Oncogene (2005) 24, 6133–6142

www.nature.com/onc

In silico whole-genome scanning of cancer-associated nonsynonymous SNPs and molecular characterization of a dynein light chain tumour variant

Abdel Aouacheria*[,1,3], Vincent Navratil[1,3], Wenyu Wen[2], Ming Jiang[2], Dominique Mouchiroud[1], Christian Gautier[1], Manolo Gouy[1] and Mingjie Zhang[2]

[1]*Laboratoire de Biométric et Biologie Evolutive, CNRS UMR 5558, Université Claude Bernard Lyon 1, F-69622 Villeurbanne Cedex, France;* [2]*Department of Biochemistry, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, PR China*

Last decade has led to the accumulation of large amounts of data on cancer genetics, opening an unprecedented access to the mapping of cancer genes in the human genome. Single-nucleotide polymorphisms (SNPs), the most common form of DNA variation in humans, emerge as an invaluable tool for cancer association studies. These genotypic markers can be used to assay how alleles of candidate genes correlate with the malignant phenotype, and may provide new clues into the genetic modifications that characterize cancer onset. In this cancer-oriented study, we detail an SNP mining strategy based on the analysis of expressed sequence tags among publicly available databases. Our whole-genome approach provides a comprehensive and unbiased description of nonsynonymous SNPs (nsSNPs) in tumoral versus normal tissues. To gain further insights into the possible relationships between genetic variation and altered phenotype, locations of a subset of nsSNPs were mapped onto protein domains known to be critical for protein function. Computational methods were also used to predict the potential impact of these cancer-associated nsSNPs on protein structure and function. We illustrate our approach through the detailed biochemical and structural characterization of a previously unknown cancer-associated mutation (G79C) affecting the 8 kDa dynein light chain (DNCL1).
Oncogene (2005) **24,** 6133–6142. doi:10.1038/sj.onc.1208745; published online 16 May 2005

Keywords: single-nucleotide polymorphism; DNCL1; cancer genomics; cancer association study; dynein light chain; expressed sequence tags

Introduction

A promising application of the large amounts of genetic data currently available for analysis lies in developing a better understanding of complex diseases such as cancer. Current efforts towards this end attempt to understand the molecular signatures of cancer through the identification of genes whose products are deregulated in malignant cells. Since each susceptibility gene does not function in isolation, cancer being a polygenic disorder, systematic searches of genes with small effect are of increasing interest.

Every gene contains some level of polymorphism, with single-nucleotide polymorphisms (SNPs) occurring every ~2000 bp throughout the human genome (Sachidanandam *et al*., 2001). Owing to large-scale SNP discovery, genetic variation in the human genome is now an emerging resource for the study of cancer-related genes (Strausberg *et al*., 2003; Imyanitov *et al*., 2004; Qiu *et al*., 2004). Since cancer is at least in part caused by the accumulation of inherited and/or somatic mutations, knowledge of these molecular changes is of invaluable importance towards dissection of complex biological pathways contributing to cancer phenotype. In this respect, SNPs localized in coding regions of candidate genes and modifying the amino-acid sequence of gene products (i.e. nonsynonymous SNPs (nsSNPs)) are of particular interest, because nsSNPs may affect protein structure and functions (Collins *et al*., 1997; Chakravarti, 1998; Syvanen *et al*., 1999). Base substitutions in the coding sequence can activate proto-oncogenes (for instance, by enabling ligand-independent proliferative signalling) or inactivate tumour suppressor genes that contribute to cancer mainly through loss-of-functions mutations. Alternatively, SNPs associated with cancer could represent interesting markers (e.g. haplotype tags), which could be useful in linkage disequilibrium studies. Expressed sequence tags (ESTs) are partial single-pass 400–600 bp sequences generated from cDNA libraries that provide an opportunity to detect these single-nucleotide differences among sequences derived from a same gene (Buetow *et al*., 2001). Since cDNA libraries are generated from a wide range of cancerous and normal tissues, some variations in genes might be directly related to the cancer phenotype.

In this report, we first present our results concerning the identification of human genetic variants (SNPs) using EST sequences from different libraries. Among those, we have detected genetic variants associated with

*Correspondence: A Aouacheria;
E-mail: aouacher@biomserv.univ-lyon1.fr
[3]These authors contributed equally to this work
Received 14 September 2004; revised 24 March 2005; accepted 11 April 2005; published online 16 May 2005

cancer (i.e. those that are statistically over-represented in ESTs derived from cancerous libraries). This whole-genome scanning strategy had the advantage of being a completely hypothesis-free approach that allowed the *ab initio* detection of cancer-associated SNPs present on EST sequences. Next, we analysed functionally the nsSNPs that were found to be associated with cancer. In particular, we mapped the locations of individual nsSNPs onto functionally relevant protein features. Predictive tools were also used to provide further insights into how these changes could be translated into biochemical events that could lead to the development and maintenance of cancer. Lastly, we provide experimental characterization of a novel polymorphism (G79C) affecting the 8 kDa dynein light chain (DNCL1/DLC8/LC8/DLC1), a multifunctional regulatory protein that plays important roles in fundamental processes such as cell proliferation, apoptosis, cytoskeleton organization and whose deregulation could influence tumour progression.

Results

Preselection of candidates for cancer association studies

One way to identify genes of cancer relevance is through the identification and characterization of genetic variants. In this study, we have used an EST-based *in silico* pipeline to detect cancer-associated coding SNPs (Table 1a). We chose to detail the SNP spectrum limited to polymorphisms introducing changes in the amino-acid sequence (nsSNPs). Our cancer association procedure led to the identification of a total of 267 nsSNPs (on 206 transcripts) that were present at significantly ($P<0.01$) higher allele frequencies in tumour compared to normal tissues (Supplementary file 1). With respect to the delineation of nsSNPs from EST data, we estimated how large the fraction of *bona fide* SNPs was expected to be after filtering using sets of verified SNPs from dbSNPs. We found that a percentage of 25% of the cancer-associated nsSNPs contained in our data set corresponds to validated nsSNPs (last column of Supplementary file 1). Next, three approaches were used for controlling the false discovery rate in our data set: Bonferroni and Benjamini and Hochberg multiple testing corrections, and a resampling procedure. The candidate SNPs positive after these stringent multiple testing corrections (23/267 after Bonferroni and 76/267 after Benjamini and Hochberg, $n=8336$) are highlighted in Supplementary file 1. By the resampling procedure, we found that 98 observed *P*-values fell below the fifth percentile of the empirical *P*-value distribution ($P<0.0011$). Noteworthy, we determined that our procedure identified 15% (31/206) previously studied genes involved in oncogenesis, based on a list of $\sim$2500 genes compiled as described in the Materials and methods section. Since the fraction of such reference genes in our initial data set was 7% (2401/34 091), our data mining protocol lead to a significant enrichment in cancer genes (*P*-value $=1.5\times10^{-5}$; χ^2 test). Thus, despite a non-negligible false-positive rate, our protocol seems intrinsically prone to detect cancer-associated genes. In any case, these results suggest that EST data could be successfully mined to propose a score-ranked preselection of candidate polymorphisms that may be useful for cancer association studies. This list of 267 nsSNPs presumably associated with the cancer phenotype is summarized in Table 2.

