekici AB, Beckmann MW, Radice P, Peterlongo P, Azzollini J, Barile M, Sawyer E, Tomlinson I, Kerin M, Miller N, Hopper JL, Schmidt DF, Makalic E, Southey MC, Teo SH, Yip CH, Sivanandan K, Tay WT, Shen CY, Hsiung CN, Yu JC, Hou MF, Guénel P, Truong T, Sanchez M, Mulot C, Blot W, Cai Q, Nevanlinna H, Muranen TA, Aittomäki K, Blomqvist C, Wu AH, Tseng CC, Van Den Berg D, Stram DO, Bogdanova N, Dörk T, Muir K, Lophatananon A, Stewart-Brown S, Siriwanarangsan P, Mannermaa A, Kataja V, Kosma VM, Hartikainen JM, Shu XO, Lu W, Gao YT, Zhang B, Couch FJ, Toland AE; TNBCC, Yannoukakos D, Sangrajrang S, McKay J, Wang X, Olson JE, Vachon C, Purrington K, Severi G, Baglietto L, Haiman CA, Henderson BE, Schumacher F, Le Marchand L, Devilee P, Tollenaar RA, Seynaeve C, Czene K, Eriksson M, Humphreys K, Darabi H, Ahmed S, Shah M, Pharoah PD, Hall P, Giles GG, Benítez J, Dunning AM, Chenevix-Trench G, Easton DF. Common non-synonymous SNPs associated with breast cancer susceptibility: Findings from the Breast Cancer Association Consortium. Hum Mol Genet (in press).

Pflueger D, Sboner A, Storz M, Roth J, Compérat E, Bruder E, Rubin MA, Schraml P, Moch H. 2013. Identification of molecular tumor markers in renal cell carcinomas with TFE3 protein expression by RNA sequencing. Neoplasia 15:1231–1240.

Rao Q, Williamson SR, Zhang S, Eble JN, Grignon DJ, Wang M, Zhou XJ, Huang W, Tan PH, Maclennan GT, Cheng L. 2013. TFE3 break-apart FISH has a higher sensitivity for Xp11.2 translocation-associated renal cell carcinoma compared with TFE3 or cathepsin K immunohistochemical staining alone: Expanding the morphologic spectrum. Am J Surg Pathol 37:804–815.

Ribeiro IP, Marques F, Caramelo F, Ferrão J, Prazeres H, Julião MJ, Rifi W, Savola S, de Melo JB, Baptista IP, Carreira IM. 2014. Genetic imbalances detected by multiplex ligation-dependent probe amplification in a cohort of patients with oral squamous cell carcinoma-the first step toward clinical personalized medicine. Tumour Biol 35:4687–4695.

Sato Y, Yoshizato T, Shiraishi Y, Maekawa S, Okuno Y, Kamura T, Shimamura T, Sato-Otsubo A, Nagae G, Suzuki H, Nagata Y, Yoshida K, Kon A, Suzuki Y, Chiba K, Tanaka H, Niida A, Fujimoto A, Tsunoda T, Morikawa T, Maeda D, Kume H, Sugano S, Fukayama M, Aburatani H, Sanada M, Miyano S, Homma Y, Ogawa S. 2013. Integrated molecular analysis of clear-cell renal cell carcinoma. Nat Genet 45:860–867.

Shima Y, Kitabayashi I. 2011. Deregulated transcription factors in leukemia. Int J Hematol 94:134–141.

Sobin LH, Gospodarowicz M, Wittekind C. 2009. International Union Against Cancer (UICC). TNM classification of malignant tumors, 7th ed. New York: Wiley-Liss.

Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H, Kurashina K, Hatanaka H, Bando M, Ohno S, Ishikawa Y, Aburatani H, Niki T, Sohara Y, Sugiyama Y, Mano H. 2007. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 448:561–566.

Sugawara E, Togashi Y, Kuroda N, Sakata S, Hatano S, Asaka R, Yuasa T, Yonese J, Kitagawa M, Mano H, Ishikawa Y, Takeuchi K. 2012. Identification of anaplastic lymphoma kinase fusions in renal cancer: Large-scale immunohistochemical screening by the intercalated antibody-enhanced polymer method. Cancer 118:4427–4436.

Takeuchi K, Soda M, Togashi Y, Suzuki R, Sakata S, Hatano S, Asaka R, Hamanaka W, Ninomiya H, Uehara H, Lim Choi Y, Satoh Y, Okumura S, Nakagawa K, Mano H, Ishikawa Y. 2012. RET, ROS1 and ALK fusions in lung cancer. Nat Med 18:378–381.

Tanuma N, Nomura M, Ikeda M, Kasugai I, Tsubaki Y, Takagaki K, Kawamura T, Yamashita Y, Sato I, Sato M, Katakura R, Kikuchi K, Shima H. 2009. Protein phosphatase Dusp26 associates with KIF3 motor and promotes N-cadherin-mediated cell-cell adhesion. Oncogene 28:752–761.

The Cancer Genome Atlas Research Network. 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499:43–49.

Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM. 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310:644–648.

van Haaften G, Dalgliesh GL, Davies H, Chen L, Bignell G, Greenman C, Edkins S, Hardy C, O'Meara S, Teague J, Butler A, Hinton J, Latimer C, Andrews J, Barthorpe S, Beare D, Buck G, Campbell PJ, Cole J, Forbes S, Jia M, Jones D, Kok CY, Leroy C, Lin ML, McBride DJ, Maddison M, Maquire S, McLay K, Menzies A, Mironenko T, Mulderrig L, Mudie L, Pleasance E, Shepherd R, Smith R, Stebbings L, Stephens P, Tang G, Tarpey PS, Turner R, Turrell K, Varian J, West S, Widaa S, Wray P, Collins VP, Ichimura K, Law S, Wong J, Yuen ST, Leung SY, Tonon G, DePinho RA, Tai YT, Anderson KC, Kahnoski RJ, Massie A, Khoo SK, Teh BT, Stratton MR, Futreal PA. 2009. Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. Nat Genet 41:521–523.

Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, Davies H, Jones D, Lin ML, Teague J, Bignell G, Butler A, Cho J, Dalgliesh GL, Galappaththige D, Greenman C, Hardy C, Jia M, Latimer C, Lau KW, Marshall J, McLaren S, Menzies A, Mudie L, Stebbings L, Largaespada DA, Wessels LF, Richard S, Kahnoski RJ, Anema J, Tuveson DA, Perez-Mancera PA, Mustonen V, Fischer A, Adams DJ, Rust A, Chan-on W, Subimerb C, Dykema K, Furge K, Campbell PJ, Teh BT, Stratton MR, Futreal PA. 2011. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature 469:539–542.

Wang L, Darling J, Zhang JS, Liu W, Qian J, Bostwick D, Hartmann L, Jenkins R, Bardenhauer W, Schutte J, Opalka B, Smith DI. 2000. Loss of expression of the DRR 1 gene at chromosomal segment 3p21.1 in renal cell carcinoma. Genes Chromosomes Cancer 27:1–10.

Yuan H, Qin F, Movassagh M, Park H, Golden W, Xie Z, Zhang P, Sklar J, Li H. 2013. A chimeric RNA characteristic of rhabdomyosarcoma in normal myogenesis process. Cancer Discov 3: 1394–1403.

Yu G, Tseng GC, Yu YP, Gavel T, Nelson J, Wells A, Michalopoulos G, Kokkinakis D, Luo JH. 2006. CSR1 suppresses tumor growth and metastasis of prostate cancer. Am J Pathol 168:597–607.

Zhang J, Wan L, Dai X, Sun Y, Wei W. 2014. Functional characterization of Anaphase Promoting Complex/Cyclosome (APC/C) E3 ubiquitin ligases in tumorigenesis. Biochim Biophys Acta 1845:277–293.

Zhu ZH, Yu YP, Shi YK, Nelson JB, Luo JH. 2009. CSR1 induces cell death through inactivation of CPSF3. Oncogene 28:41–51.

OPEN ACCESS Freely available online

PLOS | ONE

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

Anna Takahashi[1⊛], Robert Nakayama[2,3⊛], Nanako Ishibashi[4⊛], Ayano Doi[2,5], Risa Ichinohe[2,5], Yoriko Ikuyo[7], Teruyoshi Takahashi[7], Shigetaka Marui[7], Koji Yasuhara[7], Tetsuro Nakamura[7], Shintaro Sugita[8], Hiromi Sakamoto[2], Teruhiko Yoshida[2], Tadashi Hasegawa[8,9*], Hiro Takahashi[1,2,6*⊛]

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 Graduate School of Horticulture, Chiba University, Matsudo, Chiba, Japan, 7 Graduate School of Bioscience and Biotechnology, Chubu 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*, *SCD1*, *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 2.99×10^{-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.

Citation: Takahashi A, Nakayama R, Ishibashi N, Doi A, Ichinohe R, et al. (2014) Analysis of Gene Expression Profiles of Soft Tissue Sarcoma Using a Combination of Knowledge-Based Filtering with Integration of Multiple Statistics. PLoS ONE 9(9): e106801. doi:10.1371/journal.pone.0106801

Editor: Guy Brock, University of Louisville, United States of America

Received March 29, 2014; **Accepted** August 1, 2014; **Published** September 4, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported in part by the Ministry of Education, Culture, Sports, Science, and Technology of Japan (MEXT): Grants-in-Aid for Scientific Research for Young Scientists (B) (nos. 21710211 and 24710222 to H.T.) and Grant-in-Aid for Scientific Research on Innovative Areas (no. 26114703 to H.T.). This work was also supported by the Advanced Research for Medical Products Mining Program of the National Institute of Biomedical Innovation (NIBIO ID10-41), the Futaba Electronics Memorial Foundation, the Research Foundation for the Electrotechnology of Chubu, and the Nakajima Foundation. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of this manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: hiro.takahashi@chiba-u.jp (HT); hasetada@sapmed.ac.jp (TH)

⊛ These authors contributed equally to this work.

