nature medicine

Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity

Frédéric Chibon[1,2], Pauline Lagarde[1,2], Sébastien Salas[1], Gaëlle Pérot[3], Véronique Brouste[4], Franck Tirode[3], Carlo Lucchesi[3], Aurélien de Reynies[5], Audrey Kauffmann[6], Binh Bui[1,2], Philippe Terrier[7], Sylvie Bonvalot[7], Axel Le Cesne[7], Dominique Vince-Ranchère[8], Jean-Yves Blay[8], Françoise Collin[9], Louis Guillou[10], Agnès Leroux[11], Jean-Michel Coindre[1,2,12] & Alain Aurias[3]

Sarcomas are heterogeneous and aggressive mesenchymal tumors. Histological grading has so far been the best predictor for metastasis-free survival, but it has several limitations, such as moderate reproducibility and poor prognostic value for some histological types. To improve patient grading, we performed genomic and expression profiling in a training set of 183 sarcomas and established a prognostic gene expression signature, complexity index in sarcomas (CINSARC), composed of 67 genes related to mitosis and chromosome management. In a multivariate analysis, CINSARC predicts metastasis outcome in the training set and in an independent 127 sarcomas validation set. It is superior to the Fédération Francaise des Centres de Lutte Contre le Cancer grading system in determining metastatic outcome for sarcoma patients. Furthermore, it also predicts outcome for gastrointestinal stromal tumors (GISTs), breast carcinomas and lymphomas. Application of the signature will permit more selective use of adjuvant therapies for people with sarcomas, leading to decreased iatrogenic morbidity and improved outcomes for such individuals.

Adult soft tissue sarcomas (STSs) are rare heterogeneous tumors in terms of location, histology, molecular profile and prognosis. Non–translocation-related sarcomas are the most frequent malignant tumors in adults and represent roughly 50% of pathological diagnoses. They mainly comprise sarcomas with a complex karyotype, such as leiomyosarcomas, undifferentiated sarcomas or so-called malignant fibrous histiocytomas, and sarcomas with amplification, such as dedifferentiated liposarcomas[1]. Individuals with these tumors have a 40% to 50% risk of developing distant metastases, primarily in the lung, within 5 years of diagnosis[2,3].

Clinical management of STS consists mainly in surgical resection, with adjuvant therapies that depend in their timing and nature on surgical margins, tumor histological type and histological grade. Histological grading has so far been the best predictor of metastasis-free survivals. The Fédération Francaise des Centres de Lutte Contre le Cancer (FNCLCC) grading system was defined more than 20 years ago and is still the most commonly used system[4,5]. It is based on semiquantitative evaluation of tumor differentiation, necrosis and mitotic index. Histological grading has several limitations; it is an indirect measure of the underlying oncogenic changes in the tumor, its reproducibility from one pathologist to another is questionable, it is not applicable to all types of sarcomas[6] and it is poorly informative for grade 2 sarcoma (which represents about 40% of cases). We postulated that direct assessment, with microarray technology, of the genetic alterations underlying the tumor phenotype would provide a more precise estimate of tumor aggressiveness.

At the genetic level, non–translocation-related sarcomas can be divided into two main groups, one with a complex genomic profile (80%) including undifferentiated sarcomas, leiomyosarcomas, pleomorphic rhabdomyosarcomas and pleomorphic liposarcomas associated with very complex but recurrent genomic imbalances[7–9]; and a second group with a simple genetic profile (20%) based on many limited amplifications and exclusively composed of dedifferentiated liposarcomas[10,11].

In most expression profiling studies in sarcomas, the purpose has been to identify new diagnostic markers or to obtain better understanding of oncogenesis and its relationship with tumor differentiation[12–19]. A few studies to date have tried to correlate expression profiles with outcome, but the clinical impact was limited as a result of low robustness[20,21]. Here we have performed genomic and expression profiling of 183 primary non–translocation-related sarcomas from the French Sarcoma Group tumor bank. By analyzing expression profiles according to genome complexity and histological grade, we have identified a gene expression signature that is able to predict metastatic

[1]Département de Pathologie, Institut Bergonié, Bordeaux, France. [2]Institut National de la Santé et de la Recherche Médicale (INSERM) U916, Institut Bergonié, Bordeaux, France. [3]INSERM U830, Institut Curie, Paris, France. [4]Département de Biostatistique, Institut Bergonié, Bordeaux, France. [5]Carte d'Identité des Tumeurs, Ligue Nationale Contre le Cancer, Paris, France. [6]Bioinformatique, INSERM U916, Institut Bergonié, Bordeaux, France. [7]Institut Gustave Roussy, Villejuif, France. [8]Centre Léon Bérard, Lyon, France. [9]Centre G. F. Leclerc, Dijon, France. [10]Institut Universitaire de Pathologie, Lausanne, Switzerland. [11]Centre Alexis Vautrin, Vandoeuvre-les-Nancy, France. [12]Université Victor Segalen Bordeaux 2, Bordeaux, France. Correspondence should be addressed to F.C. (chibon@bergonie.org).

Received 6 August 2009; accepted 1 June 2010; published online 27 June 2010; doi:10.1038/nm.2174

outcome in non–translocation-related sarcomas. We then validated the prognostic value of this signature in a second independent cohort of 127 non–translocation-related sarcomas.

RESULTS

Genomic profiling

We performed genomic profiling of the training set (cohort 1: 183 non–translocation-related sarcomas, **Table 1**) with a homemade bacterial artificial chromosome (BAC)–comparative genomic hybridization (CGH) array containing 3,803 clones. According to both the number and the type of alterations, we identified three types of recurrent profiles (**Fig. 1**). A first group of 28 tumors (16%) with a simple amplicon profile based on coamplifications corresponded almost exclusively to dedifferentiated liposarcomas. A second group of 40 tumors (23%) had few alterations (less than 30) mainly involving the full chromosome arm or entire chromosomal gain or loss. We termed this the 'arm' profile. Finally a third group of 106 tumors (61%) was characterized by a high level of chromosomal complexity, with between 30 and 85 alterations. This we called the 'rearranged' profile. These various profiles are associated with tumor characteristics (**Supplementary Table 1**), as almost all (89%) amplicon profiles correspond to dedifferentiated liposarcomas mainly located in the internal trunk, and rearranged profiles correspond essentially to sarcomas developed in the external trunk (94%), which are mainly undifferentiated sarcomas (49%) and leiomyosarcomas (27%).