Association with tumour development

Our list of 267 cancer-associated variants contains a number of genes possibly involved in the cellular capabilities that might be acquired by cancer cells (Hanahan and Weinberg, 2000), for example, transforming protein RhoA, translationally controlled tumour protein TCTP, chk1 kinase, HLA class I and class II histocompatibility antigens, galectin-3, squamous cell carcinoma antigen 1, prostate-specific antigen, kallikrein-1, CD29, CD99, cathepsin D and metastasis-associated protein MTA1.

Noteworthy, approximately 10% of nsSNPs identified in our collection of 267 cancer-associated variants were represented by ribosomal proteins. Although we cannot formally rule out the possibility that this proportion may be a consequence of the high expression levels of ribosomal genes, it has been previously reported that regulation of ribosome function was often lost in tumour cells (Ruggero and Pandolfi, 2003). In this regard, we found a variant causing a Gly to Val change at residue 165 in ribosomal protein S19, a protein that has been associated to cancer predisposition in Diamond–Balckfan anaemia (Draptchinskaia *et al*., 1999). Another ribosomal protein, S3a, which is able to induce transformation (Naora *et al*., 1998) also displayed genetic variations associated with tumoral context in our analysis. The ability of the MYC oncogene to regulate genes involved in ribosome biogenesis and translation control is now well documented (Coller *et al*., 2000; Boon *et al*., 2001; Menssen and Hermeking, 2002). Interestingly, we identified a first nsSNP in mitotic-arrest-deficient MAD-1 protein, a transcriptional repressor of MYC target genes, and another in the product of one of the MYC target genes, namely the translation initiation factor eIF-2. Translation initiation factor IF-2 was shown to be involved in transformation when its function is upregulated (Rosenwald *et al*., 1993; Rosenwald, 1996), while MYC antagonist MAD-1 acts as a tumour suppressor in a variety of cell lines (Ayer *et al*., 1993; Roussel *et al*., 1996; Cerni *et al*., 2002). Notably, the Arg to His change at position 558 identified in the MAD-1 protein has also been reported as an uncommon polymorphism in a case study of lung cancer (Nomoto *et al*., 1999).

Functional proteomics

In this part, we wished to study the genetic variations that can alter the functions of the cancer-associated candidate proteins. Indeed, although the majority of the SNPs identified in our experiment are expected to

constitute markers associated with cancer (as inferred from the analysis of the nsSNPs/synonymous SNPs ratio, see Table 1b), it is probable that at least a fraction represent functional SNPs. To understand the relationship between genetic and phenotypic variation, it may be useful to assess the putative structural and functional consequences of the respective nonsynonymous mutations in candidate proteins. Therefore, the tumour-associated SNP data were analysed in combination with more global approaches such as functional proteomics. However, the analysis was limited to the subset of candidate nsSNPs, which were positive after the resampling procedure ($n=98$, $P<0.0011$), in order to reduce the false-positive bias ($\sim 2/10$ in this data set according to the Benjamini and Hochberg correction).

Nature of the polymorphisms Nonconservative amino-acid substitutions, including premature stop codons, that is, those that are likely to be more significant for the protein function, were over-represented (91.8%) in our set of cancer-associated nsSNPs compared to conservative substitutions. Although such a percentage was expected since the theoretical fraction of nonconservative substitutions is approximately 86%, the fraction of nonconservative amino-acid variants was found to be higher in the set of cancer-associated nsSNPs than in the set of nsSNPs not associated with cancer (91.8 versus 88.2%). These results prompted us to analyse more precisely the protein features that were affected by cancer-associated nsSNPs.

Protein domains affected by cancer-related nsSNPs We wished to distinguish nsSNPs that lead to amino-acid changes in the functional sites or domains of proteins since these variants are more likely to affect protein function. To determine which protein signatures are affected by genetic variation in our list of tumour-related candidates, we have mapped the identified mutations onto protein domains extracted from the Ensembl database. We used Interpro to define protein domain families present in our set of cancer-associated nsSNPs (Mulder *et al*., 2003). Estimates of the human proteome coverage indicate that approximately three-quarters of all human proteins have an assignment to at least one Interpro domain (available at http://www.ebi.ac.uk/proteome).

A fraction of 65 nsSNPs out of 98 (66.3%) had an assignment to known protein domains, including less informative regions, namely segments of low complexity, signal peptides, coiled coils and transmembrane regions (not shown). Table 3 shows an ordered list of the most frequently targeted Interpro entries from our set of 98 cancer-associated nsSNPs. These domains include a lectin typical signature, modules involved in control of proteolytic activation and motifs of the immunoglobulin superfamily, which have been implicated in tumour progression, metastasis as well as tumour angiogenesis (Sass, 1998; Johnson, 1999; Bassen *et al*., 2000; Gorelik *et al*., 2001; Wall *et al*., 2003; Jin and Varner, 2004; Turk *et al*., 2004).

Thus, these results suggest that at least a fraction of the tumour-associated nsSNPs may probably have some effect on protein function and phenotype. Nevertheless, localization of nsSNPs in relevant protein domains does not directly imply that they dramatically change protein function.

Possible impact of cancer-associated nsSNPs on protein structure and function The selection of nonsynonymous polymorphisms likely to cause the most severe effects on the function of the protein and on the phenotype could be facilitated considering other particular criteria. First, nsSNPs affecting amino-acid residues that are not substituted between closely related homologues are likely to display the highest impact on protein function (Poteete *et al*., 1992). Moreover, some amino-acid replacements are more likely to alter the three-dimensional (3D) structure of the candidate proteins than others. The possible impact of amino-acid allelic variants on protein activity is thus a function of

Table 1 Details on the data mining procedure. (A) SNPs counts in each analytical step. Coding SNPs and nonsynonymous SNPs are referred to as cSNPs and nsSNPs, respectively. (B) Categorization of the ns/synonymous SNP ratio nsSNPs/sSNPs based on allele frequencies. The overall nsSNPs/sSNPs ratio is 1.27 after algorithm filtering versus 0.78 after association test and 0.59 after Benjamini and Hochberg multiple testing correction ($n=14\,867$). Most of this discrepancy arises from a bias in the association procedure that focuses on relatively frequent alleles (corresponding to lower nsSNPs/sSNPs ratio in our data set, see italics)

A

Total nsSNPs/total cSNPs (after algorithm filtering)	17 628/31 456
Total nsSNPs/total cSNPs (for association test)	8 336/14 867
nsSNPs/cSNPs (with $P<0.01$)	267/609

