Identification of Osteosarcoma Driver Genes by Integrative Analysis of Copy Number and Gene Expression Data

Marieke L. Kuijjer,[1] Halfdan Rydbeck,[2,3] Stine H. Kresse,[4] Emilie P. Buddingh,[5] Ana B. Lid,[3,4] Helene Roelofs,[6] Horst Bürger,[7] Ola Myklebost,[3,4] Pancras C. W. Hogendoorn,[1] Leonardo A. Meza-Zepeda,[3,4] and Anne-Marie Cleton-Jansen[1]*

[1]Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands
[2]Department of Informatics, University of Oslo, Oslo, Norway
[3]Norwegian Microarray Consortium, Institute for Molecular Bioscience, University of Oslo, Oslo, Norway
[4]Department of Tumor Biology, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway
[5]Department of Pediatrics, Leiden University Medical Center, Leiden, The Netherlands
[6]Department of Immunohematology and Blood Transfusion, Leiden University Medical Center, Leiden, The Netherlands
[7]Institute of Pathology, University of Münster, Münster, Germany

High-grade osteosarcoma is a tumor with a complex genomic profile, occurring primarily in adolescents with a second peak at middle age. The extensive genomic alterations obscure the identification of genes driving tumorigenesis during osteosarcoma development. To identify such driver genes, we integrated DNA copy number profiles (Affymetrix SNP 6.0) of 32 diagnostic biopsies with 84 expression profiles (Illumina Human-6 v2.0) of high-grade osteosarcoma as compared with its putative progenitor cells, i.e., mesenchymal stem cells ($n = 12$) or osteoblasts ($n = 3$). In addition, we performed paired analyses between copy number and expression profiles of a subset of 29 patients for which both DNA and mRNA profiles were available. Integrative analyses were performed in Nexus Copy Number software and statistical language R. Paired analyses were performed on all probes detecting significantly differentially expressed genes in corresponding *LIMMA* analyses. For both nonpaired and paired analyses, copy number aberration frequency was set to >35%. Nonpaired and paired integrative analyses resulted in 45 and 101 genes, respectively, which were present in both analyses using different control sets. Paired analyses detected >90% of all genes found with the corresponding nonpaired analyses. Remarkably, approximately twice as many genes as found in the corresponding nonpaired analyses were detected. Affected genes were intersected with differentially expressed genes in osteosarcoma cell lines, resulting in 31 new osteosarcoma driver genes. Cell division related genes, such as *MCM4* and *LATS2*, were overrepresented and genomic instability was predictive for metastasis-free survival, suggesting that deregulation of the cell cycle is a driver of osteosarcomagenesis. © 2012 Wiley Periodicals, Inc.

INTRODUCTION

High-grade osteosarcoma is an aggressive primary bone tumor, which mostly occurs during adolescence, with a second peak at middle age, at the metaphysis of long bones. The tumor is characterized by aberrant production of osteoid matrix and by very complex karyotypes (Raymond et al., 2002; Cleton-Jansen et al., 2005). Since the introduction of DNA microarray technology, recurrent DNA copy number changes in human osteosarcoma tumor tissues have been identified by comparative genomic hybridization (CGH) and high-density single nucleotide polymorphisms (SNP) microarray analysis. There is a general consensus about gain of chromosome arms 6p, 8q, and 17p, but many additional regions are reported as well (Squire et al., 2003; Man et al., 2004; Atiye et al., 2005; Yen et al., 2009; Kresse et al., 2010). The effects of copy number alterations may be reflected by changes in expression of genes in the affected chromosomal regions. There are various publications on human osteosarcoma gene expression, but few show robust bioinformatics (as described by Kuijjer et al.,

Additional Supporting Information may be found in the online version of this article.

Supported by: EuroBoNeT, Grant number: LSHC-CT-2006-018814; The Dutch Cancer Society, Grant number: KWF, 2008-4060; The Netherlands Organization for Health Research and Development, Grant number: 92003-399; Norwegian Research Council, Grant number: 183426; Norwegian Cancer Society, Grant number: 71572—PR-2006-0396, PK01-2007-0319, 107359—PR-2007-0163; Andraa's Legacy; FUGE Program of the Norwegian Research Council.

*Correspondence to: Anne-Marie Cleton-Jansen, Department of Pathology, Leiden University Medical Center, P.O. box 9600, L1-Q, 2300RC Leiden, The Netherlands.
E-mail: A.M.Cleton-Jansen@lumc.nl

Received 7 December 2011; Accepted 2 March 2012
DOI 10.1002/gcc.21956
Published online 27 March 2012 in
Wiley Online Library (wileyonlinelibrary.com).

2011). Often, small sample sizes and heterogeneity within groups result in only a small number of significant genes, on which usually no correction for multiple testing is applied. Another problem when studying osteosarcoma gene expression data is the lack of an osteosarcoma benign precursor lesion and its debated cell of origin—although it becomes clearer that the mesenchymal stem cell or its derivative is the progenitor of osteosarcoma (Mohseny et al., 2009; Mohseny and Hogendoorn, 2011). The disease seems to develop suddenly as a full-blown tumor, rendering it difficult to detect early drivers of osteosarcomagenesis. We have previously determined differential expression related to specific clinical parameters (Buddingh et al., 2011; Kuijjer et al., 2011). In addition, we have compared osteosarcoma with osteoblastoma—a benign tumor which develops at the same site as osteosarcoma, but which does not progress into the latter. This comparison of human osteosarcoma with a control tissue showed that cell cycle regulation is the most significantly altered pathway in osteosarcoma (Cleton-Jansen et al., 2009).

There are advantages of integrating copy number and expression data when aiming to identify driver genes. First, copy number data analysis of tumors with complex genomic profiles may return numerous bystander or hitch-hiker genes, as copy number alterations may occur not only because they are advantageous for the tumor but also as a result of general genomic instability. Regions of copy number alteration may therefore encompass no driver gene at all, or may include additional genes. Also, some genes with altered copy numbers may not be expressed in the tumor due to tissue-specific expression. These aspects hamper the identification of drivers of tumorigenesis, especially when the number of recurrent genes in such tumors is high. Second, at the mRNA level, drivers affect downstream genes and switch on feedback mechanisms, again rendering it difficult to determine the real osteosarcoma drivers in a pool of differentially expressed genes (Lee et al., 2008). Integration of DNA copy number and gene expression data filters out at least part of such bystanders and of genes that act downstream of drivers of tumorigenesis, because most of these genes have altered copy numbers, but no change in expression, or vice versa, while drivers are both amplified and upregulated, or deleted and downregulated. Particularly osteosarcoma is genetically extremely instable and therefore genomic data analysis of this tumor type would benefit from an approach that distinguishes driver genes from the numerous more random genetic events.