Table 1 Cohort 1 and 2 subject characteristics

Characteristics	Cohort 1 (*n* = 183)	Cohort 2 (*n* = 127)
Median follow-up (months) [CI]	84 [69.5–100]	60.16 [48.1–72.2]
Median age (years)	63	63
S.d.	15	15
Male sex (%)	98 (53)	62 (49)
FNCLCC grade (%)		
1 and 2	69 (38)	47 (37)
3	102 (56)	70 (55)
ND	12 (6)	10 (8)
Histological type (%)		
Undifferentiated sarcomas	71 (39)	65 (51)
Leiomyosarcomas	52 (28)	33 (26)
Dedifferentiated liposarcomas	44 (24)	18 (14)
Others	16 (9)	11 (9)
Location (%)		
External trunk	144 (79)	100 (79)
Trunk wall	28 (16)	26 (20)
Extremities	114 (62)	73 (58)
Head and neck	2 (1)	1 (1)
Internal trunk	39 (21)	27 (21)
Median size (cm)	10	10
Deep-seated (%)		
Yes	173 (95)	111 (87)
No	10 (5)	12 (10)
ND		43
Vasculonervous or bone involvement (%)		
Yes	27 (15)	25 (20)
No	156 (85)	102 (80)
Relapse events (%)		
Metastasis	79 (43)	43 (33)
Local recurrences	60 (33)	27 (21)
Therapeutic management (%)		
Surgery	50 (27)	61 (48)
Surgery + radiotherapy	80 (44)	35 (28)
Surgery + chemotherapy	10 (5)	12 (9)
Surgery + radiotherapy + chemotherapy	40 (22)	17 (13)
Missing data	3 (2)	2 (2)

Characteristics listed are mainly those known for having a potential impact on tumor outcome, such as FNCLCC grade, histological type, tumor location, size, deepness and vasculonervous or bone involvement. FNCLCC G1 and G2 tumors are considered together because of too many G1 tumors (7 and 14 in cohorts 1 and 2, respectively). Other histological types are pleomorphic liposarcomas and pleomorphic rhabdomyosarcomas. CI, confidence interval; ND, not determined.

To determine whether the genomic profile was associated with clinical outcome, we split the tumors into two groups according to the genomic profile (arm versus rearranged). However, this simple classification failed to predict metastatic outcome, as metastasis-free survival was not significantly different between the two groups (log-rank, $P = 0.17$). Notably, we found a positive correlation between number of genomic alterations and histological grade ($P = 0.001$). As histological grade is an indirect marker of tumor aggressiveness, this means that genomic complexity is related to tumor aggressiveness, even if no correlation with poor outcome was obtained. We therefore wondered whether gene expression related to genomic complexity, tumor grade or both could predict metastasis outcome.

Construction of the CINSARC prognostic signature

We reexamined gene expression profiles of the 183 sarcomas (cohort 1) to test the hypothesis that gene expression in the tumor correlates with genome complexity or metastatic outcome. We first assigned the 183 samples to prognostic classes according to a previously reported signature of 70 genes selected for their relation to chromosomal instability[22]. This resulted in a nonsignificant trend for prediction of metastasis-free survival (**Supplementary Fig. 1**). We then applied a three-step approach to find a sarcoma-specific prognostic gene set (**Fig. 2**). First, we performed three *t* tests to compare the expression profiles of tumors classified according to number of CGH imbalances (fewer than 20 versus more than 35 imbalances), FNCLCC grade 3 versus grade 2 and the 70-gene Carter chromosome instability signature[22]. For the first two comparisons, we selected genes showing a more than threefold CGH difference ($P < 0.01$) or a more than twofold grade difference ($P < 0.01$). This yielded 86 significant genes (118 probe sets) for CGH imbalance and 73 significant genes (92 probe sets) for grade. Second, we performed gene ontology analysis to identify the underlying pathways related to CGH imbalance or histologic grade. Of note, the overrepresented pathways were extremely similar in both groups and were mainly involved in chromosome integrity and mitotic control (**Supplementary Table 2**). Third, we selected the genes involved in the most significantly overrepresented pathways ($P < 1 \times 10^{-5}$): 37 and 18 genes were related to CGH imbalance and histologic grade, respectively. The Carter signature added 22 genes that were not already identified by our CGH imbalance and grade tests. Among the 70 genes in the Carter signature, 39 were differentially expressed in our samples at $P < 1 \times 10^{-5}$. The three selection techniques combined led to a final gene set named CINSARC containing 67 genes (**Supplementary Table 3**).

Prediction of metastatic outcome with CINSARC

We evaluated the correlation between CINSARC and metastatic outcome in two cohorts of sarcomas, the training set used to construct the signature (183 cases, cohort 1) and an independent validation set (127 cases, cohort 2; **Table 1**). Subjects from each cohort were assigned to

