testicular, pancreatic, thyroid, and cervical carcinomas as well as in malignant melanoma and has been associated with tumour aggressiveness (Ino *et al*, 2002; Tsuiji *et al*, 2002; Aoki *et al*, 2003; Sato *et al*, 2003; Shimamura *et al*, 2003, 2004; Nakanishi *et al*, 2004; Shimada *et al*, 2004a, b; Wu *et al*, 2004; Batistatou *et al*, 2005; Nishizawa *et al*, 2005; Batistatou *et al*, 2006; Kyzas *et al*, 2006). A recent *in vitro* study has demonstrated that dysadherin has prometastatic effects that are independent of E-cadherin expression (Nam *et al*, 2006). The aim of the present study was to investigate further the expression of dysadherin in breast carcinoma, with particular emphasis to the acquisition of a lobular or a ductal phenotype, in combination with E-cadherin expression.

MATERIALS AND METHODS

One hundred formalin-fixed, paraffin-embedded archival tissue blocks of breast carcinomas were included in the current study and represented an equal number of female patients (mean age 54.5 years, range 35–79). The material consisted of 70 invasive ductal carcinomas, no special type, NST (10 Grade 1, 45 Grade 2 and 15 Grade 3, graded using the modified Bloom and Richardson method), in 30 of which an adjacent *in situ* ductal carcinoma was identified, and 30 invasive lobular carcinomas, in 15 of which an adjacent *in situ* lobular carcinoma was identified.

Immunohistochemistry

We performed immunostaining on formalin-fixed, paraffin-embedded tissue sections using the EnVision System (DAKO Corp., Netherlands), and the monoclonal antibodies: NCC-M53 against dysadherin and E-cadherin (CM170B, Biocare Medical, CA, USA). Briefly, 4-μm-thick tissue sections were deparaffinised in xylene; rehydrated through graded concentrations of alcohol and heated in a microwave oven for two cycles of 15 min each at 300 W, in citrate buffer, for antigen retrieval. Endogenous peroxidase activity was blocked with H_2O_2 solution in methanol (0.01 M), for 30 min. After washing with phosphate-buffered saline (PBS) for 5 min, the primary antibodies NCC-M53 (dilution 1:1000) and CM170B (dilution 1:50) were applied for incubation (30 min at room temperature and overnight at 4°C respectively). Then the slides were washed for 10 min with PBS and were visualised with the EnVision system using diaminobezidine tetrahydrochloride as a chromogen. Finally, all sections were counterstained with haematoxylin. Positive staining of endothelial cells and lymphocytes was used as an internal positive control for dysadherin. As an internal positive control for E-cadherin, positive staining of non neoplastic ductal epithelial cells was used. As a negative control the first antibody was substituted with normal mouse immunoglobulin of the same class.

Evaluation of the staining

For each sample, at least 1000 neoplastic cells were counted, and the percentage of cancer cells with positive membranous immunostaining as well as the staining intensity were recorded. For the purposes of statistical analysis, as described previously (Shimada *et al*, 2004b; Batistatou *et al*, 2005), when more than 50% of tumour cells were stained for dysadherin, the tumour was evaluated as 'positive dysadherin expression (Dys(+))'. When less than 50% of tumour cells were stained for dysadherin, the tumour was evaluated as 'negative dysadherin expression (Dys(−))'. Regarding E-cadherin, when more than 50% of tumour cells showed complete membranous staining, the tumour was evaluated as 'preserved E-cadherin expression (E-cad(+))', while when less than 50% of tumour cells were positive, the tumour was evaluated as 'reduced E-cadherin expression (E-cad(−))'. Cytoplasmic immunostaining was considered as aberrant expression and was not included in the immunopositive cases.

Statistical analysis

Analyses were conducted in SPSS software version 11.0 (SPSS, Inc, Chicago, IL, USA). For comparisons between antibodies' expression with clinicopathological variables we used the χ^2 test. The level of statistical significance was $P<0.05$.

RESULTS

Ductal carcinoma

Membranous E-cadherin expression was detected in epithelial cells of non-neoplastic ducts and acini and this served as internal positive control. In neoplastic cells there was some variation in distribution, depending on the grade and the pattern of stroma infiltration. Specifically, all 10 (100%) Grade 1, 37 out of 45 (82.2%) Grade 2 and six out of 15 (40%) Grade 3 neoplasms showed preserved E-cadherin expression (Table 1, Figure 1A). In immunopositive Grade 2 and Grade 3 tumours the expression of E-cadherin was more heterogeneous, with variations in intensity and distribution of positive cells. Thus, cells in clusters or in tubular structures exhibited higher percentage and more intense membranous staining than individual cells infiltrating the stroma. In the periphery of the invasive ductal carcinoma an intraductal component was observed in several cases. In this *in situ* ductal component the expression of E-cadherin was similar to the non-neoplastic epithelial cells, homogeneous and stronger that the adjacent invasive component (Figure 1B).

Dysadherin expression was detected in myoepithelial cells of ducts and acini, but not in non-neoplastic epithelial cells, as well as in endothelial cells of vessels and lymphocytes, as described previously (Batistatou *et al*, 2005, 2006). Dysadherin immunostaining was observed in the membranes of the neoplastic cells and it was heterogeneous throughout the neoplasm (Figure 1C). In particular, preferential expression in diffuse than in compact infiltrative areas was detected. Overall, 'positive dysadherin expression' was found in six out of 10 (60%) Grade 1, 34 out of 45 (75.5%) Grade 2 and all 15 (100%) Grade 3 neoplasms (Table 1, Figure 1B). Interestingly, in the adjacent *in situ* ductal carcinoma a small proportion of neoplastic cells (<10%) exhibited membranous immunostaining for dysadherin (Figure 1D). Dysadherin expression was not correlated with E-cadherin expression in IDC ($P>0.05$).

Lobular carcinoma

None of the 30 infiltrating lobular carcinomas showed preserved E-cadherin expression (Table 1, Figure 2A). The vast majority

Table 1 E-cadherin and dysadherin expression in invasive breast carcinomas

Histologic type	Preserved E-cadherin expression	'Positive' dysadherin expression
Invasive ductal carcinoma		
Grade 1 (10)	10 (100 %)	6 (60%)
Grade 2 (45)	37 (82.2%)	34 (75.5%)
Grade 3 (15)	6 (40%)	15 (100%)
Total (70)	53 (75.7%)	55 (78.6%)
Invasive lobular carcinoma		
Total (30)	0 (0%)	30 (100%)

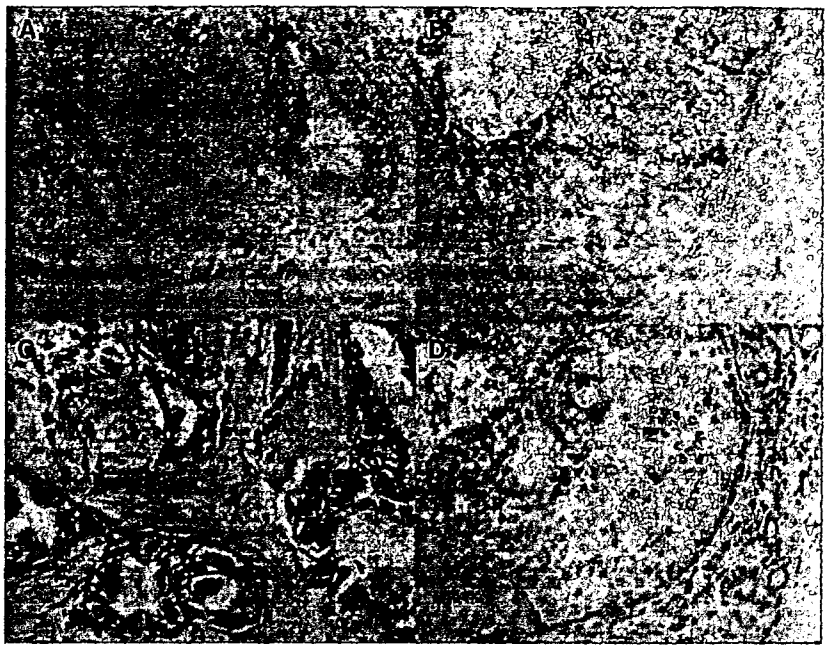

Figure 1 A case of invasive ductal carcinoma, grade II, with adjacent *in situ* component. (**A**) E-cadherin expression is significantly reduced in invasive ductal carcinoma (DABX400). (**B**) Membranous expression of E-cadherin is retained in the adjacent *in situ* component (DABX400). (**C**) Strong, membranous expression of dysadherin is evident in invasive ductal carcinoma (DABX400). (**D**) Dysadherin is not expressed in the adjacent *in situ* component (DABX400).

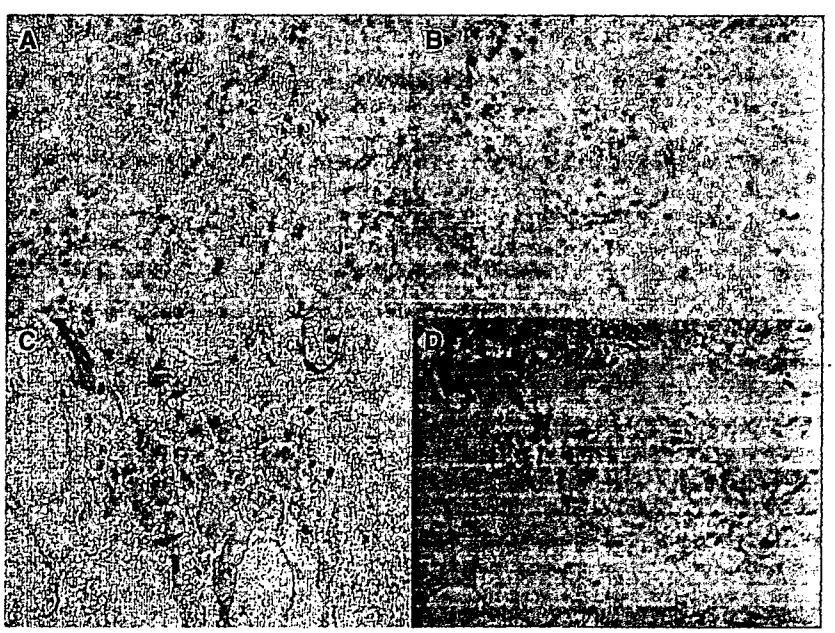

Figure 2 A case of invasive lobular carcinoma, with adjacent *in situ* component. (**A**) E-cadherin expression is lost in invasive lobular carcinoma (DABX400). (**B**) E-cadherin expression is lost in the adjacent *in situ* component (DABX400). (**C**) Membranous expression of dysadherin is evident in invasive lobular carcinoma, (DABX400). (**D**) Dysadherin is not expressed in the adjacent *in situ* component, in contrast with the infiltrating tumour (DABX400).

was completely negative, while only in two of them, <20% of neoplastic cells showed weak membranous and cytoplasmic immunopositivity. Interestingly, the adjacent *in situ* lobular carcinoma was completely negative, as well (Figure 2B).

All the 30 infiltrating lobular carcinomas exhibited 'positive dysadherin expression' (Figure 2C). In this *in situ* lobular component the expression of dysadherin was limited to a small proportion (<10%) of neoplastic cells (Figure 2D).

DISCUSSION

Two of the most important characteristics of neoplastic cells are their abilities to grow locally and to metastasise. For both of these processes tumour cells must initially dissociate from each other, either singly or in small nests and invade the surrounding stroma. Today it is generally accepted that at least for carcinomas, adhesion molecules, in particular E-cadherin,

play a pivotal role in this process by being downregulated (Hirohashi, 1998; Charalabopoulos *et al*, 2002; Hirohashi and Kanai, 2003). In general, there is an association between aberrant E-cadherin expression, tumour dedifferentiation and poor clinical outcome.

Regarding breast cancer, E-cadherin expression varies depending on the histological subtype. Thus, in ductal carcinoma E-cadherin is expressed, albeit in reduced levels and aberrant cellular locations. Although E-cadherin correlates inversely with the grade of the tumour, reduced E-cadherin expression is not adequate to predict clinical outcome and there are contradictory studies on the association between E-cadherin and survival (Knudsen and Wheelock, 2005). Moreover, it has been reported that in other breast cancers with known poor prognosis, such as inflammatory breast cancer, there is overexpression of E-cadherin (Knudsen and Wheelock, 2005). An interesting concept is that possibly the reduction of E-cadherin expression in breast carcinomas, other than lobular, is transient, due to epigenetic modifications. Several mechanisms for reversible reduction of E-cadherin expression in human neoplasms have been reported (Hirohashi, 1998; Charalabopoulos *et al*, 2002; Hirohashi and Kanai, 2003). Among them, recently, a novel cell membrane glycoprotein named 'dysadherin' (from the Greek prothema dys-, which means difficulty, or aberration, or reversibility) has been shown to downregulate E-cadherin in a post-transcriptional manner and reduce cell–cell adhesiveness in *in vitro* studies and in animal models. Dysadherin is a member of the FXYD family (FXYD5 or Related to Ion Channel). It is located at chromosome 19 and has a single transmembrane domain. It interacts with and modulates the properties of the Na^+, K^+ ATPase (Ino *et al*, 2002; Lubarski *et al*, 2005). In human tissues increased dysadherin expression has been correlated with the development of metastasis and poor prognosis in gastric, pancreatic, colorectal, oesophageal, thyroid, tongue and cervical carcinomas, as well as in malignant melanoma (Aoki *et al*, 2003; Sato *et al*, 2003; Shimamura *et al*, 2003, 2004; Nakanishi *et al*, 2004; Shimada *et al*, 2004a, b; Wu *et al*, 2004; Batistatou *et al*, 2005, 2006; Nishizawa *et al*, 2005; Kyzas *et al*, 2006). Furthermore, in a small pilot series of breast cancer patients, dysadherin expression was correlated with poor prognosis (Ino *et al*, 2002). In most of these neoplasms, as well as in testicular tumours and lymph node metastases of colorectal adenocarcinoma, increased dysadherin expression was correlated with reduced E-cadherin expression (Aoki *et al*, 2003; Sato *et al*, 2003; Shimamura *et al*, 2003, 2004; Wu *et al*, 2004; Batistatou *et al*, 2005, 2006). In invasive ductal carcinoma, as reported in this study, there is an increase in dysadherin expression, which is not related to E-cadherin expression. This lack of association has also been reported in pancreatic, primary colorectal and gastric carcinomas (Shimamura *et al*, 2003; Shimada *et al*, 2004b; Batistatou *et al*, 2005). Furthermore, in the *in situ* ductal carcinoma dysadherin was not expressed. On the basis of these data we would like to propose that in ductal carcinomas, dysadherin can promote invasion independently of the E-cadherin expression.