Frequency range	*nsSNPs/sSNPs (Cargill et al., 1999)*	*nsSNPs/sSNPs (% of data set) after filtering*	*nsSNPs/sSNPs (% of data set) after association test*	*nsSNPs/sSNPs (% of data set) after Benjamini and Hochberg correction*
B				
0–5%	1.20	1.58 (55.6%)	1.17 (33.2%)	1.147 (28.4)
5–15%	0.72	1.21 (18.7%)	0.72 (20.4%)	0.45 (23.7)
15–50%	0.61	0.84 (25.7%)	0.59 (*46.4%*)	0.447 (*47.9*)
Total	0.89	1.27 (100%)	0.78 (100%)	0.59 (100%)

Table 2 Summary of cancer-associated nsSNPs

Description	*Ref*	*Position*	*Var*	P-*value*	*Polyphen*	*SIFT*
Ig alpha-1 chain C region	T*	105	N	***5,85E-26***	+	–
40S ribosomal protein S2	T*	247	N	***3,40E-11***	+/–	+
60S ribosomal protein L3 (HIV-1 TAR RNA-binding protein B)	G*	272	S	***3,34E-09***	+	+
Trafficking protein particle complex subunit 4 (Synbindin)	M*	110	T	***3,81E-09***	+/–	–
Xaa-Pro dipeptidase	G*	159	D	***4,65E-08***	–	+
Single-stranded DNA-binding protein MSSP-1	I*	198	T	***6,33E-08***	–	+
Lithostathine 1 alpha precursor	T*	77	N	***1,16E-07***	–	–
Pulmonary surfactant-associated protein C precursor (SP-C)	T*	138	N	***2,20E-07***	+/–	+
Alpha-1-acid glycoprotein 1 precursor (AGP 1)	R*	38	Q	***5,77E-07***	–	–
Zona pellucida sperm-binding protein 3 precursor (ZP3)	S	315	P	***8,58E-07***	–	–
40S ribosomal protein S9	R*	45	M	***2,98E-06***	+/–	+
Transforming acidic coiled-coil-containing protein 2 (AZU-1)	Q*	978	K	**1,09E-05**	–	–
Thioredoxin-dependent peroxide reductase	R	170	Q*	**1,18E-05**	–	–
Fibulin-1 precursor	N*	456	D	**1,24E-05**	+/–	–
60S ribosomal protein L5	Y*	209	C	**1,45E-05**	+/–	–
Beta-2-microglobulin precursor (HDCMA22P)	H*	104	Y	**1,50E-05**	+	+
Trypsin I precursor	C*	48	F	**1,89E-05**	+	+
SH3 domain GRB2-like protein B2 (Endophilin B2)	G*	336	E	**1,90E-05**	–	–
Beta crystallin B2	R	145	W	**3,50E-05**	+	+
Transforming protein RhoA	R	5	Q	**3,69E-05**	–	–
Lithostathine 1 beta precursor	A*	85	G	**5,63E-05**	–	+
Septin 7 (CDC10 protein homologue)	E	318	A*	**7,36E-05**	+/–	+
Carboxypeptidase B precursor (PASP)	D	255	N*	**8,10E-05**	–	–
8 kDa dynein light chain (DNCL1)	**G***	**79**	**C**	**1,12E-04**	**+**	**+**
Placental ribonuclease inhibitor	P*	169	L	**1,13E-04**	+/–	–
adipocyte enhancer-binding protein 1 precursor	K	1133	E	**1,16E-04**	–	+
Prostate-specific antigen precursor (Kallikrein 3)	P*	97	T	**1,40E-04**	–	+
HLA class I histocompatibility antigen, B-8 alpha chain	G*	324	E	**1,46E-04**	+/–	+
lethal giant larvae homologue 1.	S	148	G*	**1,71E-04**	–	–
IGFBP-2-binding protein, IIp45	K	167	E*	**1,71E-04**	–	–
Mitochondrial ribosomal protein L47 isoform a	F*	200	L	**1,90E-04**	+/–	+
Tubulin alpha-1 chain	S*	107	N	**2,07E-04**	+/–	+
40S ribosomal protein S3	F*	152	L	**2,56E-04**	+	+
Polyadenylate-binding protein 3 (PABP 3)	G	165	D	**2,86E-04**	–	–
SET and MYND domain containing 2 (HSKM-B protein)	G	165	E	**3,54E-04**	+/–	–
ADP, ATP carrier protein (adenine nucleotide translocator 2)	F*	130	L	**4,10E-04**	–	+
Glucagon precursor	P*	167	Q	**4,64E-04**	+/–	–
40S ribosomal protein S26	T*	112	K	**4,70E-04**	+/–	+
Cathepsin D precursor	A	58	V	**4,71E-04**	–	–
DNA-binding protein inhibitor ID-3 (HLH-protein HEIR-1)	T	122	A*	**4,86E-04**	–	–
Immunoglobulin gamma Fc receptor III-A precursor	G	183	D	**4,89E-04**	–	+
McKusick–Kaufman/Bardet–Biedl syndromes putative chaperonin	G*	532	V	6,40E-04	+	–
Interferon-induced transmembrane protein 3	E*	21	V	6,70E-04	+/–	+
Kallikrein 1 precursor	E	145	Q	6,96E-04	–	–
Ubiquitin	S	665	F	7,73E-04	–	+
Interleukin 17 receptor C isoform 4 precursor	S*	111	L	8,68E-04	–	+
Kunitz-type protease inhibitor 2 precursor	V	200	L	8,93E-04	–	–
Prostaglandin-H2 D-isomerase precursor	P*	98	T	9,84E-04	+	+
Interferon regulatory factor 7	Q	412	R	1,01E-03	–	–
Williams–Beuren syndrome chromosome region 16 protein	R	30	G*	1,08E-03	+/–	+

A selection of 50 nsSNPs (out of 267) with significantly different allele frequency in normal versus tumoral tissues (exact Fisher's test; $P<0.01$). nsSNPs are ranked by decreasing *P*-value. Positive candidates after the multiple testing corrections are set in italics (Bonferroni), in bold (Benjamini and Hochberg) or underlined (resampling procedure). The allele matching the chimpanzee sequence at the nsSNP position in high-quality alignments was designated as the ancestral allele (asterisk in 'reference' or 'variant' columns). Alleles associated with cancer correspond to variant alleles. Our data indicate that most of the variant alleles were not ancestral (see also Supplementary file 1). For Polyphen predictions of putative SNP impact on protein function, (–) means 'benign', (+/–) possibly damaging and (+) probably damaging. For SIFT predictions, (–) 'tolerated' and (+) means 'affect protein function'. Information concerning the SNP present on DNCL1 appears in bold. For full data access, see Supplementary file 1

both the structural locations of nsSNPs and phylogenetic conservation (Sunyaev *et al*., 2000). A number of algorithms have been implemented to predict the potential of amino-acid substitutions to impact protein structure and function (Sunyaev *et al*., 2000; Chasman and Adams, 2001; Ng and Henikoff, 2002). The basic criteria for these computational methods are sequence homology, physicochemical properties of the substituted residues and structural information. The SIFT (Sorting Intolerant From Tolerant) program focuses more on sequence conservation over evolutionary time and the nature of amino acids in predicting the effect of residue substitutions on function (Ng and Henikoff, 2003). In addition to these parameters, the Polyphen (Polymorph-

ism phenotyping) tool also evaluates the location of the replacement within identified functional domains and 3D structures (Sunyaev *et al*., 2001b; Ramensky *et al*., 2002).