Nonpaired integrative analysis may be performed by determining recurrent regions of copy number alterations which have higher than expected numbers of differentially expressed genes. Paired integrative analysis is a more powerful method, because the relationship between copy number alterations and gene expression can be inferred in each specific sample, instead of being based on averaged quantities. A statistically correct method for paired integrative analysis of these different data types has not yet been defined. Paired integrative analysis is usually performed by selecting genes based on the correlation between gene expression and copy number levels, such as is performed by the recently published methods DR-Integrator (Salari et al., 2010) and Regularized dual Canonical Correlation Analysis (Soneson et al., 2010). However, gains and losses may not necessarily directly translate to the same quantity of change in expression levels (Lee et al., 2008), and important genes may be overlooked this way. A method where paired integrative analysis is detected for specific chromosomal regions with altered genomic and transcriptional status does exist (Bicciato et al., 2009), but this method is not optimal for tumors such as osteosarcoma with highly unstable genomes, since copy number values are normalized to the mean copy number over each array, and this mean value may be altered in such tumors. Two methods, PARADIGM and CNAmet, combine different types of data on a gene-based level. In PARADIGM, integration of different data types is used to detect patient-specific pathway activities (Vaske et al., 2010). CNAmet returns genes that show differential expression between samples with and without methylation and/or copy number alteration (Louhimo and Hautaniemi, 2011). This elegant approach may however hamper the identification of genes that are regulated by other frequently altered genes, such as *TP53* and *MDM2* in osteosarcoma.

Aiming to identify osteosarcoma driver genes, we performed both nonpaired and paired integrative analyses on high-grade osteosarcoma prechemotherapy biopsy data. We combined results from analyses as compared with different control sets—mesenchymal stem cells (MSCs) and osteoblasts, so that we did not exclude one of these proposed progenitors as the cell of origin of osteosarcoma. We show that the paired integrative analysis returns more affected genes than the nonpaired integrative analysis. There is an overrepresentation of genes involved in genomic stability in osteosarcoma samples. The identified genes may be important drivers in osteosarcomagenesis.

MATERIALS AND METHODS

Ethics Statement

All biological material was handled in a coded fashion. Ethical guidelines of the individual European partner institutions were followed and samples and clinical data were handled in a coded fashion and stored in the EuroBoNeT biobank.

Patient Material and Cell Lines

Genome-wide expression profiling was performed on pretreatment diagnostic biopsies of 84 resectable high-grade osteosarcoma patients from the EuroBoNeT consortium (www.eurobonet.eu). Clinicopathological details of these samples can be found in Table 1. Human bone-marrow-derived MSCs were obtained from five osteosarcoma patients and seven healthy donors. Osteoblasts ($n = 3$) were derived from MSCs on osteogenic differentiation. MSCs and osteoblasts were characterized and handled as described (Cleton-Jansen et al., 2009). Copy number analysis was performed on 32 pretreatment diagnostic biopsies, of which 29 overlapped with the 84 samples used for expression analysis.

TABLE 1. Clinicopathological Details

Category	Patient characteristics	Number of biopsies (%)
Institution	LUMC, Netherlands	36 (42.9)
	IOR, Italy	12 (14.3)
	LOH, Sweden	3 (3.6)
	Radiumhospitalet, Norway	1 (1.2)
	WWUM, Germany	32 (38.1)
Location primary tumor	Femur	40 (47.6)
	Tibia/Fibula	28 (33.3)
	Humerus	11 (13.1)
	Axial skeleton	1 (1.2)
	Unknown/other	4 (4.8)
Histological subtype	Osteoblastic	52 (61.9)
	Chondroblastic	9 (10.7)
	Fibroblastic	7 (8.3)
	Telangiectatic	4 (4.8)
	Minor subtype	11 (13.1)
	Unknown	1 (1.2)
Huvos grade	1 or 2	38 (45.2)
	3 or 4	33 (39.3)
	Unknown/NA	14 (16.7)
Metastasis at diagnosis	Yes	14 (16.7)
	No	69 (82.1)
	Unknown	1 (1.2)
Sex	Male	54 (64.3)
	Female	29 (34.5)
	Unknown	1 (1.2)
Age	<20 years	64 (76.2)
	>=20 years	19 (22.6)
	Unknown	1 (1.2)

Copy Number Microarray Data Analysis

Affymetrix Genome-Wide Human SNP 6.0 arrays (Affymetrix, Santa Clara, CA) were used for SNP data analysis. Genomic DNA preparation, labeling, hybridization, and scanning were performed as described by Kresse et al. (2010). Microarray data preprocessing was performed as described previously (Pansuriya et al., 2011). Hybridization quality was estimated by the genotype call rate using the Birdseed genotype calling algorithm in Genotyping Console (version 4.0, Affymetrix). Samples of poor quality were excluded from further analyses. We performed copy number analysis in Nexus software version 5 (Biodiscovery, El Segundo, CA) using CNCHP log-ratio files generated by Genotyping Console using 27 controls as a baseline, which is a subset of the reference set of 29 samples which was used by Pansuriya et al., 2011. We rejected two samples based on results from the quality control analysis in Genotyping Console. Hidden Markov model- (HMM-) based SNP-FASST segmentation was used to identify aberrant genomic regions. To be included as frequently aberrant, a copy number alteration was called when detected in at least 35% of all cases. Correlation of copy number alterations with clinical data was performed in Nexus software, with correction for multiple testing.

Genome-Wide Gene Expression Microarray Data Analysis

Osteosarcoma tissue handling, RNA isolation, synthesis of cDNA, cRNA amplification, hybridization of cRNA onto the Illumina Human-6 v2.0 Expression BeadChips (Illumina, San Diego, CA), and microarray data processing and quality control in the statistical language R version 2.10 (R Development Core Team, 2005) were performed as described previously (Buddingh et al., 2011). High correlations between these microarray data and corresponding qPCR results have been demonstrated previously (Buddingh et al., 2011). Unsupervised hierarchical cluster analysis was performed using R package *pvclust* with default settings (Suzuki and Shimodaira, 2006).

Data Deposition

MIAME-compliant copy number and gene expression data have been deposited in the GEO database (www.ncbi.nlm.nih.gov/geo/, superseries accession number GSE33383).

Detection of Significantly Differentially Expressed Genes

We performed a factorial *LIMMA* analysis (Smyth, 2004) in order to determine differential expression between high-grade osteosarcoma samples ($n = 84$) and control tissues—MSCs ($n = 12$) and osteoblasts ($n = 3$). Also, gene expression differences between MSCs and osteoblasts were determined. We used a Benjamini and Hochberg False Discovery Rate (FDR) of 0.05 as cut-off for significance.