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 Francaise 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]. The molecular markers specific to each tumor type are useful for tumor classification [8]. 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 [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

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
	MPNST	5
Histological grade	1	14
	2	23
	3	51
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

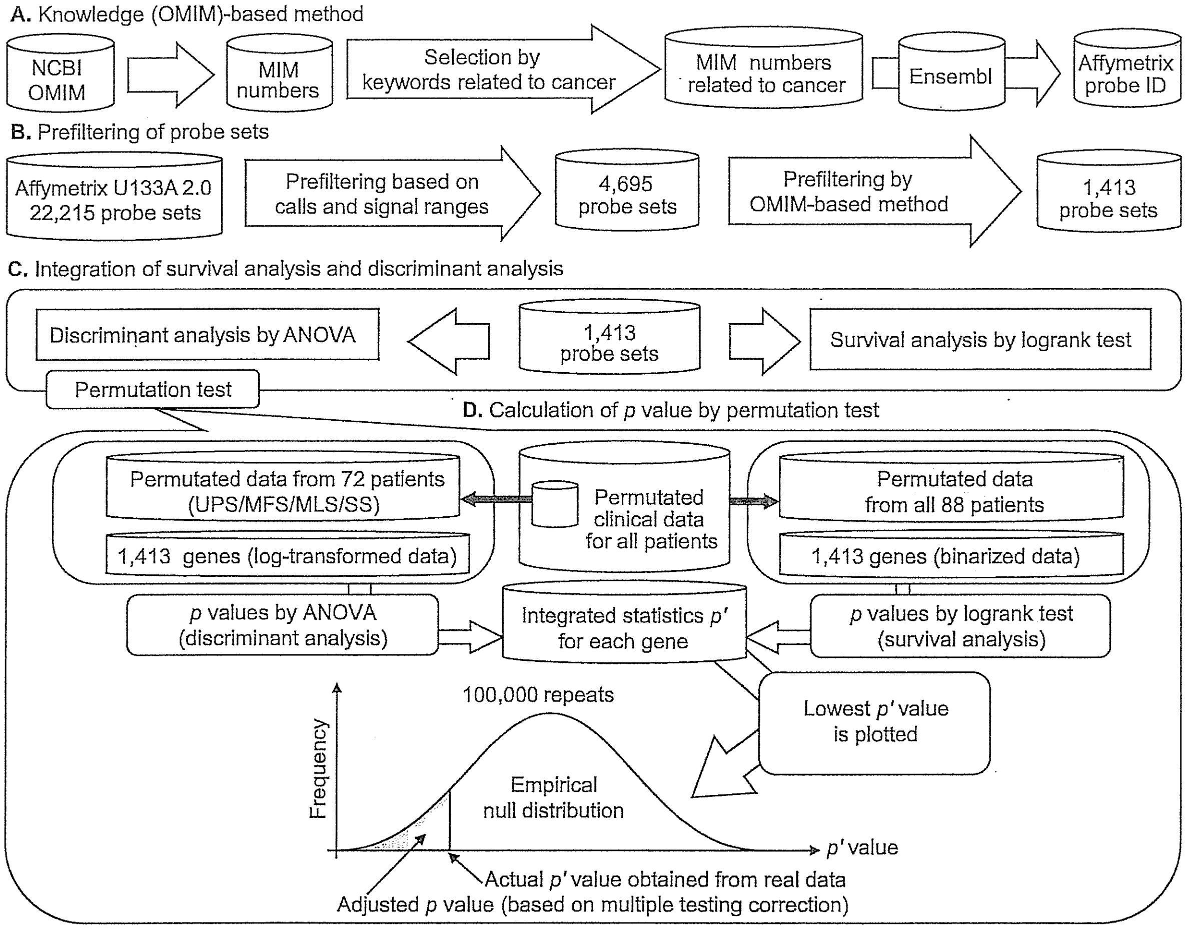
A. Knowledge (OMIM)-based method
NCBI OMIM
MIM numbers
Selection by keywords related to cancer
MIM numbers related to cancer
Ensembl
Affymetrix probe ID
B. Prefiltering of probe sets
Affymetrix U133A 2.0 22,215 probe sets
Prefiltering based on calls and signal ranges
4,695 probe sets
Prefiltering by OMIM-based method
1,413 probe sets
C. Integration of survival analysis and discriminant analysis
Discriminant analysis by ANOVA
1,413 probe sets
Survival analysis by logrank test
Permutation test
D. Calculation of p value by permutation test
Permutated data from 72 patients (UPS/MFS/MLS/SS)
1,413 genes (log-transformed data)
Permutated clinical data for all patients
Permutated data from all 88 patients
1,413 genes (binarized data)
p values by ANOVA (discriminant analysis)
Integrated statistics p' for each gene
p values by logrank test (survival analysis)
100,000 repeats
Lowest p' value is plotted
Frequency
Empirical null distribution
p' value
Actual p' value obtained from real data
Adjusted p value (based on multiple testing correction)