Polyphen analysis of our set of cancer-associated nsSNPs indicates that one-third (33.7%) of nsSNPs are likely to affect function (20 'possibly damaging' and 11 'probably damaging' over 92 predictions). When the SIFT algorithm is applied, the percentage of replacements predicted to impact protein function increases to 36.7% (34 of 94).

Taken together, these results suggest that a subset of tumour-related nonconservative mutations could affect important functional features on proteins. However, in spite of its usefulness in the selection of individual candidate SNPs, the global approach we conducted does not provide a mechanistic explanation for each potentially deleterious variant. So, in a last part of this work, we wanted to exemplify our approach through the description of a presumably deleterious genetic variation affecting a protein that would be of widespread distribution, strong evolutionary conservation and interacting with a variety of proteins.

Example of genetic variation: DNCL1

Rationale for DNCL1(G79C) characterization We focused our attention on the 8 kDa light chain of dynein (DNCL1), one of the most conserved proteins throughout evolution. DNCL1 expression is ubiquitous in various cell types and a growing number of proteins have been reported as DNCL1 interaction partners (Naisbitt *et al*., 2000; Schnorrer *et al*., 2000; Puthalakath *et al*., 2001; Fuhrmann *et al*., 2002). Structural studies showed that DNCL1 forms a homodimer containing two symmetric target-binding channels, which could accommodate several peptides derived from interacting partners (for instance, Bcl-2 family proapoptotic proteins Bim or Bmf or neuronal nitric oxide synthase; nNos) (Jaffrey and Snyder, 1996; Puthalakath *et al*., 1999; Fan *et al*., 2001; Puthalakath and Strasser, 2002). The present study identified a particular nsSNP corresponding to a Gly to Cys substitution at amino-acid position 79, that is, a residue located in a β turn structure at the very end of the target-binding channels (Figure 1a). Analysis of the EST data showed that this variant was exclusively found in cancer tissues (corresponding to eight different libraries prepared from ovary, colon, kidney and larynx tumours). Furthermore, this G79C substitution was predicted to affect protein function according to the SIFT algorithm and was categorized as 'probably damaging' according to the Polyphen program. From an evolutionary perspective, comparison of homologous proteins indicates that four replacements involving amino acids Q, D, E or N (but never C) occur at this position, those replacements being categorized as benign using Polyphen and SIFT prediction tools. Structurally, Gly79 is located in a void at the surface of the molecule and has been depicted as one of the $\sim$12 residues contributing to the groove and exclusively devoted to peptide binding (Liang *et al*., 1999) (Figure 1b). Since this amino acid normally establishes hydrogen bonds with bound peptides, we reasoned that the alteration we identified would likely result in loss of binding affinity and/or altered specificity.

Substitution of conserved Gly79 with Cys perturbs DNCL1's target binding To test experimentally the functional impact of the G79C mutation, we compared the interactions of the wild-type DNCL1 and the G79C mutant with its target proteins including BS69, dynein intermediate chain and proapoptotic protein Bim. We found that the G79C mutant of DNCL1 displayed significantly weaker bindings to all three targets that we have tested (Figure 2a, and data not shown).

We next investigated the conformational impact of the G79C mutation to DNCL1 using NMR spectroscopy. Figure 2b compares the backbone ^{1}H-^{15}N HSQC spectra of the wild-type (red) and the mutant (green) forms of DNCL1. The nicely dispersed HSQC spectrum of the G79C mutant indicated that the protein is well folded. However, one can notice that the mutation induced significant chemical shift changes to a number of amino-acid residues throughout the protein. The mutation-induced chemical shift changes were mapped to the 3D structure of the DNCL1 dimer using the minimal shift perturbation method (Farmer *et al*., 1996) (Figure 2c). We noticed that the mutation of Gly79 in the $\beta3/\beta4$-loop induced particularly large chemical shift changes to the residues in the $\beta1/\beta2$-loop and the N-terminus of $\alpha2$ helix of the protein, both of which are known to be intimately involved in the target binding (Liang *et al*., 1999; Fan *et al*., 2001). The large chemical shift perturbations observed for the $\beta1/\beta2$-loop and the N-terminus of the $\alpha2$ helix are presumably due to the close vicinities between the $\beta3/\beta4$-loop to these two regions in the 3D structure of the protein. Earlier on, we showed that the $\beta1/\beta2$-loop is particularly important for maintaining the dimer conformation and for the target binding of DNCL1 (Wang *et al*., 2003). Alteration of the $\beta1/\beta2$-loop by insertion of two residues (Gly-Ser) completely disrupted DNCL1's target-binding capacity. It is possible that the conformational changes of the $\beta1/\beta2$-loop induced by the G79C mutation may partially account for the weakened target binding of DNCL1.

Discussion

In this work, we have presented a list of SNPs that could constitute relevant genetic markers for cancer diagnosis, and could be useful for cancer association and linkage disequilibrium studies. Since completeness of the currently available SNP databases is obviously not reached (Reich *et al*., 2003), concerted efforts for detection of genetic variations and the use of different algorithms and resources are needed to assemble a more complete list of genes likely to be mutated in cancer. In this regard, our whole-genome approach provides a comprehensive description of nsSNPs in cancer versus

Table 3 Protein domains from Interpro affected by nsSNPs

Rank	*Total* (n = *11092*)	*Tumour* (n = *78*)	*Interpro*	*Name*
1 (**)	215	7	IPR007110	Immunoglobulin like
2 (***)	73	6	IPR001254	Serine protease, trypsin family
3 (**)	102	4	IPR000504	RNA-binding region RNP-1 (RNA recognition motif)
4 (***)	6	3	IPR001064	Beta and gamma crystallin
5 (***)	18	3	IPR003990	Pancreatitis-associated protein
6 (**)	25	3	IPR001304	C-type lectin
7 (**)	15	2	IPR001461	Aspartic protease A1, pepsin
8 (*)	21	2	IPR000566	Lipocalin-related protein and Bos/Can/Equ allergen
9 (*)	1	1	IPR000434	Polycystic kidney disease type 1 protein
10 (*)	1	1	IPR000892	Ribosomal protein S26E

Top 10 Interpro domains affected by cancer-associated nsSNPs (positive after the resampling procedure) are listed. Asterisks correspond to domains significantly enriched in the tumoral set (exact Fisher's test): $*P<0.05$, $**P<0.01$, $***P<0.001$. The *P*-values were computed for each Interpro domain based on the number of nsSNP associated with cancer ('tumour') versus total number of nsSNPs ('total'). The analysis was limited to the subset of 98 candidate nsSNPs positive after the resampling procedure, which corresponds to 65 nsSNPs (total hits = 78, some nsSNPs affecting more than one domain)

normal tissues, unbiased by what is expected to be associated with the tumoral context. Moreover, while most of the works published to date have used previously known SNPs from public databases, our approach was designed to allow the *de novo* identification of uncharacterized SNPs using EST data.