Lobular breast carcinomas, typically exhibit loss of E-cadherin expression, but they tend to have a more favourable clinical outcome than the more common ductal carcinomas. This loss is an early event affecting not only lobular carcinoma *in situ* but even atypical lobular neoplasia (Mastracci *et al*, 2005). This silencing of E-cadherin is attributed to genetic as well as epigenetic events (Knudsen and Wheelock, 2005; Mastracci *et al*, 2005). In approximately 50% of lobular carcinomas loss of E-cadherin involves LOH at the chromosomal region of 16q, which includes the E-cadherin gene CDH1 locus and mutations in the remaining allele (Kanai *et al*, 1994; Vos *et al*, 1997; Huiping *et al*, 1999; Knudsen and Wheelock, 2005; Mastracci *et al*, 2005). This LOH definitely accompanies mutations in cases of invasive lobular carcinoma, however, the classic pattern of LOH coupled with inactivating mutations in lobular carcinoma *in situ* has not been confirmed. In the other 50% loss of E-cadherin is attributed to epigenetic events, with hypermethylation of the E-cadherin promoter region at CpG islands being one of the most important ones and possibly occurring very early, even at the stage of atypical lobular hyperplasia (Sarrio *et al*, 2004; Shibata *et al*, 2004; Knudsen and Wheelock, 2005; Mastracci *et al*, 2005).

In this study we have confirmed the loss of E-cadherin expression, in *in situ* and invasive breast carcinoma. On the basis of the rare expression of dysadherin in lobular carcinoma *in situ* we can conclude that dysadherin is not responsible for E-cadherin downregulation in lobular carcinoma. An interesting finding from our study is the difference in dysadherin expression between *in situ* and invasive lobular carcinoma. In breast carcinoma the progression from *in situ* to invasive disease is not clearly defined and the specific events that mark the transition to an invasive tumour are under intense investigation. Lack of E-cadherin expression cannot be associated with an invasive phenotype, since it is also evident in the *in situ* component. On the other hand, we have shown that dysadherin is specifically and constantly expressed in invasive lobular carcinomas. On the basis of this finding we propose that dysadherin is a possible causative player in the process of acquiring an invasive phenotype, as well as a possible marker for invasiveness. It has also been proposed that loss of E-cadherin expression is responsible for the distinct pattern of invasion observed in lobular neoplasms (Knudsen and Wheelock, 2005; Mastracci *et al*, 2005). We would like to add that another major contributor to this characteristic invasion pattern, with single cells arranged in cords is the expression of dysadherin. The latter possibly acts, either alone or in conjunction with loss of E-cadherin, by allowing cells to dissociate from each other. Studies on the function of dysadherin are available by experimental data, where dysadherin appears to play an important role in neoplastic cell invasion and metastasis (Ino *et al*, 2002; Nam *et al*, 2006). The exact molecular mechanisms of these effects have not been elucidated yet. Recently, it has been shown that, besides downregulating E-cadherin, dysadherin can promote invasion at least in breast cancer cells *in vitro*, through an E-cadherin-independent mechanism. This mechanism involves enhanced signaling through the NF-κB pathway, which leads to increased production of the tumour-promoting (C-C motif) ligand 2 (CCL2) (Nam *et al*, 2006).

In conclusion, in this study we have investigated the role of specific adhesion/dysadhesion molecules in the development of breast carcinoma. We have selected ductal carcinoma which is by far the most common type, and lobular carcinoma which has a distinctive microscopic appearance. We have shown similarities and differences between these two types. Interestingly, in ductal as well as in lobular carcinoma, dysadherin was expressed only in the invasive and not in the *in situ* component, and this expression was independent of E-cadherin. Thus, dysadherin may play an important role in breast cancer progression by promoting invasion and, particularly in lobular carcinomas, it might also be used as a marker of invasion.

REFERENCES

Aoki S, Shimamura T, Shibata T, Nakanishi Y, Moriya Y, Sato Y, Kitajima M, Sakamoto M, Hirohashi S (2003) Prognostic significance of dysadherin expression in advanced colorectal carcinoma. *Br J Cancer* 88: 726–732

Batistatou A, Charalabopoulos AK, Scopa CD, Nakanishi Y, Kappas A, Hirohashi S, Agnantis NJ, Charalabopoulos K (2006) Expression patterns of dysadherin and E- cadherin in lymph node metastases of colorectal carcinoma. *Virchows Arch* 448: 763–777

Batistatou A, Scopa CD, Ravazoula P, Nakanishi Y, Peschos D, Agnantis NJ, Hirohashi S, Charalabopoulos KA (2005) Involvement of dysadherin and E-cadherin in the development of testicular tumors. *Br J Cancer* 93: 1382–1387

Becker KF, Atkinson MJ, Reich U, Becker I, Nekarda H, Siewert JR, Hofler H (1994) E- cadherin gene mutations provide clues to diffuse type gastric carcinomas. *Cancer Res* 54: 3845–3852

Chang HW, Chow V, Lam KY, Wei WI, Yuen A (2002) Loss of E-cadherin expression resulting from promoter hypermethylation in oral tongue carcinoma and its prognostic significance. *Cancer* 94: 386–392

Charalabopoulos K, Binolis J, Karkabounas S (2002) Adhesion molecules in carcinogenesis. *Exp Oncol* 24: 249–257

Charalabopoulos K, Gogali A, Kostoula OK, Constantopoulos H (2004) Cadherin superfamily of adhesion molecules in primary lung cancer. *Exp Oncol* 16: 256–260

Charpin C, Garcia S, Bouvier C, Devictor B, Andrac L, Choux R, Lavaut M (1997) E-cadherin quantitative immunocytochemical assays in breast carcinomas. *J Pathol* 431: 317–321

Elzagheid A, Kuopio T, Ilmen M, Collan Y (2002) Prognostication of invasive ductal breast cancer by quantification of E-cadherin immunostaining:the methodology and clinical relevance. *Histopathology* 41: 127–133

Gillett CE, Miles DW, Ryder K, Skilton D, Liebman RD, Springall RJ, Barnes DM, Hanby AM (2001) Retention of the expression of E-cadherin and catenins is associated with shorter survival in grade III ductal carcinoma of the breast. *J Pathol* 193: 433–441

Gupta A, Desphande CG, Badve S (2003) Role of E-cadherins in development of lymphatic tumor emboli. *Cancer* 97: 2341–2347

Heimann R, Lan F, McBride R, Hellman S (2000) Separating favourable from unfavourable prognostic markers in breast cancer: the role of E-cadherin. *Cancer Res* 60: 298–304

Hirohashi S (1998) Inactivation of E-cadherin-mediated cell adhesion system in human cancers. *Am J Pathol* 153: 333–339

Hirohashi S, Kanai Y (2003) Cell adhesion system and human cancer morphogenesis. *Cancer Sci* 94: 575–581

Huiping C, Sigurgeirsdottir JR, Jonasson JG, Eiriksdottir G, Johannsdottir JT, Egilsson V, Ingvarsson S (1999) Chromosome alterations and E-cadherin gene mutations in human lobular breast cancer. *Br J Cancer* 81: 1003–1010

Ino Y, Gotoh M, Sakamoto M, Tsukagoshi K, Hirohashi S (2002) Dysadherin, a cancer-associated cell membrane glycoprotein, downregulates E-cadherin and promotes metastasis. *Proc Natl Acad Sci USA* 99: 365–370

Kanai Y, Oda T, Tsuda H, Ochiai A, Hirohashi S (1994) Point mutation of the E-cadherin gene in invasive lobular carcinoma of the breast. *Jpn J Cancer Res* 85: 1035–1039

Knudsen KA, Wheelock MJ (2005) Cadherins and the mammary gland. *J Cel Biol* 95: 488–496

Kyzas PA, Stefanou D, Batistatou A, Agnantis NJ, Nakanishi Y, Hirohashi S, Charalabopoulos K (2006) Dysadherin expression in head and neck squamous cell carcinoma: association with lymphangiogenesis and prognostic significance. *Am J Surg Pathol* 30: 185–193

Lubarski I, Pihakaski-Maunshbach K, Karlish SJ, Maunsbach AB, Garty H (2005) Interaction with the Na,K-Atpase and tissue distribution of FXYD5 (related to ion channel). *J Biol Chem* 280: 37717–37724

Massarelli E, Brown E, Tran NK, Liu DD, Izzo JG, Lee JJ, El-Naggar AK, Hong WK, Papadimitrakopoulou VA (2005) Loss of E-cadherin and p27 expression is associated with head and neck squamous tumorigenesis. *Cancer* 103: 952–959

Mastracci TL, Tjan S, Bane AL, O'Malley FP, Andrulis IL (2005) E-cadherin alterations in atypical lobular hyperplasia and lobular carcinoma *in situ* of the breast. *Mod Pathol* 18: 741–751

Nakanishi Y, Akimoto S, Sato Y, Kanai Y, Sakamoto M, Hirohashi S (2004) Prognostic significance of dysadherin expression in tongue cancer: immunohistochemical analysis of 91 cases. *Appl Immunohistochem Mol Morphol* 12: 323–328

Nam J-S, Kang M-J, Suchar AM, Shimamura T, Kohn EA, Michalowska AM, Jordan VC, Hirohashi S, Wakefield LM (2006) Chemokine (C-C motif) ligand 2 mediates the prometastatic effect of dysadherin in human breast cancer cells. *Cancer Res* 66: 7176–7184

Nishizawa A, Nakanishi Y, Yoshimura K, Sasajima Y, Yamazaki N, Yamamoto A, Hanada K, Kanai Y, Hirohashi S (2005) Clinicopathologic significance of dysadherin expression in cutaneous malignant melanoma. *Cancer* 103: 1693–1700

Oka H, Shiozaki H, Kabayashi K, Inoue M, Tahara H, Kobayashi T, Takatsuka Y, Matsuyoshi N, Hirano S, Takeichi M (1993) Expression of E-cadherin cell adhesion molecules in human breast cancer tissues and its relationship to metastasis. *Cancer Res* 53: 1696–1701

Parker C, Rampaul RS, Pinder SE, Bell JA, Wencyk PM, Blamey RW, Nicholson RI, Robertson JF (2001) E-cadherin as a prognostic indicator in primary breast cancer. *Br J Cancer* 85: 1958–1963

Rakha EA, Abd El, Rehim D, Pinder SE, Lewis Sa, Ellis IO (2005) E-cadherin expression in invasive non-lobular carcinoma of the breast and its prognostic significance. *Histopathology* 46: 685–693

Sarrio D, Perez-Mies B, Hardisson D, Moreno-Bueno G, Suarez A, Cano A, Martin-Perez J, Gamallo C, Palacios J (2004) Cytoplasmic localization of p120ctn and E_cadherin loss characterize lobular breast carcinoma from preinvasive to metastatic lesions. *Oncogene* 23: 3272–3283

Sato H, Ino Y, Miura A, Abe Y, Sakai H, Ito K, Hirohashi S (2003) Dysadherin: expression and clinical significance in thyroid carcinoma. *J Clin Endocrinol Metab* 88: 4407–4412

Shibata T, Kokubu A, Sekine S, Kanai Y, Hirohashi S (2004) Cytoplasmic p120ctn regulates the invasive phenotypes of E-cadherin deficient breast cancer. *Am J Pathol* 164: 2269–2278

Shimada Y, Hashimoto Y, Kan T, Kawamura J, Okumura T, Soma T, Kondo K, Teratani N, Watanabe G, Ino Y, Sakamoto M, Hirohashi S, Imamura M (2004a) Prognostic significance of dysadherin expression in esophageal squamous cell carcinoma. *Oncology* 67: 73–80

Shimada Y, Yamasaki S, Hashimoto Y, Ito T, Kawamura J, Soma T, Ino Y, Nakanishi Y, Sakamoto M, Hirohashi S, Imamura M (2004b) Clinical significance of dysadherin expression in gastric cancer patients. *Clin Cancer Res* 10: 2818–2823

Shimamura T, Sakamoto M, Ino Y, Sato Y, Shimada K, Kosuge T, Sekihara H, Hirohashi S (2003) Dysadherin overexpression in pancreatic ductal adenocarcinoma reflects tumor aggressiveness: relationship to E-cadherin expression. *J Clin Oncol* 21: 659–667

Shimamura T, Yasuda J, Ino Y, Gotoh M, Tsuchiya A, Nakajima A, Sakamoto M, Kanai Y, Hirohashi S (2004) Dysadherin expression facilitates cell motility and metastatic potential of human pancreatic cancer cells. *Cancer Res* 64: 6989–6995

Tsuiji H, Takasaki S, Sakamoto M, Irimura T, Hirohashi S (2002) Aberrant O-glycosylation inhibits stable expression of dysadherin, a carcinoma-associated antigen, and facilitates cell-cell adhesion. *Glycobiology* 13: 521–527

Vos CB, Cleton-Jansen AM, Berx G, de Leeuw WJ, ter Haar NT, van Roy F, Cornelisse CJ, Peterse JL, van de Vijver MJ (1997) E-cadherin inactivation in lobular carcinoma *in situ* of the breast:an early event in tumorigenesis. *Br J Cancer* 76: 1131–1133