Nonpaired Integrative Analysis

Nonpaired integrative analysis was performed by importing lists of differentially expressed genes into the Copy Number module of Nexus software version 5. Based on the length of the gene list, Nexus software performs a Fisher's exact test in order to determine whether the number of differentially expressed genes in a specific region with a significant copy number alteration is larger than expected by chance. Genes present in such regions of copy number alteration with FDR-adjusted *P*-values (Q-bounds in Nexus software) < 0.05 were returned from this integrative analysis. We did not apply any restrictions on the size of copy number aberrations. A few small altered regions that did not encompass an entire gene were detected, but these regions did not return genes upon integration with expression data. Nexus software only reports genes which are both gained and overexpressed, or both deleted and downregulated.

Paired Integrative Analysis

For the paired integrative analysis, copy number data of all autosomal overlapping genes between the copy number and gene expression data were exported from Nexus software, and converted into a binary matrix containing all genes with a gain (1) and no gain (0), and a similar binary matrix for losses. As in the nonpaired integrative analysis, we did not apply any restrictions on the size of copy number alterations. Gene expression data of each probe for each sample were normalized against average gene expression values of the corresponding probes over all control samples (either expression data from 12 MSCs or from three osteoblasts)—this was performed by subtracting the average expression of the control samples from the expression levels of the sample of interest, since these are log-transformed expression values. For both analyses, only genes that were significantly differentially expressed between the 84 osteosarcoma samples and the specific control set were analyzed, in order to make sure that all genes returned from the integrative analysis were significantly differentially expressed. Subsequently, genes that overlapped between the copy number binary matrices and that matched the fold change of expression (upregulation for genes with gains, and downregulation for genes with losses) were returned as two-way contingency tables using scripts in R. Genes that were altered in two types of data were further filtered by applying the sample recurrence criterion of 35%. This generated lists of recurrent two-way altered genes. The odds ratios for having both copy number and expression changes were calculated for different combinations, for instance gain and upregulation. We used Bonferroni corrected Chi-square or Fisher's exact *P*-values <0.05 to determine significance.

Gene Set Enrichment

GO term enrichment was tested using Bioconductor package *topGO* (Alexa et al., 2006). Lists of significantly affected genes were compared with all genes eligible for the analysis. GO terms with Fisher's exact *P*-values < 0.0001, as calculated by the *weight01* algorithm from *topGO*, were defined significant.

Genomic Instability Scores and Survival Analysis

We calculated genomic instability scores for 83 (out of the 84) osteosarcoma biopsies (for one sample no follow-up data were available) and all controls, as well as for two normal bone samples (obtained from cancer patients at the Norwegian Radium Hospital), 20 osteosarcoma xenografts (Kresse et al., unpublished data, and Kuijjer et al., 2011), 19 osteosarcoma cell lines (Ottaviano et al., 2010), and the HeLa cervical cancer cell line. For calculation of the genomic instability scores, we

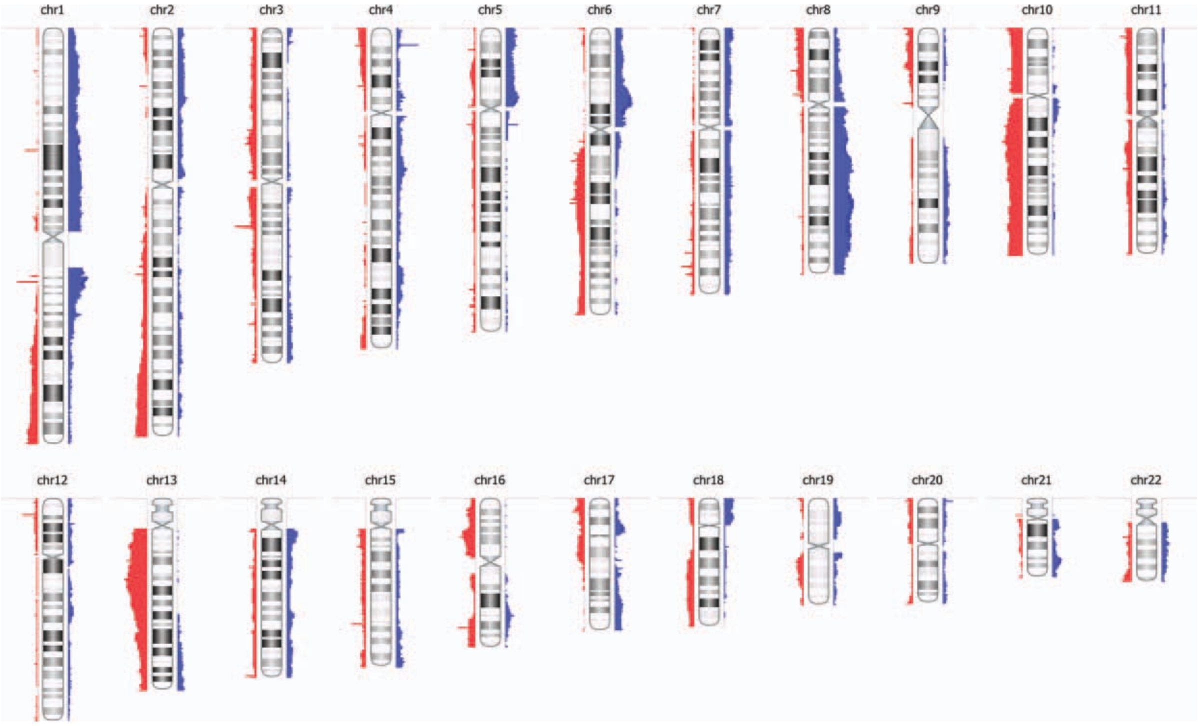

Figure 1. Genome-wide frequency plot of copy number alterations on chromosomes 1–22 in 32 high-grade osteosarcoma prechemotherapy biopsies. Left of the chromosomes, loss; right, gain. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

refer to the article by Carter et al. (2006). In short, this method calculates per sample per probe the expression of that particular probe minus the mean expression of that probe over all samples. For each sample, the sum of these values for all probes present in the genomic instability signature is calculated. This value is then compared between all samples and thus gives a relative measure of genomic instability. We used 24 genes of the CIN25 signature, because for one gene no probe was present on the Illumina v2.0 BeadChip. For genes with multiple probes, we used the probe that showed the highest variation in expression levels. We determined metastasis-free survival using the Kaplan-Meier method and performed a Logrank test for trend using GraphPad Software (La Jolla, CA, www.graphpad.com).