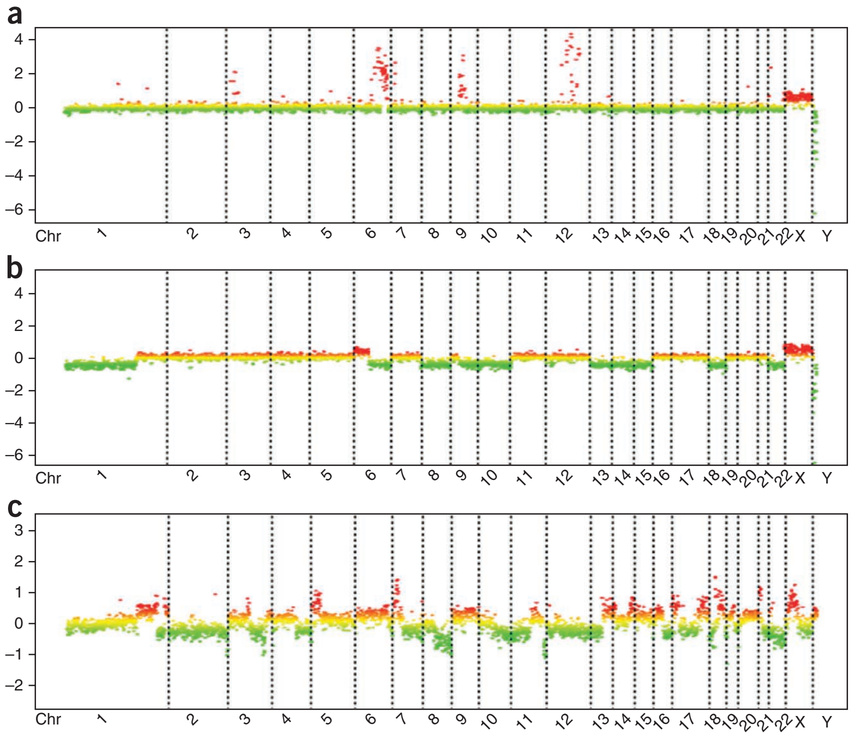


Figure 1 The three main types of genomic profile established by BAC array-CGH. The *x* axis represents human chromosomes from 1 to Y, and the log ratios of tumor versus reference are indicated on the *y* axis. (**a**) Genomic profile from amplified, type (16% of the cases), where almost exclusively amplifications are identified. (**b**) Genomic profile from arm type (23% of the cases), where losses and gains involve a chromosome arm or a full chromosome. (**c**) Genomic profile from rearranged type (61% of the cases), composed of many gains and losses with breakpoints within chromosome arm. Chr, chromosome.

two groups using the nearest centroid method (see Methods). The centroids in cohort 1 define the signature and are therefore fixed. CINSARC grade is assigned by Spearman correlation to the nearest centroid. CINSARC grade 1 (C1) corresponds to lower CINSARC score and good prognosis. CINSARC grade 2 (C2) corresponds to higher CINSARC score and poor prognosis. In cohort 1, Kaplan-Meier analysis revealed that subjects in C1 (low scores, **Fig. 3a**) and C2 (high scores, **Fig. 3a**) have 5-year metastasis-free survival (MFS) rates of 75% and 35%, respectively ($P = 1 \times 10^{-7}$). In cohort 2, subjects in C1 and C2 have 5-year MFS rates of 84% and 48%, respectively ($P = 5 \times 10^{-4}$, **Fig. 3a**). **Table 2** shows a multivariate analysis taking into consideration other standard prognostic factors that were significant in univariate analysis, that is, histological type, FNCLCC tumor grade and vasculonervous or bone involvement. The risk of metastasis was higher in both cohorts in tumors with high CINSARC grade (cohort 1: hazard ratio = 3.7; 95% confidence interval = [2.2–6.3]; cohort 2: hazard ratio = 2.7; 95% confidence interval = [1.02–7.2]). These results contrast with those obtained with the FNCLCC grading system (**Fig. 3b**) to stratify tumors, as this system predicts metastatic outcome in cohort 2 (hazard ratio = 2.35; 95% confidence interval = [1.13–4.9]) but not in cohort 1 ($P = 0.4$).

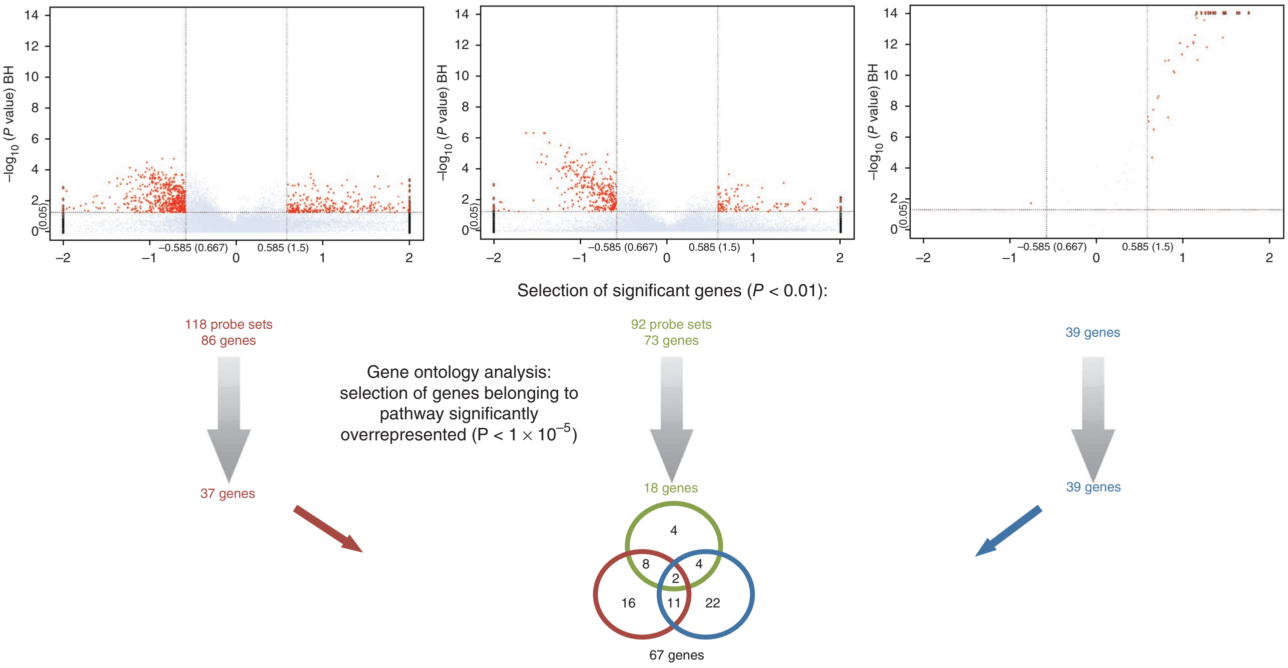
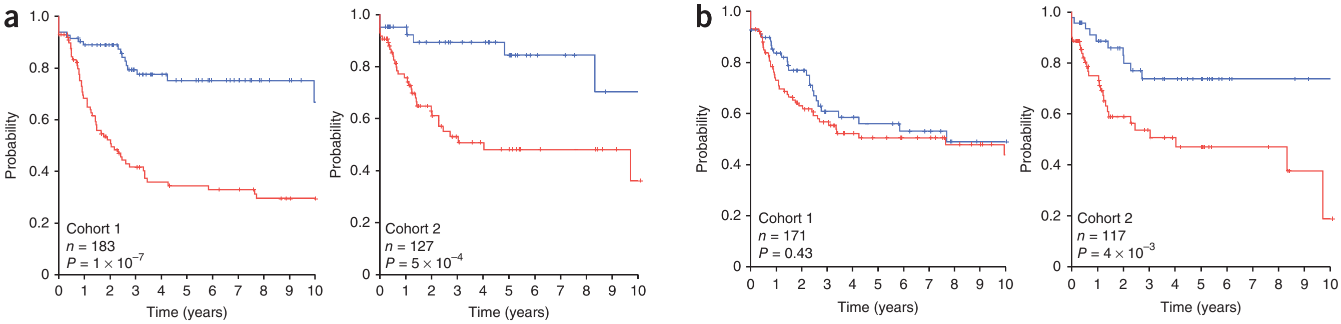