Wu D, Qiao Y, Kristensen GB, Li S, Troen G, Holm R, Nesland JM, Suo Z (2004) Prognostic significance of dysadherin expression in cervical squamous cell carcinoma. *Pathol Oncol Res* 10: 212–218

Modern Pathology (2007) 20, 749–759

www.modernpathology.org

Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma

Robert Nakayama[1,2,3], Takeshi Nemoto[2,4], Hiro Takahashi[2], Tsutomu Ohta[2], Akira Kawai[5], Kunihiko Seki[6], Teruhiko Yoshida[2], Yoshiaki Toyama[3], Hitoshi Ichikawa[1] and Tadashi Hasegawa[6,7]

[1]*Cancer Transcriptome Project, National Cancer Center Research Institute, Tokyo, Japan;* [2]*Genetics Division, National Cancer Center Research Institute, Tokyo, Japan;* [3]*Department of Orthopaedic Surgery, Keio University School of Medicine, Tokyo, Japan;* [4]*Department of Dermatology, Tokyo Medical and Dental University, Tokyo, Japan;* [5]*Orthopedics Division, National Cancer Center Hospital, Tokyo, Japan;* [6]*Pathology Division, National Cancer Center Hospital, Tokyo, Japan and* [7]*Department of Surgical Pathology, Sapporo Medical University School of Medicine, Sapporo, Japan*

In soft tissue sarcomas, the diagnosis of malignant fibrous histiocytoma (MFH) has been a very controversial issue, and MFH is now considered to be reclassified into pleomorphic subtypes of other sarcomas. To characterize MFH genetically, we used an oligonucleotide microarray to analyze gene expression in 105 samples from 10 types of soft tissue tumors. Spindle cell and pleomorphic sarcomas, such as dedifferentiated liposarcoma, myxofibrosarcoma, leiomyosarcoma, malignant peripheral nerve sheath tumor (MPNST), fibrosarcoma and MFH, showed similar gene expression patterns compared to other tumors. Samples from those five sarcoma types could be classified into respective clusters based on gene expression by excluding MFH samples. We calculated distances between MFH samples and other five sarcoma types (dedifferentiated liposarcoma, myxofibrosarcoma, leiomyosarcoma, MPNST and fibrosarcoma) based on differentially expressed genes and evaluated similarities. Three of the 21 MFH samples showed marked similarities to one of the five sarcoma types, which were supported by histological findings. Although most of the remaining 18 MFH samples showed little or no histological resemblance to one of the five sarcoma types, 12 of them showed moderate similarities in terms of gene expression. These results explain the heterogeneity of MFH and show that the majority of MFHs could be reclassified into pleomorphic subtypes of other sarcomas. Taken together, gene expression profiling could be a useful tool to unveil the difference in the underlying molecular backgrounds, which leads to a rational taxonomy and diagnosis of a diverse group of soft tissue sarcomas.
Modern Pathology (2007) **20,** 749–759; doi:10.1038/modpathol.3800794; published online 27 April 2007

Keywords: gene expression; malignant fibrous histiocytoma; myxofibrosarcoma; soft tissue sarcoma; reclassification; undifferentiated pleomorphic sarcoma

Malignant soft tissue tumors are a diverse group of tumors of mesenchymal origin, which have generally been classified according to their histological resemblance to normal tissue. Understanding of molecular pathology gained in recent decades shows that some soft tissue tumors exhibit single recurrent genetic aberrations, such as chromosomal translocations resulting in gene fusions (*SYT-SSX* in synovial sarcoma, *TLS-CHOP* in myxoid/round cell liposarcoma) or somatic mutations (*KIT* in gastrointestinal stromal tumors), and they are now classified by these molecular markers specific to each tumor.[1] In contrast, other malignant soft tissue tumors, such as malignant fibrous histiocytoma (MFH), fibrosarcoma and leiomyosarcoma, are characterized by numerous, nonrecurrent complex chromosomal aberrations, and frequently show overlapping histological appearance and immunohistochemical phenotypes that are often difficult to

Correspondence: Dr H Ichikawa, PhD, Cancer Transcriptome Project, National Cancer Center Research Institute, 5–1–1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan.
E-mail: hichikaw@ncc.go.jp
Received 9 January 2007; revised 20 March 2007; accepted 27 March 2007; published online 27 April 2007

interpret by pathologists.[2] Among them, diagnosis of MFH has been the most controversial issue.[3–5] MFH has been considered the most common soft tissue sarcoma of adults; it is manifested by a broad range of histological appearances and consists of four subtypes: storiform and pleomorphic type, myxoid type, giant cell type and inflammatory type. Recent clinicopathological, ultrastructural and immunohistochemical studies revealed that MFH shows no evidence of true histiocytic differentiation and that it is not a single entity but rather a heterogeneous collection of pleomorphic subtypes of other sarcomas. Since each type of sarcoma other than MFH shows distinct biological behavior, particularly in local recurrence or metastasis rate, MFH showing a variety of clinicopathological characteristics should be further reclassified to correctly evaluate the malignant potential of each case. In the latest edition of the WHO classification, myxoid type MFH was classified as myxofibrosarcoma in the fibroblastic category, and other subtypes of MFH without any evidence of differentiation were classified as undifferentiated high grade pleomorphic sarcoma.[6] WHO classification also suggested that the term 'MFH' might disappear when criteria for the diagnosis of pleomorphic sarcomas showing a distinct differentiation state can be reproducibly defined.[6] In this work, we used the term 'MFH' to identify tumors diagnosed as storiform and pleomorphic type MFH, and the term 'myxofibrosarcoma' for so-called MFH with predominant (>50%) myxoid features conventionally diagnosed as myxoid type MFH.

Recently, several studies report gene expression profiling of soft tissue tumors using microarray technologies to provide new insights into the tumor characterization. They described distinct patterns of gene expression in respective tumors with single, recurrent genetic aberrations, such as synovial sarcoma, myxoid/round cell liposarcoma, clear cell sarcoma or gastrointestinal stromal tumors, and heterogeneous patterns in spindle cell and pleomorphic sarcomas which are generally characterized by complex chromosomal aberrations.[7–12] No further detailed analysis of gene expression in spindle cell and pleomorphic sarcomas have been reported so far.

In this study, we analyzed gene expression profile of total 105 cases representing 10 types of soft tissue tumors to identify their molecular characteristics. We observed similarity in gene expression among spindle cell and pleomorphic sarcomas, forming a relatively loose cluster, which is separated from the distinct clusters of synovial sarcoma, myxoid/round cell liposarcoma and lipoma + well-differentiated liposarcoma. Next, we primarily analyzed 64 samples of spindle cell and pleomorphic sarcomas and showed heterogeneity of MFH in terms of gene expression. We selected genes that could clearly distinguish between dedifferentiated liposarcoma, myxofibrosarcoma, leiomyosarcoma, malignant peripheral nerve sheath tumor (MPNST) and fibrosarcoma and quantified similarities as distances between MFH samples and the five sarcoma types.

Materials and methods

Patients and Tumor Samples

Characteristics of 105 soft tissue tumors used in this study are shown in Supplementary data 1. Among them, 35 samples were previously analyzed in a different method.[13] All patients received histological diagnosis of primary soft tissue tumor at National Cancer Center Hospital, Tokyo, from 1996 to 2002. In this paper, we use the term 'MFH' to describe samples diagnosed as storiform and pleomorphic type MFH showing predominant pleomorphic features without immunohistochemical phenotypes characteristic of specific differentiation, and the term 'myxofibrosarcoma' to describe MFH with predominant (>50%) myxoid features conventionally diagnosed as myxoid type MFH. Before the gene expression analysis pathologists confirmed the diagnosis of MFH was appropriate at the time of primary diagnosis. Tumor samples were collected from the part with macroscopically high tumor content by pathologists immediately after surgical excision and cryopreserved in liquid nitrogen until use. This study was approved by the ethics committee of National Cancer Center and conducted according to tenets of the Declaration of Helsinki.

Gene Expression Profiling

Total RNA was isolated using TRIzol (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instruction. Samples were analyzed with a GeneChip Human Genome U133A array (Affymetrix, Santa Clara, CA, USA) containing 22 283 probe sets. Target cRNA preparation from total RNA, hybridization to the microarray, washing and staining with an antibody amplification procedure and scanning were all carried out according to the manufacturer's instructions. The expression value (Signal) of each probe set was calculated using GeneChip Operating Software (GCOS) ver. 1.3 (Affymetrix), so that the mean of expression values in each experiment was set at 100 to adjust for minor differences between experiments.

Statistical Analysis

Gene expression data were subsequently imported into GeneSpring GX7.2 software (Agilent Technologies, Santa Clara, CA, USA) and normalized to the median of all samples enrolled in the analysis and log-transformed for each gene. Hierarchical clustering analysis was performed using Pearson's correlation. To select appropriate probe sets defining five

types of spindle cell and pleomorphic sarcomas (dedifferentiated liposarcoma, myxofibrosarcoma, leiomyosarcoma, MPNST and fibrosarcoma), we performed Student's *t*-tests between one and the other four sarcoma types. The top 50 probe sets with low *P*-values in each *t*-test were summed. The centroid of each sarcoma type was determined by calculating the average of the selected probe sets. The distance (*D*) from a centroid to a sample was defined as $D=1-r$, using Pearson's correlation coefficient (r, $-1 \leq r \leq 1$). Inter-centroid distances were also calculated using Pearson's correlation coefficient.

Histological Analysis

Histological sections of the tumors were stained with hematoxylin and eosin and reviewed for all samples, and representative sections were examined immunohistochemically using the labeled streptavidin–biotin method. Sections were dewaxed, rehydrated and moistened with phosphate-buffered saline (pH 7.4), autoclaved at 121°C for 10 min in 10 mM citrate buffer (pH 6.0) and incubated with antibodies to the following molecules on an automated immunostaining system i6000 (BioGenex, San Ramon, CA, USA) for 30 min, as described previously:[14] vimentin, desmin, α-smooth muscle actin (αSMA), muscle-specific actin, h-caldesmon, CD34, S-100 protein, epithelial membrane antigen, cytokeratin and neurofilament. Heat-induced epitope retrieval was not undertaken when sections were stained with antibodies to S-100 protein and epithelial membrane antigen.

Quantitative RT-PCR

Real-time quantitative reverse transcription (RT)-PCR was carried out using the 7500 Fast Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) with FastStart TaqMan Probe Master (Rox) and Universal ProbeLibrary (Roche Applied Science, Mannheim, Germany). One microgram of total RNA from 17 tumor samples (myxofibrosarcoma ($n=5$), MFH ($n=7$), leiomyosarcoma ($n=2$) and MPNST ($n=3$)) was reverse-transcribed to synthesize single-stranded cDNAs using SuperScript III (Invitrogen), and 1/100 of the cDNA was used for each PCR. Probes and primers were designed using Probe Finder software (Roche Applied Science) (Supplementary data 2). Transcript levels were normalized to that of the *ACTB* transcript.

Results

Overview of Gene Expression in Soft Tissue Sarcomas

Gene expression data of 105 soft tissue tumor samples consisting of synovial sarcoma ($n=16$), myxoid/round cell liposarcoma ($n=19$), lipoma ($n=3$), well-differentiated liposarcoma ($n=3$), dedifferentiated liposarcoma ($n=15$), myxofibrosarcoma ($n=15$), leiomyosarcoma ($n=6$), MPNST ($n=3$), fibrosarcoma ($n=4$) and MFH ($n=21$) were obtained using an oligonucleotide microarray containing 22 283 probe sets. Among them, 12 599 probe sets whose expression values were not less than 100 in at least 3 of 105 samples were analyzed. To overview the transcriptome of sarcomas in our data set, we first performed principal component analysis with 12 599 probe sets (Figure 1a), which is a decomposition technique to reduce multidimensional data into several specialized dimensions. The *x* and *y* axes in Figure 1a indicate the first and second principal components, respectively, representing the top and second largest fractions of the overall variability. In this analysis, 105 samples were roughly classified into four groups based on their position relative to the first and second principal components. Both synovial sarcoma and myxoid/round cell liposarcoma samples were located on the negative side of the first principal component, while well-differentiated liposarcoma, dedifferentiated liposarcoma and other spindle cell and pleomorphic sarcoma samples were on the positive side. On the negative side of the second principal component were myxoid/round cell liposarcoma, well-differentiated liposarcoma and lipoma samples, all of which are adipocytic tumors. Interestingly, some dedifferentiated liposarcoma samples were distributed close to well-differentiated liposarcoma samples, while others were midway between well-differentiated liposarcoma and other spindle cell and pleomorphic sarcoma samples. These results suggest that the first principal component was associated with the difference between synovial sarcoma + myxoid/round cell liposarcoma and spindle cell and pleomorphic sarcomas, and that the second principal component was associated with adipocytic differentiation. Probe sets contributing significantly to the first and second principal components are listed in Supplementary data 3.

