Oncogene (2004) 23, 2484–2498

www.nature.com/onc

Identification of genes associated with tumorigenesis and metastatic potential of hypopharyngeal cancer by microarray analysis

Anne Cromer[1,3], Annaïck Carles[1,3], Régine Millon[2,3], Gitali Ganguli[1], Frédéric Chalmel[1], Frédéric Lemaire[1], Julia Young[1], Doulaye Dembélé[1], Christelle Thibault[1], Danièle Muller[2], Olivier Poch[1], Joseph Abecassis[2] and Bohdan Wasylyk*[,1]

[1]*Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, 1 Rue Laurent Fries, BP 10142, 67404 Illkirch cedex, France;* [2]*UPRES EA 34-30, Centre Paul Strauss, 3 rue de la Porte de l'Hôpital, 67085 Strasbourg, France*

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common cancer among men in the developed world. There is a need, for both clinical and scientific reasons, to find markers to identify patients with aggressive disease as early as possible, and to understand the events leading to malignant transformation and susceptibility to metastasis. We report the first large-scale gene expression analysis of a unique HNSCC location, the hypopharynx. Four normal and 34 tumour samples were analysed with 12 600 gene microarrays. Clusters of differentially expressed genes were identified in the chromosomal regions 3q27.3, 17q21.2–q21.31, 7q11.22–q22.1 and 11q13.1–q13.3, which, interestingly, have already been identified by comparative genomic hybridization (CGH) as major regions of gene amplification. We showed that six overexpressed genes (EIF4G1, DVL3, EPHB4, MCM7, BRMS1 and SART1) located in these regions are indeed amplified. We report 119 genes that are highly differentially expressed between 'early' tumours and normal samples. Of these, we validated by quantitative PCR six novel poorly characterized genes. These genes are potential new markers of HNSCC. Comparing patients with relatively nonaggressive and aggressive tumours (without or with clinical evidence of metastasis 3 years after surgery), we identified 164 differentially expressed genes potentially involved in the acquisition of metastatic potential. This study contributes to the understanding of HNSCC, staging patients into prognostic groups and identifying high-risk patients who may benefit from more aggressive treatment.
Oncogene (2004) **23,** 2484–2498. doi:10.1038/sj.onc.1207345
Published online 15 December 2003

Keywords: HNSCC; chromosomal location; prognosis; gene amplification

*Correspondence: B Wasylyk; E-mail: boh@igbmc.u-strasbg.fr
[3]These authors contributed equally to this work
Received 22 July 2003; revised 22 October 2003; accepted 11 November 2003

Introduction

Head and neck squamous cell carcinoma (HNSCC), which affects the oral cavity, the oropharynx, the larynx and the hypopharynx, is the sixth most common cancer among men in the developed world. Well-known risk factors include tobacco and alcohol. Over the last decades, diagnosis and management have improved, but not long-term survival rates. The prognosis of HNSCC is influenced by many factors, such as TNM staging and pathological grading of differentiation. However, since these factors are not sufficient to evaluate outcome, there is a need to identify molecular biomarkers that will help to stage patients in prognostic groups and to identify high-risk patients who may benefit from different treatments. Some genes have been suggested to be potential prognostic markers in HNSCC (Quon *et al*., 2001). However, molecular markers do not yet contribute to the clinical decision-making process.

Cancer appears to result from the progressive accumulation of genetic aberrations. Amplified chromosomal regions may contain dominant oncogenes, whereas deleted regions may harbour tumour suppressor genes. HNSCC is frequently associated with specific chromosomal aberrations, including amplification of 3q, 8q, 9q, 20q, 7p, 11q13 and 5p, and deletion of 3p, 9p, 21q, 5q, 13q, 18q and 8p (Gollin, 2001). mRNA expression levels often reflect these changes. However, the deregulation of gene expression in cancer is more complex than a simple relationship between genomic aberration and gene expression (Platzer *et al*., 2002; Pollack *et al*., 2002).

Expression-array profiling has recently been used to distinguish between cancer subtypes (Golub *et al*., 1999) and stages of progression (Bittner *et al*., 2000). The capacity of tumour cells to metastasize could be acquired early in tumorigenesis (Bernards and Weinberg, 2002), and patterns of gene expression can predict the metastatic outcome of patients with breast (van't Veer *et al*., 2002) and prostate (Singh *et al*., 2002) cancer. Microarray analysis of gene expression has been reported for HNSCC (Leethanakul *et al*., 2000; Alevizos *et al*., 2001; Al Moustafa *et al*., 2002; El-Naggar *et al*., 2002; Mendez *et al*., 2002), but did not concern

well-defined stages of tumour evolution in a precise head and neck site.

Here we describe the first large-scale gene expression analysis of the hypopharynx, a localization associated with particularly aggressive behaviour (Genden *et al*., 2003). The aim of our study was to identify genes involved in tumorigenesis, as well as gene expression patterns that will distinguish tumours that will metastasize from those that will not. We have identified by microarray analysis genes whose expression differs between tumours and normal tissues, as well as between tumours with similar initial stage and histopathological features but with different clinical outcomes.

Results and discussion

We determined with Affymetrix HG-U95A microarrays the expression profiles of 34 hypopharyngeal cancer samples (including 31 individual tumours, two pools of two and one pool of three additional tumours) and four

Table 1 Patient and sample characteristics

Class	*Patient*	*Sample*	*Localiz.*	*T*	*N*	*M*	*Diff*	*Sex*	*Age*	*Treatment after surgery (4)*	*Evolution*	*Actual state*	*Overall survival (5)*	*Disease free survival (5)*
Normal														
N	pool(1)*	N	U											
N	1047(2)*	N	U											
N (3)	1102*	N	U											
N (3)	1107*	N	U											
'Early' tumours														
Pool_E	PE1*	T	H	2	0	0	2	M	53	RX	0	D	15	11
	PE2*	T	H	1	0	0	2	M	61	N	0	D	25	15
	PE3*	T	H	2	0	0	2	M	44	N	0	A	41	41
E	1047(2)*	T	H	2	0	0	1	M	68	RX	0	A	12	12
E	435*	T	H	1	0	0	1	M	72	RX	SC	A	92	91
E	834*	T	H	1	0	0	2	M	52	N	SC	D	61	42
'Late' tumours														
LR	118	T	H	4	2b	0	1	M	51	RX	LR	D	39	35
LR	429	T	H	3	3	0	2	M	58	RX	LR	D	77	37
LR	684	T	H	3	1	0	3	M	65	RX	LR	D	40	29
LR	702	T	H	3	2c	0	3	M	70	RX	LR	D	5	4
Pool_NM	PNM1*	T	H	3	2b	0	3	M	66	RX	0	D	58	54
	PNM2*	T	H	3	2b	0	2	M	52	RX	SC	A	44	39
NM	065*	T	H	4	2c	0	1	M	67	RX	0	A	118	118
NM	103	T	H	3	1	0	2	M	55	RX	0	A	143	95
NM	167	T	H	4	2c	0	2	M	61	RX	0	A	140	86
NM	190*	T	H	1	2b	0	1	M	70	RX	0	A	63	39
NM	279*	T	H	3	2c	0	1	M	53	RX	0	A	104	104
NM	357	T	H	3	2c	0	2	M	45	RX	0	A	101	101
NM	409	T	H	4	1	0	2	M	43	RX	0	A	116	116
NM	423	T	H	2	2b	0	2	M	64	RX	0	A	83	83
NM	847	T	H	2	1	0	3	F	63	RX	0	A	67	67
NM	856*	T	H	3	2b	0	1	M	59	RX	0	A	66	66
Pool_M	PM1*	T	H	2	2c	0	3	M	54	RX	M	D	25	16
	PM2*	T	H	3	2b	0	2	M	54	RX	M	D	7	5
M	135*	T	H	3	2c	0	2	M	51	RX	M	D	84	14
M	165	T	H	4	2c	0	2	M	71	RX	M	D	22	9
M	203	T	H	3	2b	0	2	M	54	RX	M	D	20	11
M	209*	T	H	3	2c	0	2	M	62	RX	M	D	18	8
M	215	T	H	4	3	0	2	M	43	RX + CT	M	D	12	3
M	218	T	H	3	3	0	2	M	56	RX	M	D	12	8
M	330	T	H	3	2b	0	3	M	63	RX	M	D	7	7
M	408	T	H	1	2c	0	3	M	49	RX + CT	M	A	16	1
M	621	T	H	3	3	0	3	M	41	RX + CT	M	D	34	28
M	629	T	H	3	3	0	2	F	57	RX + CT	M	D	8	8
M	744	T	H	3	2c	0	2	M	51	RX	M	D	9	8
M	829	T	H	4	3	0	2	M	69	RX	M	D	5	1
M	888*	T	H	3	2b	0	2	M	64	RX	M	D	26	21
M	933*	T	H	2	2b	0	2	M	70	RX	M	D	23	22

Sample type: T = tumour, N = normal; localization (localiz.): H = hypopharynx, U = uvula; T, N and M reflect the TNM nomenclature for tumour stage; evolution: 0 = no evolution, LR = local recurrence, M = metastasis, SC = secondary cancer; last follow-up status: D = dead, A = alive; *samples from batch 1. (1) The N pool contains the normal samples corresponding to the tumours in the TE, TNM and TM pools, the E pool consists of three tumours (PE1, PE2 and PE3), the NM pool two (PNM1 and PNM2) and the M pool two (PM1 and PM2); (2) 1047T and 1047N are tumour and normal samples from the same patient; (3) samples were macrodissected to avoid nonepithelial cells; (4) RX = radiotherapy treated, RX + CT = radiotherapy and chemotherapy treated, N = no treatment; (5) in months

normal samples (including one pool of seven normal tissues from patients whose tumours are in the tumour pools). A full description of the clinical data, including diagnosis and outcome, is shown in Table 1. We used three different approaches to analyse the data, in order to identify gene expression signatures of tumorigenesis and metastatic evolution.

Global gene expression profile

The data from 34 hypopharyngeal cancer and four normal samples were filtered to exclude genes with low expression, resulting in a working set of 3962 probe sets (see Materials and methods, Preprocessing of the data). We were interested to know if there are patterns of gene expression that cluster samples and single out particular subgroups. Unsupervised hierarchical clustering was used to group samples according to similarity in gene expression, without prior knowledge of sample identity and without any gene selection (Figure 1a). The four normal (N) samples cluster together, indicating a strong similarity in their expression profiles. The tumour (T) samples are less similar among themselves than the normal samples, as represented by the length of the vertical lines linking them. Cluster analysis can find coherent patterns of gene expression, but provides little information about statistical significance. In order to identify genes that have significant changes in expression, we used significance analysis of microarray (SAM) (Tusher *et al*., 2001). In a three-step process of paired comparisons (see Materials and methods), we selected 2377 probe sets that are differentially expressed (DE) between tumour and normal tissues. Unsupervised hierarchical clustering with the SAM selected genes resulted in tighter clustering of the normal samples and clearer distinction from the tumour samples (Figure 1b). The expression profiles remained more heterogeneous among tumours than among normal samples. Heterogeneity of expression between tumours can be used for classification (Liu, 2003).