Figure 2 Schematic representation of CINSARC gene set establishment procedure. Expression profiles were compared with Welch tests after grouping of the tumors according to CGH alteration number (left), FNCLCC grade (middle) or chromosome instability expression signature (right). After gene ontology analysis allowing identification of the pathways involved, the 67 genes composing the CINSARC signature were selected. The Venn diagram at the bottom indicates the number genes present in only one analysis or common.



		0	1	2	3	4	5
	Pts at risk	83	69	60	45	36	30
C1	Events	5	9	9	15	16	17
	MFS	0.94	0.89	0.89	0.79	0.78	0.75
	Pts at risk	99	60	41	31	25	23
C2	Events	7	29	43	51	55	56
	MFS	0.93	0.68	0.52	0.42	0.36	0.35

		0	1	2	3	4	5
	Pts at risk	42	34	28	24	22	17
C1	Events	2	2	4	4	4	5
	MFS	0.95	0.95	0.90	0.90	0.90	0.84
	Pts at risk	85	54	33	24	20	17
C2	Events	6	19	27	32	33	34
	MFS	0.93	0.76	0.63	0.53	0.50	0.48

		0	1	2	3	4	5
	Pts at risk	69	53	41	27	25	21
G1	Events	5	11	15	23	24	25
	MFS	0.93	0.84	0.77	0.61	0.59	0.56
	Pts at risk	102	67	53	43	31	27
G3	Events	7	26	35	40	43	44
	MFS	0.93	0.73	0.63	0.57	0.52	0.51

		0	1	2	3	4	5
	Pts at risk	47	37	28	21	19	15
G2	Events	1	5	7	10	10	10
	MFS	0.98	0.89	0.83	0.74	0.74	0.74
	Pts at risk	70	41	24	18	14	12
G3	Events	7	16	24	26	27	28
	MFS	0.90	0.75	0.59	0.54	0.51	0.47

Figure 3 Metastasis-free survival analysis according to CINSARC signature and to FNCLCC grading system. (**a**) Metastasis-free survival analysis according to CINSARC signature. CINSARC-stratified subjects in two groups with significantly different metastasis-free survival rate (MFS rate; *y* axis) during the time after diagnosis (*x* axis). Subjects with the lowest CINSARC scores (Centroid C1) are in blue, and those with the highest ones (Centroid C2) are in red. (**b**) MFS analysis according to FNCLCC grading system. Grade (G) 1–2 subjects are in blue, and G3 subjects are in red. *P* values correspond to the log-rank test comparing the survival curves. Pts = Patients. Events = Cumulated events.

CINSARC grade is therefore an independent prognostic factor that is strongly associated with the development of metastases.

We tested CINSARC in sarcoma subgroups based on clinical and pathological presentation, that is, histological type, localization and FNCLCC grade (**Fig. 4** and **Supplementary Table 4**). Considering only undifferentiated sarcomas and leiomyosarcomas, CINSARC grading split subjects into two groups with significantly different outcomes in cohorts 1 ($P = 1.05 \times 10^{-7}$) and 2 ($P = 0.02$). Similar results were obtained excluding tumors developed in the internal trunk, which are known to be more aggressive sarcomas (data not shown). The same analysis within separated histological groups, that is, undifferentiated sarcomas and leiomyosarcomas, confirmed these results (**Fig. 4**). The results are notable for leiomyosarcomas, as 5-year MFS rates ($P = 2.7 \times 10^{-4}$) for C1 and C2 were 73% and 19%, respectively, in cohort 1. In cohort 2, the test did not reach significance, probably owing to the small size of the group ($n = 33$). We also investigated the performance of CINSARC among subjects of the same histological grade regardless of histological type (**Fig. 4**). In grade 2 tumors, CINSARC allowed identification of two groups of subjects, C1 and C2, with clearly different outcomes in cohorts 1 ($P = 0.05$) and 2 ($P = 0.009$). Similarly, in grade 3 tumors, C1 and C2 from cohort 1 present highly distinct outcomes ($P = 7.6 \times 10^{-6}$), but in cohort 2 significance was not reached because of too many cases with good prognosis (12 versus 58 with poor prognosis). Metastasis-free survival according to CINSARC was not significantly different in the two groups of dedifferentiated liposarcomas. This can essentially be explained by the small size of the groups (44 and 18 cases) and the very low metastatic relapse rate of dedifferentiated liposarcomas.