To identify genes whose expression differed in a statistically significant manner among all sarcoma types, we performed an analysis of variance (ANOVA) among 10 tumor types and selected 2590 probe sets with *P*-values of less than 1.0×10^{-5}. Two-dimensional hierarchical clustering analysis using those 2590 probe sets showed that synovial sarcoma and myxoid/round cell liposarcoma samples displayed distinct gene expression profiles and formed robust clusters (Figure 1b). On the other hand, myxofibrosarcoma, leiomyosarcoma, MPNST, fibrosarcoma and MFH samples did not show distinct gene expression profiles, but rather formed a single loose cluster and shared a similar expression profile. We also found that lipoma and well-differentiated liposarcoma samples and some of the dedifferentiated liposarcoma samples displayed similar gene expression profiles and formed a cluster, whereas

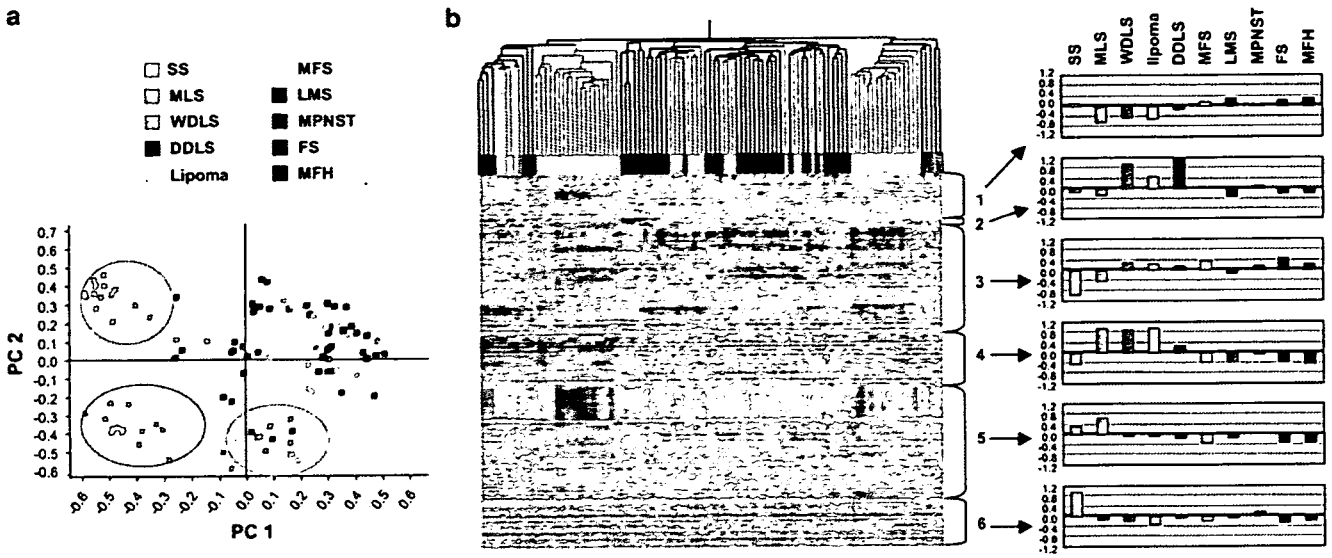

Figure 1 Gene expression overview of 105 soft tissue tumors. (**a**) Principal component analysis. A total of 12 599 probe sets with expression values not less than 100 in at least three samples were used in this analysis. *x* and *y* axes represent the first and second principal components (PC1 and PC2), respectively. Each dot represents a sample colored by its histological type. (**b**) Two-dimensional hierarchical clustering analysis. A total of 2590 probe sets differentially expressed among histological types ($P<1.0\times10^{-5}$ by ANOVA) were used. Columns represent samples and rows represent probe sets. Red and green indicate high and low expression, respectively. The 2590 probe sets were roughly divided into six clusters (clusters 1–6). The six graphs on the right show averages of normalized expression values of those clusters for each histological type. Note that spindle cell and pleomorphic sarcomas, such as dedifferentiated liposarcoma, myxofibrosarcoma, leiomyosarcoma, MPNST, fibrosarcoma and MFH, form a loose cluster and share a similar expression profile compared with synovial sarcoma, myxoid/round cell liposarcoma, well-differentiated liposarcoma and lipoma. SS, synovial sarcoma; MLS, myxoid/round cell liposarcoma; WDLS, well-differentiated liposarcoma; DDLS, dedifferentiated liposarcoma; MFS, myxofibrosarcoma; LMS, leiomyosarcoma and FS, fibrosarcoma.

the other dedifferentiated liposarcoma samples did not share that profile but instead formed a loose cluster with fibrosarcoma, myxofibrosarcoma and MFH samples.

The 2590 probe sets were classified into six clusters according to their expression patterns (Figure 1b and Supplementary data 4). Interestingly, we found two major clusters (clusters 3 and 5) whose expression patterns were similar between synovial sarcoma and myxoid/round cell liposarcoma samples. Cluster 3, whose expression was low in synovial sarcoma and myxoid/round cell liposarcoma, contained many HLA genes, and cluster 5, whose expression was high in both synovial sarcoma and myxoid/round cell liposarcoma, contained many genes encoding ribosomal proteins and cancer testis antigens, such as *CTAG1B*, *CTAG2* and *PRAME*. Of note, these genes contributed largely to the first principal component (see Supplementary data 3). On the other hand, cluster 1, whose expression was low in myxoid/round cell liposarcoma, well-differentiated liposarcoma and lipoma samples, included cell cycle associated genes such as *CCNB1*, *CDKN3*, and *CDC20*, while cluster 4, whose expression was high in myxoid/round cell liposarcoma, well-differentiated liposarcoma and lipoma samples, included adipocytic differentiation-associated genes such as *LPL*, *ACACB* and *PLIN*. These genes contributed largely to the second principal component (see Supplementary data 3). Cluster 6, whose expression was high in synovial sarcoma, included *COL2A1*, *COL9A3*, *SSX1* and *SSX2*. The small but robust cluster, cluster 2, consisted of *MDM2*, *CDK4* and other genes located in 12q13-15, which are known to be amplified in both well-differentiated liposarcoma and dedifferentiated liposarcoma.

Heterogeneity of MFH in Gene Expression and Classification of Spindle Cell and Pleomorphic Sarcomas

Spindle cell and pleomorphic sarcomas frequently display overlapping histological appearance and immunohistochemical phenotypes. Samples from these types of sarcoma did not separate into distinct histological types in the analysis using whole samples (Figure 1). To determine whether they could be grouped by gene expression, we analyzed 64 samples of spindle cell and pleomorphic sarcomas (dedifferentiated liposarcoma, myxofibrosarcoma, leiomyosarcoma, MPNST, fibrosarcoma and MFH). We performed principal component analysis with 11 300 probe sets whose expression values were not less than 100 in at least three of 64 samples, and two-dimensional hierarchical clustering analysis using 1671 probe sets selected by ANOVA among six sarcoma types ($P<0.01$) (Supplementary data 5). In the clustering analysis,

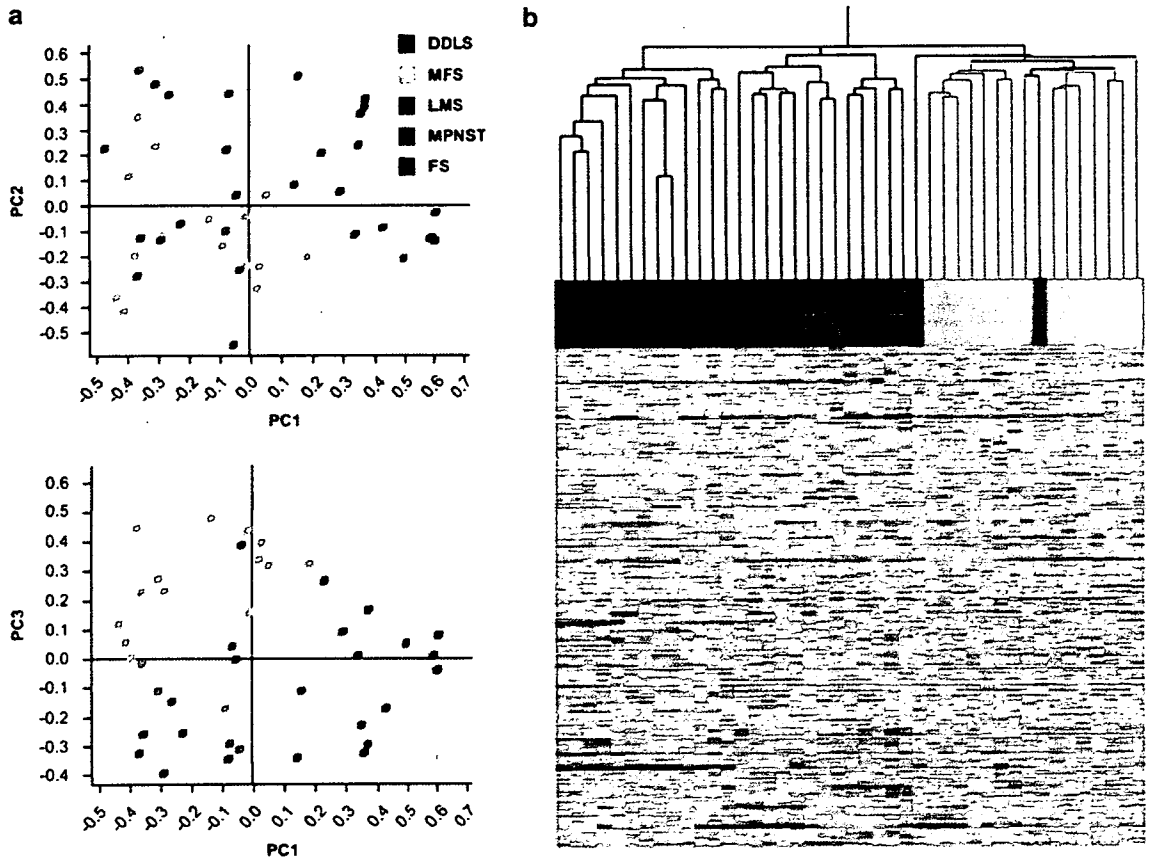

Figure 2 Classification of spindle cell and pleomorphic sarcomas without MFH. (**a**) Principal component analysis. A total of 11 300 probe sets with expression values not less than 100 in at least three of 64 spindle cell and pleomorphic sarcoma samples including MFH were used in this analysis. *x* and *y* axes in the upper panel represent the first and second principal components (PC1 and PC2), and *x* and *y* axes in the lower panel represent the first and third principal components (PC1 and PC3), respectively. (**b**) Two-dimensional hierarchical clustering analysis. A total of 1457 probe sets differentially expressed among five types of spindle cell and pleomorphic sarcomas ($P<0.01$ by ANOVA) were used. Columns represent samples and rows represent probe sets. Red and green indicate high and low expression, respectively. Note that most samples formed clusters corresponding to their histology. DDLS, dedifferentiated liposarcoma; MFS, myxofibrosarcoma; LMS, leiomyosarcoma and FS, fibrosarcoma.

dedifferentiated liposarcoma, myxofibrosarcoma, leiomyosarcoma, MPNST and fibrosarcoma samples appeared to form their own clusters, whereas those of MFH partitioned into several groups, some close to clusters of other sarcomas. These results suggest that MFH is heterogeneous in terms of gene expression as observed histologically.

Next, we analyzed 43 samples of five spindle cell and pleomorphic sarcomas (dedifferentiated liposarcoma, myxofibrosarcoma, leiomyosarcoma, MPNST and fibrosarcoma) and excluded MFH samples. In principal component analysis with 11 300 probe sets, samples of the same tumor type appeared to cluster (Figure 2a). We then performed two-dimensional hierarchical clustering analysis with 1457 probe sets selected from the 11 300 probe sets by ANOVA among five sarcoma types ($P<0.01$) (Figure 2b). Although we found three exceptions (one leiomyosarcoma and two dedifferentiated liposarcoma samples), almost all dedifferentiated liposarcoma, myxofibrosarcoma, leiomyosarcoma, MPNST and fibrosarcoma samples formed their own respective clusters suggesting that each type of spindle cell and pleomorphic sarcoma formed a homogeneous group in terms of gene expression by excluding MFH samples.

Distances of MFH Samples from Other Spindle Cell and Pleomorphic Sarcomas

Since MFH samples did not form a clearly distinctive cluster, we next addressed a question whether MFH could be reclassified into other types of spindle cell and pleomorphic sarcomas by gene expression and quantified similarities between MFH samples and those sarcoma types using differentially expressed genes. To select appropriate probe sets defining spindle cell and pleomorphic sarcomas, we performed the Student's *t*-test between one and the other four of the five sarcoma types, namely, dedifferentiated liposarcoma, myxofibrosarcoma, leiomyosarcoma, MPNST and fibrosarcoma. In this analysis, we excluded three exceptional samples

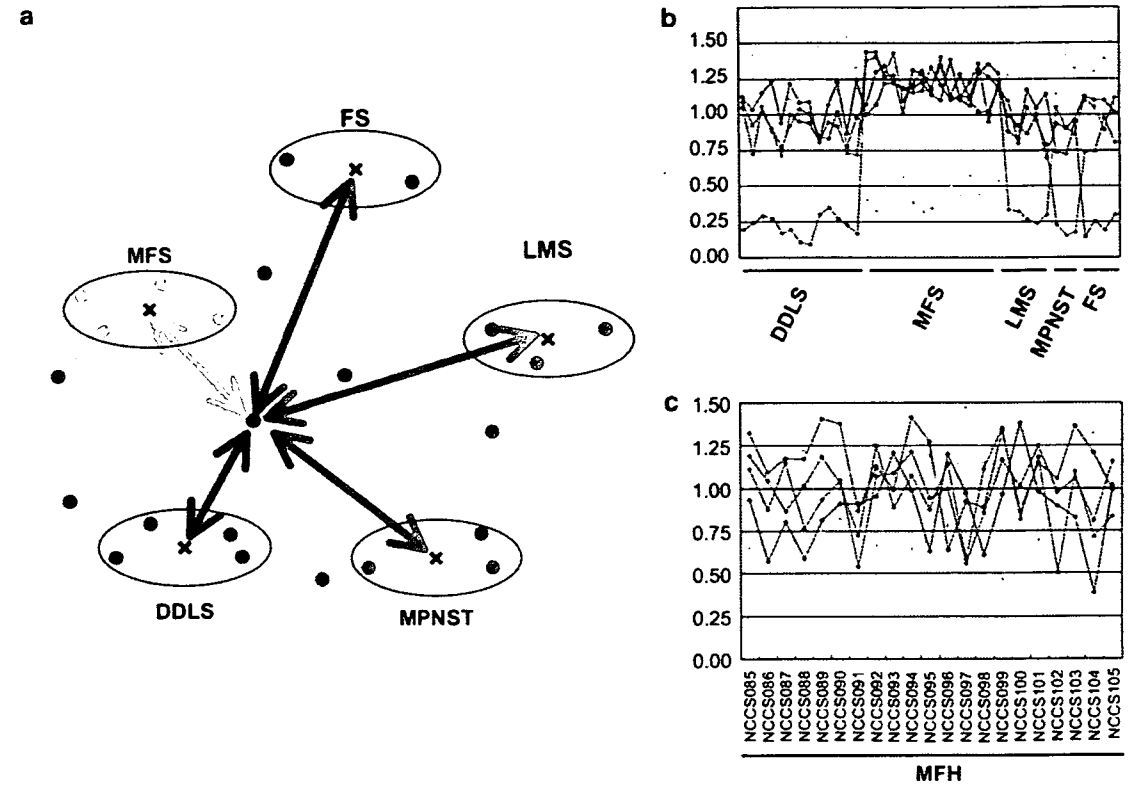

Figure 3 Distance evaluation of spindle cell and pleomorphic sarcoma samples from five sarcoma types. (**a**) Scheme of distance calculation. Each dot represents a sample colored according to its histology. Each × -mark represents the centroid of each histological type of sarcoma. Each arrow indicates the distance from a sample to a centroid colored by the histology. (**b**) Distances of 40 control samples from the five centroids. Note that the closest centroids matched their histology. (**c**) Distances of 21 MFH samples from the five centroids. DDLS, dedifferentiated liposarcoma; MFS, myxofibrosarcoma; LMS, leiomyosarcoma and FS, fibrosarcoma.