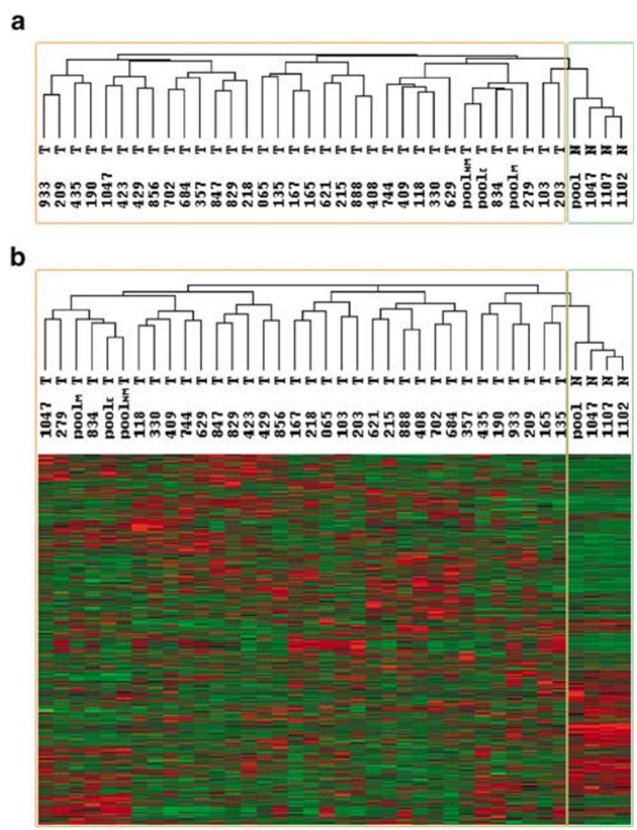

Figure 1 Classification of hypopharyngeal cancer by global gene expression profiling. Unsupervised hierarchical clustering of 38 tumour and normal HNSCC samples using 3962 working probe sets with significant levels of expression (**a**), and 2377 probe sets selected for being differentially expressed between tumour and normal by SAM (**b**). The dendrogram indicates the degree of relatedness of the samples, the shorter the vertical branch linking two samples, the closer their expression profiles. The light green box surrounds the normal (N) cluster, and orange box the tumour (T) cluster. Gene expression profiles in (**b**) are represented with a green to red (lower to higher expression) colour scale

Chromosomal localization of differentially expressed genes

Genomic aberrations associated with malignant transformation can affect gene expression. Gene expression profiles of cell lines show that chromosomal modifications can lead to regional biases in gene expression (Monni *et al*., 2001; Phillips *et al*., 2001). We searched for similar regional biases with the SAM selected DE genes using three parameters: local density (LD), the normalized relative density (NRD) and the nearest neighbour (NN) score (Materials and methods; Figure 2). We first showed that the chromosomal distribution of the total gene set on the microarrays is representative of the distribution of known genes (data not shown), which assured us that there is no underlying bias. We found global high densities of DE sequences on chromosomes 19, 17 and 22 and very low densities on chromosomes 13 and Y (Figure 2), which reflects the overall densities of genes on these chromosomes. We can distinguish three groups of chromosomal regions (Figure 2), with different LD, NRD and NN score characteristics.

The first group, which has the characteristics high LD, high NRD and small NN scores, consists of four chromosomal regions (cytobands 2p23.3, 12q13.12, 22q12.1 and 22q13.1). The high density of small genes in these regions could explain their high LD and NRD with low NN scores. A total of 40 genes are located on the 28th Mb of chromosome HS02 (2p23.3) and 39 on the 57th Mb of HS12 (12q13.12; http://www.ncbi.nlm.nih.gov/mapview/), which is higher than average (30 genes per Mb). Most of these genes are less than 10 000 base pairs in length. Apparently, the first group distinguishes four regions with numerous small genes.

The second group, which has the characteristics low LD, low NRD and high NN scores, consists of 18 chromosomal regions (only the NN score is visible in Figure 2): cytobands 1p36.33, 2q14.3–q21.2, 2q35, 4q32.3, 5q31.3, 6p25.1, 6p21.32, 6q13, 7p15.3, 7p15.1, 7q22.1, 7q31.1, 7q32.2–q32.3, 9q22.1, 10q26.13–q26.2, 12q23.3–q24.11, 14q11.1–q11.2 and 16q22.1–q23.1. The

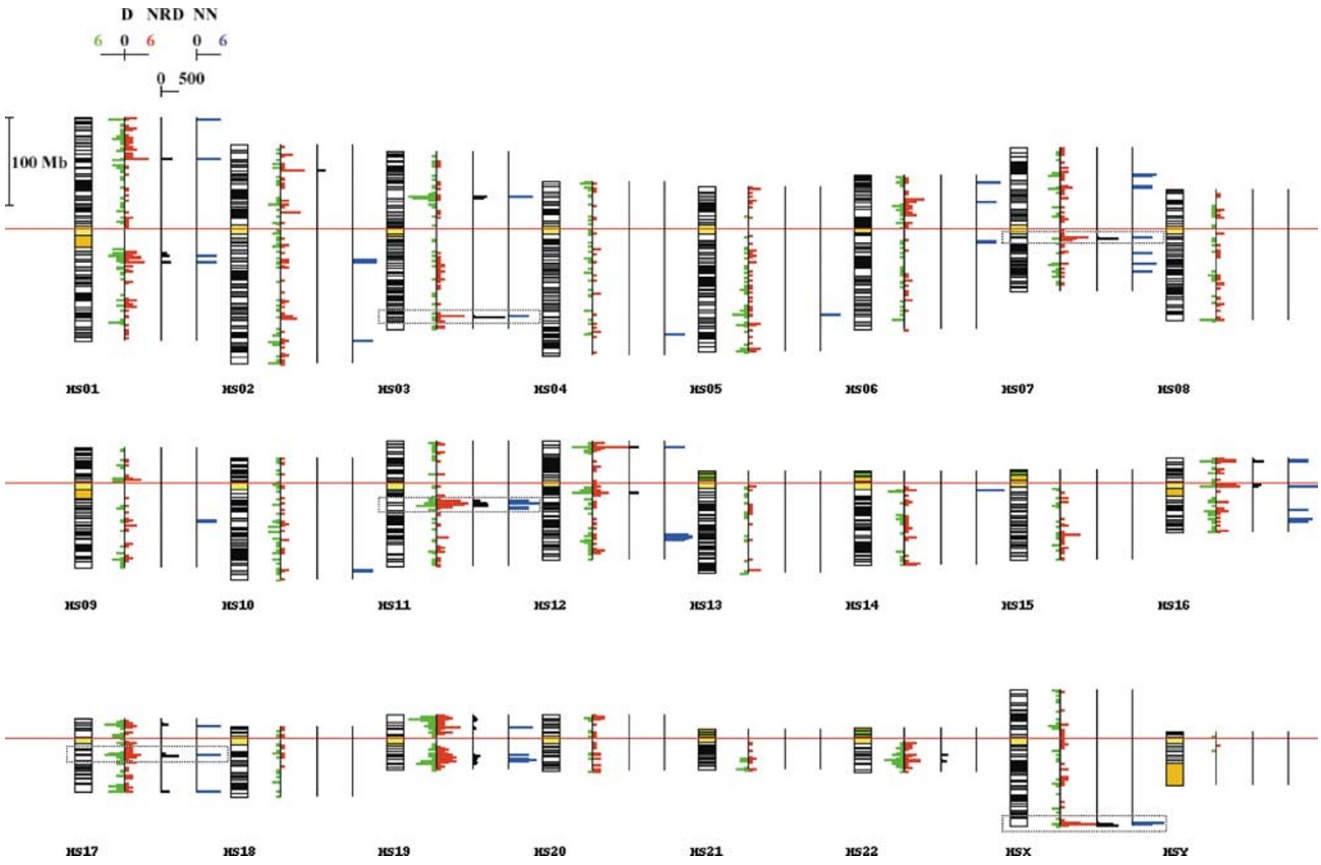

Figure 2 Chromosomal distribution of sequences differentially expressed (DE) in HNSCC. Three vertical histograms are shown on the right of each chromosome. The first histogram represents the local density (LD, labelled D), that is, the number of DE sequences per million base pairs (Mb; red, overexpressed; green, underexpressed). The second histogram shows the normalized relative density (NRD) calculated for each Mb where more than six DE sequences are located. The third histogram shows in blue the Mb areas where at least two consecutive DE sequences have nearest neighbour (NN) scores greater than four (see Materials and methods). Black dotted boxes surround the strongest clusters. Chromosomes and histograms are drawn to the same scale, and the cytobands are drawn according to ftp://ftp.ncbi.nih.gov/genomes/H_sapiens

low density of large genes in these regions could explain the low LD, low NRD and high NN scores (LD and NRD are affected by low gene density, but not the NN score). For example, there is only 1 gene per Mb at 14q11.1–q11.2 (http://www.ncbi.nlm.nih.gov/mapview/). Apparently, the second group points to 18 chromosomal regions with long genes.