Figure 1. A schematic of gene selection and the simulation based on the permutation test. (A) The knowledge (OMIM)-based method. The list of OMIM numbers related to cancer (e.g., cancer, carcinoma, sarcoma, tumor, and neoplasm) was selected and converted into Affymetrix probe IDs in Ensembl. (B) Prefiltering of probe sets. This procedure was based on the number of absent calls and the range of signals. A signal range (95th percentile to 5th percentile) of >2000 was used as a percentile filter. Furthermore, we excluded probe sets for which the number of absent calls was >50% (44/88). Probe sets related to cancer were selected using the OMIM-based method. (C) Integration of survival analysis and discriminant analysis. (D) Clinical data from all patients were permutated. Permutated data for 72 STS patients (20 UPS, 15 MFS, 20 MLS, and 17 SS patients) were extracted from the permutated data of all patients. For these data, p values (p_1) were calculated by applying ANOVA to the log-transformed gene expression data to discriminate among UPS, MFS, MLS, and SS. In addition, permutated data from 88 patients were used for survival analysis. For these data, p values (p_2) were calculated by applying the logrank test to the binarized gene expression data to analyze the outcomes in the STS group. The integrated statistic p' was defined as $p_1 \times p_2$. The lowest p' value was selected for each repetition. This procedure was repeated 100,000 times, and an empirical null distribution was constructed. Using the distribution, the actual p' value obtained from the real data was converted to the adjusted p value (based on the correction for multiple testing problems).
doi:10.1371/journal.pone.0106801.g001

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 non-tumorous 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

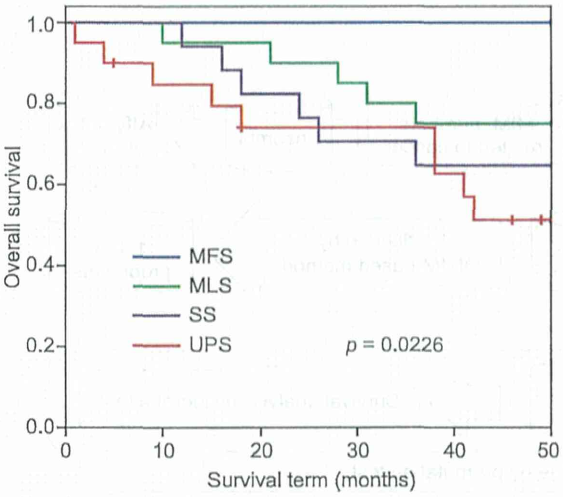

Figure 2. Kaplan-Meier curves for 4 histological types of STS. *P* value was calculated by logrank test. UPS: undifferentiated pleomorphic sarcoma, MLS: myxoid liposarcoma, SS: synovial sarcoma, MFS: myxofibrosarcoma.
doi:10.1371/journal.pone.0106801.g002

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

Table 2. Genes extracted using the simulation based on the permutation test.

Affymetrix probe ID	Accession no.	Gene symbol	*p* value		Integrated statistics *p*′	Adjusted *p* value
			ANOVA	Log-rank test		
200832_s_at	AB032261	*SCD1*	2.47E-06	6.06E-03	1.50E-08	6.70E-04
200887_s_at	NM_007315	*STAT1*	1.17E-04	1.91E-02	2.24E-06	3.59E-02
201231_s_at	NM_001428	*ENO1/MBP1*	2.27E-08	1.06E-03	2.40E-11	<1.00E-05
201508_at	NM_001552	*IGFBP4*	3.21E-06	4.01E-02	1.29E-07	3.76E-03
202236_s_at	NM_003051	*SLC16A1/MCT1*	1.12E-04	6.93E-04	7.77E-08	2.34E-03
202870_s_at	NM_001255	*CDC20*	9.26E-07	6.28E-03	5.81E-09	2.90E-04
203065_s_at	NM_001753	*CAV1*	1.33E-10	3.28E-02	4.35E-12	<1.00E-05
203323_at	BF197655	*CAV2*	5.67E-10	2.35E-02	1.33E-11	<1.00E-05
203554_x_at	NM_004219	*PTTG1*	7.33E-09	5.64E-03	4.13E-11	<1.00E-05
207011_s_at	NM_002821	*PTK7*	2.57E-07	1.89E-02	4.86E-09	2.70E-04
207168_s_at	NM_004893	*H2AFY/H2AX*	2.83E-05	1.80E-02	5.11E-07	1.19E-02
207543_s_at	NM_000917	*P4HA1*	1.06E-08	5.73E-04	6.06E-12	<1.00E-05
208680_at	L19184	*PRDX1*	5.73E-08	1.64E-02	9.37E-10	6.00E-05
208694_at	U47077	*PRKDC/DNA-PKcs*	1.71E-04	1.31E-02	2.25E-06	3.60E-02
208767_s_at	AW149681	*LAPTM4B*	5.47E-05	1.65E-02	9.04E-07	1.81E-02
209030_s_at	NM_014333	*CADM1/TSLC1*	1.80E-10	4.20E-02	7.59E-12	<1.00E-05
209031_at	AL519710	*CADM1/TSLC1*	2.10E-11	5.68E-03	1.19E-13	<1.00E-05
209543_s_at	M81104	*CD34*	2.66E-06	1.54E-02	4.10E-08	1.33E-03
210495_x_at	AF130095	*FN1*	3.90E-08	1.78E-02	6.96E-10	2.00E-05
210559_s_at	D88357	*CDK1/CDC2*	7.69E-07	4.30E-02	3.31E-08	1.14E-03
212097_at	AU147399	*CAV1*	1.54E-09	2.95E-03	4.53E-12	<1.00E-05
212464_s_at	X02761	*FN1*	1.93E-08	1.78E-02	3.44E-10	1.00E-05
217294_s_at	U88968	*ENO1/MBP1*	8.81E-08	2.33E-02	2.05E-09	1.50E-04
217871_s_at	NM_002415	*MIF*	5.67E-08	1.46E-02	8.29E-10	5.00E-05
218308_at	NM_006342	*TACC3*	2.82E-05	2.26E-02	6.38E-07	1.40E-02
218502_s_at	NM_014112	*TRPS1*	1.48E-18	3.99E-02	5.90E-20	<1.00E-05
218755_at	NM_005733	*KIF20A/MKlp2*	3.01E-06	2.02E-02	6.08E-08	1.94E-03
219918_s_at	NM_018123	*ASPM*	1.22E-05	1.64E-02	2.00E-07	5.51E-03
220942_x_at	NM_014367	*FAM162A/HGTD-P*	4.44E-05	3.21E-02	1.42E-06	2.56E-02