that did not fall into the appropriate cluster (Figure 2b). The top 50 probe sets with low *P*-values in each *t*-test were summed to obtain 248 probe sets (Supplementary data 6). On the basis of the expression of these 248 probe sets, the centroids of those five sarcoma types were calculated in advance, and inter-centroid distances and distances from five centroids to each control sample ($n=40$) were evaluated (Supplementary data 7 and Figure 3a and b). All inter-centroid distances were greater than 0.77 and the closest centroids for 40 control samples matched their histological types (Figure 3b), indicating that the evaluated distances were good indicators of sarcoma classification. We then evaluated the distances of each MFH sample from the five centroids (Figure 3c) and focused on determining the minimum (D_{min}) of the five distances. Small D_{min} values indicate high similarity to one of the five histological types in terms of gene expression. We used two cutoff values of 0.5 and 0.75 to evaluate similarity, because the majority of D_{min} values in control samples were less than 0.5 and most of the remaining four distances in each control sample were greater than 0.75. Among 21 samples, 3 showed marked similarity ($D_{min} \leq 0.5$), 12 showed moderate similarity ($0.5 < D_{min} \leq 0.75$) and the remaining 6 showed little similarity ($D_{min} > 0.75$). Among 15 MFH samples showing high or moderate similarity ($D_{min} \leq 0.75$), 6 were similar to myxofibrosarcoma, 5 to fibrosarcoma, 2 to MPNST and 1 each to dedifferentiated liposarcoma and leiomyosarcoma.

Histological Reviews

We re-examined the histology of 21 MFH samples with the knowledge of similarity to other types of spindle cell and pleomorphic sarcomas based on gene expression. Three MFH samples that showed high gene expression similarity ($D_{min} \leq 0.5$) displayed marked pleomorphism, indicating that a diagnosis of MFH was appropriate at the time of diagnosis. However, these samples also showed histological signatures of relevant subtypes. The NCCS099 sample, which was significantly close to the myxofibrosarcoma centroid ($D_{min} = 0.46$), showed prominent myxoid features very similar to myxofibrosarcoma in one third of the tumor (Figure 4a). The NCCS102 sample which was very close to the leiomyosarcoma centroid ($D_{min} = 0.50$) was positive for desmin and αSMA (Figure 4b–d). The NCCS104 sample, which was very close to the fibrosarcoma centroid ($D_{min} = 0.39$), showed focal

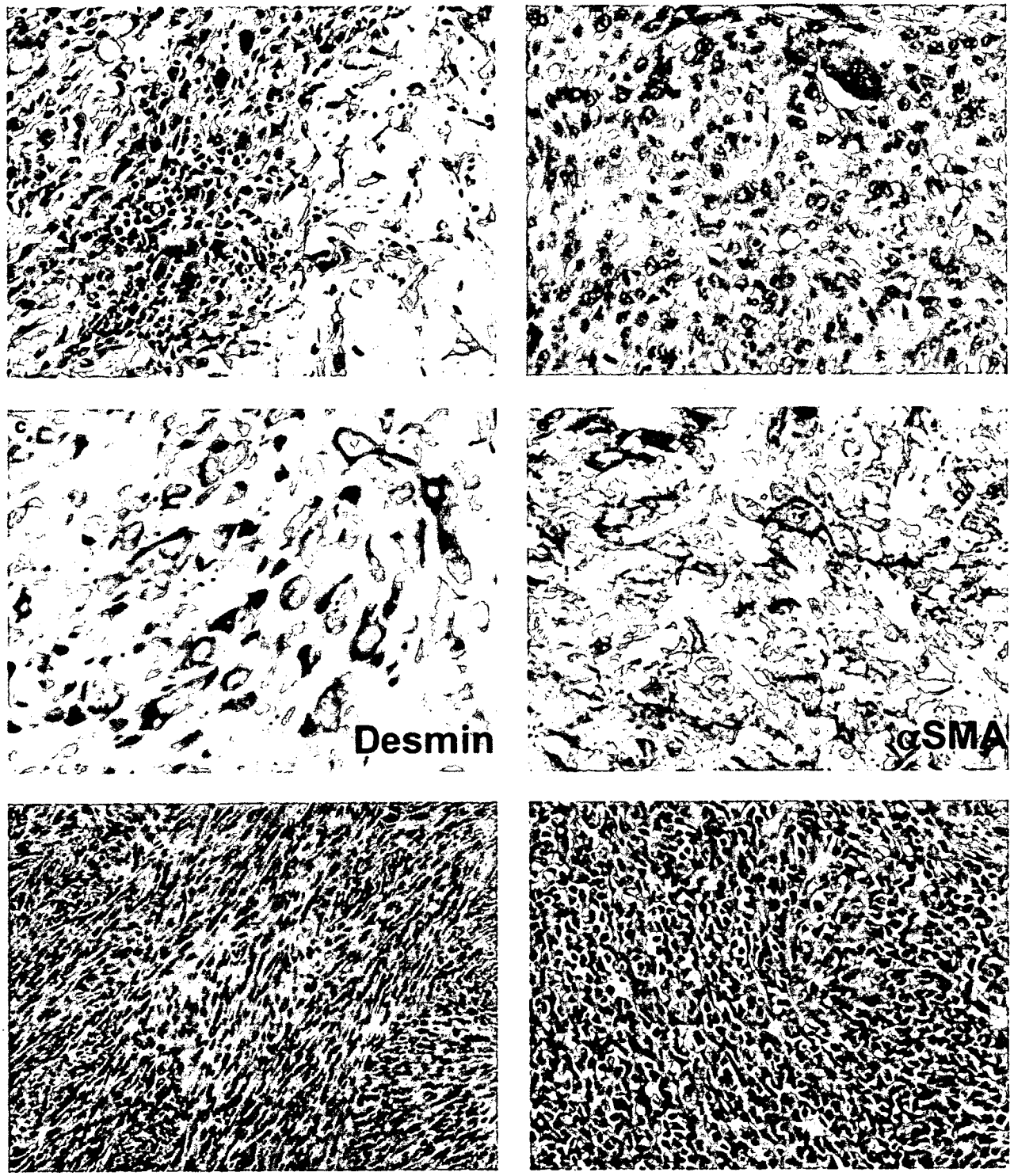

Figure 4 Histological review. (**a**) The border between pleomorphic area and myxoid area observed in the NCCS099 sample ($D_{min} = 0.46$ to myxofibrosarcoma) (hematoxylin and eosin stain). (**b–d**) Histology of the NCCS102 sample ($D_{min} = 0.50$ to leiomyosarcoma). This tumor showed marked pleomorphism (**b**) hematoxylin and eosin stain, but tumor cells were positive for desmin (**c**) and αSMA (**d**). (**e**) Fibrosarcomatous fascicular area seen in the NCCS104 sample ($D_{min} = 0.39$ to fibrosarcoma) (hematoxylin and eosin stain). (**f**) Epithelioid structure observed in the NCCS097 sample ($D_{min} = 0.56$ to MPNST) (hematoxylin and eosin stain).

fibrosarcoma-like herringbone and fascicular patterns by microscopic analysis (Figure 4e).

Among twelve samples showing moderate similarity to other types of sarcomas ($0.50 < D_{min} \leq 0.75$), the NCCS096 sample close to the dedifferentiated liposarcoma centroid ($D_{min} = 0.64$) was obtained from a recurrent sarcoma in the retroperitoneum. Although microscopic findings did not show

evidence of adipocytic differentiation or features of well-differentiated liposarcoma in regions adjacent to the tumor, the site of involvement suggested the possibility that the tumor originated from dedifferentiated liposarcoma. All five samples close to myxofibrosarcoma ($0.50 < D_{min} \leq 0.75$) showed scattered myxoid areas, but these findings were not sufficient to reclassify them as myxofibrosarcoma histologically. The NCCS097 sample, another pleomorphic sarcoma close to the MPNST centroid ($D_{min} = 0.56$), exhibited scattered whorled and epithelioid structures (Figure 4f) as well as tumor cells positive for cytokeratin, neurofilament and αSMA, indicating that this tumor had neuroectodermal differentiation. Its similarity to leiomyosarcoma ($D = 0.58$) would be reflected in αSMA positivity. For the NCCS091 sample close to MPNST and the other four close to fibrosarcoma, we did not observe any significant histological similarity to MPNST or fibrosarcoma, respectively. In summary, although more than half of the MFH samples ($n = 12$) were moderately similar in terms of gene expression to other sarcomas ($0.50 < D_{min} \leq 0.75$), only little resemblance was detectable by histological examination. Finally, the remaining six samples with high D_{min} values ($D_{min} > 0.75$) showed no identifiable histological similarity to the five sarcoma types (dedifferentiated liposarcoma, myxoid/round cell liposarcoma, leiomyosarcoma, MPNST and fibrosarcoma).

Genes Overexpressed in Myxofibrosarcoma

Diagnostically useful markers for myxofibrosarcoma are not well known. To search for candidate markers that genetically characterize myxofibrosarcoma, we selected upregulated genes by comparing myxofibrosarcoma samples ($n = 15$) with other spindle cell and pleomorphic sarcoma samples ($n = 25$). Three samples excluded from the previous analysis and 21 samples of MFH were not used for the marker search. From 11 300 probe sets, we selected 10 probe sets (five genes) with P-values < 0.001 based on the Student's t-test and more than five-fold greater expression (Table 1). Among them, expression of four probe sets (four genes) in respective spindle cell and pleomorphic sarcomas are shown in Figure 5a by the box-and-whisker plots. Since *ANK1* expression in MFH was much higher than that seen in myxofibrosarcoma (data not shown), its upregulation was not considered to be specific to myxofibrosarcoma. We performed quantitative RT-PCR with three other genes, *WISP2*, *GPR64* and *TNXB*, to verify the microarray findings (Figure 5b). Quantitative RT-PCR data confirmed consistent high expression of *GPR64* and *TNXB* in myxofibrosarcoma samples and in some MFH samples showing similarity to myxofibrosarcoma in terms of gene expression.

Discussion

An important aim of this study was to obtain new insights to classify a diverse group of soft tissue sarcomas. Our data showed that soft tissue sarcomas examined roughly fell into four groups (Figure 1a) (1) synovial sarcoma; (2) myxoid/round cell liposarcoma; (3) lipoma, well-differentiated liposarcoma with part of dedifferentiated liposarcoma and (4) spindle cell and pleomorphic sarcomas. Six histological types of spindle cell and pleomorphic sarcomas (dedifferentiated liposarcoma, myxofibrosarcoma, leiomyosarcoma, MPNST, fibrosarcoma and MFH) did not display distinct profiles but they shared a similar gene expression profile, forming a loose cluster in the hierarchical clustering analysis (Figure 1b). These results were broadly consistent with previous reports,[7,10] and histological similarity among spindle cell and pleomorphic sarcomas could be explained by similarities in gene expression. We could find some MPNST samples were located adjacent to the robust synovial sarcoma cluster in the hierarchical clustering analysis (Figure 1b), indicating that those MPNST samples shared similar expression patterns with synovial sarcoma as reported by Nagayama *et al.*[8] Our data also showed a common gene expression signature in synovial sarcoma and myxoid/round cell liposarco-

Table 1 Genes highly expressed in myxofibrosarcoma

Gene symbol	*Fold change*	*P-value*	*Description*	*Probe set ID*
WISP2	10.99	1.6×10^{-5}	WNT1 inducible signaling pathway protein 2	205792_at
GPR64	10.56	7.7×10^{-5}	G protein-coupled receptor 64	206002_at
TNXB	8.30	3.7×10^{-5}	Tenascin XB	208609_s_at
ANK1	7.03	1.8×10^{-5}	Ankyrin 1, erythrocytic	208352_x_at
S100A3	6.41	3.0×10^{-7}	S100 calcium binding protein A3	206027_at
ANK1	5.99	5.5×10^{-5}	Ankyrin 1, erythrocytic	205391_x_at
TNXB	5.96	2.9×10^{-5}	Tenascin XB	213451_x_at
TNXB	5.86	6.5×10^{-4}	Tenascin XB	216339_s_at
TNXB	5.74	1.3×10^{-4}	Tenascin XB	206093_x_at
TNXB	5.74	4.7×10^{-5}	Tenascin XB	216333_x_at

The top 10 probe sets with high fold changes were selected from 321 probe sets differentially expressed ($P < 0.001$ by Student's t-test) between myxofibrosarcoma samples ($n = 15$) and other spindle cell and pleomorphic sarcoma samples ($n = 25$) analyzed in Figure 3b.