Finally, the third group, which has high LD, high NRD and high NN scores, is particularly interesting because it potentially identifies clusters of over- and underexpressed genes. It consists of 17 chromosomal regions (cytobands 1p34.1, 1q21.3, 1q23.1, 3p21.31, 3q27.3, 7q11.22, 11q13.1–q13.3, 12p13.32, 16p13.3, 16q11.1, 17p13.1, 17q21.2, 17q25.3, 19p13.13, 19q13.2, 19q13.32–q13.33 and Xq28). The strongest clusters of differentially expressed sequences, with the highest LD, NRD and NN scores, are framed with a dotted line in Figure 2 (3q27.3, 17q21.2, Xq28, 7q11.22 and 11q13.1–q13.3). Interestingly, four of these regions (all except Xq28) have been shown to have strong DNA copy number gains by comparative genomic hybridization (CGH) in pharyngeal squamous cell carcinoma (PSCC) (Huang *et al*., 2002). The correlation is particularly good on chromosome 3. The cluster in 3p21, consisting of a maximal density of underexpressed sequences combined with a high NRD and a high NN score (Figure 2), corresponds to low DNA copy number by CGH. On the contrary, the cluster of overexpressed sequences located in 3q27.3 corresponds to high DNA copy number by CGH. Interestingly, +3q (gain at 3q) is known to be particularly important for the progression of HNSCC (Redon *et al*., 2001; Huang *et al*., 2002) and +7q and +17q for PSCC (Huang *et al*., 2002). Furthermore, +11q12–13 is one of the smallest recurrent chromosomal regions with a high-level amplification (Huang *et al*., 2002), particularly in hypopharyngeal tumours (Muller *et al*., 1997). These results suggest that overexpression is attributable to DNA amplification in these regions. Some of the genes located in these clusters (3q27.3, 17q21.2–q21.31, 7q11.22–22.1 and 11q13.1–q13.3) are known to be involved in tumorigenesis, such as PAI1 and PPFIA1. We were interested in investigating whether 'poorly known' genes with potentially interesting functions in these regions are amplified. We investigated EIF4G1, DVL3, KRT17 and KRT16, EPHB4, MCM7, BRMS1 and SART1 (Table 2), which

Table 2 Eight selected genes located in the four chromosomal regions with the highest gene-density bias

NRD	*Chromosomal location*	*Genomic location (Mb)*	*Gene*	*AC*	*E/N*
1228	3q27.3	180.6	Eukaryotic translation initiation factor 4 gamma (EIF4G1)	Q04637	1.78
	3q27.3	180.7	Segment polarity protein dishevelled homolog DVL-3 (DVL3)	U75651	3.65
950	17q21.31	39.2	Keratin, type I cytoskeletal 17 (Cytokeratin 17) (KRT17)	BC011901	7.57
	17q21.31	39.2	Keratin, type I cytoskeletal 16 (Cytokeratin 16) (KRT16)	AF061812	8.45
760	7q22.1	98.9	Ephrin type-B receptor 4 precursor (EC 2.7.1.112) (EPHB4)	U07695	3.12
	7q22.1	98.2	DNA replication licensing factor MCM7	D55716	4.22
556	11q13.1	68.6	Breast cancer metastasis-suppressor 1 BRMS1	AF159141	1.93
	11q13.1	68.3	Squamous cell carcinoma Antigen Recognized by T cells (SART1)	AB006198	1.81

NRD = normalized relative density, Mb = million base pairs, AC = accession number, E/N, 'early' tumour (E)/normal (N) expression ratio; all the information on locations and genes was obtained automatically using the Gscope bioinformatics platform (Ripp *et al*, in preparation; Materials and methods)

are involved in translation, signal transduction, epidermal differentiation, cell growth, DNA replication, tumour suppression and regulation of proliferation, respectively.

Quantitative PCR on genomic DNA showed that EIF4G1 and DVL-3 from 3q27 are amplified (three copies or more) in 5/12 and 6/12 patients, respectively, EPHB4 and MCM7 from 7q22 in 5/12 and 3/12, and BRMS1 and SART1 from 11q13 in 5/12 and 5/12 (Figure 3). The two genes from 11q13 were amplified in the same five patients, indicating that the amplification covers at least the 350 000 base pairs that separate these two genes. In contrast, KRT17 and KRT16 from 17q21 were not amplified in the patients analysed, indicating that gene overexpression is not systematically associated with DNA amplification. There are several large-scale studies that compared DNA and RNA levels. About 10% of overexpressed genes were reported to be amplified (Hyman *et al*., 2002), whereas about 50% of amplified genes were overexpressed (Hyman *et al*., 2002; Pollack *et al*., 2002). However, in another study, only about 4% of amplified genes were overexpressed (Platzer *et al*., 2002). Other approaches to identify amplified genes based on their genomic distribution have been reported, which however did not normalize for gene density (Crawley and Furge, 2002; Kano *et al*., 2003). Our approach efficiently highlights genes whose altered expression may result from changes in copy number.

Genes differentially expressed in tumorigenesis

We searched for genes involved in tumorigenesis by comparing gene expression profiles in the four normal and four 'early' tumour samples (Table 1). The 'early' tumours are small in size, well or moderately differentiated, with no lymph node involvement. In order to select differentially expressed genes, we used SAM, a method for identifying genes with statistically significant changes in expression (Tusher *et al*., 2001). In all, 1595 probe sets that are differentially expressed between E and N were selected, of which 136 probe sets (119 unique genes) exhibit a greater than fivefold change (55 genes up in normal and 64 up in tumours; Tables 3 and 4). The duplicate probe sets, for the same gene, gave similar results. These 136 probe sets improve the degree of relatedness of the samples compared to the working set of 3962 genes (compare the dendrograms in Figure 4a and b).

Among these differentially expressed genes, there are new or uncharacterized sequences. We verified the differential expression of six of these novel genes with quantitative PCR (Q-PCR) on 12 additional hypopharyngeal tumours and their matching normal samples (Figure 5). There was a very good correlation between the microarray and Q-PCR results. For all six genes, 8–10/12 patients (67–83%) had >2-fold differences by Q-PCR, whereas 27–34/34 (79–100%) had >2-fold differences on the arrays. These results validate the gene profiling approach, identify potential new markers for diagnosis, and new genes for the study of the early molecular changes leading to tumour formation. The potential functions of these novel genes were investigated by comparing them with neighbouring co-clustering genes (Figure 4c). AB011112 clusters with two mucins, and a blast analysis defines it as 'transmembrane activator and CAML interactor', but it has not been described in the literature. Y09538 (LIM-domain containing) and AF091087 cluster with unknown sequences (U51712, N74607), a serine protease inhibitor (AJ228139) and an extracellular matrix protein (U68186), raising the possibility that they are involved in extracellular matrix remodelling. Furthermore, the 5′ part of AF091087 resembles the *Xenopus laevis* mitotic phosphoprotein 22 (AAM33244). AA418080 clusters with BMP-1, OSF-2, stromelysin 3 and collagen, linking it to the extracellular matrix. M69199, a putative G0/G1 switch protein, clusters with IL-1 beta, GLUT3 and PLOD3, and coincidently IL-1 upregulates GLUT3 in the ovary (Kol *et al*., 1997). Finally, the U61836 cluster does not propose a function, but blast analysis suggests that it encodes a polyamine oxidase. The study of these new genes could lead to new diagnosis tools, therapeutic approaches and mechanistic insights.

Some of the genes revealed by our analysis have already been reported to be differentially expressed in HNSCC. We found 12/55 decreased sequences in common with previous studies, and 35/64 increased (Leethanakul *et al*., 2000; Alevizos *et al*., 2001; Al Moustafa *et al*., 2002; El-Naggar *et al*., 2002; Mendez *et al*., 2002) (Tables 3 and 4). Our results correlated very well with these other studies, except for the one that

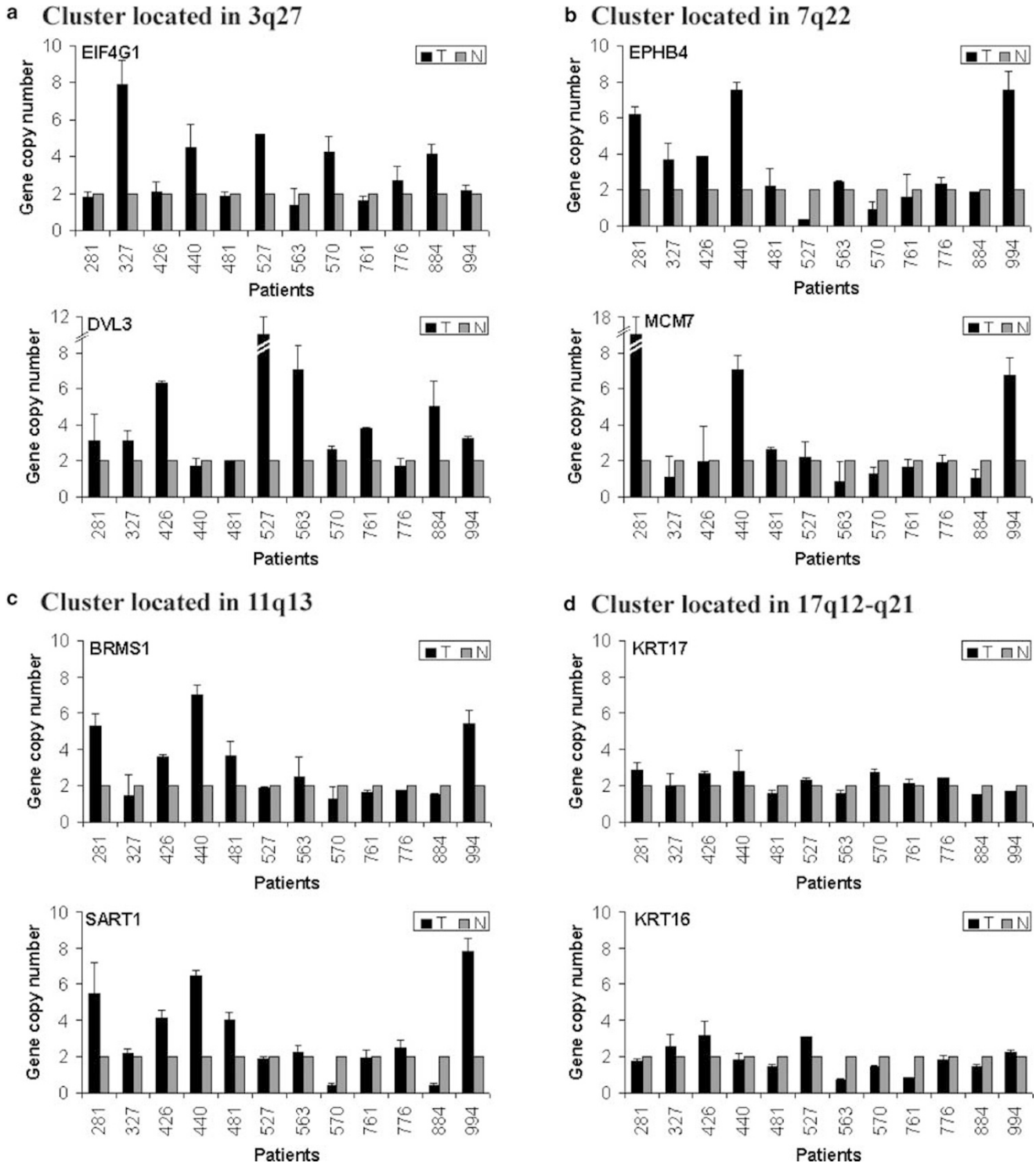

Figure 3 DNA levels measured by quantitative PCR for eight genes with biased chromosomal distributions that are overexpressed in tumours. DNA from 12 patients was analysed for EIF4G1 and DVL3 in 3q27 (**a**), EPHB4 and MCM7 in 7q22 (**b**), BRMS1 and SART1 in 11q13 region (**c**) and KRT17 and KRT16 in 17q12–q21(**d**). DNA levels were normalized to the GAPDH control. The normal sample DNA levels were adjusted to a copy number of two, and the relative copy numbers of the tumour samples compared to their matching normal samples are shown

compared primary cell cultures of tumours and normal matching samples (Al Moustafa *et al*., 2002), suggesting that cell culture influences expression. Our results correlate well with a laser capture microdissection study (Alevizos *et al*., 2001), in that all of the common genes have the same pattern of expression, suggesting that the strong differences we observe come from epithelial components of the samples. The genes in the 129 gene