RESULTS

Recurrent Chromosomal Regions with Copy Number Aberrations in High-Grade Osteosarcoma

Thirty two high-grade osteosarcoma prechemotherapy biopsies were hybridized to Affymetrix SNP 6.0 arrays in order to determine recurrent copy number alterations. In total, 67 regions with recurrent alterations were detected, of which 35 regions had copy number gain, and 32 copy number loss (see Supporting Information Table 1). Recurrent gains were present on chromosome arms 1p, 1q, 4p, 5p, 6p, and 8q, and losses on chromosome arms 1p, 1q, 2q, 3q, 6q, 7q, 8p, 10p, 10q, 12p, 13q, 15q, 16p, and 16q. A genome-wide frequency plot of copy number alterations is shown in Figure 1. No significant correlation was detected for specific regions with copy number alterations and clinical information (tested clinical parameters are shown in Table 1).

Comparison of Osteoblasts and MSCs

Unsupervised hierarchical cluster analysis resulted in separate clusters for biopsies and cell lines. Within the cell line cluster, osteosarcoma cell lines formed one subcluster, whereas MSCs and osteoblasts formed a second subcluster (Supporting Information Fig. 1). This indicates that the control cell lines are more similar to one another than to osteosarcoma cells. On the basis of hierarchical clustering of gene expression data, we cannot determine the cell of origin of osteosarcoma. A total of 1,382 genes were differentially expressed between osteoblasts and MSCs. GO term enrichment resulted in seven significant GO terms, which are represented in Supporting Information Figure 2. In summary, GO term enrichment showed differences in cellular structure, proliferation, and apoptosis. Genes showing significant differences between both control cell types, however, can nonetheless be differentially expressed between osteosarcoma samples and both control cell types, thus can still be important drivers of osteosarcomagenesis. We therefore set out to select genes that showed differential expression in osteosarcoma as compared with both MSCs and osteoblasts.

Gene Expression Signature of High-Grade Osteosarcoma

We detected 12,542 and 2,939 probes encoding for genes that were significantly differentially expressed between the 84 osteosarcoma biopsies and MSCs and osteoblasts, respectively. MA plots, showing log-intensity ratios and log-intensity averages for both analyses, are depicted in Supporting Information Figure 3. A total of 1,679 probes overlapped between both analyses, of which 1,639 were either up- or downregulated in both. GO term analysis on the genes represented by these 1,639 probes showed an enrichment of apoptosis and signal transduction genes. Antigen processing and presentation, as well as angiogenesis were also over-represented (Supporting Information Fig. 4).

Paired Integrative Analysis Is More Sensitive Than Nonpaired Integrative Analysis

Nonpaired integrative analysis was performed on data from 32 samples hybridized on SNP arrays and from 84 samples hybridized on gene expression arrays, whereas paired analysis was performed on a subset of 29 samples for which both SNP and expression data were available. In total, 16,105 autosomal genes were represented both on SNP and on gene expression arrays. Nonpaired integrative analysis resulted in 253 significantly affected genes in osteosarcoma biopsies versus mesenchymal stem cells, whereas 71 genes were detected when osteoblasts were used as a control. A total of 45 genes were identified in both analyses versus MSCs and versus osteoblasts (Fig. 2). Of these 45 genes, 23 were also detected in expression analyses of a panel of 19 osteosarcoma cell lines (Ottaviano et al., 2010) versus MSCs and osteoblasts (Supporting Information Fig. 5A). For the paired integrative analyses, we determined whether the number of genes with gain combined with overexpression and with loss combined with downregulation was higher than expected per sample, based on the numbers of copy number alterations and gene expression changes in the whole genome. This was true for most samples, as depicted in Figure 3, where the odds ratios and significance of data dependencies are shown. Paired integrative analysis resulted in 445 and 138 genes when compared with MSCs and osteoblasts, respectively. A total of 101 genes overlapped between these different analyses (Fig. 2), and of this set, 31 genes were also detected in the cell line expression data (Supporting Information Fig. 5B, Table 2). Hence, paired analyses detected >90% of all genes found with corresponding nonpaired analyses. In addition, approximately twice as many genes as found in the corresponding nonpaired analyses were detected (Fig. 2, Supporting Information Fig. 6). Note that in the paired analysis fewer samples are included. Thus, paired analysis gives more robust results despite the lower sample size. Changing the threshold of FDR-adjusted *P*-values in the nonpaired integrative analysis from 0.05 to 0.15 (data not shown) did not alter this ratio.

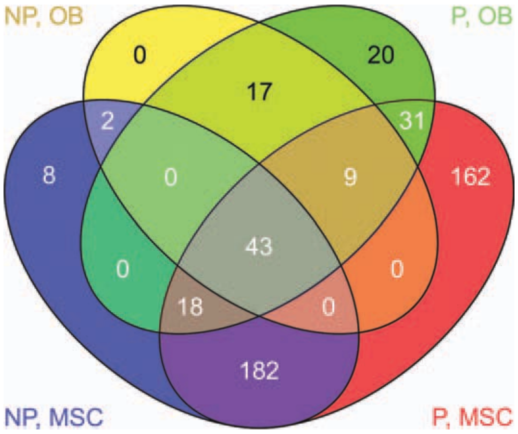

Figure 2. Venn diagram with numbers of affected genes in both nonpaired and paired analyses, and in osteosarcoma biopsies versus MSCs and versus osteoblasts. NP, nonpaired integrative analysis; P, paired integrative analysis; OB, analysis of osteosarcoma biopsies versus osteoblasts; MSC, analysis of osteosarcoma samples versus mesenchymal stem cells. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Genomic Instability Genes Play a Role in Osteosarcoma Progression

We calculated genome instability scores using the method of Carter et al. (2006), which compares levels of gene expression of a previously defined genomic instability signature between