Prognostic value of CINSARC in other cancers

Because CINSARC was prognostic in subjects with non–translocation-related sarcomas, we tested it in other sarcomas and other tumor types. In a series of 32 GISTs[23], CINSARC grading split the tumors into two groups of highly different metastatic outcome ($P = 0.003$). No subjects developed metastasis in C1, whereas the MFS rates in C2 were 61% and 30% at 5 and 10 years, respectively (**Supplementary Fig. 2**).

Because the CINSARC is composed of genes involved in chromosome integrity and because expression is correlated to chromosomal imbalances, we hypothesized that the CINSARC could also have prognostic value in highly rearranged tumors such as breast carcinomas. We therefore applied it to two series of breast cancers from the Netherlands Cancer Institute[24,25]. For each series, the subjects were split by CINSARC grade into groups with significantly different outcome ($P = 2.76 \times 10^{-4}$ and $P = 1.34 \times 10^{-5}$, respectively; **Supplementary Fig. 2**).

Because CINCARC discriminates between low-risk and high-risk subjects with breast cancer and sarcomas, we tested whether it could also predict outcome in a third family of tumors, diffuse large B cell lymphomas[26]. Because of the large difference in expression profile of these tumors, we first recalibrated the CINSARC centroids on a training set of 136 lymphomas and then validated it on a test set of 278 lymphomas. The CINSARC signature identified two groups with clearly distinct overall survival in the two sets of subjects ($P = 0.01$ and $P = 5.34 \times 10^{-4}$, respectively; **Supplementary Fig. 2**).

DISCUSSION

Soft tissue sarcomas are aggressive tumors prone to local recurrence and metastasis leading to death in 40% to 50% of cases[2,3]. Patient management depends mostly on the stage of the disease. Although histological staging provides valuable information about the clinical behavior of certain types of sarcomas, it has limited predictive value for other types, especially non–translocation-related, poorly differentiated and unclassified sarcomas. To increase the predictive value of histology in terms of prognosis, several grading systems have been developed[4,27–30]. Among them, the US National Cancer Institute[30] and the FNCLCC systems[4] are widely used. The latter slightly increases the ability to predict distant metastases and should be considered as the 'gold standard'[5]. Nevertheless, as stated above, histological grading has several limitations.

Despite these limitations, for more than 20 years there has been no reliable alternative to histological grading. To our knowledge, only two studies performed on 30 leiomyosarcomas[20] and 89 pleomorphic sarcomas[21] reported a prognostic molecular signature. In both reports, the signatures were composed of a large number

Table 2 Multivariate analysis

	Cohort 1 (n = 183)		Cohort 2 (n = 127)	
	Univariate (P value, log-rank)	Multivariate (HR [95% CI] Cox model)	Univariate (P value, log-rank)	Multivariate (HR [95% CI] Cox model)
Median age (years)	0.07		0.99	
Male sex (%)	0.76		0.095	
FNCLCC grade (%)				
1 and 2	0.43		**4×10^{-3}**	2.35 [1.13–4.9]
3				
ND				
Histological type (%)				
Undifferentiated sarcomas	**0.028**	Not retained	**9×10^{-5}**	Reference
Leiomyosarcomas				2.7 [1.42 –5.37]
Dedifferentiated liposarcomas				Nonsignificant
Others				ND
Location (%)				
External trunk	0.24		**0.93**	
Trunk wall				
Extremities				
Head and neck				
Internal trunk				
Median size (cm)	0.33		0.48	
Deep-seated (%)				
Yes	0.49		0.26	
No				
ND				
Vasculonervous or bone involvement (%)	**8×10^{-7}**	3.34 [1.9–5.7]	0.069	
Yes				
No				
CINSARC				
C1	**1×10^{-7}**	3.7 [2.2–6.3]	**5×10^{-4}**	2.71 [1.02–7.2]
C2				

The prognostic value of CINSARC signature and FNCLCC grading system as well as significant clinical parameters (log rank $P < 0.05$, in bold) have been evaluated in a multivariate analysis with the Cox proportional hazard model. HR, hazard ratio.

of genes (335 and 244, respectively) but without any clearly emerging biologic pathway. Both studies involved relatively small series of specific sarcoma subtypes, and neither of these two signatures has been compared to the FNCLCC grading system. In addition, they remain to be validated in an independent group, therefore limiting their clinical usefulness. We have now identified and validated a gene expression signature composed of 67 genes associated with genome alterations number, tumor aggressiveness and metastatic outcome.

This gene expression signature assigns patient outcome better than the FNCLCC grading system. MFS demonstrates that in both cohorts, which comprise different histological types, CINSARC grade identified a group of tumors with a poor outcome, whereas the FNCLCC system did not significantly separate tumors according to outcome in cohort 1 (P = 0.4) and was less discriminating than CINSARC in cohort 2. CINSARC grading also has the advantage, as compared to FNCLCC grading, of stratifying tumor prognosis into two groups instead of the three with the FNCLCC system, which are difficult to use for clinical management. In line with this, CINSARC splits into two groups with different metastatic outcome those tumors previously considered with the FNCLCC system as having the same metastatic potential. In other words, with CINSARC grading, we can reassign tumors with intermediate FNCLCC grade 2 to good or poor prognosis groups. This is a major step forward, as it suggests that CINSARC grade can be used to assign treatments. From our results with GISTs, breast carcinomas and lymphomas, it also seems that CINSARC may be a transversal signature predicting clinical outcome in a broad array of cancers, meaning that CINSARC probably reflects a fundamental biological property of tumors that are likely to undergo metastasis.