Table 3 Sequences decreased in tumours

Rank	*N/E*	*Unigene*	*Accession*	*Description*	*References*
Antioncogene and antiproliferative proteins					
24	5.2	Hs.81134	X52015	Interleukin-1 receptor antagonist	(1); (2)
41	12.4	Hs.103505	X99977	ARS gene, component B	
44	88.3	Hs.76422	M22430	RASF-A PLA2	
53	13.0	Hs.75736	J02611	Apolipoprotein D	
39	6.3	Hs.75106	M25915	Complement cytolysis inhibitor (CLI)	(2)
5	11.8	Hs.65424	X64559	Tetranectin	(5)
59	42.2	Hs.183752	AA532495	Clone = IMAGE-996282 (MSMB)	
34	6.8	Hs.2962	AA131149	Clone = IMAGE-587049 (S100P)	
Cancer-associated proteins					
12	71.9	Hs.80395	X76220	MAL gene exon 1	(3)
3	15.9	Hs.79368	Y07909	Progression associated protein (PAP)	(1); (5)
18	6.1	Hs.50964	X16354	Transmembrane carcinoembryonic antigen BGPa	
23	6.2	Hs.220529	M29540	Carcinoembryonic antigen (CEA)	
30	8.0	Hs.20166	AF043498	Prostate stem cell antigen (PSCA)	
22	9.2	Hs.13775	U51712	cDNA/gb = U51712 (SMAP31)	
33	7.5	Hs.73848	M18728	Nonspecific crossreacting antigen	
28	6.3	Hs.44	M57399	Nerve growth factor (HBNF-1)	
46	6.1	Hs.7306	AF056087	Secreted frizzled related protein	
Metabolism					
25	5.4	Hs.75888	U30255	Phosphogluconate dehydrogenase (hPGDH)	
26	5.1	Hs.105435	AF042377	GDP-mannose 4,6 dehydratase	
15	5.3	Hs.2533	U34252	Gamma-aminobutyraldehyde dehydrogenase	(1); (5)
20	12.0	Hs.233441	M57951	Bilirubin UDP-glucuronosyltransferase isozyme 2	
31	8.5	Hs.389	X76342	ADH7	(5)
38	8.4	Hs.575	M74542	Aldehyde dehydrogenase type III (ALDHIII)	
8	16.4	Hs.2022	L10386	Transglutaminase E3 (TGASE3)	(3)
29	7.1	Hs.334841	U29091	Selenium-binding protein (hSBP)	
42	8.3	Hs.5920	AJ238764	*N*-acetylmannosamine kinase	
16	9.1	Hs.81071	U68186	Extracellular matrix protein 1	(5)
Structural proteins					
4	7.6	Hs.80342	X07696	Cytokeratin 15	(1); (2)
40	5.9	Hs.74070	X14640	Keratin 13	(2); (3); (5)
35	16.7	Hs.3235	X07695	Cytokeratin 4 C-terminal region	(1); (3); (5)
6	7.2	Hs.74304	AF001691	195 kDa cornified envelope precursor (PPL)	
56	8.6	Hs.158295	X05451	Myosin light chain 3 (MLC-3f)	
48	36.3	Hs.931	S73840	Type IIA myosin heavy chain	
51	11.8	Hs.78344	AF013570	Smooth muscle myosin heavy chain SM2	
37	23.0	Hs.115166	AF045941	Sciellin (SCEL)	
1	6.7	Hs.3164	X76732	NEFA protein	
58	8.5	Hs.334629	U96094	Sarcolipin (SLN)	
Mucous-related					
11	6.4	Hs.234642	AB001325	Aquaporine 3 (AQP3)	(2)
52	7.0	Hs.221986	U46569	Aquaporin-5 (AQP5)	
7	5.4	Hs.89603	J05582	Pancreatic mucin	
50	12.6	Hs.198267	AJ010901	MUC4 gene, 3 flanking region	
2	5.9	Hs.89603	HG371-HT26388	Mucin 1, epithelial, alt. splice 9	
60	66.2	Hs.82961	AI985964	Clone = IMAGE-2493903 (TFF3)	
55	107.2	Hs.169224	L08044	Intestinal trefoil factor	
9	23.8	Hs.64867	AJ228139	LETKI precursor	
13	7.4	Hs.8272	AI207842	Clone = IMAGE-1953089 (prostaglandin D2)	
Other					
47	14.8	Hs.73931	M16276	MHC class II HLA-DR2-Dw12 mRNA DQw1-beta	
49	6.1	Hs.204040	AF004230	Monocyte/macrophage Ig-related receptor MIR-7	
45	10.0	Hs.99918	M54994	Bile salt-activated lipase (BAL)	
54	5.1	Hs.250760	F27891	Clone = s4000025D03 (COX6A2)	
27	6.0	Hs.75329	AF063002	LIM protein SLIMMER	
57	14.4		AF001548	Clone CIT987SK-A-815A9	
17	8.2	Hs.16622	Y09538	ZNF185 gene	
32	5.1	Hs.64742	AB011112	KIAA0540 protein	
36	5.4	Hs.206501	AF091087	Clone 643 unknown	

The rank represents the order of the genes according to SAM score. Numbers refer to: (1) Alevizos *et al*. (2001); (2) Al Moustafa *et al*. (2002); (3) El-Naggar *et al*. (2002); (5) Mendez *et al*. (2002). Genes with an underlined reference have an inverted expression pattern

Table 4 Sequences increased in tumours

Rank	*E/N*	*Unigene*	*Accession*	*Description*	*References*
Tumour invasion and extracellular matrix morphology					
3	87.6	Hs.83169	M13509	Skin collagenase (MMP-1)	(1)
15	15.5	Hs.83326	X05232	Stromelysin (MMP-3)	(5)
12	9.0	Hs.151738	J05070	Type IV collagenase (MMP-9)	
16	25.2	Hs.2258	X07820	Stromelysin-2 (MMP-10)	(4)
37	10.6	Hs.155324	X57766	Stromelysin-3 (MMP-11)	
6	16.3	Hs.1695	L23808	Human metalloproteinase HME (MMP-12)	(5)
20	6.8	Hs.1274	M22488	Bone morphogenetic protein 1 (BMP-1)	
61	5.8	Hs.153357	AF046889	Lysyl hydroxylase isoform 3 (PLOD3)	
41	36.4	Hs.82085	M14083	Beta-migrating plasminogen activator inhibitor I	(5)
22	9.6	Hs.77274	X02419	uPA	(5)
40	13.1	Hs.179657	U09937	Urokinase-type plasminogen receptor	
Extracellular matrix proteins					
1	10.6	Hs.179573	J03464	Collagen alpha-2 type I	(1); (5)
47	7.3	Hs.82985	Y14690	Procollagen alpha 2(V)	(2); (5)
11	7.1	Hs.75617	X05610	Type IV collagen alpha (2) chain	(5)
5	5.2	Hs.119571	X14420	Pro-alpha-1 type 3 collagen	(5)
9	7.1	Hs.119129	M26576	Alpha-1 collagen type IV	(5)
45	18.3	Hs.82772	J04177	Alpha-1 type XI collagen (COL11A1)	
35	18.9	Hs.172928	Y15915	Collagen (type1 alpha1)/PDGF beta (chimaeric)	
57	5.9	Hs.79914	U21128	Lumican	(5)
17	6.7	Hs.111779	J03040	SPARC/osteonectin	(1); (5)
28	11.6	Hs.204133	X78565	Tenascin-C	(5)
46	36.1	Hs.313	AF052124	Osteopontin	
49	5.4	Hs.83551	U19718	Microfibril-associated glycoprotein (MFAP2)	
26	5.5	Hs.287820	M10905	Cellular fibronectin	(2)
Structural proteins					
2	7.6	Hs.2785	Z19574	Cytokeratin 17	(2); (5)
19	10.0	Hs.115947	AF061812	Keratin 16 (KRT16A)	(2); (5); (6)
36	5.1	Hs.119000	M95178	Non-muscle alpha-actinin	(5)
23	7.8	Hs.75517	U17760	Laminin S B3 chain (LAMB3)	(2); (5)
42	19.7	Hs.54451	Z15008	Laminin (LAMB2)	(2); (6)
66	16.0	Hs.83450	L34155	Laminin-related protein (LamA3)	(2)
4	5.4	Hs.121576	AJ001381	Mutated allele of a myosin class I	
Growth and stress response					
73	39.4	Hs.624	M28130	Interleukin 8 (IL8)	(5)
52	12.7		X04500	Prointerleukin 1 beta	(2); (5)
54	8.4	Hs.126256	M15330	Interleukin 1-beta (IL1B)	(2)
39	5.6	Hs.93913	X04430	Interferon-beta-2 (IL6)	
65	11.3	Hs.789	X54489	Melanoma growth stimulatory activity (MGSA)	(1); (2); (5)
25	5.2	Hs.211600	M59465	Tumour necrosis factor alpha-inducible protein A20	(2)
63	6.0	Hs.265827	U22970	Interferon-inducible peptide (6–16)	(3); (5)
Other					
21	6.2	Hs.136348	D13666	Osteoblast specific factor 2 (OSF-2)	(1); (5)
30	5.0	Hs.90572	U33635	Colon carcinoma kinase-4 (CCK4)	(5)
31	6.5	Hs.418	U09278	Fibroblast activation protein (FAP)	(5)
33	6.7	Hs.135150	AF030428	Lung type-I cell membrane-associated protein (T1A-2)	
51	5.8	Hs.125359	AA704137	Clone = IMAGE-1119984 (similar to THY1)	
59	6.5	Hs.7594	M20681	Glucose transporter-like protein-III (GLUT3)	
74	10.2	Hs.72879	M77481	MAGE-1 antigen	
75	10.8	Hs.36978	U03735	MAGE-3 antigen	
58	5.1	Hs.183648	U22815	LAR-interacting protein 1a gene (PPFIA1)	
53	7.4	Hs.73817	D90144	LD78 alpha precursor (CCL3)	
24	5.5	Hs.118893	D86983	KIAA0230 (peroxidasin)	(1)
29	5.7	Hs.373503	M27826	Endogenous retroviral protease	(2)
34	5.0	Hs.91747	AL096719	cDNA DKFZp566N043 (profilin)	
71	10.3	Hs.75212	X16277	Ornithine decarboxylase ODC	
60	9.9	Hs.79389	D83018	nel-related protein 2	
72	5.2	Hs.226307	AL022318	Phorbolin 3	
76	8.6	Hs.80962	U91618	Proneurotensin/proneuromedin N	
44	7.3	Hs.101850	M11433	Cellular retinol-binding protein	(5)
50	11.4	Hs.373503	AA151971	Clone = IMAGE-588365 (similar to S100P)	
55	8.9	Hs.8786	AB014679	*N*-acetylglucosamine-6-*O*-sulphotransferase (GlcNAc6ST)	