The limitations owing to our EST-based approach have been well explained in two recent publications (Imyanitov *et al*., 2004; Qiu *et al*., 2004). Briefly, large-scale analyses are tributary to the quality and quantity of data present in the used databases. Moreover, several biases could account for the allelic imbalances that could be observed (or not), particular the uncertainty concerning the origin of samples. Besides, there is a need for more information pointing altered function of candidate variants before attempting a correlation with cancer. In addition, a candidate nsSNP may contribute to cancer in certain genetic and environmental contexts, but not in others. To avoid overinterpretation of data, the proteins affected by nonsynonymous amino-acid changes related to tumoral context should not be taken as direct targets for cancer therapy. However, our approach combining mutation detection in EST data and functional proteomic description of the SNPs could highlight protein features and molecular processes to prioritize in further studies.

For instance, our results support the view that deregulation of ribosome biogenesis, translation and protein synthesis may be a hallmark of tumour progression, a notion illustrated by the genetic variations discovered in the MYC-Mad-eIF-2 network. Analysis of protein domain distribution between normal and tumoral state allowed us to identify typical protein signatures likely to be targeted during the transformation process. In this regard, lectin module is shared by proteins that control tumour cell survival, adhesion to extracellular matrix, as well as tumour vascularization and other processes that are crucial for metastatic spread and growth (Sass, 1998; Bassen *et al*., 2000; Gorelik *et al*., 2001). Recent evidence shows that tumour antigens exploit C-type lectins to escape intracellular degradation resulting in abortive immunity. Proteolytic enzymes released in the pericellular microenvironment are other key players in cancer cell–stroma interactions and constitutes biomarkers for malignant tumours, with roles in migration and invasive growth of tumour cells (Wall *et al*., 2003; Turk *et al*., 2004). The serine protease motif was present in our analysis together with the aspartic protease module, suggesting that deregulation of proteolytic enzymes or an abnormal proteinase/antiproteinase balance could exist in cancer cell environment.

We have also tried to characterize the consequences of genetic variations in coding sequences using predictive tools (Polyphen and SIFT). These programs were approximately 80% successful in benchmarking experiences referring to deleterious amino-acid replacements (Chasman and Adams, 2001; Sunyaev *et al*., 2001a; Ng and Henikoff, 2002). These algorithms are roughly concordant in their predictions (Xi *et al*., 2004) and have been used in a growing number of reports (Iida *et al*., 2002; Mohrenweiser *et al*., 2002; Fleming *et al*., 2003). However, most mutations cannot be fully understood in terms of structural, proteomic or phylogenetic features, but require experiments including mutagenesis studies to analyse the effect of changing an amino acid on function, stability, solubility and interactions with other molecules, catalysis, post-translational modifications, allosteric regulation and subcellular localization. This is the case for the polymorphism identified in DNCL1 that we described in more details in the present study. DNCL1 is a highly conserved protein found in species as distant as *Aspergillus*, *Chlamydomonas*, nematode and human. DNCL1 interacts with a number of proteins involved in a variety of functions, including myosin V and dynein, nNOS, the *Drosophila* mRNA localization protein Swallow, the transcriptional regulator ikappaB and the postsynaptic scaffold protein GKAP (Naisbitt *et al*., 2000; Schnorrer *et al*., 2000; Puthalakath *et al*., 2001; Fuhrmann *et al*., 2002). DNCL1 also interacts with Bim and Bmf, two proapoptotic members of the Bcl-2 family of apoptotic regulators (Puthalakath *et al*.,

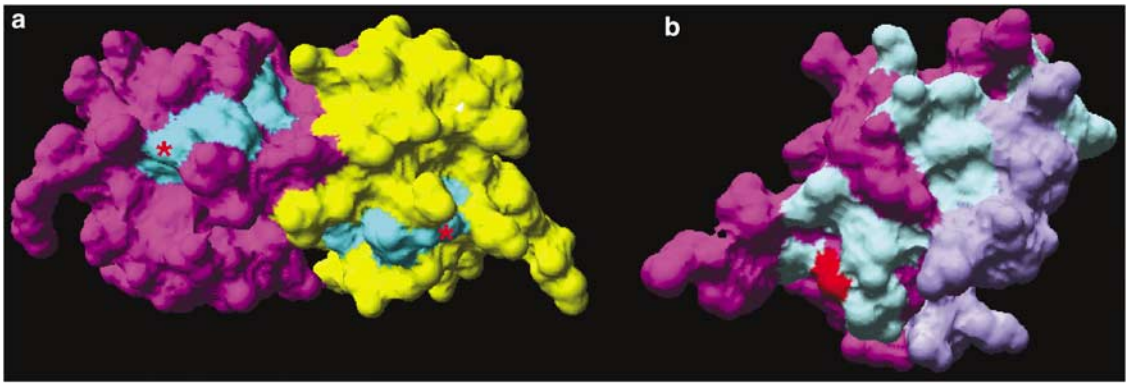

Figure 1 Structure of DNCL1 and location of the G79C amino-acid substitution. (**a**) Molecular surface representation of the DNCL1 homodimer (PDB code: 1F3C). The peptide-binding grooves of the DNCL1 dimer are highlighted (light blue). Residue Gly79 (red asterisks) affected by genetic variation in tumoral context is located in a pocket at the top of the target-binding channel. (**b**) Surface diagram highlighting the location of the Gly79 residue (red) using the structure of the DNCL1 monomer (PDB code: 1PWJ). Colour-coded amino acid surface denote residue functions: dark blue denotes both dimer interface and peptide binding and light blue denotes peptide binding. Substitution G79C (red) occurs in a β turn located at the top of the binding groove. This replacement is nonconservative and predicted to severely affect protein function (see text)

1999, 2001). Based on these observations, DNCL1 is likely to constitute a multifunctional regulatory protein involved in various biological processes. In *Drosophila*, partial mutation of the *DNCL1* gene leads to severe developmental defects including abnormal axonal projections, while total loss-of-function mutations are lethal due to apoptosis (Dick *et al*., 1996; Phillis *et al*., 1996), further emphasizing the role of DNCL1 in cytoskeletal dynamics and cell death/survival decisions. In spite of numerous data suggesting a crucial cellular role for DNCL1, very few reports have focused on possible relationships between DNCL1 alterations and cancer.