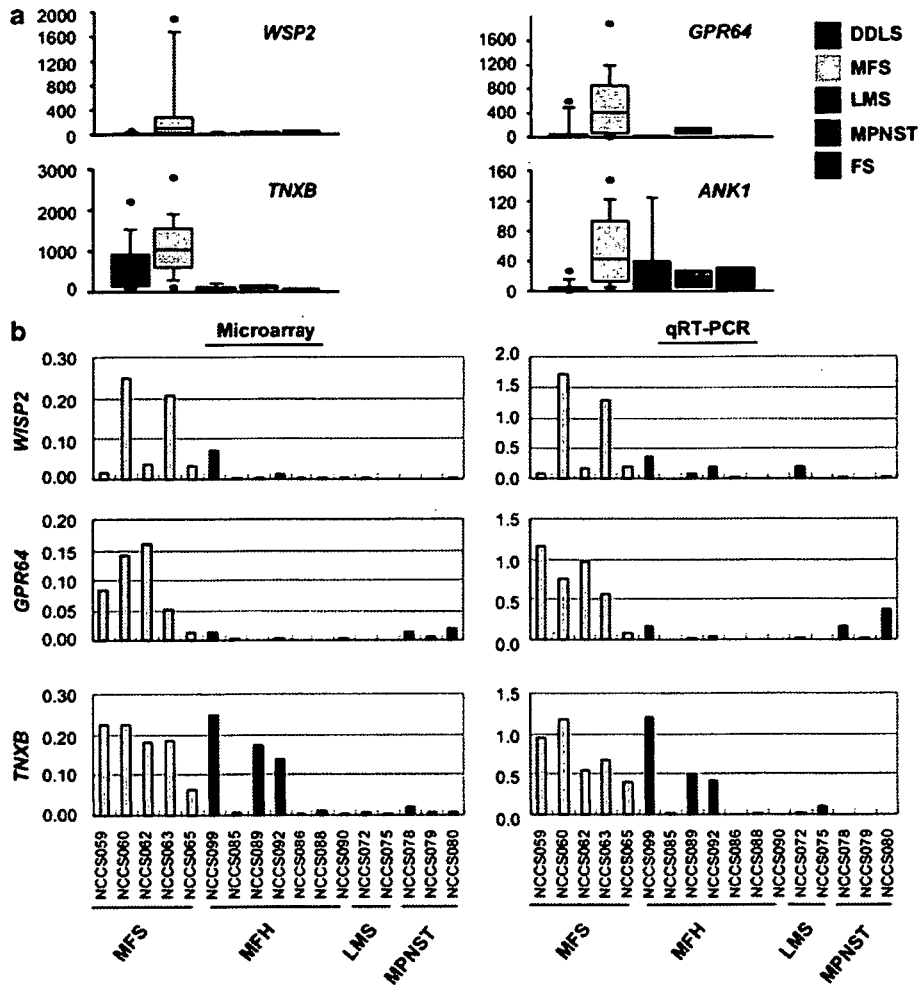

Figure 5 Genes highly expressed in myxofibrosarcoma. (**a**) Box-and-whisker plots indicating expression values for each histological type of spindle cell and pleomorphic sarcomas. A total of 40 control samples were analyzed. (**b**) Comparison between microarray analysis and quantitative RT-PCR of *WISP2*, *GPR64* and *TNXB* expression. Expression levels were normalized to that of *ACTB* in both microarray and RT-PCR data. NCCS099 ($D_{min} = 0.46$), NCCS085 ($D_{min} = 0.66$), NCCS089 ($D_{min} = 0.75$) and NCCS092 ($D_{min} = 0.66$) were similar to myxofibrosarcoma in terms of gene expression (see Figure 3c). DDLS, dedifferentiated liposarcoma; MFS, myxofibrosarcoma; LMS, leiomyosarcoma and FS, fibrosarcoma.

ma samples, distinguishing them from other tumors. Overexpression of genes encoding ribosomal proteins in myxoid/round cell liposarcoma was reported previously.[11] Another report showed that *SOX11*, *CTAG1*, *CTAG2* and *PRAME* were overexpressed in liposarcomas and absent or minimally expressed in all other tumors examined.[15] Among those genes, *CTAG1* and *PRAME* are both categorized as cancer testis antigens, and their expression in synovial sarcoma has also been reported.[16] Consistent with those reports, we found that *SOX11*, *CTAG1*, *CTAG2* and *PRAME* are highly expressed in both synovial sarcoma and myxoid/round cell liposarcoma. These similarities in gene expression may correlate with biological characteristics of synovial sarcoma and myxoid/round cell liposarcoma and suggest that these two sarcomas may share a common oncogenic pathway.

The so-called MFH was thought to be the most common soft tissue sarcoma in adults, and currently, it is widely accepted as a common morphological manifestation of a variety of poorly differentiated sarcomas. Re-evaluation of 'MFH' by different methods has been undertaken. Fletcher *et al*[5] reclassified 100 tumors primarily diagnosed as 'MFH' by histological methods and showed that the most common diagnosis was myxofibrosarcoma ($n = 29$), followed by leiomyosarcoma ($n = 20$). Hasegawa *et al*[14] examined immunoreactivity for smooth muscle markers from 100 samples of 'MFH' and reported that a large subset showed poorly differentiated smooth muscle or myofibroblastic features and should be regarded as pleomorphic leiomyosarcoma or pleomorphic myofibrosarcomas. Using comparative genomic hybridization, Derre *et al*[17] showed similar recurrent genomic imbalances in 'MFH' and leiomyosarcoma, and Coindre *et al*[18] reported that most inflammatory types of MFH developing in the retroperitoneum are identical to dedifferentiated liposarcoma. Here, we discussed

the possibility that 21 MFH samples could be reclassified into other types of spindle cell and pleomorphic sarcomas based on similarities in gene expression. For convenience of evaluation, we separated MFH samples into three groups according to the level of similarity to other sarcoma types. MFH with marked similarity ($D_{min} \leq 0.5$), MFH with moderate similarity ($0.50 < D_{min} \leq 0.75$) and MFH with no similarity ($D_{min} > 0.75$). Three samples very similar in gene expression to other sarcoma types ($D_{min} \leq 0.5$) resembled the corresponding histological types of spindle cell and pleomorphic sarcomas, and we concluded that these samples could probably be diagnosed as pleomorphic subtypes of those respective sarcomas based on current histological criteria. We then found that despite only marginal histological resemblance, more than half of the MFH samples (12/21) showed gene expression profiles similar to other sarcoma types ($0.50 < D_{min} \leq 0.75$). We considered that these moderate similarities in gene expression could correspond with pleomorphic change in each sarcoma type. Thus, although the samples cannot be diagnosed based on current histological criteria, it is possible to reclassify them as a pleomorphic subtype of those sarcomas based on gene expression. In this study, 40% (6/15) of reclassified MFH samples ($D_{min} \leq 0.75$) were similar to myxofibrosarcoma and 33% (5/15) were similar to fibrosarcoma, suggesting that a large subset of 'MFH' represents pleomorphic subtypes of fibroblastic sarcomas. Among the six cases of MFH similar to myxofibrosarcoma, five other than NCCS089 had deep-seated lesions, four (NCCS085, NCCS092, NCCS094 and NCCS101) had distant metastasis, and one (NCCS094) suffered local recurrence after surgery. Although the local recurrence rate (1/6) was unexpectedly low and distant metastasis rate (4/6) was high compared to canonical myxofibrosarcoma, these data could be consistent with the report showing that deep-seated lesions of myxofibrosarcoma were higher-grade, pleomorphic and large and increased the incidence of distant metastases.[19] About 30% of the MFH samples (6/21) did not show similarities to other sarcoma types ($D_{min} > 0.75$). One possibility is that '*de novo* undifferentiated pleomorphic sarcomas' truly exist. It is also possible that these samples represent advanced stage of dedifferentiation, which is beyond the analytical power of our study design. Another possibility is that the samples were derived from sarcomas of other differentiation not examined in this study. Extraskeletal osteosarcoma, rhabdomyosarcoma and other sarcomas could be the candidate. Reclassification accuracy should be improved by examining additional histological types of spindle cell and pleomorphic sarcomas.

Given that almost one third of MFH samples shared similar gene expression patterns ($D_{min} \leq 0.75$) with myxofibrosarcoma, we hypothesize that a large subset of 'MFH' may be pleomorphic subtype of myxofibrosarcoma. Myxofibrosarcoma is one of the most frequent sarcomas seen in late adults. However, little is known about its normal tissue counterparts, or factors underlying its extremely high local recurrence rate,[19] nor are there any good markers available for histological diagnosis. Identification of genes highly expressed in myxofibrosarcoma would offer an important clue to address these problems. Here, we found *WISP2*, *GPR64* and *TNXB* were upregulated in myxofibrosarcoma compared with other spindle cell and pleomorphic sarcomas. WISP2 is a member of the WNT1 inducible signaling pathway (WISP) protein subfamily, which belongs to the connective tissue growth factor family. WISP family members are secreted, cell- and matrix-associated proteins that play critical roles in cell differentiation and survival, wound repair, vascular disease, fibrosis and progression of certain cancers.[20–22] GPR64 is a highly conserved, tissue-specific heptahelical receptor of the human epididymis,[23–25] and there are no reports on the relationship of GPR64 to any type of cancer. TNXB is the largest member of the tenascin family of extracellular matrix proteins, which have anti-adhesive effects as opposed to the adhesion activity of fibronectin. It is expressed in musculoskeletal, cardiac and dermis tissue, and its deficiency is associated with the connective tissue disorder Ehlers-Danlos syndrome.[26–28] Although it is not clear if these genes play a role in myxofibrosarcoma, they may serve as novel diagnostic markers.

In this study, we primarily analyzed gene expression of MFH and other types of spindle cell and pleomorphic sarcomas (dedifferentiated liposarcoma, myxofibrosarcoma, leiomyosarcoma, MPNST and fibrosarcoma). Although these sarcomas showed a similar gene expression pattern and formed a relatively loose cluster, samples from five types of spindle cell and pleomorphic sarcomas were classified into respective histological types by excluding MFH samples. We identified genes that were differentially expressed among the five sarcoma types and could reclassify more than 70% of MFH samples into the five sarcoma types based on their similarities in gene expression using a combination of simple statistical analysis. These results suggest that gene expression profiling will be a useful tool to reclassify MFH and to aid histological diagnosis of a diverse group of soft tissue sarcomas. Although we cannot currently predict differences in clinical behavior of reclassified MFH due to the limited number of samples analyzed, accumulation of gene expression data should improve prediction of clinically important events, such as local recurrence, metastasis or therapeutic responses.

Acknowledgement

We are grateful to Ms Rie Ito and Ms Sachiyo Mitani for technical assistance.

Disclosure/conflict of interest

This work was supported by the program for promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation (NiBio) and by Grants-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology, Japan. There is no conflict of interest to declare.

References

1 Helman LJ, Meltzer P. Mechanisms of sarcoma development. Nat Rev Cancer 2003;3:685–694.
2 Hasegawa T, Yamamoto S, Nojima T, *et al.* Validity and reproducibility of histologic diagnosis and grading for adult soft-tissue sarcomas. Hum Pathol 2002;33: 111–115.
3 Fletcher CD. Pleomorphic malignant fibrous histiocytoma: fact or fiction? A critical reappraisal based on 159 tumors diagnosed as pleomorphic sarcoma. Am J Surg Pathol 1992;16:213–228.
4 Hollowood K, Fletcher CD. Malignant fibrous histiocytoma: morphologic pattern or pathologic entity? Semin Diagn Pathol 1995;12:210–220.
5 Fletcher CD, Gustafson P, Rydholm A, *et al.* Clinicopathologic re-evaluation of 100 malignant fibrous histiocytomas: prognostic relevance of subclassification. J Clin Oncol 2001;19:3045–3050.
6 Fletcher CD, van den Berg E, Molenaar WM. Pleomorphic malignant fibrous histiocytoma/undifferentiated high grade pleomorphic sarcoma. In: Fletcher CD, Unni KK, Mertens F (eds). World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of Soft Tissue and Bone. IARC Press: Washington, DC, USA, 2002, pp 120–122.
7 Nielsen TO, West RB, Linn SC, *et al.* Molecular characterisation of soft tissue tumours: a gene expression study. Lancet 2002;359:1301–1307.
8 Nagayama S, Katagiri T, Tsunoda T, *et al.* Genome-wide analysis of gene expression in synovial sarcomas using a cDNA microarray. Cancer Res 2002;62:5859–5866.
9 Lee YF, John M, Edwards S, *et al.* Molecular classification of synovial sarcomas, leiomyosarcomas and malignant fibrous histiocytomas by gene expression profiling. Br J Cancer 2003;88:510–515.
10 Segal NH, Pavlidis P, Antonescu CR, *et al.* Classification and subtype prediction of adult soft tissue sarcoma by functional genomics. Am J Pathol 2003; 163:691–700.
11 Skubitz KM, Skubitz AP. Characterization of sarcomas by means of gene expression. J Lab Clin Med 2004; 144:78–91.
12 Baird K, Davis S, Antonescu CR, *et al.* Gene expression profiling of human sarcomas: insights into sarcoma biology. Cancer Res 2005;65:9226–9235.
13 Takahashi H, Nemoto T, Yoshida T, *et al.* Cancer diagnosis marker extraction for soft tissue sarcomas based on gene expression profiling data by using projective adaptive resonance theory (PART) filtering method. BMC Bioinform 2006;7:399.
14 Hasegawa T, Hasegawa F, Hirose T, *et al.* Expression of smooth muscle markers in so called malignant fibrous histiocytomas. J Clin Pathol 2003;56:666–671.
15 Skubitz KM, Cheng EY, Clohisy DR, *et al.* Differential gene expression in liposarcoma, lipoma, and adipose tissue. Cancer Invest 2005;23:105–118.
16 Segal NH, Blachere NE, Guevara-Patino JA, *et al.* Identification of cancer-testis genes expressed by melanoma and soft tissue sarcoma using bioinformatics. Cancer Immun 2005;5:2.
17 Derre J, Lagace R, Nicolas A, *et al.* Leiomyosarcomas and most malignant fibrous histiocytomas share very similar comparative genomic hybridization imbalances: an analysis of a series of 27 leiomyosarcomas. Lab Invest 2001;81:211–221.
18 Coindre JM, Hostein I, Maire G, *et al.* Inflammatory malignant fibrous histiocytomas and dedifferentiated liposarcomas: histological review, genomic profile, and MDM2 and CDK4 status favour a single entity. J Pathol 2004;203:822–830.
19 Mentzel T, Calonje E, Wadden C, *et al.* Myxofibrosarcoma: clinicopathologic analysis of 75 cases with emphasis on the low-grade variant. Am J Surg Pathol 1996;20:391–405.
20 Pennica D, Swanson TA, Welsh JW, *et al.* WISP genes are members of the connective tissue growth factor family that are up-regulated in wnt-1-transformed cells and aberrantly expressed in human colon tumors. Proc Natl Acad Sci USA 1998;95:14717–14722.
21 Brigstock DR. The CCN family: a new stimulus package. J Endocrinol 2003;178:169–175.
22 Perbal B. CCN proteins: multifunctional signalling regulators. Lancet 2004;363:62–64.
23 Kierszenbaum AL. Epididymal G protein-coupled receptor (GPCR): two hats and a two-piece suit tailored at the GPS motif. Mol Reprod Dev 2003;64:1–3.
24 Obermann H, Samalecos A, Osterhoff C, *et al.* HE6, a two-subunit heptahelical receptor associated with apical membranes of efferent and epididymal duct epithelia. Mol Reprod Dev 2003;64:13–26.
25 Kirchhoff C, Obermann H, Behnen M, *et al.* Role of epididymal receptor HE6 in the regulation of sperm microenvironment. Mol Cell Endocrinol 2006;250: 43–48.
26 Burch GH, Gong Y, Liu W, *et al.* Tenascin-X deficiency is associated with Ehlers-Danlos syndrome. Nat Genet 1997;17:104–108.
27 Schalkwijk J, Zweers MC, Steijlen PM, *et al.* A recessive form of the Ehlers-Danlos syndrome caused by tenascin-X deficiency. N Engl J Med 2001; 345:1167–1175.
28 Mao JR, Taylor G, Dean WB, *et al.* Tenascin-X deficiency mimics Ehlers-Danlos syndrome in mice through alteration of collagen deposition. Nat Genet 2002;30:421–425.