Adjusted *p* values were calculated using the permutation test (100,000 repeats).
doi:10.1371/journal.pone.0106801.t002

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	
			ρ	*p* value	ρ	*p* value
200832_s_at	AB032261	*SCD1*	−0.0191	8.60E-01	0.0237	8.26E-01
200887_s_at	NM_007315	*STAT1*	−0.146	1.73E-01	−0.177	9.95E-02
201231_s_at	NM_001428	*ENO1/MBP1*	0.356	6.66E-04	0.247	2.01E-02
201508_at	NM_001552	*IGFBP4*	−0.247	2.04E-02	−0.211	4.87E-02
202236_s_at	NM_003051	*SLC16A1/MCT1*	0.400	1.12E-04	0.341	1.17E-03
202870_s_at	NM_001255	*CDC20*	0.413	6.27E-05	0.204	5.65E-02
203065_s_at	NM_001753	*CAV1*	−0.250	1.87E-02	−0.159	1.39E-01
203323_at	BF197655	*CAV2*	−0.363	5.11E-04	−0.094	3.82E-01
203554_x_at	NM_004219	*PTTG1*	0.402	1.05E-04	0.132	2.20E-01
207011_s_at	NM_002821	*PTK7*	0.265	1.26E-02	0.232	2.95E-02
207168_s_at	NM_004893	*H2AFY/H2AX*	0.411	7.03E-05	0.161	1.35E-01
207543_s_at	NM_000917	*P4HA1*	0.449	1.12E-05	0.424	3.89E-05
208680_at	L19184	*PRDX1*	0.258	1.51E-02	0.111	3.05E-01
208694_at	U47077	*PRKDC/DNA-PKcs*	0.409	7.64E-05	0.229	3.21E-02
208767_s_at	AW149681	*LAPTM4B*	0.329	1.75E-03	0.130	2.27E-01
209030_s_at	NM_014333	*CADM1/TSLC1*	0.196	6.70E-02	0.136	2.05E-01
209031_at	AL519710	*CADM1/TSLC1*	0.231	3.03E-02	0.143	1.85E-01
209543_s_at	M81104	*CD34*	−0.363	5.11E-04	−0.239	2.52E-02
210495_x_at	AF130095	*FN1*	0.286	6.99E-03	0.096	3.73E-01
210559_s_at	D88357	*CDK1/CDC2*	0.435	2.34E-05	0.259	1.50E-02
212097_at	AU147399	*CAV1*	−0.237	2.64E-02	−0.163	1.28E-01
212464_s_at	X02761	*FN1*	0.286	6.99E-03	0.0944	3.82E-01
217294_s_at	U88968	*ENO1/MBP1*	0.387	1.97E-04	0.187	8.03E-02
217871_s_at	NM_002415	*MIF*	0.421	4.41E-05	0.308	3.47E-03
218308_at	NM_006342	*TACC3*	0.333	1.52E-03	0.136	2.05E-01
218502_s_at	NM_014112	*TRPS1*	0.276	9.23E-03	0.242	2.31E-02
218755_at	NM_005733	*KIF20A/MKlp2*	0.407	8.35E-05	0.162	1.31E-01
219918_s_at	NM_018123	*ASPM*	0.399	1.16E-04	0.204	5.71E-02
220942_x_at	NM_014367	*FAM162A/HGTD-P*	0.151	1.60E-01	0.239	2.47E-02

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.