The polymorphism we identified affects a Gly residue located in the close vicinity of a β (namely β3) strand that has been shown to contact binding partners (Liang *et al*., 1999; Jin and Varner, 2004). Moreover, it is believed that an nsSNP has an increased probability of functional alteration when it causes a mutation in an important protein surface pocket or void (Stitziel *et al*., 2003). Thus, as many disease-associated nsSNPs, Gly79 is located in a void at the surface of DNCL1. Since the β3/β4-loop is also of structural role for forming the sharp turn, any change in this region was suspected to change directly the folding of the entire molecule and the interaction between DNCL1 and its interactors. We demonstrated here that the G79C mutation effectively induced a clear conformational change to DNCL1 and significantly reduced *in vitro* target-binding capabilities compared to the wild-type version. Potential hindrance of ligand binding introduced by the G79C polymorphism resembles the G75D mutation on retinol-binding protein, which interferes with retinol binding both electrostatically and sterically (Wang and Moult, 2001). These findings therefore validate the computational predictions that were made for this candidate, and provide a biochemical clue as to the possible molecular alteration of DNCL1 in tumours. Recent reports suggest that DNCL1 function is regulated by phosphorylation and that its deregulation could promote tumorigenesis (Vadlamudi *et al*., 2004). Since DNCL1 is a linker protein that brings together proteins with appropriate target sequences, genetic variation on DNCL1 could either disrupt normal interactions or provoke unusual interactions, favouring the emergence of an abnormal interactome that could contribute to tumour development.

Future studies may deal with insertion–deletion patterns or chromosomal translocations that could be common events in tumoral cells. In addition, complementary analyses focusing on SNPs present in 5′- or 3′-UTR of transcripts can be undertaken. Such tumour-specific noncoding SNPs might further be tentatively correlated with expression rates in genes that are differentially expressed in cancer. It is hoped that mined data together with molecular biology will help to elucidate the biological pathways associated with cancer, with the ultimate goal of biomedical applications in drug design, useful biomarkers for diagnostics and improved patient health.

Materials and methods

Protocol for SNP data mining

Data classification Human ESTs from dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) were extracted using the ACNUC sequence retrieval system (Gouy *et al*., 1985). ESTs were classified according to their UNIGENE library accession number (http://ftp.ncbi.nih.gov/repository/UniGene/). The Evoke ontology (http://www.egenetics.com/evoke.html) was used to classify the libraries through a number of criteria such as tissue origin and pathological context including tumoral state. A total of 5135 'tumoral' and 2503 'normal' (i.e. nonpathological) libraries were catalogued. Our approach to EST clustering used the human genome as a reliable guide. ENSEMBL RNAs annotated on human genome assembly (http://www.ensembl.org/release 16.3) were used as a backbone for the clustering of dbEST sequences using BLASTN (Altschul *et al*., 1997) (alignment length $\geqslant$100 bp and