Gene ontology analysis of the 67 CINSARC genes showed that all annotated genes are involved in the same biological processes, that is,

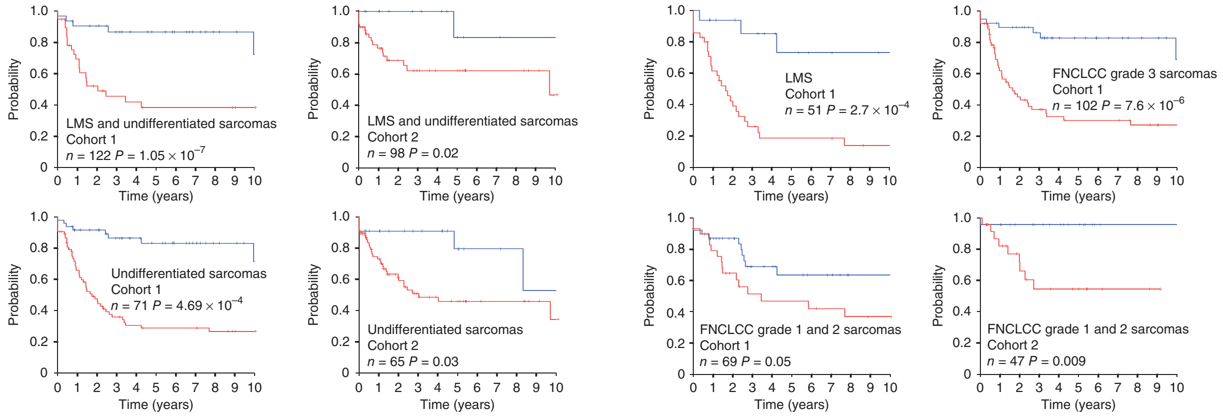


Figure 4 Metastasis-free survival analysis in clinically relevant groups. Because the CINSARC grade is fixed once established, tumors can be divided into homogeneous and clinically relevant groups. The CINSARC signature has been applied to undifferentiated sarcomas (cohort 1 and 2), to leiomyosarcomas (cohort 1) and to FNCLCC G1–2 subjects (cohorts 1 and 2) and FNCLCC G3 subjects (cohort 1). P values correspond to the log-rank test comparing the survival curves.

mitosis and control of chromosome integrity (**Supplementary Table 4**). These genes can be sorted into five main groups according to their role in mitosis: cell cycle and mitosis checkpoint (12 genes), chromosome biogenesis, condensation, alignment and segregation (26 genes), mitotic spindle and centrosome (12 genes) and microtubule motor, kinesin complex (8 genes) and cytokinesis (4 genes). Among the five others, three are experimentally related to chromosome instability[22], and two are related to histological grade in our study.

Some of these genes have already been implicated in prognostic signatures in breast cancer[31–33], where they are usually classified as proliferation genes. This is clearly the case, but the term 'proliferation' is inadequate, as these genes are strongly enriched in genes implicated in mitosis and chromosome management. Here we demonstrate that their expression is correlated with the complexity level of the tumor genome. In line with this, it was previously shown that genes involved in their signature are more related to chromosome instability than to proliferation[22]. Therefore, the molecular mechanisms leading to distant metastases might be related to the potentiality of tumor cells to induce or allow chromosome instability. Actually, cancer metastasis consists of a long series of sequential, interrelated steps, a process by which tumor cells disseminate, migrate, survive in the circulatory system, invade into a secondary site and start to proliferate[34,35]. We can hypothesize from our results that, in nonspecific translocation–related tumors, the more rearranged a genome, the higher the probability to obtain a gene expression profile permitting cells to complete the process of allowing dissemination and distant metastasis development.

At present, the benefit of chemotherapy in sarcomas is controversial, with recent studies and meta-analyses tending to demonstrate an effect on local and distant relapses[36–38]. Nevertheless, the efficacy of chemotherapy is marginal (from 3% to 10% depending on the endpoint[38]). One of the explanations could be patient selection by histologic grading. As demonstrated here, the CINSARC signature is a powerful metastatic predictor that could improve patient selection and increase the benefit of chemotherapy. Moreover, the biological meaning of the CINSARC genes defines them as potential targets for new therapeutic approaches targeting the early steps of metastatic development.

Further validation of the CINSARC signature is needed before it can replace the histological grading system in routine practice, but the fact that the CINSARC signature is validated in sarcomas as a stronger prognosis factor and is associated with poor outcome across such heterogeneous group of tumors (sarcomas to lymphomas and carcinomas) means that it should be seriously considered for selection of subjects in the context of prospective clinical trials. To validate the CINSARC signature as a decisional criterion for eligibility to adjuvant therapy in all sarcomas and particularly in GISTs where a targeted therapy already exists (Gleevec), we are currently leading a European project aiming to evaluate the predictive value of the CINSARC signature in 400 new sarcomas from all other histological subtypes, including 100 GISTs.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemedicine/.

Accession codes. Minimum Information About a Microarray Experiment–compliant data have been deposited at: Gene Expression Omnibus with accession number GSE21050.

Note: Supplementary information is available on the Nature Medicine website.

ACKNOWLEDGMENTS

This work was supported by grants from the French Institut National du Cancer, the French Projets Hospitaliers de Recherche Clinique program, the European Connective Tissue Cancer Network, the Curie Institute, the Bergonie Cancer Institute and l'INSERM. The construction of the human BAC array was supported by grants from the Carte d'Identité des Tumeurs program of the French Ligue Nationale Contre le Cancer. We are grateful to members of the French Sarcoma Group for providing tumor samples : Y.-M. Robin, X. Leroy, X. Saster-Garau, B. Marquès, M-C. Château, J-P. Ghnassia and F. Mishelany, J-B. Courrège, V. Dapremont and C. Ferreira. We would like to thank R. Maki and R. Iggo for critical reading of the manuscript. We thank R. Cook for English editing.