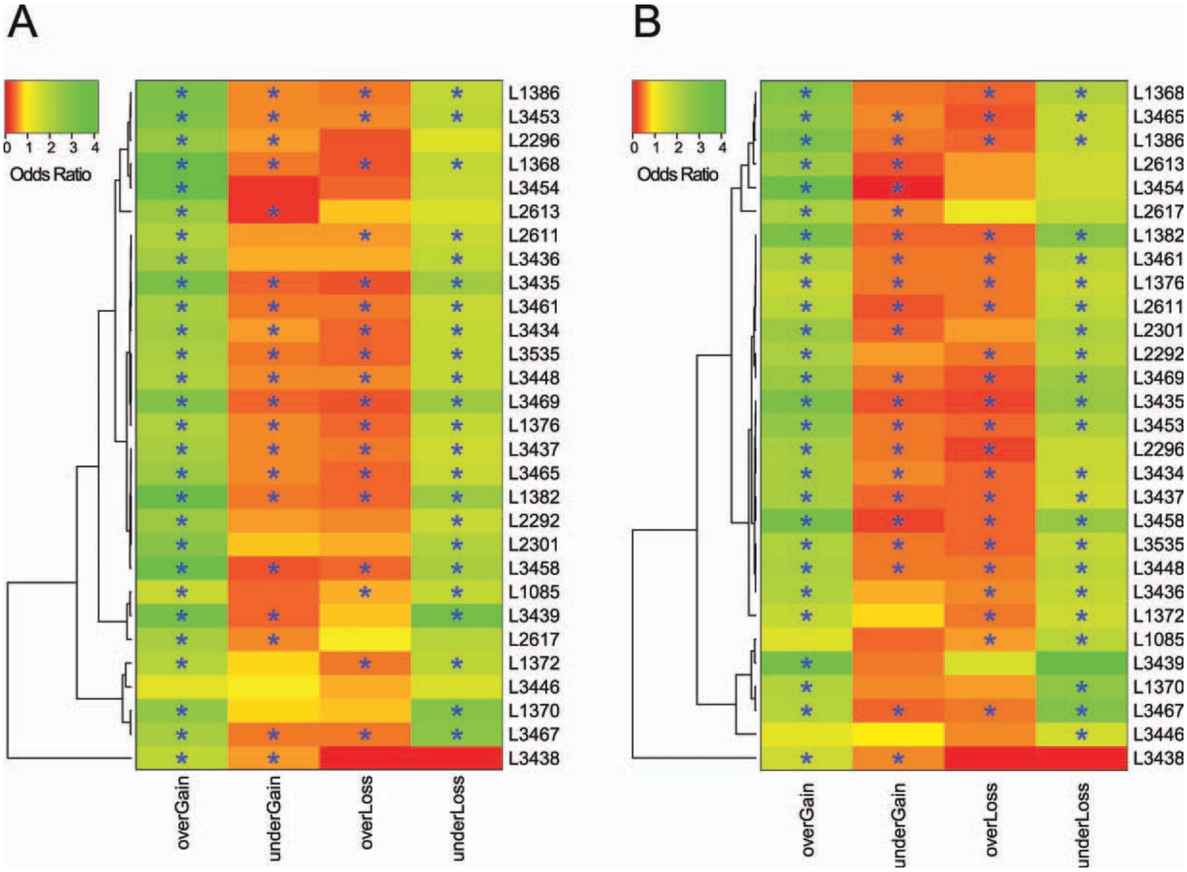

Figure 3. Dependence of gene copy number and gene expression data. The heatmaps depict odds ratios for the numbers of genes per sample which show gain and overexpression (overGain), gain and underexpression (underGain), loss and overexpression (overLoss), and loss and underexpression (underLoss). Chi-square tests, or, in case a group contained <10 genes, Fisher's exact tests, were performed in order to evaluate whether the number of genes reported from the integrative analysis was higher than expected by chance. * Bonferroni-corrected *P*-values <0.05. A: osteosarcoma biopsies versus MSCs; B: versus osteoblasts. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

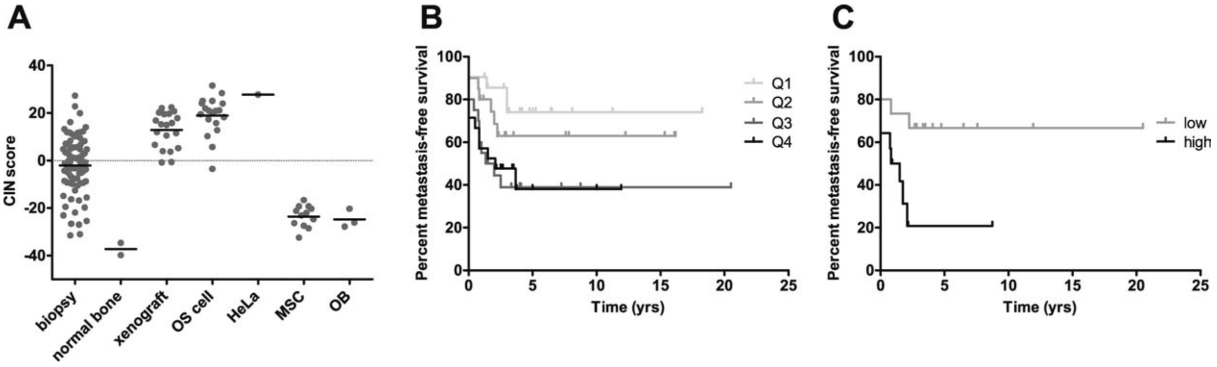

Figure 4. Genomic instability scores and metastasis-free survival. A: Genomic instability scores for high-grade osteosarcoma biopsies, normal bone, osteosarcoma xenografts and cell lines, the HeLa cell line, and mesenchymal stem cells (MSC) and osteoblasts (OB), as calculated by the method of Carter et al., 2006. B: Metastasis-free survival Kaplan-Meier curves for four quartiles of genomic instability scores. C: Metastasis-free survival Kaplan-Meier curves for the total amount of genes with copy number gains and losses, using a cut-off based on the median amount of genes per sample showing copy number aberration.

samples in a dataset, for all osteosarcoma biopsies and different control tissues and cell lines (Fig. 4A). The osteosarcoma biopsies showed highly variable scores, whereas genomic instability scores for the controls, normal bone, MSCs, and osteoblasts were relatively low. High instability scores were detected for osteosarcoma xenografts, cell lines, and the HeLa cell line, in increasing order. This signature predicted for metastasis-free survival in osteosarcoma samples as well (Fig. 4B), with high scores correlating with shorter metastasis-free survival (Logrank test for trend $P = 0.0112$). As expected, the total number of genes with copy number gains or losses, which is a direct measure of genomic instability from the SNP data, was predictive for progression as well (Logrank test $P = 0.018$, Fig. 4C).

Candidate Osteosarcoma Driver Genes

The 31 genes returned by the paired integrative analysis on clinical samples that also were differentially expressed in osteosarcoma cell lines are shown in Table 2, together with their chromosomal locations, aberration frequencies, and log fold changes. A total of 22/31 genes have been described to play a role in cancer. Interestingly, one third of these 22 genes have a role in cell cycle regulation, matching the importance of cell cycle and replication in osteosarcomagenesis as was found both using the genomic instability scores of the expression data and the overall chromosomal instability as detected in the copy number data (Fig. 4).