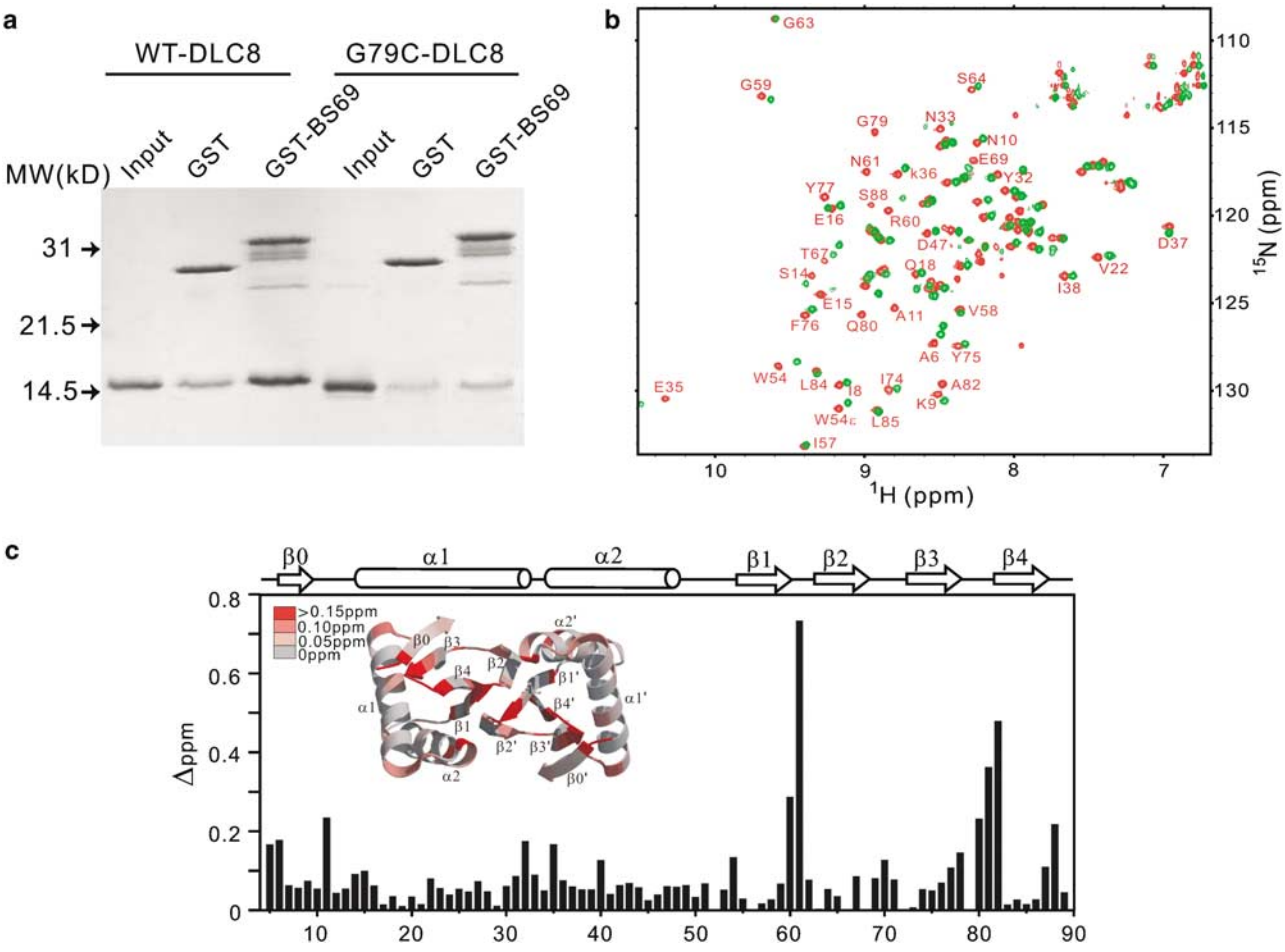

Figure 2 Characterization of DNCL1/G79C amino-acid substitution. (**a**) Coomassie-blue staining of the SDS–PAGE gel showing the binding of GST-BS69 to the wild-type DNCL1 and the G79C mutant, respectively. Purified DNCL1 and its G79C mutant were used as input in each pull-down assay, respectively. Purified GST was used as the negative control of the binding. One can notice that GST showed some background level of binding to DNCL1. (**b**) Superposition plot of the ^{1}H, ^{15}N HSQC spectra of the wild type (red) and the G79C mutant (green) forms of DNCL1. The assignment of the wild-type DNCL1 is labelled with each amino-acid residue name and number. (**c**) Plot of chemical shift changes as a function of the residue number of DNCL1 induced by the G79C mutation. The combined ^{1}H and ^{15}N chemical shift changes are defined as: $\Delta_{ppm} = [(\Delta\delta_{HN})^2 + (\Delta\delta_N \times \alpha_N)^2]^{1/2}$ Where $\Delta\delta_{HN}$ and $\Delta\delta_N$ represent chemical shift differences of amide proton and nitrogen chemical shifts of the wild-type DNCL1 and the G79C mutant, respectively. The scaling factor (α_N) used to normalize the ^{1}H and ^{15}N chemical shifts is 0.17. The secondary structure of DNCL1 is indicated on the top of the plot. The inset shows the amplitude (in pseudocolour scale) of the mutation-induced chemical shift changes of DNCL1 mapped onto the 3D structure of the DNCL1 dimer

similarity $\geqslant$95%). Best hit matches were subsequently selected in order to reduce false assignment to paralogous sequences.

SNP detection We have developed an algorithm to identify exonic SNPs in multiple alignments of various ESTs associated to a particular annotated transcript. This algorithm takes advantage of EST library redundancy and performs four filters to reduce the effect of sequencing inaccuracies at each position. The first filter required that each position within a multiple alignment of ESTs should have an exact match with the reference RNA (windows length = 10 bp around each variant position). The second filter considered a position as informative if the number of libraries in the multiple alignment was superior to a fixed minimum threshold (library number $\geqslant$5). The third filter of the algorithm required the variant to be found at least two times independently, that is, in two different libraries. A last independent filter that required a minimum of two variant ESTs in one of the libraries was subsequently added in order to increase further the stringency of the mining strategy.

SNP information

We associated detection method information (reference and variant EST depth coverage) and nucleotide substitution features (codon conservation analysis: synonymous/nonsynonymous replacement, transition/transversion, position in codon) for the SNPs that have been filtered out. In order to provide information on protein features: amino-acid position of the variant, conservative/nonconservative amino-acid modification and protein family domains (including Interpro) were extracted from Ensembl (http://www.ensembl.org/).

Cancer association

Finally, nsSNPs for which EST counts were available for testing variant over-representation in tumoral context

($n=8336$) were retained for a one-sided Fisher's exact test ($P<0.01$). We privileged the counting of ESTs rather than a count per library because of the frequent lack of precision concerning the origin of the source tissues and the use of pooled samples. To adjust *P*-values produced by Fisher's exact test for multiple testing, three approaches were used: (a) Bonferroni and (b) Benjamini and Hochberg corrections, which are very conservative methods for controlling the false discovery rate, and (c) a resampling procedure. The standard Bonferroni correction multiplies the uncorrected *P*-value by the number of statistical tests. The Benjamini and Hochberg correction consists of ranking all *P*-values in increasing order and adjusting each by multiplying by the total number of tests and dividing by the rank of that *P*-value. The resampling procedure simulates the distribution of the minimum *P*-value that we would expect if there was no association with cancer. To do this, reference and variant margins were fixed at each SNP; Fisher's exact test was then performed for 1000 resampled data sets, and the smallest *P*-value was recorded. This resampling procedure was repeated for $n=8336$ SNPs, from which an empirical distribution of the minimum *P*-value was obtained. From this distribution, we estimated the *P*-value that corresponded to the conventional 5% threshold.

The intensity of the bias of tumoral versus normal allele frequency was calculated according to the following formula:

$$Ib = (a/V - [T-a]/R)$$

where '*a*' is the number of tumoral variants, '*V*' the total number of variants, '*T*' the sum of tumoral counts (variant plus reference) and *R* is the total number of reference alleles (*Ib* being close to 1 in case of strong association).

Predicted protein activity

Polyphen (http://tux.embl-heidelberg.de/ramensky/polyphen.cgi) and SIFT (http://blocks.fhcrc.org/sift/) algorithms were employed to predict the impact of amino-acid substitutions on protein activity. Default parameters values of these programs were used.

Cancer gene list

The selection of cancer genes has been drawn from review of the published literature on the molecular determinants of cancer and from the following links: http://www.cancerindex.org/geneweb/genes_a.htm/; http://bit.fmrp.usp.br/jamborestes/ and http://caroll.vjf.cnrs.fr/cancergene/Aonco_consult.html/. The complete list is available at ftp://pbil.univ-lyon1.fr/pub/GeM/oncogene/cancergenes.txt.

DNCL1(G79C) characterization

Expression and purification of recombinant proteins Preparation of pure recombinant DNCL1 was described in our earlier work (Fan *et al*., 1998). The G79C mutation of DNCL1 was generated using standard PCR-based mutagenesis method. The mutant DNCL1 was prepared using the identical method as for the preparation of the wild-type protein. GST-fused, DNCL1-binding proteins including BS69 and Bim were prepared as described earlier (Lo *et al*., 2001).

Pull-down and peptide competition experiments Direct interactions between DNCL1 and various GST-fused proteins were assayed in PBS buffer, pH 7.4. Equal molar amounts of DNCL1 and one of the GST-fusion proteins ($\sim$2 nM each) were mixed in 750 μl of the assay buffer. The GST-fusion protein/DNCL1 complexes were pelleted by 30 μl of fresh GSH–Sepharose beads. The pellets were washed three times with 1.0 ml of the assay buffer, and subsequently boiled with 30 μl of 2 $\times$ SDS–PAGE sample buffer. The intensity of the DNCL1 band on SDS–PAGE gels was used to judge the strength of the interaction between DNCL1 and various GST-fusion proteins.

NMR experiments ^{1}H-^{15}N HSQC spectra of ^{15}N-labelled DNCL1 and its G79C mutant were acquired on a Varian Inova 750 MHz spectrometer equipped with a *z*-gradient shielded triple resonance probe. All NMR spectra were recorded at 30°C, with a protein concentration of $\sim$0.3 mM dissolved in 100 mM potassium phosphate buffer, pH 7.0.

Acknowledgements
We thank Pr Germain Gillet, Adel Khelifi and Cristina Vieira-Heddi for their valuable comments. VN is supported by a grant from INRA. AA is recipient of a fellowship from the ARC. Research in MZ's laboratory was supported by grants from the Research Grants Council of Hong Kong. MZ is a Croucher Foundation Senior Research Fellow.

References

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

Ayer DE, Kretzner L and Eisenman RN. (1993). *Cell*, **72,** 211–222.