AUTHOR CONTRIBUTIONS

F. Collin, L.G., P.T., D.V.-R., A.L.C., B.B., S.B., A.L., J.-Y.B. and J.-M.C. supplied tumor tissues, did the central pathology review and collected the clinical follow-up data. F. Chibon supervised the laboratory experiments. P.L., G.P. and F. Chibon performed laboratory experiments. F.T. and C.L. developed the statistical software. A.d.R. and A.K. applied the centroid method. V.B. performed survival analysis. F. Chibon, S.S., A.d.R. and A.A. analyzed the data. F. Chibon, J.-M.C. and A.A. designed the study. F. Chibon, J.-M.C. and A.A. wrote the report. All investigators reviewed the final report.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemedicine/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Fletcher, C.D.M., Unni, K.K. & Mertens, F. Pathology and genetics of tumors of soft tissue and bone. in *World Health Organization Classification of Tumors* (Lyon, IARC Press, 2002).
2. Weitz, J., Antonescu, C.R. & Brennan, M.F. Localized extremity soft tissue sarcoma: improved knowledge with unchanged survival over time. *J. Clin. Oncol.* **21**, 2719–2725 (2003).
3. Zagars, G.K. *et al.* Prognostic factors for patients with localized soft-tissue sarcoma treated with conservation surgery and radiation therapy: an analysis of 225 patients. *Cancer* **97**, 2530–2543 (2003).
4. Trojani, M. *et al.* Soft-tissue sarcomas of adults; study of pathological prognostic variables and definition of a histopathological grading system. *Int. J. Cancer* **33**, 37–42 (1984).
5. Guillou, L. *et al.* Comparative study of the National Cancer Institute and French Federation of Cancer Centers Sarcoma Group grading systems in a population of 410 adult patients with soft tissue sarcoma. *J. Clin. Oncol.* **15**, 350–362 (1997).
6. Coindre, J.M. *et al.* Predictive value of grade for metastasis development in the main histologic types of adult soft tissue sarcomas: a study of 1240 patients from the French Federation of Cancer Centers Sarcoma Group. *Cancer* **91**, 1914–1926 (2001).
7. Idbaih, A. *et al.* Myxoid malignant fibrous histiocytoma and pleomorphic liposarcoma share very similar genomic imbalances. *Lab. Invest.* **85**, 176–181 (2005).
8. Chibon, F. *et al.* The use of clustering software for the classification of comparative genomic hybridization data. An analysis of 109 malignant fibrous histiocytomas. *Cancer Genet. Cytogenet.* **141**, 75–78 (2003).
9. Derré, J. *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.* **81**, 211–215 (2001).
10. Chibon, F. *et al.* A subgroup of malignant fibrous histiocytomas is associated with genetic changes similar to those of well-differentiated liposarcomas. *Cancer Genet. Cytogenet.* **139**, 24–29 (2002).
11. Coindre, J.M. *et al.* Most malignant fibrous histiocytomas developed in the retroperitoneum are dedifferentiated liposarcomas: a review of 25 cases initially diagnosed as malignant fibrous histiocytoma. *Mod. Pathol.* **16**, 256–262 (2003).
12. Nielsen, T.O. *et al.* Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet* **359**, 1301–1307 (2002).
13. Baird, K. *et al.* Gene expression profiling of human sarcomas: insights into sarcoma biology. *Cancer Res.* **65**, 9226–9235 (2005).
14. Fritz, B. *et al.* Microarray-based copy number and expression profiling in dedifferentiated and pleomorphic liposarcoma. *Cancer Res.* **62**, 2993–2998 (2002).
15. Matushansky, I. *et al.* A developmental model of sarcomagenesis defines a differentiation-based classification for liposarcomas. *Am. J. Pathol.* **172**, 1069–1080 (2008).
16. Segal, N.H. *et al.* Classification and subtype prediction of adult soft tissue sarcoma by functional genomics. *Am. J. Pathol.* **163**, 691–700 (2003).
17. Lee, Y.F. *et al.* Molecular classification of synovial sarcomas, leiomyosarcomas and malignant fibrous histiocytomas by gene expression profiling. *Br. J. Cancer* **88**, 510–515 (2003).

18. Nakayama, R. *et al.* Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma. *Mod. Pathol.* **20**, 749–759 (2007).
19. Singer, S. *et al.* Gene expression profiling of liposarcoma identifies distinct biological types/subtypes and potential therapeutic targets in well-differentiated and dedifferentiated liposarcoma. *Cancer Res.* **67**, 6626–6636 (2007).
20. Lee, Y.F. *et al.* A gene expression signature associated with metastatic outcome in human leiomyosarcomas. *Cancer Res.* **64**, 7201–7204 (2004).
21. Francis, P. *et al.* Diagnostic and prognostic gene expression signatures in 177 soft tissue sarcomas: hypoxia-induced transcription profile signifies metastatic potential. *BMC Genomics* **8**, 73 (2007).
22. Carter, S.L., Eklund, A.C., Kohane, I.S., Harris, L.N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* **38**, 1043–1048 (2006).
23. Yamaguchi, U. *et al.* Distinct gene expression-defined classes of gastrointestinal stromal tumor. *J. Clin. Oncol.* **26**, 4100–4108 (2008).
24. van 't Veer, L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
25. van de Vijver, M.J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
26. Lenz, G. *et al.* Stromal gene signatures in large-B-cell lymphomas. *N. Engl. J. Med.* **359**, 2313–2323 (2008).
27. Broders, A.C., Hargrave, R. & Meyerding, H.W. Pathologic features of soft tissue fibrosarcoma with special reference to the grading of its malignancy. *Surg. Gynecol. Obstet.* **69**, 267–280 (1939).
28. Russell, W.O. *et al.* A clinical and pathological staging system for soft tissue sarcomas. *Cancer* **40**, 1562–1570 (1977).
29. Markhede, G., Angervall, L. & Stener, B. A multivariate analysis of the prognosis after surgical treatment of malignant soft-tissue tumors. *Cancer* **49**, 1721–1733 (1982).
30. Costa, J., Wesley, R.A., Glatstein, E. & Rosenberg, S.A. The grading of soft tissue sarcomas. Results of a clinicohistopathologic correlation in a series of 163 cases. *Cancer* **53**, 530–541 (1984).
31. Sotiriou, C. *et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* **98**, 262–272 (2006).
32. Sotiriou, C. & Pusztai, L. Gene-expression signatures in breast cancer. *N. Engl. J. Med.* **360**, 790–800 (2009).
33. Desmedt, C. & Sotiriou, C. Proliferation: the most prominent predictor of clinical outcome in breast cancer. *Cell Cycle* **5**, 2198–2202 (2006).
34. Fidler, I.J. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nat. Rev. Cancer* **3**, 453–458 (2003).
35. Stafford, L.J., Vaidya, K.S. & Welch, D.R. Metastasis suppressors genes in cancer. *Int. J. Biochem. Cell Biol.* **40**, 874–891 (2008).
36. Sarcoma Meta-analysis Collaboration. Adjuvant chemotherapy for localised resectable soft-tissue sarcoma of adults: meta-analysis of individual data. *Lancet* **350**, 1647–1654 (1997).
37. Frustaci, S. *et al.* Adjuvant chemotherapy for adult soft tissue sarcomas of the extremities and girdles: results of the Italian randomized cooperative trial. *J. Clin. Oncol.* **19**, 1238–1247 (2001).
38. Pervaiz, N. *et al.* A systematic meta-analysis of randomized controlled trials of adjuvant chemotherapy for localized resectable soft-tissue sarcoma. *Cancer* **113**, 573–581 (2008).