TABLE 2. Candidate Osteosarcoma Driver Genes

Symbol	Cytoband[a]	CNA[b]	CNA freq (%)[c]	logFC[d]
CLCC1	1p13.3	Gain	41.4	1.24
MCM4	8q11.21	Gain	37.9	1.35
AKR1C3	10p15.1	Loss	37.9	−1.94
AKR1C4	10p15.1	Loss	37.9	−1.34
ARHGAP22	10q11.22	Loss	37.9	−0.45
PGBD3	10q11.23	Loss	41.4	−0.82
ARID5B	10q21.2	Loss	48.3	−2.33
REEP3	10q21.3	Loss	48.3	−0.51
HERC4	10q21.3	Loss	51.7	−1.31
PBLD	10q21.3	Loss	48.3	−0.29
RUFY2	10q21.3	Loss	48.3	−0.20
KIAA1279	10q22.1	Loss	43.1	−0.57
SRGN	10q22.1	Loss	43.1	−2.26
AIFM2	10q22.1	Loss	44.8	−0.52
CHST3	10q22.1	Loss	48.3	−1.17
FAS	10q23.31	Loss	44.8	−0.42
PCGF5	10q23.32	Loss	37.9	−0.34
PPP1R3C	10q23.32	Loss	37.9	−2.89
AVPI1	10q24.2	Loss	37.9	−2.35
BLOC1S2	10q24.31	Loss	37.9	−0.51
CASC2	10q26.11	Loss	44.8	−0.18
FAM45A	10q26.11	Loss	39.7	−0.78
ERCC6	13q11.23	Loss	41.4	−0.52
WASF3	13q12.13	Loss	44.8	−2.43
C13orf33	13q12.3	Loss	48.3	−2.26
LHFP	13q14.11	Loss	48.3	−1.89
WBP4	13q14.11	Loss	55.2	−0.93
TSC22D1	13q14.11	Loss	58.6	−1.39
RCBTB1	13q14.2	Loss	58.6	−0.25
LATS2	13q21.11	Loss	44.8	−0.96
DCUN1D3	16p12.3	Loss	37.9	−1.39

All frequencies and fold changes are mean values of both integrative analyses—osteosarcoma biopsies versus MSCs and osteosarcoma biopsies versus osteoblasts. For genes for which more than one probe was present on the array, the probe with the highest fold change was used.

[a] UCSC cytogenetic band.

[b] Copy number aberration.

[c] Copy number aberration frequency ($n = 29$).

[d] log fold change in biopsies (negative means downregulation, positive means upregulation).

DISCUSSION

In this study, we report copy number and gene expression alterations in high-grade osteosarcoma prechemotherapy biopsies, and then integrate these data in order to detect osteosarcoma driver genes. Copy number analyses, which were obtained with high-density SNP microarrays, showed very high genomic instability in the osteosarcoma biopsies. The pattern of aberrations is in line with previous studies using aCGH and SNP arrays, which show recurrent gains in chromosome arms 1p, 6p, and 8q, and losses in chromosome 10. The previously reported recurrent amplification on chromosome arm 17p (Squire et al., 2003; Man et al., 2004; Atiye et al., 2005; Yen et al., 2009) is not listed, because we used a very strict cut-off for aberration frequency (35%). Aberration frequencies of 17% (Man et al., 2004) and 26% (Yen et al., 2009) were previously found on chromosome arm 17p, and a distinct amplification in 17p with an aberration frequency of 21% can be seen in Figure 1. We chose such a high cut-off for recurrent aberrations in order to enrich for selected genetic events and exclude the numerous haphazard alterations that can be attributed to the high genomic instability of high-grade osteosarcoma. In addition, we previously determined that this cut-off, as compared with cut-offs of 15% and 50%, showed the most consistent results in subsequent network and pathway analyses on osteosarcoma cell line SNP data (data not shown). For genome-wide gene expression analyses, both MSCs and osteoblasts were used as control cells, and we only considered overlapping genes between both comparisons, in order to make sure the affected genes were differentially regulated in osteosarcoma when compared with its putative progenitor cells. This analysis identified a large number ($n = 1{,}639$) of probes encoding for differentially expressed genes. Many of these genes encode tissue type-specific proteins, as is shown in the GO term enrichment analysis, and appear as upregulated in osteosarcoma biopsies because the in vitro grown control cells, MSCs and osteoblasts, lack surrounding stroma and are nurtured under other conditions. Antigen processing and presentation as well as angiogenesis pathways were expected to be upregulated, as macrophages and other infiltrating cells are present in osteosarcoma tissue (Buddingh et al., 2011), and as angiogenesis plays a role in osteosarcoma progression (Lee et al., 1999). Nevertheless, most stroma-derived gene expression is filtered out by integration with copy number data, as this expression is not a result of underlying copy number changes. In addition to stroma-related gene sets, GO term analysis showed enrichment in apoptosis and signal transduction genes, which are probably altered in the osteosarcoma tumor cells and not in the stroma. Because genes with concordant changes in copy number and gene expression are likely to be enriched in drivers of tumorigenesis, we performed integrative analyses on both types of data.

We found a remarkable increase in significant differential genes in paired compared with nonpaired analysis, i.e., 101 versus 45. In general, paired integrative analysis was advantageous over nonpaired integrative analysis, identifying roughly twice as many genes, also when different aberration frequency cut-offs or less stringent cut-offs for significance were used in the nonpaired analysis. Nonpaired analysis as performed in Nexus software compares the number of differentially expressed genes in a region of copy number aberration with the expected number of differentially expressed genes, which is based on the total number of differentially expressed genes over the whole genome. This method may be too rigorous, because an altered copy number region may encompass tissue-specific genes, which may not be expressed in the particular tumor tissue. These genes then have altered copy number, but no difference in expression. If an altered copy number region contains a relatively large number of such genes plus only a few candidate drivers, the entire region will be removed from the output of the analysis, which increases the amount of false negatives. Moreover, in the cancer gene expression profile, a large number of genes downstream of drivers, i.e., directly or indirectly regulated by drivers, or present in feedback loops will be differentially expressed. This increases the total number of differentially expressed genes, which again lowers the chance that a specific altered region is returned from the nonpaired integrative analysis as significantly affected. Furthermore, a single differentially expressed gene in a certain region of copy number alteration may still exert its driving function, and this driving function usually does not depend on the proportion of differentially expressed genes in the same region. Because of this, and because our method of paired integrative analysis is gene-based and not region-based, we did not perform a correction based on the total number of differentially expressed genes when compared with the affected copy number regions in the paired analysis in R, and this may be an additional reason why more genes are returned from the paired analysis. However, in all samples, except for one (L3438), the number of genes showing both copy number alteration and differential expression was higher than expected when compared with the numbers of copy number alterations and differentially expressed genes over the whole genome. This was significant for the vast number of samples (28/29, 23/29, 27/29, and 23/29, for combinations gain and overexpression, loss and underexpression in biopsies versus MSCs, and gain and overexpression, loss and underexpression in biopsies versus osteoblasts, respectively, as shown in Fig. 3).