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 $p_1 \times p_2$, 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

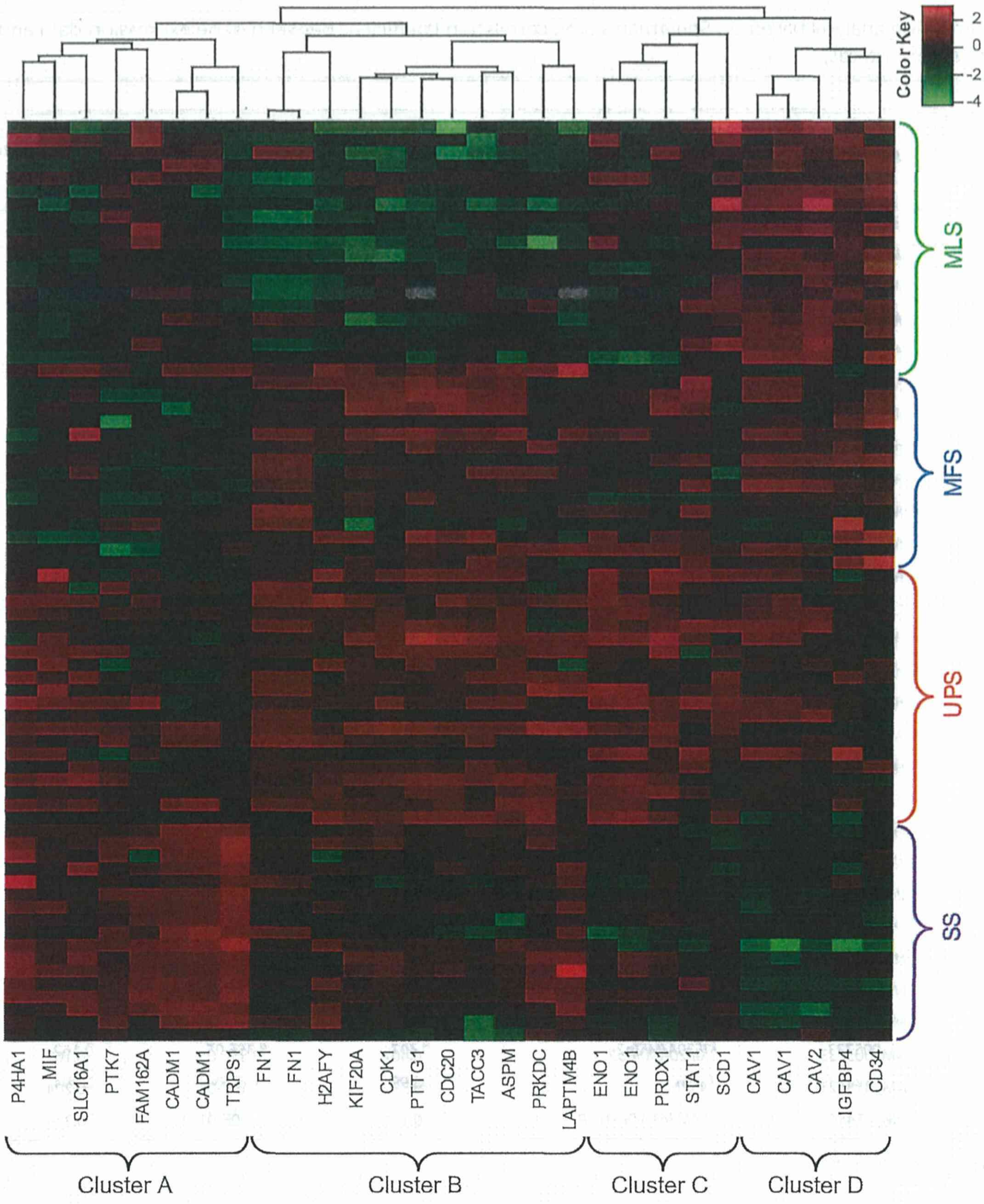

Figure 3. Heatmap and hierarchical clustering analyses. Twenty-nine probe sets were extracted using a simulation based on the permutation test (with adjusted $p<0.05$). The 29 probe sets were roughly divided into 4 clusters (clusters A–D). Columns represent probe sets, and rows represent samples. Red 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. 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

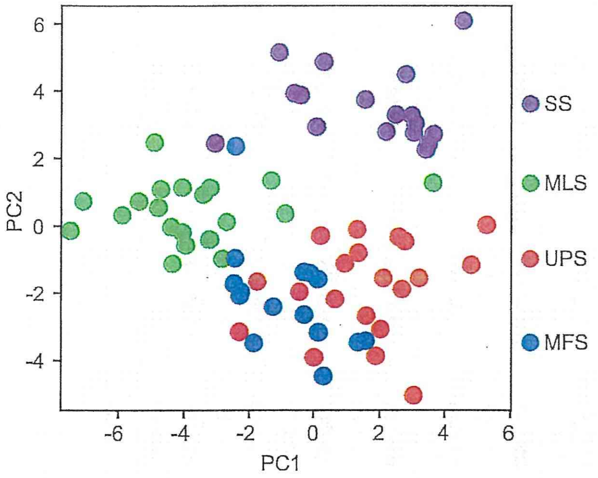

Figure 4. Principal component analysis using 29 probe sets for 4 histological types. The x-axis and y-axis represent the first and second principal components (PC1 and PC2), respectively. Each dot represents a sample colored according to its histological type. UPS: undifferentiated pleomorphic sarcoma, MLS: myxoid liposarcoma, SS: synovial sarcoma, MFS: myxofibrosarcoma.
doi:10.1371/journal.pone.0106801.g004

U133A 2.0 Array using information retrieved from Ensembl on January 10, 2014. There were 5155 Affymetrix probe set IDs, as shown in Table S3.