ONLINE METHODS

Subjects and samples. The FSG database, which is part of the Conticabase (www.conticabase.org), contains data from adult soft tissue sarcomas treated in 11 centers, including a description of subjects, primary tumors, treatment, follow-up and availability of tumor samples. Every case was histologically reviewed by the pathologist subgroup and classified according to the 2002 World Health Organization classification by histology, immunohistochemistry and molecular genetics and cytogenetics when needed. For this study, we selected two cohorts of soft tissue sarcomas with no recurrent chromosomal translocations and for which a frozen tissue of the primary untreated tumor was available (**Table 1**).

According to French law at the time of the study, experiments were performed in agreement with the Bioethics Law 2004 800 and the Ethics Charter from the National Institute of Cancer; all subjects signed a nonopposition statement for research use of the sample.

DNA extraction and array comparative genomic hybridization analysis. Genomic DNA was isolated with a standard phenol-chloroform extraction protocol. Array-CGH experiments were done with a DNA microarray developed in our laboratory. We spotted 3,874 BAC DNAs (BACPAC Resources Center, Children's Hospital, Oakland Research Institute) in triplicate on UltraGAPS slides (Corning). These clones cover the whole genome with a resolution of 1 Mb. The probes were prepared and hybridized as previously described[39]. The data were analyzed with software developed at Institut Curie (CAPweb, http://bioinfo-out.curie.fr/CAPweb/). Cyanine-5 / cyanine-3 ratios >2.0 were considered as amplifications, and ratios >1.2 and <0.8 were considered as gains and losses, respectively. Analysis of array-CGH (computation of genomic alterations) was provided by the VAMP interface (http://bioinfo.curie.fr/vamp)[40].

RNA extraction and expression analysis. Total RNAs were extracted from frozen tumor samples with TRIzol reagent (Life Technologies) and purified with the RNeasy Min Elute Cleanup Kit (Qiagen) according to the manufacturer's procedures. We checked RNA quality on an Agilent 2100 bioanalyzer (Agilent Technologies). Samples were then analyzed on Human Genome U133 Plus 2.0 array (Affymetrix), according to the manufacturer's procedures. We simultaneously normalized all microarray data with the GCRMA (GC-Robust Multi-Array Analysis) algorithm[41]. To identify differentially expressed probe sets in cohort 1 (training set), we performed Welch *t* tests, and, to minimize the false discovery rate, *P* values were adjusted with the Benjamini-Hochberg procedure (R-Mulltest package)[42]. The EntrezGene identification numbers corresponding to selected probe sets ($P < 1 \times 10^{-2}$) and to all other probe sets from the array were obtained (in case of redundancy, only one EntrezGene identification number was retained). Each EntrezGene identification number was mapped to GeneOntology (http://www.geneontology.org/) with the R package org.Hs.eg.db. Finally, to rank gene ontology terms according to their relative amounts of gene expression changes, the *z* score for each gene ontology term was calculated. Positive *z* scores indicate gene ontology terms with a greater number of genes in the selected group than would be expected by chance. Negative *z* scores indicate gene ontology terms with fewer genes meeting the criterion than expected by chance. A *z* score close to zero indicates that the number of genes meeting the criterion was similar to that as would be expected by chance. The statistical significance of this result was evaluated using a Fisher test.

A pathway was considered as significantly overrepresented in the selected group when the corrected *P* value was $<1 \times 10^{-5}$ and the *z* score was >5.

Statistical analyses. We performed Chi-square tests to evaluate the relationship between the various tumor features, genomic alterations and expression profiles.

To assign prognosis, we applied the nearest centroid method. Centroids represent a centered mean of expression for the signature genes for each patient outcome (metastatic and nonmetastatic). Thus, centroids were calculated from cohort 1 samples for sarcomas and GISTs and from training set samples for breast cancers and lymphomas. Each sample of the training (for example, sarcomas cohort 1) and the validation (for example, sarcomas cohort 2) sets was allocated to the prognostic class (centroid) with the highest Spearman correlation.

Metastasis-free survival was obtained by the Kaplan-Meier method and calculated from the date of initial diagnosis to the date of first metastasis, last follow-up or death of the subject within diagnosis of metastasis. Metastasis-free survival curves were compared with the log-rank test. Hazard ratios and multivariate analysis were performed with the Cox proportional hazard model. For multivariate analysis, the method of maximum likelihood with stepwise backward elimination was used. All statistical analyses were performed with SPSS software version 16.0.

39. Vincent-Salomon, A. *et al.* Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity. *Breast Cancer Res.* **9**, R24 (2007).
40. La Rosa, P. *et al.* VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles. *Bioinformatics* **22**, 2066–2073 (2006).
41. Wu, Z. *et al.* A model based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.* **99**, 909–917 (2004).
42. R Development Core Team. R: A language and environment for statistical computing. <http://www.R-project.org> (2010).