Genomic instability scores showed that the instability in osteosarcoma tissues ranges from a level comparable to that of the controls, to the high instability levels of repeatedly passaged tumors in xenografts and osteosarcoma cell lines. We demonstrated both on copy number data, as well as by applying a genomic instability gene signature to genome-wide gene expression data, that high genomic instability in osteosarcoma is correlated with poor metastasis-free survival. This suggests that genes playing a role in genomic instability may be potent drivers of osteosarcoma progression, as has been reported for various other tumor types (Carter et al., 2006). Paired integrative analysis confirmed this result, as one third of the genes with a possible role in tumorigenesis had a function connected to the cell cycle. Of these genes, *MCM4* showed gain and overexpression and was only detected by the paired integrative analysis. *MCM4* is part of the minichromosome maintenance complex, which functions as a replication helicase, with a role in maintaining genomic stability (Aguilera and

Gomez-Gonzalez, 2008). This gene has been reported overexpressed in various tumor types (Freeman et al., 1999; Alison et al., 2002; Majid et al., 2010). Genes that were detected in both nonpaired and paired analyses were all deleted and underexpressed. *AVPI1*, or arginine vasopressin-induced 1, may be involved in cell cycling (UniProt Consortium, 2011). *ERCC6* is involved in transcription-coupled nucleotide excision repair, which is a critical survival pathway protecting against cancer (Fousteri and Mullenders, 2008). *RCBTB1*, a candidate tumor suppressor, was recently shown to have growth inhibitory activity in osteosarcoma cells by regulating pathways of DNA damage/repair and apoptosis (Zhou and Munger, 2010). *LATS2*, or large tumor suppressor homolog 2, plays a critical role in centrosome duplication, maintenance of mitotic fidelity, and genomic instability (Visser and Yang, 2010). Positive feedback between the p53 and Lats2 tumor suppressors prevents tetraploidization (Aylon et al., 2006), which could be an initiating step in osteosarcomagenesis, leading to genomic instability (Ganem and Pellman, 2007a; Ganem et al., 2007b). Also, a role of Lats2 in quenching of the increased genomic instability of H-Ras-induced transformation has been identified (Aylon et al., 2009). *DCUN1D3* encodes for a UVC-responsive protein involved in cell cycle progression and cell growth (Ma et al., 2008). Additional candidate genes with no direct role in cell cycle regulation include for example genes with a role in apoptosis (*AIFM2*, *BLOC1S2*, *FAS*) and metabolism (*AKR1C3* and *-4*). Some previously reported genes with a driver role in osteosarcoma were not identified, mainly because our high cut-off for recurrence. For example, *CDKN2A*, *MDM2*, and *E2F2* had recurrence frequencies of 28%, 17%, and 34%, respectively (in the dataset of 29 samples). *CDKN2A* and *MDM2* were not significantly differentially expressed, but *E2F2* was consistently significantly overexpressed with log fold changes >1.50 in all analyses (biopsies and cell lines as compared with different controls). *TP53* and *RB1* aberrations were present in >35% of all samples (38% and 69%, respectively). *TP53* was significantly downregulated in biopsies as compared with both controls, but not in the osteosarcoma cell line dataset. *RB1* showed significant downregulation when compared with MSCs, but not with osteoblasts, indicating a difference between these controls in *RB1* signaling. We set our cut-off for recurrence to 35% and only selected genes present both in osteosarcoma biopsies as well as in cell lines as compared with two different control sets, in order to select for the most important osteosarcoma drivers. Using this method, we were able to detect previously unreported driver genes.

In summary, we have shown that an individual gene-based paired integrative analysis of copy number and gene expression data performs better than a region-based nonpaired analysis. Several osteosarcoma candidate driver genes, especially genes playing a role in cell cycle progression, have been identified. Additional research, particularly functional studies, should reveal whether these genes are early or late drivers in osteosarcomagenesis.

ACKNOWLEDGMENTS

The authors like to thank Marcel Winter and Ronald Duim for gene expression microarray experiments; Konstantin Agelopoulos, Massimo Serra, and Fredrik Mertens for providing clinical samples; Heidi Namløs for control bone samples; and Bert Eussen for providing Nexus Copy Number software.

REFERENCES

Aguilera A, Gomez-Gonzalez B. 2008. Genome instability: A mechanistic view of its causes and consequences. Nat Rev Genet 9:204–217.

Alexa A, Rahnenfuhrer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22:1600–1607.

Alison MR, Hunt T, Forbes SJ. 2002. Minichromosome maintenance (MCM) proteins may be pre-cancer markers. Gut 50:290–291.

Atiye J, Wolf M, Kaur S, Monni O, Bohling T, Kivioja A, Tas E, Serra M, Tarkkanen M, Knuutila S. 2005. Gene amplifications in osteosarcoma-CGH microarray analysis. Genes Chrom Cancer 42:158–163.

Aylon Y, Michael D, Shmueli A, Yabuta N, Nojima H, Oren M. 2006. A positive feedback loop between the p53 and Lats2 tumor suppressors prevents tetraploidization. Genes Dev 20:2687–2700.

Aylon Y, Yabuta N, Besserglick H, Buganim Y, Rotter V, Nojima H, Oren M. 2009. Silencing of the Lats2 tumor suppressor overrides a p53-dependent oncogenic stress checkpoint and enables mutant H-Ras-driven cell transformation. Oncogene 28:4469–4479.

Bicciato S, Spinelli R, Zampieri M, Mangano E, Ferrari F, Beltrame L, Cifola I, Peano C, Solari A, Battaglia C. 2009. A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets. Nucleic Acids Res 37:5057–5070.

Buddingh EP, Kuijjer ML, Duim RA, Burger H, Agelopoulos K, Myklebost O, Serra M, Mertens F, Hogendoorn PCW, Lankester AC, Cleton-Jansen AM. 2011. Tumor-infiltrating macrophages are associated with metastasis suppression in high-grade osteosarcoma: A rationale for treatment with macrophage-activating agents. Clin Cancer Res 17: 2110–2119.

Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z. 2006. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat Genet 38:1043–1048.

Cleton-Jansen AM, Buerger H, Hogendoom PCW. 2005. Central high-grade osteosarcoma of bone: Diagnostic and genetic considerations. Curr Diagn Pathol 11:390–399.