Table 4 *continued*

Rank	*E/N*	*Unigene*	*Accession*	*Description*	*References*
56	6.5	Hs.118633	AJ225089	2–5 oligoadenylate synthetase 59 kDa isoform	
69	9.6	Hs.105924	AF071216	Beta defensin 2 (HBD2)	
67	18.0	Hs.76118	X04741	Protein gene product (PGP) 9.5	(2)
43	6.0	Hs.288467	AA418080	Clone = IMAGE-767773	
68	7.2	Hs.95910	M69199	G0S2 protein	(2)
48	7.4	Hs.92374	U61836	Putative cyclin G1 interacting protein (C20orf16)	

The rank represents the order of the genes according to the SAM score. Numbers refer to: (1) Alevizos *et al*. (2001); (2) Al Moustafa *et al*. (2002); (3) El-Naggar *et al*. (2002); (4) Leethanakul *et al*. (2000); (5) Mendez *et al*. (2002); (6) Villaret *et al*. (2000). Genes with underlined references have an opposite expression pattern

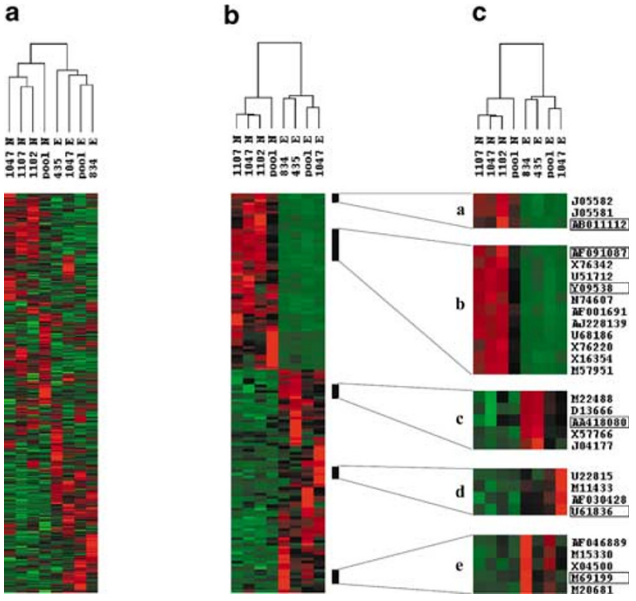

Figure 4 Clustering of 'early' tumours by gene expression profiling. Unsupervised hierarchical clustering of the 'early' tumours and normal samples with the 3962 working probe sets, with significant signals (**a**), and with the 136 probe sets (119 unique genes) selected by SAM that have a fold change greater than 5 (**b**). The neighbouring genes (accession numbers), with the same expression profiles as the six poorly characterized sequences selected for further study (boxed), are highlighted in (**c**). Gene expression profiles are represented with a green to red (lower to higher expression) colour scale

set fall into a number of functional classes. Many of the genes overexpressed in tumour are involved in the formation and remodelling of the extracellular matrix, an important process in tumorigenesis (Kerkela and Saarialho-Kere, 2003). A number of growth and stress response genes are also overexpressed in tumours (Table 4). Genes with decreased expression are involved in inhibition of proliferation, metabolism, structure and mucous-related. Interestingly, IL-1 is increased and the IL-1 receptor antagonist decreased in tumours, a misregulation associated with predisposition to cancer (Arend, 2002) and abnormal cell proliferation in acute myeloblastic leukaemia (Tao *et al*., 2000).

Potential prognostic genes

We selected genes with potential prognostic value by analysing the 30 'late' tumour samples. Since tumour location is an independent predictor of survival, our study included only hypopharyngeal tumours with similar histological characteristics (T1–T4 tumours with lymph node involvement and no metastasis at the time of surgery). Surgical resection was complete for all patients included in this study (evaluated by a pathologist for absence of tumour cells in the resection margins). All the patients received adjuvant treatment by radiotherapy and four also by chemotherapy. The samples were divided into three subtypes according to clinical behaviour 3 years or more after surgery: patient who did not (NM) or did (M) develop metastases, or had local recurrence (LR) within 3 years (see Table 1). The data were normalized by singular value decomposition (SVD) (Nielsen *et al*., 2002), which eliminates potential bias due to the use of two batches of experiments (see Materials and methods). Unsupervised clustering using the normalized data for the 3962 working gene set distinguished NM or M from normal (Figure 6Aa and b), but did not distinguish M from NM (Figure 6Ac). In order to identify markers of metastatic propensity, genes with different expression patterns between M and NM were selected by two methods. Firstly, 80 probe sets were identified with the Student's *t*-test ($P<0.02$). Secondly, 121 probe sets were selected that are differentially expressed at least 1.5-fold in at least half of the samples. The first method selects for homogenous differences, the second for subgroups. Low fold changes could arguably be significant if they arise from relatively few metastatic cells diluted in the tumour mass. Combining the two approaches gives 168 probe sets (164 genes; supplementary table) and 33 common probe sets (32 genes; Tables 5 and 6). Clustering with the 168 probe sets defines two groups of samples and correctly assigns 24/26 (92%) of the M and NM tumours (Figure 6Ba). We noticed that clustering with the 33 common probe sets of the 24 correctly assigned samples defined a set of six M samples that segregate together (on the right in Figure 6Ca). It is perhaps worth noting that these patients have a longer overall survival (Table 1), but the significance of this observation remains to be assessed with additional samples.

The biological significance of the 168 probe sets was assessed by adding other samples to the clustering. When the N group is added, it clusters with the NM group (Figure 6Bb), suggesting that NM tumours could be closer to the normal samples than the M tumours. Similarly, the E samples cluster within NM (Figure 6Bc),

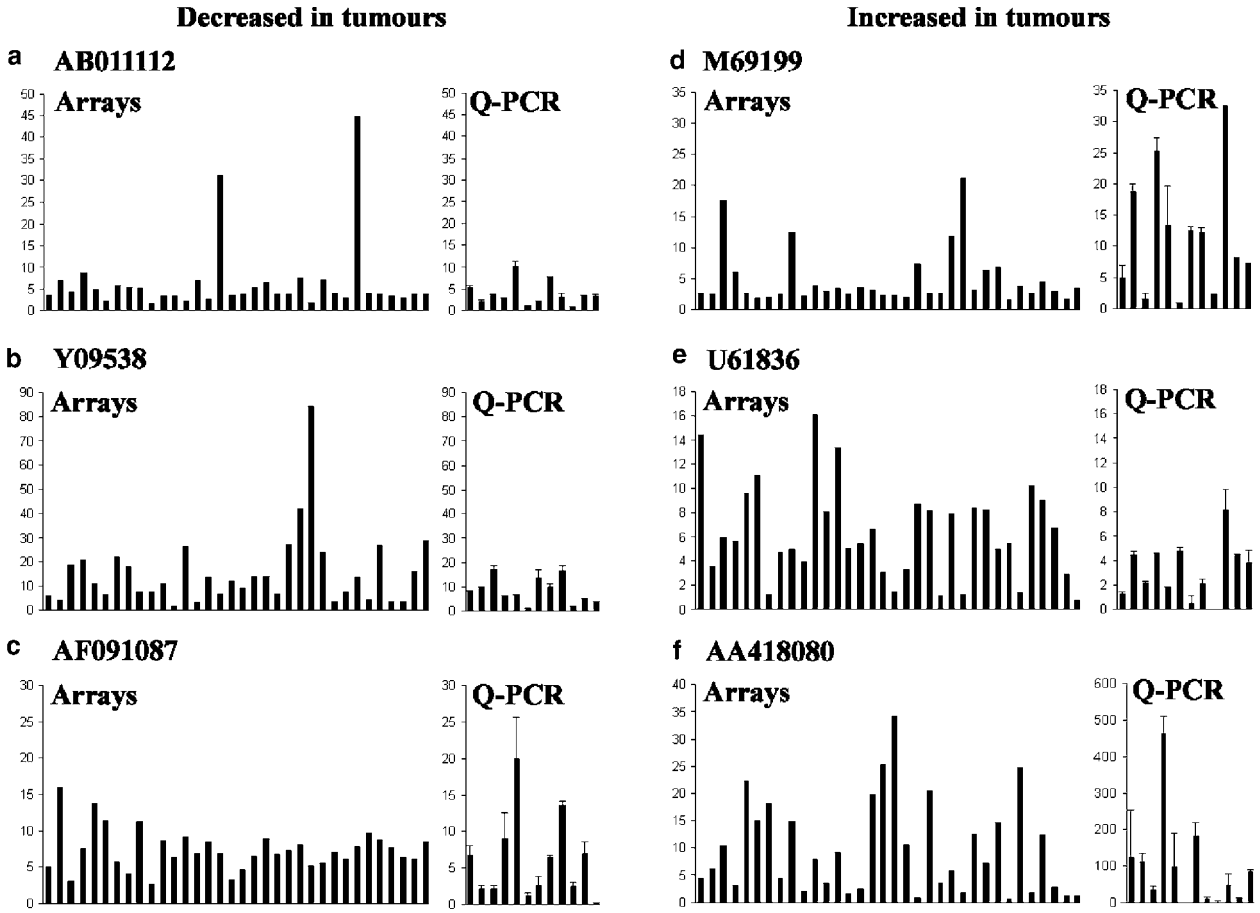

Figure 5 Differential expression in HNSCC of six poorly characterized sequences. The N/T ratio for underexpressed (**a–c**) and the T/N ratio for overexpressed (**d–f**) sequences are shown. The graphs on the left show the fold changes observed on the microarrays (normalized to the average of the normal samples), and the graphs on the right the fold changes observed by Q-PCR on hypopharyngeal samples from additional patients. Expression levels are normalized to ubiquitin, and fold changes are relative to the matching normal samples

suggesting that NM tumours could be closer to the 'early' tumours. With LR, 2/4 cluster with NM and 2/4 in a separate subgroup (Figure 6Bd). Similar results are obtained with the 80, 121 and 33 probe subsets (data not shown). These results raise the possibility that the 164 unique genes (that correspond to the 168 probe sets) may be characteristic of the tumours that will metastasize, but do not seem to be associated with local recurrence.