Principal component analysis (PCA)

We used PCA to reduce the gene expression profile data to a two-dimensional dataset. PCA was first proposed in 1901 by Pearson [57]. This method is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (PCs). The number of PCs is less than or equal to the number of original variables. This transformation is defined in such a way that the first PC has the greatest possible variance.

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 [58], 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 *q* 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.

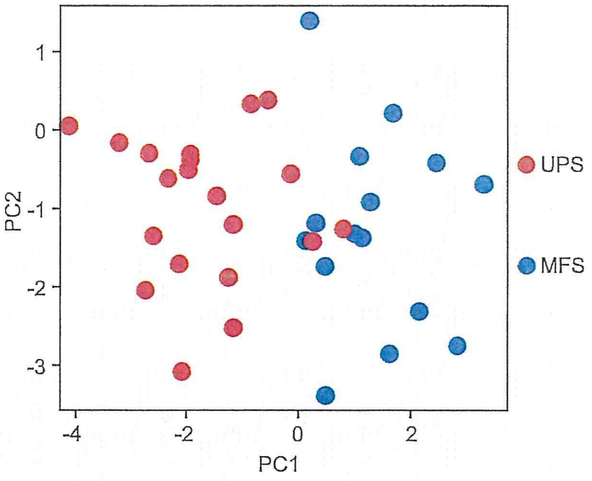

Figure 5. Principal component analysis using 9 probe sets for UPS and MFS. The x-axis and y-axis represent the first and second principal components (PC1 and PC2), respectively. Each dot represents a sample colored according to its histological type. UPS: undifferentiated pleomorphic sarcoma, MLS: myxoid liposarcoma, SS: synovial sarcoma, MFS: myxofibrosarcoma.
doi:10.1371/journal.pone.0106801.g005

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 ρ were upregulated in highly malignant tumors, whereas genes with negative ρ 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 (*SCD1*: $\rho=-0.0191$, $p=0.860$; *STAT1*: $\rho=-0.146$, $p=0.173$) or metastasis (*SCD1*: $\rho=0.0237$, $p=0.826$; *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	*	1.13E-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	6.49E-04	*	1.25E-03
202870_s_at	NM_001255	*CDC20*	2.10E-01		3.58E-01	1.23E-03	*	2.74E-03	6.26E-06	*	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
203323_at	BF197655	*CAV2*	8.45E-01		8.75E-01	6.14E-05	*	1.98E-04	1.26E-03	*	2.15E-03
203554_x_at	NM_004219	*PTTG1*	3.76E-01		4.96E-01	8.95E-05	*	2.60E-04	1.59E-08	*	2.31E-07
207011_s_at	NM_002821	*PTK7*	6.14E-03	*	2.23E-02	4.21E-03	*	6.78E-03	9.19E-01		9.19E-01
207168_s_at	NM_004893	*H2AFY/H2AX*	4.37E-02		1.15E-01	1.18E-01		1.37E-01	6.75E-06	*	1.78E-05
207543_s_at	NM_000917	*P4HA1*	1.22E-04	*	8.87E-04	2.64E-02	*	3.48E-02	2.51E-03	*	4.05E-03
208680_at	L19184	*PRDX1*	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	*	2.06E-03
208767_s_at	AW149681	*LAPTM4B*	4.20E-01		5.30E-01	3.73E-02	*	4.60E-02	8.30E-03	*	1.27E-02
209030_s_at	NM_014333	*CADM1/TSLC1*	2.49E-01		3.80E-01	2.81E-07	*	1.82E-06	6.43E-01		6.66E-01
209031_at	AL519710	*CADM1/TSLC1*	6.04E-02		1.35E-01	2.67E-07	*	1.82E-06	2.71E-01		3.42E-01
209543_s_at	M81104	*CD34*	8.73E-03	*	2.81E-02	1.78E-01		1.91E-01	3.97E-05	*	8.22E-05
210495_x_at	AF130095	*FN1*	4.83E-01		5.61E-01	2.50E-03	*	4.27E-03	3.53E-06	*	1.42E-05
210559_s_at	D88357	*CDK1/CDC2*	7.05E-02		1.46E-01	2.35E-02	*	3.24E-02	3.57E-06	*	1.42E-05
212097_at	AU147399	*CAV1*	6.43E-01		6.91E-01	3.14E-07	*	1.82E-06	4.16E-01		4.83E-01
212464_s_at	X02761	*FN1*	5.22E-01		5.83E-01	2.33E-03	*	4.22E-03	2.07E-06	*	1.20E-05
217294_s_at	U88968	*ENO1/MBP1*	4.24E-04	*	2.46E-03	4.07E-05	*	1.69E-04	1.55E-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
218308_at	NM_006342	*TACC3*	2.36E-01		3.80E-01	7.67E-04	*	2.02E-03	2.91E-05	*	6.49E-05
218502_s_at	NM_014112	*TRPS1*	3.64E-01		4.96E-01	5.21E-11	*	1.51E-09	1.85E-02	*	2.68E-02
218755_at	NM_005733	*KIF20A/MKlp2*	4.44E-01		5.37E-01	9.97E-03	*	1.45E-02	4.41E-06	*	1.42E-05
219918_s_at	NM_018123	*ASPM*	1.11E-01		2.15E-01	2.25E-03	*	4.22E-03	7.89E-07	*	5.72E-06
220942_x_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