Cleton-Jansen AM, Anninga JK, Briaire-de Bruijn I, Romeo S, Oosting J, Egeler RM, Gelderblom H, Taminiau AH, Hogendoorn PCW. 2009. Profiling of high-grade central osteosarcoma and its putative progenitor cells identifies tumourigenic pathways. Br J Cancer 101:2064.

Fousteri M, Mullenders LH. 2008. Transcription-coupled nucleotide excision repair in mammalian cells: Molecular mechanisms and biological effects. Cell Res 18:73–84.

Freeman A, Morris LS, Mills AD, Stoeber K, Laskey RA, Williams GH, Coleman N. 1999. Minichromosome maintenance proteins as biological markers of dysplasia and malignancy. Clin Cancer Res 5:2121–2132.

Ganem NJ, Pellman D. 2007a. Limiting the proliferation of polyploid cells. Cell 131:437–440.

Ganem NJ, Storchova Z, Pellman D. 2007b. Tetraploidy, aneuploidy and cancer. Curr Opin Genet Dev 17:157–162.

Kresse SH, Szuhai K, Barragan-Polania AH, Rydbeck H, Cleton-Jansen AM, Myklebost O, Meza-Zepeda LA. 2010. Evaluation of high-resolution microarray platforms for genomic profiling of bone tumours. BMC Res Notes 3:223.

Kuijjer ML, Namlos HM, Hauben EI, Machado I, Kresse SH, Serra M, Llombart-Bosch A, Hogendoorn PC, Meza-Zepeda LA, Myklebost O, Cleton-Jansen AM. 2011. mRNA expression profiles of primary high-grade central osteosarcoma are preserved in cell lines and xenografts. BMC Med Genomics 4:66.

Lee H, Kong SW, Park PJ. 2008. Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. Bioinformatics 24:889–896.

Lee YH, Tokunaga T, Oshika Y, Suto R, Yanagisawa K, Tomisawa M, Fukuda H, Nakano H, Abe S, Tateishi A, Kijima H, Yamazaki H, Tamaoki N, Ueyama Y, Nakamura M. 1999. Cell-retained isoforms of vascular endothelial growth factor (VEGF) are correlated with poor prognosis in osteosarcoma. Eur J Cancer 35:1089–1093.

Louhimo R, Hautaniemi S. 2011. CNAmet: An R package for integrating copy number, methylation and expression data. Bioinformatics 27:887–888.

Ma T, Shi T, Huang J, Wu L, Hu F, He P, Deng W, Gao P, Zhang Y, Song Q, Ma D, Qiu X. 2008. DCUN1D3, a novel UVC-responsive gene that is involved in cell cycle progression and cell growth. Cancer Sci 99:2128–2135.

Majid S, Dar AA, Saini S, Chen Y, Shahryari V, Liu J, Zaman MS, Hirata H, Yamamura S, Ueno K, Tanaka Y, Dahiya R. 2010. Regulation of minichromosome maintenance gene family by microRNA-1296 and genistein in prostate cancer. Cancer Res 70:2809–2818.

Man TK, Lu XY, Jaeweon K, Perlaky L, Harris CP, Shah S, Ladanyi M, Gorlick R, Lau CC, Rao PH. 2004. Genome-wide array comparative genomic hybridization analysis reveals distinct amplifications in osteosarcoma. BMC Cancer 4:45.

Mohseny AB, Szuhai K, Romeo S, Buddingh EP, Briaire-de Bruijn I, de Jong D, van Pel M, Cleton-Jansen AM, Hogendoorn PCW. 2009. Osteosarcoma originates from mesenchymal stem cells in consequence of aneuploidization and genomic loss of Cdkn2. J Pathol 219:294–305.

Mohseny AB, Hogendoorn PC. 2011. Concise review: Mesenchymal tumors: When stem cells go mad. Stem Cells 29:397–403.

Ottaviano L, Schaefer KL, Gajewski M, Huckenbeck W, Baldus S, Rogel U, Mackintosh C, de Alava E, Myklebost O, Kresse SH, Meza-Zepeda LA, Serra M, Cleton-Jansen AM, Hogendoorn PCW, Buerger H, Aigner T, Gabbert HE, Poremba C. 2010. Molecular characterization of commonly used cell lines for bone tumor research: A Trans-European EuroBoNet Effort. Genes Chrom Cancer 49:40–51.

Pansuriya TC, Oosting J, Krenacs T, Taminiau AH, Verdegaal SH, Sangiorgi L, Sciot R, Hogendoorn PC, Szuhai K, Bovee JV. 2011. Genome-wide analysis of Ollier disease: Is it all in the genes? Orphanet J Rare Dis 6:2.

R Development Core Team. 2005. R: A Language and Environment for Statistical Computing, Reference Index Version 2.10.0. Vienna, Austria: R Foundation for Statistical Computing.

Raymond AK, Ayala AG, Knuutila S. 2002. Conventional osteosarcoma. In: Fletcher CDM, Unni KK, Mertens F, editors. World Health Classification of Tumours. Pathology and Genetics of Tumours of Soft Tissue and Bone. Lyon: IARC Press, pp 264–270.

Salari K, Tibshirani R, Pollack JR. 2010. DR-Integrator: A new analytic tool for integrating DNA copy number and gene expression data. Bioinformatics 26:414–416.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:Article 3.

Soneson C, Lilljebjorn H, Fioretos T, Fontes M. 2010. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. BMC Bioinformatics 11:191.

Squire JA, Pei J, Marrano P, Beheshti B, Bayani J, Lim G, Moldovan L, Zielenska M. 2003. High-resolution mapping of amplifications and deletions in pediatric osteosarcoma by use of CGH analysis of cDNA microarrays. Genes Chrom Cancer 38:215–225.

Suzuki R, Shimodaira H. 2006. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22:1540–1542.

UniProt Consortium. 2011. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res 39:D214–D219.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26:i237–i245.

Visser S, Yang X. 2010. LATS tumor suppressor: A new governor of cellular homeostasis. Cell Cycle 9:3892–3903.

Yen CC, Chen WM, Chen TH, Chen WY, Chen PC, Chiou HJ, Hung GY, Wu HT, Wei CJ, Shiau CY, Wu YC, Chao TC, Tzeng CH, Chen PM, Lin CH, Chen YJ, Fletcher JA. 2009. Identification of chromosomal aberrations associated with disease progression and a novel 3q13.31 deletion involving LSAMP gene in osteosarcoma. Int J Oncol 35:775–788.

Zhou X, Munger K. 2010. Clld7, a candidate tumor suppressor on chromosome 13q14, regulates pathways of DNA damage/repair and apoptosis. Cancer Res 70:9434–9443.