We assigned the 164 unique genes to general functional categories with a gene ontology resource (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL91). The largest functional group in the poor prognosis M tumours is cell growth and signalling (supplementary table), which includes, for example, cyclin D3 involved in cell cycle control. Genes worthy of note include autotaxin, fusin (CXCR4) and IL8. Autotaxin (ATX) is a tumour motility-stimulating protein (Stracke *et al*., 1992; Umezu-Goto *et al*., 2002), which is also angiogenic (Nam *et al*., 2001). Fusin has been implicated in metastasis in prostate cancer, neuroblastoma, ovarian cancer and melanoma (Geminder *et al*., 2001; Murakami *et al*., 2002; Taichman *et al*., 2002). IL-8 contributes to carcinoma cell invasion in the colon, breast and oral cavity (Youngs *et al*., 1997; Li *et al*., 2001; Watanabe *et al*., 2002). NM tumours, with good prognosis, overexpress genes with opposite functions, such as cell adhesion, intercellular junctions and cell shape (supplementary table). Several of the genes selected from the comparison of M and NM tumours have been identified in related studies. Aldehyde dehydrogenase A (van't Veer *et al*., 2002), hypothetical protein HSPC111 (LaTulippe *et al*., 2002) and integrin alpha3 (MacDonald *et al*., 2001) are overexpressed in good prognosis tumours and in NM. However, only one out of three genes that are in common with another HNSCC study falls into a potentially related category (visinin-like 1 is up in NM, and both ribophorin II and hypothetical protein FLJ10097 are up in M, whereas all three are in the 'better predictor of outcome' group; Belbin *et al*., 2002).

We investigated whether we could predict if a tumour is M or NM, using a leave-one-out strategy. We were unable to predict correctly the subgroup of all the samples (data not shown). We found that the number of samples analysed is important for prediction accuracy. Using five M and five NM samples (microarray batch 1) for training and selecting genes, and using the remaining six NM and 10 M (microarray batch 2) for blind validation, we

correctly predicted 62% of blind samples. In the converse experiment, with more samples for training, the prediction was improved to 80% (supplementary figure). A total of 98 breast tumours were needed to generate a predictive signature (van't Veer *et al*., 2002). An analysis of a larger collection of tumours will be needed to establish whether or not a useful predictive signature can be identified for hypopharyngeal carcinoma.

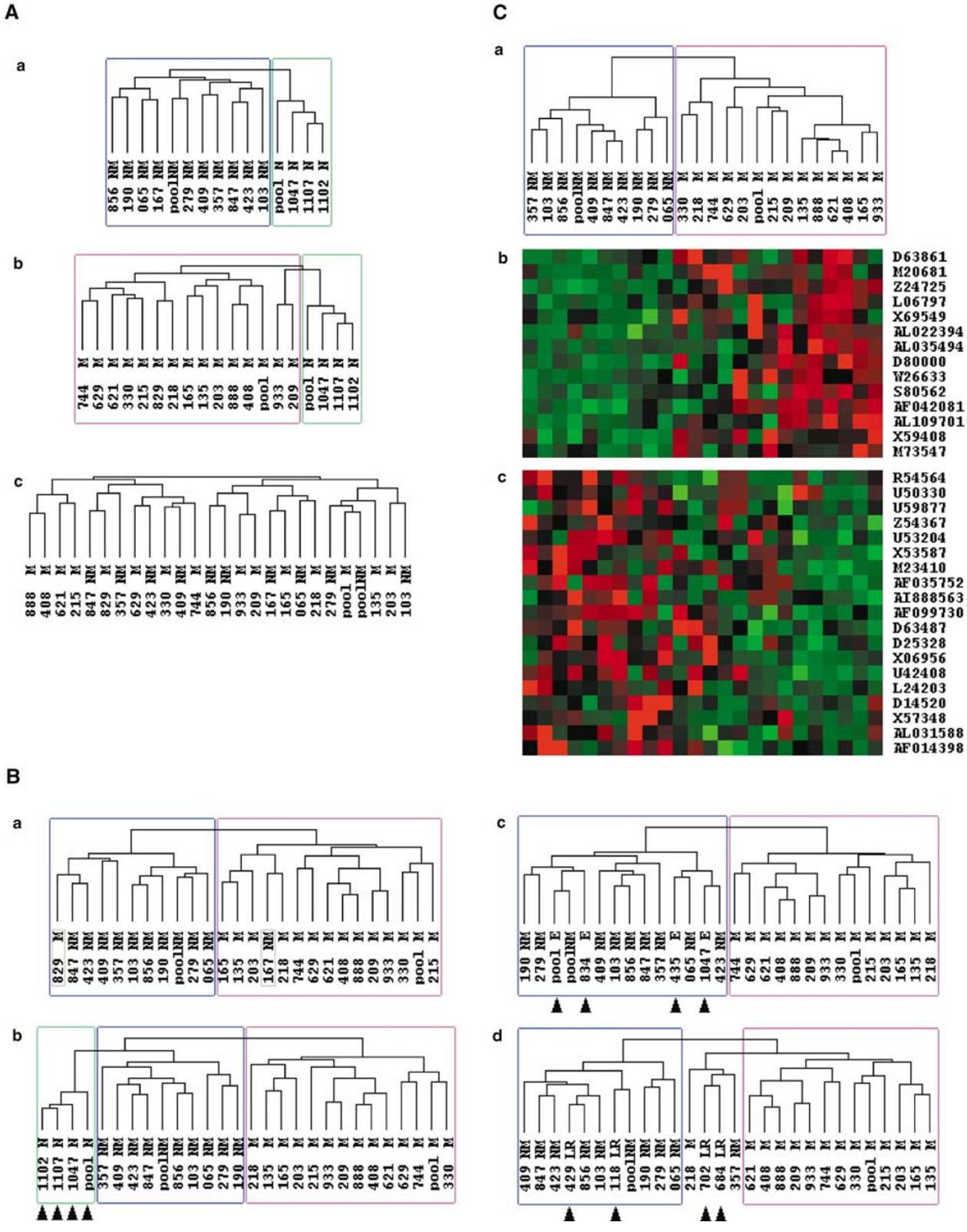

Table 5 Genes increased in M tumours

M/NM ratio	*Unigene*	*Accession*	*Description*
2.46	Hs.184736	AL035494	Hypothetical protein FLJ10097
1.94	Hs.89414	L06797	Chemokine (C-X-C motif), receptor 4 (fusin)
1.73	Hs.211602	D80000	SMC1 (structural maintenance of chromosomes 1. yeast)-like 1
1.65	Hs.83656	X69549	Rho GDP dissociation inhibitor (GDI) beta
1.62	Hs.143482	D63861	Peptidylprolyl isomerase D (cyclophilin D)
1.60	Hs.177556	W26633	Melanoma antigen, family D, 1
1.57	Hs.75260	Z24725	Mitogen inducible 2
1.55	Hs.178112	M73547	DNA segment, single copy probe LNS-CAI/LNS-CAII
1.52	Hs.14368	AF042081	SH3 domain binding glutamic acid-rich protein like
1.51	Hs. 83532	X59408	MCP: membrane cofactor protein
1.51	Hs.75847	AL109701	CREBBP/EP300 inhibitory protein 1
1.50	Hs.194662	S80562	Calponin 3, acidic
1.50	Hs.7594	M20681	Solute carrier family 2 (facilitated glucose transporter), member 3
1.48		AL022394	Sequence from 20q

Table 6 Genes increased in NM tumours

NM/M ratio	*Unigene*	*Accession*	*Description*
2.33	Hs.84728	D14520	Kruppel-like factor 5 (intestinal)
2.25	Hs.85266	X53587	Integrin, beta 4
2.13	Hs.5753	AF014398	Inositol(myo)-1(or 4)-monophosphatase 2
2.03	Hs. 98485	AF099730	GJB3: gap junction protein, beta 3, 31 kDa (connexin 31)
1.88	Hs.75318	X06956	Tubulin, alpha 1 (testis specific)
1.83	Hs.139851	AF035752	Caveolin 2
1.81	Hs.82237	L24203	Tripartite motif-containing 29
1.81	Hs.184510	X57348	Stratifin
1.77	Hs.79706	Z54367	Plectin 1, intermediate filament binding protein, 500 kDa
1.76	Hs.149098	AI888563	Smoothelin
1.67	Hs.2340	M23410	Junction plakoglobin
1.66	Hs.18141	U42408	Ladinin 1
1.63	Hs.223025	U59877	RAB31, member RAS oncogene family
1.63	Hs.122552	AL031588	G-2 and S-phase expressed 1
1.59	Hs.99910	D25328	Phosphofructokinase, platelet
1.51	Hs.112028	R54564	Misshapen/NIK-related kinase
1.50	Hs.1274	U50330	Bone morphogenetic protein 1
1.35	Hs.82563	D63487	KIAA0153 protein

Conclusion

We have found a pattern of gene expression that distinguishes hypopharyngeal tumours from related normal tissue. In addition, our study has revealed new uncharacterized sequences implicated in tumour formation. In a parallel large-scale differential display analysis of hypopharyngeal cancer, 1200 sequences were identified that differ in expression between tumour and normal (data not shown; Lemaire *et al*., 2003). In all, 223 are common between the two studies, indicating that the two approaches are not redundant. Chromosomal aberrations have been shown to occur in HNSCC. We have identified six new genes that are amplified as well as overexpressed in hypopharyngeal carcinoma, using a new analysis method based on genomic clustering of overexpressed genes. Our study has defined a set of 164 genes that classify similar histological and clinical tumours with different outcome on the basis of expression levels. This signature is an 'indicator' but not a 'predictor' of HNSCC outcome, since the number of samples is not sufficient to crossvalidate the data set. Many of these sequences are still uncharacterized and could be major determinants of the capacity to metastasize. They could contribute to the prediction of whether a patient with hypopharyngeal cancer will develop metastases.