Supplementary Information accompanies the paper on Modern Pathology website (http://www.nature.com/modpathol)

ELSEVIER

Biochemical Engineering Journal 33 (2007) 100–109

Biochemical Engineering Journal

www.elsevier.com/locate/bej

New cancer diagnosis modeling using boosting and projective adaptive resonance theory with improved reliable index

Hiro Takahashi [a,b,c], Yasuyuki Murase [a], Takeshi Kobayashi [d], Hiroyuki Honda [a,*]

[a] *Department of Biotechnology, School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan*

[b] *Research Fellow of the Japanese Society for the Promotion of Science (JSPS), Japan*

[c] *Genetics Division, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan*

[d] *School of Bioscience and Biotechnology, Chubu University, Matsumoto-cho 1200, Kasugai, Aichi 487-8501, Japan*

Received 22 April 2006; received in revised form 20 June 2006; accepted 5 August 2006

Abstract

An optimal and individualized treatment protocol based on accurate diagnosis is urgently required for the adequate treatment of patients. For this purpose, it is important to develop a sophisticated algorithm that can manage large amount of data, such as gene expression data from DNA microarray, for optimal and individualized diagnosis. Especially, marker gene selection is essential in the analysis of gene expression data.

In the present study, we developed the combination method of projective adaptive resonance theory and boosted fuzzy classifier with SWEEP operator method for model construction and marker selection. And we applied this method to microarray data of acute leukemia and brain tumor. The method enabled the selection of 14 important genes related to the prognosis of the tumor. In addition, we proposed improved reliability index for cancer diagnostic prediction of blinded subjects. Based on the index, the discriminated group with over 90% prediction accuracy was separated from the others.

PART-BFCS with improved RI_{BFCS} method does not only show high performance, but also has the feature of reliable prediction further. This result suggests that PART-BFCS with improved RI_{BFCS} method has the potential to function as a new method of class prediction for diagnosis of patients.

Keywords: Cancer diagnosis; Fuzzy classifier; Projective adaptive resonance theory; Marker gene selection; Reliability index

1. Introduction

Cancer is a major cause of disease related to human deaths in many developed countries. Frequently, the prognosis of cancer patients with the same clinical diagnosis can be different. Therefore, it is important that the prognosis of cancer patients is accurately determined, and an adequate treatment is proposed. However, the sensitivity of cancer patients to radiotherapy and/or chemotherapy is determined by complex causality involving multiple factors, and not a single factor because the mechanisms of cancer development (or malignancy) are extremely complex. Gene expression data from DNA microarray are individualized and are useful in the diagnosis and prognosis of diseases [1]. However, to conduct analysis, it is necessary to select significantly differentially expressed genes that are strongly related to diagnosis or prognosis of disease because the performance of classification analysis can decline due to such large quantities of data.

Feature selection has been performed in order to screen candidate genes for modeling. There are two types of approaches—wrapper approach and filter approach. In the former approach, features (genes) are selected as a part of mining algorithms, such as support vector machines (SVM) [2], fuzzy neural network (FNN) combined with SWEEP operator (FNN-SWEEP) method [1], and boosted fuzzy classifier with SWEEP operator (BFCS) method [3]. On the other hand, in the filter approach, features are selected by filtering methods, such as *U*-test, *t*-test, signal-to-noise statistic (S2N) [4] and projective adaptive resonance theory (PART) [5], prior to the application of mining algorithms.

These methods were often used alone in previous studies. In the present study, we combined various wrapper and filtering approaches and then, we applied these methods to gene expression profile data of leukemia and central nervous system tumor. It is necessary that specific and essential marker genes are selected for cancer classification and diagnosis. Minimum gene sets with-

* Corresponding author. Tel.: +81 52 789 3215; fax: +81 52 789 3214.
E-mail address: honda@nubio.nagoya-u.ac.jp (H. Honda).

1369-703X/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.bej.2006.08.004

out false positive ones should be extracted. Therefore, various methods were compared under the condition of small inputs. The combination method of PART and BFCS was the best under this condition.

2. Materials and methods

2.1. Data processing

We used two kinds of gene expression profiles. The first one is the gene expression profiles, obtained from http://www.genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi, reported by Golub et al. [4]. The data set comprised 7129 human genes (probe sets) and 72 patients (47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML)), which were obtained from acute leukemia patients at the first time of diagnosis. In this experiment, the data set was partitioned into one data set comprised of two groups: 38 patients (27 ALL, 11 AML) as a modeling data set for constructing the class prediction model (predictor) and 34 patients (20 ALL, 14 AML) as a blinded data set for evaluating the constructed predictor. We excluded those genes for which all the 72 patients showed an intensity of less than 1000 signals [6] prior to applying the various filtering methods. Thus, 2476 genes were selected for the present study.

The second one is gene expression data set of medulloblastoma, which is a type of central nervous system (CNS) tumor, obtained from http://www.genome.wi.mit.edu/MPR/CNS, reported by Pomeroy et al. [7]. Patients with medulloblastoma are treated by combinations of surgery, radiotherapy, and chemotherapy. In the present data set, the following three drugs are mainly used for chemotherapy: vincristine, cisplatin, and cytoxan. Therefore, by using gene selection and prognosis modeling proposed in the present study, the gene related to the treatment response can be extracted. The data set comprised 7129 human genes (probe sets) and 60 patients from whom tumor specimens were obtained by surgery. Among these 60 patients, a few patients (16) had a short follow-up period. Therefore, we used the data of the remaining 44 patients for the construction of a 4-year survival prediction model. Of these 44 patients, 26 patients remained alive after 4 years and 18 patients had died. In this experiment, the data set was randomly partitioned into three data sets consisting of two groups: 30 or 29 patients (18 or 17 survivors, 12 dead) as a modeling data set for constructing the class prediction model (predictor) and 14 or 15 patients (8 or 9 survivors, 6 dead) as a blinded data set for evaluating the constructed predictor. We excluded those genes for which all the 44 patients showed the intensity of less than 1000 signals prior to applying the various filtering methods. Thus, 2713 genes were selected for the present study.

In order to validate performance of models, 10 independent predictors were constructed from these genes by the parameter increasing method (PIM). The prediction accuracy of the blinded data set was utilized for comparison of model performance, and the accuracy was calculated as the average of 10 independent combination predictors.

A total of 1000 genes were selected by various gene screening methods, e.g. Mann—Whitney's U-test, signal-to-noise statistic (S2N), and projective adaptive resonance theory (PART), prior to the model construction step. Subsequently, various modeling methods were applied as described in the following sections.

2.2. Determination of optimal input number

When a large number of inputs are provided in the model, the model is excess fitted to the training data and the robustness is lost. Therefore, in order to construct a model with relatively high robustness, we assumed that the number of IF-THEN rules should not exceed the sample number [1]. Then, we used a stopping condition in the present study such that the total input number became $N_{\text{attribute}}$ in all the selected weak learners; $N_{\text{attribute}}$ is defined according to the following condition:

$$N_{\text{attribute}} < \log_2 N \tag{1}$$

where $N_{\text{attribute}}$ indicates the optimum selected attribute number.

Using Eq. (1), $N_{\text{attribute}}$ is 4 since N is 30 (or 29) for the CNS data set and 5 since N is 38 for the leukemia data set.

2.3. Boosted fuzzy classifier with SWEEP operator (BFCS)

Boosting was proposed by Schapire [8], and thus far, several derivative boosting algorithms [9–11] have been developed. Boosting is useful for class prediction using high dimensional inputs and is very fast algorithms.

In the previous study, we developed a boosted fuzzy classifier with SWEEP operator (BFCS) method [3] on the basis of AdaBoost [9], which is the most basic boosting algorithm. This method enables the evaluation of reliability of the predictions for each patient. On the other hand, it is difficult to evaluate the reliability of the predicted results of the conventional boosting.

Fig. 1 shows the structure of BFCS. BFCS is composed of one-input type I fuzzy neural network (FNN) models [12]. In the present study, one-input FNN models were used as weak learners in the BFCS model, and they were combined by connection weights, which were determined by the AdaBoost algorithm. FNN has three types of weight parameters (w_c, w_g, and w_f) [12]. In the present study, parameter w_g is a constant value ($=2.0\ln((1.0+0.995)/(1.0-0.995))$) [12], and w_c is a threshold that has the best odds ratio in the case that only one input was used. w_c and w_g were determined; w_f was calculated by the SWEEP operator method [12].

2.3.1. Reliability index for BFCS (old RI_{BFCS})

Reliability index (RI) based on fuzzy inference has been proposed to evaluate the result of class prediction by Huang and Li [13]. We have developed a reliability index for BFCS (RI_{BFCS}) by modifying RI for boosting.

We modified RI equations as follows:

$$\text{RI}_{\text{BFCS}} = \begin{cases} \text{INT}(\text{diff}_{\text{BFCS}} \cdot 10) + 1, & \text{if } 0 \le \text{diff}_{\text{BFCS}} < 0.9 \\ 10, & \text{if } \text{diff}_{\text{BFCS}} \ge 0.9 \end{cases} \tag{2}$$

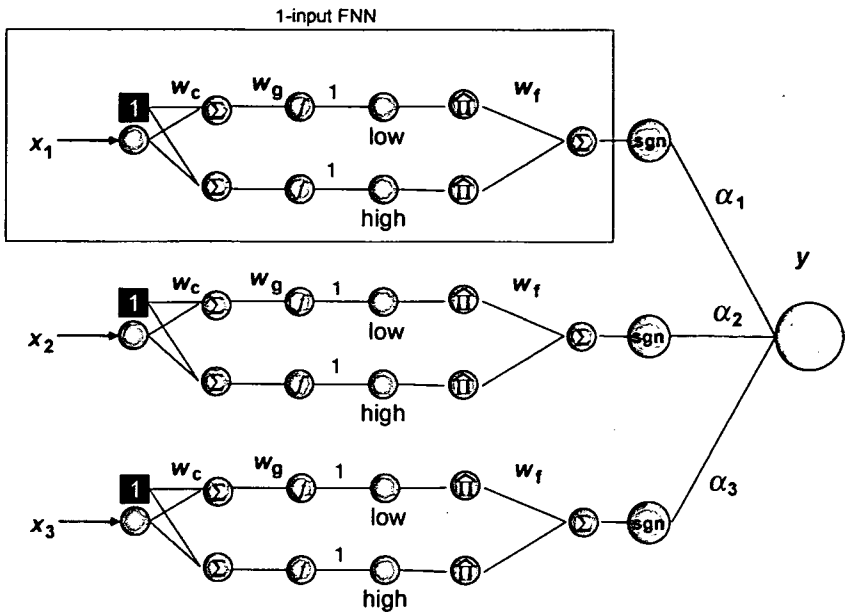

Fig. 1. Concept of the BFCS model.

where

$$\mathrm{diff}_{\mathrm{BFCS}} = \frac{\sum_t^T \left\{ \alpha_t \cdot \underset{v \in M_t}{\mathrm{argmin}}(\mathrm{diff}_v) \right\}}{\sum_t^T \alpha_t} \tag{3}$$

where T indicates the number of weak learners in the BFCS model, M_t indicates set of input variables in tth selected weak learner, α_t indicates the connection weight of the tth selected model in the construction of BFCS models, and diff_v is defined by the following equation:

$$\mathrm{diff}_v = u_{\mathrm{highest}}(X_v) - u_{\mathrm{next\ highest}}(X_v) \tag{4}$$

where v indicates the vth input in the BFCS model and $u(x_v)$ indicates the grade of the fuzzy membership function when the vth input x_v was inputted. It is defined by the following equation [12]:

$$u_v = \frac{1}{1 + \exp\{-w_g(x_v + w_c)\}} \tag{5}$$

$\mathrm{RI}_{\mathrm{BFCS}}$ is calculated for each example. Here, the greater $\mathrm{RI}_{\mathrm{BFCS}}$ the sample has, the more reliable its prediction.