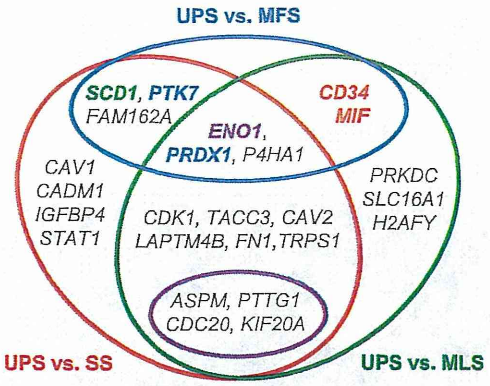

Figure 6. A Venn diagram of gene classification based on pairwise comparisons of histological types using Welch's *t* test. Genes inside the red circle were statistically significant (q <0.05 calculated using Welch's t test and the BH method) in the comparison of UPS with SS. Genes inside the green oval were statistically significant (q <0.05) in the comparison of UPS with MLS. Genes inside the blue oval were statistically significant (q <0.05) in the comparison of UPS and MFS. Genes inside the pink oval are common to CINSARC and our 25-gene set. For PCA of the 9-probe set, *MIF* and *CD34* highlighted in red were the first and third largest contributing coefficients to PC1, respectively. *PTK7* and *PRDX1* highlighted in blue were the first and second largest contributing coefficients to PC2, respectively. *ENO1/MBP1* highlighted in purple was the second largest contributing coefficient to PC1 and the third largest contributing coefficient to PC2. *SCD1* highlighted in green was the largest contributing coefficient to PC3.
doi:10.1371/journal.pone.0106801.g006

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 the 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 of the 29 probe sets, as shown in Table 4. We extracted 9 probe sets, representing 8 genes (q value <0.05): enolase 1 (*ENO1*)/c-myc-promoter binding protein-1 (*MBP1*); prolyl 4-hydroxylase subunit alpha-1 (*P4HA1*); peroxiredoxin 1 (*PRDX1*); *CD34*; family with sequence similarity 162, member A (*FAM162A*)/human growth and transformation-dependent protein (*HGTD-P*); protein tyrosine kinase 7 (*PTK7*); and macrophage migration inhibitory factor (*MIF*). We performed PCA to calculate the first and the second PCs from these 9 probe sets. Detailed information of PCA, including eigenvector, standard deviation, proportion of variance, and cumulative proportion, are shown in Table S5. The distribution of the 2 histological subtypes, UPS and MFS, on the 2 axes is shown in Fig. 5. UPS and MFS were classified into approximately 2 clusters. For the contribution of this classification, *MIF*, *ENO1/MBP1*, and *CD34* contributed to the top 3 largest coefficients for PC1, *PTK7*, *PRDX1*, and *ENO1/MBP1* contributed to the top 3 largest coefficients for PC2, and only *SCD1* contributed to the largest coefficients for PC3, as shown in Table S5. *MIF*, *ENO1/MBP1*, and *SCD1* were extracted in our previous study [51]. We also 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.

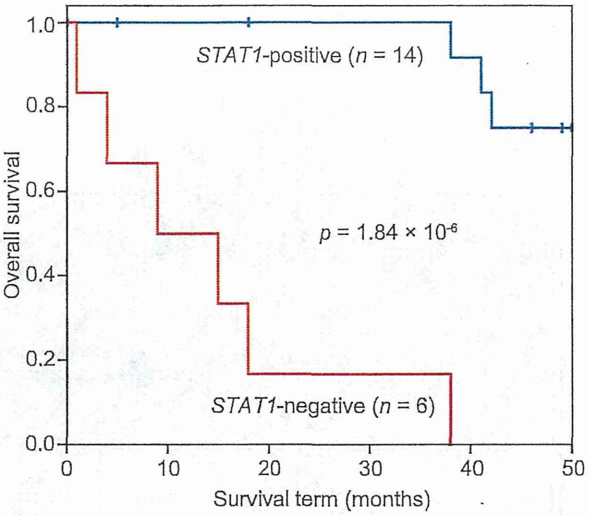

Figure 7. The Kaplan-Meier curve and the logrank test for *STAT1* in UPS patients. The *STAT1*-positive group (*STAT1* expression level >4871.5) consisted of 14 patients (blue line), and the *STAT1*-negative group consisted of 6 patients (red line). A hazard ratio (exp(B) = 30.2) was calculated using the Cox proportional hazards model.
doi:10.1371/journal.pone.0106801.g007

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