Figure 6 Classification of hypopharyngeal samples by gene expression profiling. (**A**) Dendrograms from unsupervised hierarchical clustering of M tumours with normal (**a**), NM with normal (**b**) or NM with M (**c**) using the 3962 working probe sets with significant expression. (**B**) Dendrograms from unsupervised hierarchical clustering using the 168 M versus NM selected probe sets (164 unique genes; see Results) and either the 26 M + NM tumour samples (**a**) or the 24 tumours (excluding those that misclustered in (**a**)) with the normal (**b**), the early (**c**) or the local recurrence (**d**) samples. (**C**) Dendrograms from unsupervised hierarchical clustering using the 33 M versus NM common probe set (32 genes) selection (see Results). The profiles of genes overexpressed in M (**b**) and NM (**c**) are represented on a green to red (lower to higher expression) colour scale. Blue boxes surround NM clusters, pink M clusters and light green N clusters. The two misplaced samples (829 and 167) in (**Ba**) are not included in (**Bb**, **Bc**, **Bd** or **C**). Arrowheads point to the samples that were added to the clustering: N in (**Bb**), E in (**Bc**) and LR in (**Bd**)

Materials and methods

Tissue samples

Primary tumour samples were obtained, with informed consent, from 38 patients undergoing surgery for hypopharyngeal tumours as a primary treatment without previous radiation or chemotherapy. A tumour fragment was taken near the advancing edge of the primary tumour (avoiding its necrotic centre), immediately frozen and stored in liquid nitrogen. The rest of the tumour was fixed in 6% buffered formaldehyde and embedded in paraffin for histopathological analysis. Tumour fragments were composed of at least 70% cancer cells, as assessed on adjacent histological stained sections. The TNM system of the UICC was used for tumour-node-metastasis staging (Sobin and Fleming, 1997). Tumours were classified into two groups: 'early' stage tumours (T1–T2), without lymph node involvement following histological examination of cervical lymph node sections, and 'late' stage tumours (mainly T2–T4), with lymph node involvement. None of the patients presented distant metastasis at the time of surgery. After surgery, all of the patients received adjuvant radiotherapy (RX) and four of them radiotherapy combined with chemotherapy. 'Late' tumours were subdivided into no metastatic (NM), metastatic (M) and local recurrence (LR) 'propensity' according to clinical outcome during a 3-year follow-up (Table 1). Normal samples were collected from the farthest margin of resection (usually uvula) from eight patients (seven were pooled). Normal uvula was also obtained from two nonsmokers, without cancer history, treated by uvulo-palato-pharyngo-plasty for obstructive deep apnoea syndrome. The surface epithelium of the latter samples was macrodissected, to enrich for epithelial tissue. Total RNAs were extracted using the RNAeasy* kit (Qiagen, France) with DNAseI treatment. The quality of the RNA preparation was examined by agarose gel electrophoresis.

Hybridization

Gene expression profiles were analysed using Affymetrix HG-U95A microarrays containing probe sets representing $\sim$12 650 distinct transcription features. cDNA synthesis, cRNA synthesis and labelling, as well as array hybridization, were performed as described in the Affymetrix user's manual (Affymetrix, Santa Clara, CA, USA) using 5 μg of total RNA. The T7 RNA polymerase promoter was incorporated in the first round of double-stranded cDNA synthesis. This cDNA was used for *in vitro* transcription with the ENZO BioArray High Yield IVT kit, to amplify the RNA and incorporate the biotinylated ribonucleotides required for staining after hybridization. The yield and quality of the cRNA was checked by spectrophotometry and capillary electrophoresis using the Agilent 2100 Bioanalyser and RNA 6000 LabChip kit (Agilent technology). A measure of 10 μg of fragmented, biotinylated cRNA was hybridized to the Affymetrix arrays at 45°C for 16 h, as described in the Affymetrix user's manual. Washing and staining of arrays were performed using the GeneChip Fluidics Station 400, and scanning with the Affymetrix GeneArray Scanner.

Preprocessing of microarray data

Acquisition and quantification of array images as well as primary data analysis were performed using the Microarray Suite v5.0 and Data Mining Tool v2.0 Affymetrix software packages. Microsoft Excel was used for further statistical analyses. All arrays were globally scaled to a target value of 800, using the average signal from all gene features and Microarray Suite v5.0. The data set was filtered for low expression values prior to statistical analysis. Genes exhibiting mean values under an arbitrary threshold value of 650 (half of the mean value of all genes and all chips) among all 38 experiments were eliminated, leaving 3962 working probe sets, which can be considered to be expressed genes.

The samples were analysed in two hybridization batches, which are shown in Table 1. For the NM versus M comparison (Results, Potential prognostic genes section), the results from two experiments with different batches of microarrays were normalized by SVD (Alter *et al*., 2000), using the Matrix and Linear Algebra package for Excel v1.1 (available at http://digilander.libero.it/foxes/index.htm). SVD linearly transforms the expression data from the gene and array matrix to a condensed 'eigengene' and 'eigenarray' representative matrix. A unique eigengene was identified that correlated the best with the two array batches. The influence of this eigengene and its corresponding eigenarray was subtracted from all the data (Nielsen *et al*., 2002).

Hierarchical clustering

We applied a hierarchical clustering algorithm to the samples and genes. The algorithm organizes all the data elements into a single tree. We mean-centred genes and arrays and used complete linkage clustering (cluster and tree-view at http://rana.lbl.gov/EisenSoftware.htm; Eisen *et al*., 1998). In the dendrograms, shorter branches connect more similar samples. The red to green colour scale represents the mean-adjusted expression values, where red corresponds to higher and green to lower expression.

Supervised gene selection

To compare tumour (T) and normal (N) patient samples, we used the Significance Analysis of Microarrays add-in to Microsoft Excel (Tusher *et al*., 2001; http://www-stat.stanford.edu/$\sim$tibs/SAM/index.html). The genes were selected with SAM's default parameters and no minimum fold change to englobe the largest significant set of differentially expressed genes. Comparison of E against N results in 1595 significant probe sets, M against N in 1587 and NM against N in 1590, resulting in a total of 2377 unique probe sets. To refine the E versus N comparison, we selected 136 probe sets (119 unique genes) that have an E/N or N/E ratio greater than 5. Genes differentially expressed between NM and M patients were selected by two methods. Firstly, the *t*-test ($P<0.02$) identified 80 probe sets (79 genes). Secondly, 121 probe sets (118 genes) were selected that, for at least half of the samples of the subtype, exhibit a signal ratio of more than 1.5 between the individual samples and the average of the opposite tumour subtype, but never in the opposite subtype. This gave a combined working set of 168 probe sets (164 unique genes).

Bioinformatics

Bioinformatics analysis was performed using the Gscope bioinformatics platform (Ripp *et al*., in preparation) dedicated to managing large collections of protein or nucleotide sequences. A five-step protocol was used to localize the Affymetrix sequences on the human genome and to analyse their chromosomal distribution. Firstly, 12 448 Affymetrix sequences were automatically located on the human genome using BLAST (Altschul *et al*., 1997). Secondly, the local densities (LD) of Affymetrix sequences per million base pairs (Mb) were calculated. Thirdly, these densities were related to

the number of Affymetrix sequences per chromosome. These relative densities were normalized in relation to relative density in known genes in the human genome of reference (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens) as following: NRD = [(NAffySeq per Mb)/(NAffySeq per chromosome)]/[(NgenesNCBI per Mb)/(NgenesNCBI per chromosome)] × 100, where NRD is the normalized relative density, NAffySeq number of Affymetrix sequences, Mb million base pairs and NgenesNCBI = number of genes known at NCBI. Fourthly, our analysis focused on the 2377 differentially expressed (DE) Affymetrix sequences selected by SAM (see above). Redundancy introduced by multiple probes mapping to the same gene was eliminated, leaving 2037 unique sequences. In addition to LD and NRD, we studied the neighbourhood of the DE sequences. For each of the 2037 DE sequences, the six neighbouring Affymetrix sequences on either side were identified. Out of these 12 sequences, the number of DE sequences was counted to obtain a nearest neighbours (NN) score. Fifthly, the functions of the 2037 DE sequences were established by automatically searching protein data banks.

Quantitative PCR on DNA

DNA was isolated from tissue samples by proteinase K digestion, phenol extraction and ethanol precipitation. A measure of 25 ng was used for PCR amplification with Sybr Green and the Roche Lightcycler (Roche Molecular Biochemicals). Oligonucleotide primers are designed to cross intron/exon junctions where possible with primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). Their sequences are: DVL3, 5′-TCCATTTTCTAATGGGCTGG-3′ and 5′-ACAATGGAGATGCCCAAGAA-3′; EIF4G1, 5′-GCTGGATGGATTGGGGAGGAG-3′ and 5′-TGGCCGCAGTGGTGTTTTATT-3′; KRT16, 5′-CCAGAGACCTGGAGGAACAG-3′ and 5′-CGTCTTCACATCCAGCAAGA-3′; KRT17, 5′-CTGGCCCCTACCCCACTTTA-3′ and 5′-GAGATGACCCTTGCCATCCTG-3′; EPHB4, 5′-TCCTGCAAGGAGACCTTCAC-3′ and 5′-CAGAGGCCTCGCAACTACAT-3′; MCM7, 5′-CCAGGCAACATCAACATCTG-3′ and 5′-ATTACAGGCGTGAGCAAACA-3′; BRMS1, 5′-ATTGCCAAGCTGGAGGTG-3′ and 5′-CTTTCTCTGGGCTCCTTCCT-3′; SART1, 5′-TGTCCCGTAGGCCAAGTTAC-3′ and 5′-AGAATCGGCGAGTCAGGAAC-3′; GAPDH, 5′-GGAGCCAAAAGGGTCATCAT-3′ and 5′-GGCATTGCTGCAAAGAAAGAG-3′; RLPO, 5′-AATGTTGCCAGTGTCTGTCTG-3′ and 5′-AAGGTAGAAGGCCACATCACC-3′. DNA levels were measured in two independent experiments, normalized to GAPDH DNA levels, and the matching normal samples were normalized to represent two DNA copies.