2.3.2. Improved reliability index for BFCS (new RI_{BFCS})

In the present study, we propose improved reliability index by modifying equation of $\mathrm{RI}_{\mathrm{BFCS}}$ for more practical cancer diagnosis. For previous reliability index, argmin(diff_v)s in each weak learner, that mean distance from boundary line, are multiplied by α_t and summed. For improved reliability index, argmin(diff_v) in weak learner that output opposite to integrated model, is used as negative value. It is defined by the following equation:

$$\mathrm{diff}_{\mathrm{BFCS}} = \frac{\sum_t^T \left\{ \alpha_t \cdot g_t \cdot \underset{v \in M_t}{\mathrm{argmin}}(\mathrm{diff}_v) \right\}}{\sum_t^T \alpha_t} \tag{6}$$

where

$$g_t = \begin{cases} -1, & \text{if sign } (O_t) \neq \text{sign } (O_I) \\ +1, & \text{if sign } (O_t) = \text{sign } (O_I) \end{cases} \tag{7}$$

where O_t indicates output of tth model, and O_I indicates output of integrated model.

2.4. k-Nearest neighbor (kNN)

The k-nearest neighbor (kNN) methods are based on a distance function for pairs of tumor samples, such as the Euclidean distance. The kNN proceeds as follows to classify blind data set observations on the basis of the modeling data set. For each patient in the blind data set (a) finding the k-closest patients in the modeling data set and (b) predicting the class by majority vote; that is, choosing the class that is most common among those k-neighbors. The number of neighbors $k = 3$ was used because a similar cross-validation accuracy of model was obtained in the modeling data set for various k.

2.5. Multiple regression analysis (MRA)

The multiple regression analysis (MRA) is one of conventional methods. The MRA is a concerned with describing and evaluating the relationship between a patient's outcome and gene expression. MRA models are used to help us predict patient's outcome by using gene expression data.

2.6. Weighted voting (WV)

The weighted voting (WV) method was originally proposed by Golub et al. [4] to manage microarray data. The weights of each gene were calculated by the signal-to-noise. The linear models of one gene were assembled with gene weight.

2.7. Support vector machine (SVM)

The support vector machine (SVM) was originally proposed by Vapnik and Chervonenkis [14] and is used to avoid the "curse of dimensionality". SVM is superior to many other conventional methods and is frequently used in bioinformatics. In the present study, the SVM-LIGHT software package [15] was used. It was modified, and a PIM function was added to select a combination of inputs. In the present study, the regulatory parameter c was the default value of SVM_LIGHT ((avg(input vector)^2)$^{-1}$). A linear kernel was used because a similar cross-validation accuracy of model was obtained in the modeling data set for various kernels.

2.8. Fuzzy neural network (FNN) combined with SWEEP operator method (FNN-SWEEP)

The fuzzy neural network (FNN) combined with SWEEP operator method (FNN-SWEEP method) was also applied for model construction. The FNN-SWEEP method was originally proposed by Noguchi et al. [16] and was modified by Ando et al. [1] to manage microarray data. FNN has three types of weight parameters (w_c, w_g, and w_f) [12] as shown in Fig. 2. If w_c and w_g are fixed, FNN can be treated as multiple linear regression model in which w_f is variable parameter. Therefore, w_f was easily optimized without training. In the FNN-SWEEP method, only parameter w_f was optimized by the SWEEP operator method during the feature selection step. After the input combinations were determined, FNN models with the selected input combinations were optimized using a backpropagation algorithm on model construction step. In the backpropagation algorithm, the number of epochs was set to

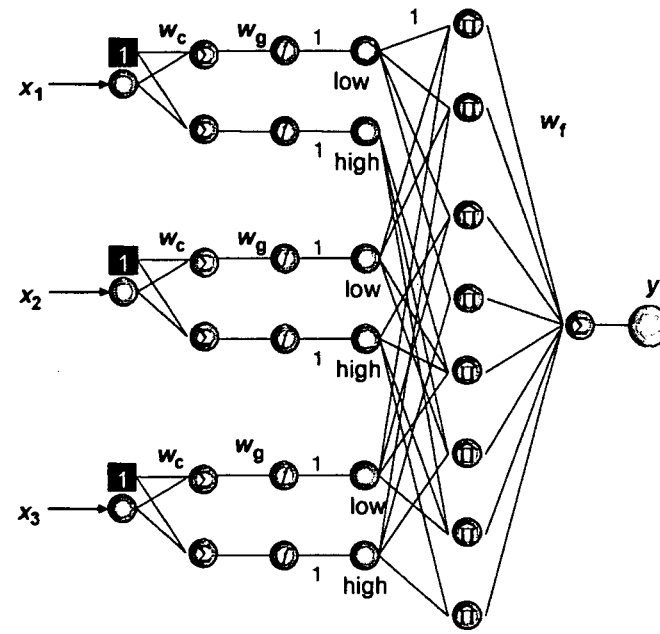

Fig. 2. Three-input type-I FNN model.

Table 1
Comparison of accuracies on various combination methods for leukemia data set (%)

	Inputs				
	1	2	3	4	5
BFCS with PART	77.9 ± 10.7	67.4 ± 7.6	84.7 ± 7.4	86.5 ± 4.4	89.1[a] ± 7.3
BFCS with S2N	78.8 ± 10.6	67.4 ± 7.6	84.4 ± 7.3	85.6 ± 5.7	83.2 ± 2.2
BFCS with *U*-test	78.8 ± 10.6	67.4 ± 7.6	84.4 ± 7.3	85.6 ± 5.7	83.2 ± 2.2
BFCS without screening	78.8 ± 10.6	67.4 ± 7.6	84.4 ± 7.3	85.6 ± 5.7	83.2 ± 2.2
SVM with PART	77.4 ± 10.0	79.4 ± 7.5	80.0 ± 8.2	80.9 ± 9.7	82.4 ± 8.4
SVM with S2N	76.2 ± 11.2	78.5 ± 7.0	81.8 ± 7.7	83.2 ± 9.0	82.4 ± 9.7
SVM with *U*-test	76.2 ± 11.2	78.5 ± 7.0	82.6 ± 6.2	84.1 ± 6.7	83.5 ± 8.0
SVM without screening	76.2 ± 11.2	78.5 ± 7.0	83.5 ± 6.2	84.7 ± 6.4	85.0 ± 7.7
FNN-SWEEP with PART	77.6 ± 12.2	77.1 ± 13.1	79.7 ± 9.1	80.3 ± 8.1	85.9 ± 7.7
FNN-SWEEP with S2N	77.9 ± 11.9	80.3 ± 7.8	81.8 ± 8.0	81.5 ± 8.2	81.5 ± 9.0
FNN-SWEEP with *U*-test	77.9 ± 11.9	80.3 ± 7.8	81.2 ± 7.5	82.6 ± 9.3	81.2 ± 8.5
FNN-SWEEP without screening	77.9 ± 11.9	80.3 ± 7.8	81.8 ± 8.0	84.4 ± 9.0	83.5 ± 8.7
*k*NN with PART	80.3 ± 11.8	75.3 ± 11.8	76.5 ± 11.8	80.0 ± 12.3	77.6 ± 12.5
*k*NN with S2N	79.1 ± 12.8	82.9 ± 12.8	82.6 ± 12.5	79.7 ± 9.8	79.4 ± 9.1
*k*NN with *U*-test	79.1 ± 12.8	84.1 ± 9.9	82.1 ± 9.0	81.5 ± 10.5	81.8 ± 10.8
*k*NN without screening	79.1 ± 12.8	79.4 ± 12.4	80.0 ± 11.3	78.8 ± 10.7	81.5 ± 9.3
MRA with PART	77.4 ± 11.2	79.4 ± 10.9	79.4 ± 10.3	75.3 ± 11.4	64.1 ± 8.2
MRA with S2N	77.9 ± 11.1	80.6 ± 8.8	83.2 ± 7.7	74.7 ± 9.6	64.7 ± 8.2
MRA with *U*-test	77.9 ± 11.1	80.6 ± 8.8	83.5 ± 8.0	76.2 ± 9.7	67.1 ± 7.0
MRA without screening	77.9 ± 11.1	80.6 ± 8.8	83.8 ± 8.2	76.2 ± 7.0	66.8 ± 6.8
WV with PART	79.7 ± 10.7	76.5 ± 12.5	82.4 ± 7.0	75.3 ± 8.6	72.4 ± 11.2
WV with S2N	78.2 ± 11.2	83.5 ± 7.5	70.9 ± 13.1	71.2 ± 12.6	70.6 ± 10.1
WV with *U*-test	78.2 ± 11.2	85.6 ± 5.8	76.2 ± 10.7	73.2 ± 14.2	76.2 ± 11.5
WV without screening	78.2 ± 11.2	78.8 ± 7.9	76.2 ± 13.6	77.1 ± 10.7	85.3 ± 9.4

The average blinded accuracies and their S.D.s were calculated from 10 combination models constructed by PIM.
[a] The highest accuracy.

5000, and the learning rate was set to 0.1, these values are the same as those reported by Ando et al. [1].

2.9. Model construction with parameter selection

The parameter increasing method (PIM) [17] was used to select input combinations for model construction of FNN-SWEEP, SVM, *k*NN, MRA, and WV. This was done as follows.

First, we predicted the subtype of each sample by using the prediction model with a single input. Prediction models for each probe were constructed in a series, and all the probes were ordered based on the accuracy of the constructed models. In the next step, the probe having the highest accuracy level was used for constructing a combination model.

Second, we selected a partner probe for the probe selected in the first step in order to increase the prediction accuracy. To accomplish this, we constructed a two-input model in which a ranked probe was designated as input 1, and input 2 (partner probe) was selected to provide the highest training accuracy while applying FNN-SWEEP (or SVM, *k*NN, MRA, and WV) and PIM to the modeling data. By repeating this step, a combination of $N_{\text{attribute}}$ candidate probes was identified for use as input probes in the model construction.

Finally, combinations of $N_{\text{attribute}}$ probes, i.e. from the first to the $N_{\text{attribute}}$th probe were evaluated. We constructed $N_{\text{attribute}}$ predictor models, beginning with one input using only the first-selected probe to $N_{\text{attribute}}$ inputs using all the $N_{\text{attribute}}$ probes. The predictor models were specifically constructed by using a backpropagation algorithm for FNN-SWEEP or quadratic programming for SVM. The performance of the prediction models was evaluated by applying them to the blinded data set.

For the two gene expression profile data, the genes with the first to the 10th highest accuracies were used as the first inputs for the construction of the 10 combination models by PIM. The S.D.s of blinded accuracies were calculated by using ones of these 10 combination models.

2.10. PART-BFCS method

Previously, we developed PART filtering method by modifying PART [18,19]. And, we developed and combined the PART filtering method as a gene filtering method and BFCS as a modeling method. In this PART-BFCS method, PART first preselects the genes that show small variances within a class. Then, BFCS rapidly selects these genes to build a highly accurate and reliable predictor.

PART has two important parameters, vigilance and distance parameters. The vigilance parameter was optimized so that modeling samples clustered well. The distance parameter was used to control the number of extracted genes. The genes extracted by PART showed low standard deviation (S.D.) in lower gene expression class. The predictor using genes with low S.D. in lower class showed high performance [5].

In BFCS model, one-input FNN models on the basis of neural network and fuzzy logic, were used as weak learners. FNN has three types of connection weights (w_c, w_g, and w_f). These parameters were optimized as mentioned in section of BFCS algorithm. The only one parameter that should be optimized is the number of input in boosting model. This parameter was optimized by using the number of samples.

Table 2
Frequency of construction of high performance model

	Methods	
	Leukemia[a]	CNS[b]
BFCS with PART	4/10	13/30
BFCS with S2N	0/10	3/30
BFCS with *U*-test	0/10	3/30
BFCS without screening	0/10	3/30
SVM with PART	2/10	2/30
SVM with S2N	1/10	2/30
SVM with *U*-test	2/10	0/30
SVM without screening	0/10	0/30
FNN-SWEEP with PART	0/10	3/30
FNN-SWEEP with S2N	0/10	0/30
FNN-SWEEP with *U*-test	0/10	0/30
FNN-SWEEP without screening	0/10	0/30
*k*NN with PART	0/10	0/30
*k*NN with S2N	0/10	0/30
*k*NN with *U*-test	0/10	0/30
*k*NN without screening	0/10	0/30
MRA with PART	0/10	0/30
MRA with S2N	0/10	0/30
MRA with *U*-test	0/10	0/30
MRA without screening	0/10	0/30
WV with PART	0/10	1/30
WV with S2N	0/10	2/30
WV with *U*-test	0/10	1/30
WV without screening	0/10	0/30

[a] Ten combination models from first to 10th models were constructed by PIM for each method in five-inputs. The accuracies of the models with first and second highest performance were 100% (=100 × 34/34) and 97.1% (=100 × 33/34), respectively. The number of the models with 100% or 97.1% accuracies were counted from 10 combination models.

[b] Ten combination models from first to 10th models were constructed by PIM for each method and each set (of three-fold cross-validation) in four-inputs. The accuracies of the models with first and second highest performance were 86.7% (=100 × 13/15) and 85.7% (=100 × 12/14), respectively. The number of the models with 86.7% or 85.7% accuracies, were counted from 30 combination models for three data sets.

3. Results and discussion

3.1. Comparison of the performance of PART-BFCS and the other methods

The performances of wrapper approaches with filter approaches as class predictors were investigated. For comparison, many combinations of various wrapper approaches, such as BFCS, SVM, FNN-SWEEP, *k*-nearest neighbor (*k*NN), multiple regression analysis (MRA), and weighted voting (WV), and various filtering approaches, such as *U*-test, S2N, PART, and no screening, were constructed. The performance of the predictors