Quantitative PCR on cDNA

Total RNAs from matching tumour and normal samples from individual patients were extracted using the RNAeasy kit (Qiagen). First-strand cDNA was synthesized with 1 μg of RNA in 20 μl of a reaction mixture containing 0.3 μg of hexanucleotide primers (Boehringer-Mannheim), 200 U of Superscript II RNase H reverse transcriptase (Invitrogen) and 40 U of RNasin (Promega), dNTP and buffer. The mixture was incubated at 42°C for 50 min and then heated to 95°C for 10 min. A measure of 2 μl of a 1/25 dilution of the reverse transcriptase reaction was used for PCR amplification with Sybr Green and the Lightcycler (Roche Molecular Biochemicals). Oligonucleotide primers were designed to cross exon/exon junctions with primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). Their sequences are: AB011112, 5′-CACGGAAGGAGTATTTGACCA-3′ and 5′-CGATACTGCAGGAGGAGAAAG-3′; Y09538, 5′-CGTGAAGGAGTACGTGAATGC-3′ and 5′-ATGGCAGCAGATACCAAGATG-3′; U61836, 5′-CACTTCTTGAGCAGGGTTTCA-3′ and 5′-ACGCCATTCTTGGAATA- GAGG-3′; AA418080, 5′-GTAGCCATGACATTGGAGCAC-3′ and 5′-GACAACATGGTGCAGAGAGGT-3′; AF091087, 5′-CAGGGAGAAGCATTGATTGAT-3′ and 5′-TTCTCTCCCTTCAACCTGTGA-3′; M69199, 5′-TCAGAGAAACCGCTGACATCT-3′ and 5′-ATGCAAAATGGTG-GTCATTGT-3′; ubiquitin B, 5′-GCTTTGTTGGGTGA- GCTTGT-3′ and 5′-CGAAGATCTGCATTTTGACCT-3′. Gene expression levels were measured in two independent experiments and normalized to ubiquitin expression levels.

Acknowledgements
We thank (a) Diemunsch F. for technical assistance; (b) the IGBMC core facilities; (c) the Ligue Regionale contre le Cancer, the Association pour la Recherche sur le Cancer, the Ministère de la Recherche et de la Technologie, and the Fondation pour la Recherche Médicale for fellowships to A Cromer, A Carles, G Ganguli, F Lemaire and J Young; (c) the Ligue Régionale (Bas-Rhin/Haut-Rhin) contre le Cancer for funding for an RT-QPCR machine; (d) The Ministère de la Recherche for the purchase of the Affymetrix arrays; and (e)ARERS Verre Espoir (no. 138.02), Aventis, the Centre National de la Recherche Scientifique, the Institut National de la Santé et de la Recherche Médicale, the Hôpital Universitaire de Strasbourg, the Association pour la Recherche sur le Cancer, the Fondation pour la Recherche Médicale, the Ligue Nationale Française contre le Cancer (Equipe labellisée), the Ligue Régionale (Haut-Rhin) contre le Cancer, the Ligue Régionale (Bas-Rhin) contre le Cancer, the European Union (FP5 project QLK6-2000-00159) and the Ministère de la Recherche (Décisions 99H0161 and 98C0372) for financial assistance.

References

Al Moustafa AE, Alaoui-Jamali MA, Batist G, Hernandez-Perez M, Serruya C, Alpert L, Black MJ, Sladek R and Foulkes WD. (2002). *Oncogene*, **21,** 2634–2640.

Alevizos I, Mahadevappa M, Zhang X, Ohyama H, Kohno Y, Posner M, Gallagher GT, Varvares M, Cohen D, Kim D, Kent R, Donoff RB, Todd R, Yung CM, Warrington JA and Wong DT. (2001). *Oncogene*, **20,** 6196–6204.

Alter O, Brown PO and Botstein D. (2000). *Proc. Natl. Acad. Sci. USA*, **97,** 10101–10106.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ. (1997). *Nucleic Acids Res.*, **25,** 3389–3402.

Arend WP. (2002). *Cytokine Growth Factor Rev.*, **13,** 323–340.

Belbin TJ, Singh B, Barber I, Socci N, Wenig B, Smith R, Prystowsky MB and Childs G. (2002). *Cancer Res.*, **62,** 1184–1190.

Bernards R and Weinberg RA. (2002). *Nature*, **418,** 823.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D and Sondak V. (2000). *Nature*, **406,** 536–540.

Crawley JJ and Furge KA. (2002). *Genome Biol.*, **3,** 751–758.

Eisen MB, Spellman PT, Brown PO and Botstein D. (1998). *Proc. Natl. Acad. Sci. USA*, **95,** 14863–14868.
El-Naggar AK, Kim HW, Clayman GL, Coombes MM, Le B, Lai S, Zhan F, Luna MA, Hong WK and Lee JJ. (2002). *Oncogene*, **21,** 8206–8219.
Geminder H, Sagi-Assif O, Goldberg L, Meshel T, Rechavi G, Witz IP and Ben-Baruch A. (2001). *J. Immunol.*, **167,** 4747–4757.
Genden EM, Ferlito A, Bradley PJ, Rinaldo A and Scully C. (2003). *Oral. Oncol.*, **39,** 207–212.
Gollin SM. (2001). *Head Neck*, **23,** 238–253.
Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD and Lander ES. (1999). *Science*, **286,** 531–537.
Huang Q, Yu GP, McCormick SA, Mo J, Datta B, Mahimkar M, Lazarus P, Schaffer AA, Desper R and Schantz SP. (2002). *Genes Chromosomes Cancer*, **34,** 224–233.
Hyman E, Kauraniemi P, Hautaniemi S, Wolf M, Mousses S, Rozenblum E, Ringner M, Sauter G, Monni O, Elkahloun A, Kallioniemi OP and Kallioniemi A. (2002). *Cancer Res.*, **62,** 6240–6245.
Kano M, Nishimura K, Ishikawa S, Tsutsumi S, Hirota K, Hirose M and Aburatani H. (2003). *Physiol Genom.*, **13,** 31–46.
Kerkela E and Saarialho-Kere U. (2003). *Exp. Dermatol.*, **12,** 109–125.
Kol S, Ben-Shlomo I, Ruutiainen K, Ando M, Davies-Hill TM, Rohan RM, Simpson IA and Adashi EY. (1997). *J. Clin. Invest.*, **99,** 2274–2283.
LaTulippe E, Satagopan J, Smith A, Scher H, Scardino P, Reuter V and Gerald WL. (2002). *Cancer Res.*, **62,** 4499–4506.
Leethanakul C, Patel V, Gillespie J, Pallente M, Ensley JF, Koontongkaew S, Liotta LA, Emmert-Buck M and Gutkind JS. (2000). *Oncogene*, **19,** 3220–3224.
Lemaire F, Millon R, Young J, Cromer A, Wasylyk C, Schultz I, Muller D, Marchal P, Zhao C, Melle D, Bracco L, Abecassis J and Wasylyk B. (2003). *Br. J. Cancer*, **89,** 1940–1949.
Li A, Varney ML and Singh RK. (2001). *Clin. Cancer Res.*, **7,** 3298–3304.
Liu ET. (2003). *Curr. Opin. Genet. Dev.*, **13,** 97–103.
MacDonald TJ, Brown KM, LaFleur B, Peterson K, Lawlor C, Chen Y, Packer RJ, Cogen P and Stephan DA. (2001). *Nat. Genet.*, **29,** 143–152.
Mendez E, Cheng C, Farwell DG, Ricks S, Agoff SN, Futran ND, Weymuller Jr EA, Maronian NC, Zhao LP and Chen C. (2002). *Cancer*, **95,** 1482–1494.
Monni O, Barlund M, Mousses S, Kononen J, Sauter G, Heiskanen M, Paavola P, Avela K, Chen Y, Bittner ML and Kallioniemi A. (2001). *Proc. Natl. Acad. Sci. USA*, **98,** 5711–5716.
Muller D, Millon R, Velten M, Bronner G, Jung G, Engelmann A, Flesch H, Eber M, Methlin G and Abecassis J. (1997). *Eur. J. Cancer*, **33,** 2203–2210.
Murakami T, Maki W, Cardones AR, Fang H, Tun Kyi A, Nestle FO and Hwang ST. (2002). *Cancer Res.*, **62,** 7328–7334.
Nam SW, Clair T, Kim YS, McMarlin A, Schiffmann E, Liotta LA and Stracke ML. (2001). *Cancer Res.*, **61,** 6938–6944.
Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, O'Connell JX, Zhu S, Fero M, Sherlock G, Pollack JR, Brown PO, Botstein D and van de Rijn M. (2002). *Lancet*, **359,** 1301–1307.
Phillips JL, Hayward SW, Wang Y, Vasselli J, Pavlovich C, Padilla-Nash H, Pezullo JR, Ghadimi BM, Grossfeld GD, Rivera A, Linehan WM, Cunha GR and Ried T. (2001). *Cancer Res.*, **61,** 8143–8149.
Platzer P, Upender MB, Wilson K, Willis J, Lutterbaugh J, Nosrati A, Willson JK, Mack D, Ried T and Markowitz S. (2002). *Cancer Res.*, **62,** 1134–1138.
Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL and Brown PO. (2002). *Proc. Natl. Acad. Sci. USA*, **99,** 12963–12968.
Quon H, Liu FF and Cummings BJ. (2001). *Head Neck*, **23,** 147–159.
Redon R, Muller D, Caulee K, Wanherdrick K, Abecassis J and du Manoir S. (2001). *Cancer Res.*, **61,** 4122–4129.
Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR and Sellers WR. (2002). *Cancer Cell*, **1,** 203–209.
Sobin LH and Fleming ID. (1997). *Cancer*, **80,** 1803–1804.
Stracke ML, Krutzsch HC, Unsworth EJ, Arestad A, Cioce V, Schiffmann E and Liotta LA. (1992). *J. Biol. Chem.*, **267,** 2524–2529.
Taichman RS, Cooper C, Keller ET, Pienta KJ, Taichman NS and McCauley LK. (2002). *Cancer Res.*, **62,** 1832–1837.
Tao M, Li B, Nayini J, Andrews CB, Huang RW, Devemy E, Song S, Venugopal P and Preisler HD. (2000). *Cytokine*, **12,** 699–707.
Tibshirani R, Hastie T, Narasimhan B and Chu G. (2002). *Proc. Natl. Acad. Sci. USA*, **99,** 6567–6572.
Tusher VG, Tibshirani R and Chu G. (2001). *Proc. Natl. Acad. Sci. USA*, **98,** 5116–5121.
Umezu-Goto M, Kishi Y, Taira A, Hama K, Dohmae N, Takio K, Yamori T, Mills GB, Inoue K, Aoki J and Arai H. (2002). *J. Cell Biol.*, **158,** 227–233.
van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R and Friend SH. (2002). *Nature*, **415,** 530–536.
Villaret DB, Wang T, Dillon D, Xu J, Sivam D, Cheever MA and Reed SG. (2000). *Laryngoscope*, **110,** 374–381.
Watanabe H, Iwase M, Ohashi M and Nagumo M. (2002). *Oral Oncol.*, **38,** 670–679.
Youngs SJ, Ali SA, Taub DD and Rees RC. (1997). *Int. J. Cancer*, **71,** 257–266.

Supplementary Information accompanies the paper on Oncogene website (http://www.nature.com/onc).