Bassen R, Brichory F, Caulet-Maugendre S, Delaval P and Dazord L. (2000). *Bull. Cancer*, **87,** 703–707.

Boon K, Caron HN, van Asperen R, Valentijn L, Hermus MC, van Sluis P, Roobeek I, Weis I, Voute PA, Schwab M and Versteeg R. (2001). *EMBO J.*, **20,** 1383–1393.

Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, Little DP, Strausberg R, Koester H, Cantor CR and Braun A. (2001). *Proc. Natl. Acad. Sci. USA*, **98,** 581–584.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ and Lander ES. (1999). *Nat. Genet.*, **22,** 231–238.

Cerni C, Skrzypek B, Popov N, Sasgary S, Schmidt G, Larsson LG, Luscher B and Henriksson M. (2002). *Oncogene*, **21,** 447–459.

Chakravarti A. (1998). *Nat. Genet.*, **19,** 216–217.

Chasman D and Adams RM. (2001). *J. Mol. Biol.*, **307,** 683–706.

Coller HA, Grandori C, Tamayo P, Colbert T, Lander ES, Eisenman RN and Golub TR. (2000). *Proc. Natl. Acad. Sci. USA*, **97,** 3260–3265.

Collins FS, Guyer MS and Charkravarti A. (1997). *Science*, **278,** 1580–1581.

Dick T, Ray K, Salz HK and Chia W. (1996). *Mol. Cell Biol.*, **16,** 1966–1977.

Draptchinskaia N, Gustavsson P, Andersson B, Pettersson M, Willig TN, Dianzani I, Ball S, Tchernia G, Klar J, Matsson H, Tentler D, Mohandas N, Carlsson B and Dahl N. (1999). *Nat. Genet.*, **21,** 169–175.

Fan J, Zhang Q, Tochio H, Li M and Zhang M. (2001). *J. Mol. Biol.*, **306,** 97–108.
Fan JS, Zhang Q, Li M, Tochio H, Yamazaki T, Shimizu M and Zhang M. (1998). *J. Biol. Chem.*, **273,** 33472–33481.
Farmer II BT, Constantine KL, Goldfarb V, Friedrichs MS, Wittekind M, Yanchunas Jr J, Robertson JG and Mueller L. (1996). *Nat. Struct. Biol.*, **3,** 995–997.
Fleming MA, Potter JD, Ramirez CJ, Ostrander GK and Ostrander EA. (2003). *Proc. Natl. Acad. Sci. USA*, **100,** 1151–1156.
Fuhrmann JC, Kins S, Rostaing P, El Far O, Kirsch J, Sheng M, Triller A, Betz H and Kneussel M. (2002). *J. Neurosci.*, **22,** 5393–5402.
Gorelik E, Galili U and Raz A. (2001). *Cancer Metastasis Rev.*, **20,** 245–277.
Gouy M, Gautier C, Attimonelli M, Lanave C and di Paola G. (1985). *Comput. Appl. Biosci.*, **1,** 167–172.
Hanahan D and Weinberg RA. (2000). *Cell*, **100,** 57–70.
Iida A, Saito S, Sekine A, Mishima C, Kitamura Y, Kondo K, Harigae S, Osawa S and Nakamura Y. (2002). *J. Hum. Genet.*, **47,** 285–310.
Imyanitov EN, Togo AV and Hanson KP. (2004). *Cancer Lett.*, **204,** 3–14.
Jaffrey SR and Snyder SH. (1996). *Science*, **274,** 774–777.
Jin H and Varner J. (2004). *Br. J. Cancer*, **90,** 561–565.
Johnson JP. (1999). *Cancer Metast. Rev.*, **18,** 345–357.
Liang J, Jaffrey SR, Guo W, Snyder SH and Clardy J. (1999). *Nat. Struct. Biol.*, **6,** 735–740.
Lo KW, Naisbitt S, Fan JS, Sheng M and Zhang M. (2001). *J. Biol. Chem.*, **276,** 14059–14066.
Menssen A and Hermeking H. (2002). *Proc. Natl. Acad. Sci. USA*, **99,** 6274–6279.
Mohrenweiser HW, Xi T, Vazquez-Matias J and Jones IM. (2002). *Cancer Epidemiol Biomarkers Prev.*, **11,** 1054–1064.
Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R and Zdobnov EM. (2003). *Nucleic Acids Res.*, **31,** 315–318.
Naisbitt S, Valtschanoff J, Allison DW, Sala C, Kim E, Craig AM, Weinberg RJ and Sheng M. (2000). *J. Neurosci.*, **20,** 4524–4534.
Naora H, Takai I and Adachi M. (1998). *J. Cell Biol.*, **141,** 741–753.
Ng PC and Henikoff S. (2002). *Genome Res.*, **12,** 436–446.
Ng PC and Henikoff S. (2003). *Nucleic Acids Res.*, **31,** 3812–3814.
Nomoto S, Haruki N, Takahashi T, Masuda A, Koshikawa T, Fujii Y and Osada H. (1999). *Oncogene*, **18,** 7180–7183.
Phillis R, Statton D, Caruccio P and Murphey RK. (1996). *Development*, **122,** 2955–2963.
Poteete AR, Rennell D and Bouvier SE. (1992). *Proteins*, **13,** 38–40.
Puthalakath H, Huang DC, O'Reilly LA, King SM and Strasser A. (1999). *Mol. Cell*, **3,** 287–296.
Puthalakath H and Strasser A. (2002). *Cell Death Differ.*, **9,** 505–512.
Puthalakath H, Villunger A, O'Reilly LA, Beaumont JG, Coultas L, Cheney RE, Huang DC and Strasser A. (2001). *Science*, **293,** 1829–1832.
Qiu P, Wang L, Kostich M, Ding W, Simon JS and Greene JR. (2004). *BMC Cancer*, **4,** 4.
Ramensky V, Bork P and Sunyaev S. (2002). *Nucleic Acids Res.*, **30,** 3894–3900.
Reich DE, Gabriel SB and Altshuler D. (2003). *Nat. Genet.*, **33,** 457–458.
Rosenwald IB. (1996). *Cancer Lett.*, **102,** 113–123.
Rosenwald IB, Rhoads DB, Callanan LD, Isselbacher KJ and Schmidt EV. (1993). *Proc. Natl. Acad. Sci. USA*, **90,** 6175–6178.
Roussel MF, Ashmun RA, Sherr CJ, Eisenman RN and Ayer DE. (1996). *Mol. Cell Biol.*, **16,** 2796–2801.
Ruggero D and Pandolfi PP. (2003). *Nat. Rev. Cancer*, **3,** 179–192.
Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES and Altshuler D. (2001). *Nature*, **409,** 928–933.
Sass PM. (1998). *Cancer Invest.*, **16,** 322–328.
Schnorrer F, Bohmann K and Nusslein-Volhard C. (2000). *Nat. Cell Biol.*, **2,** 185–190.
Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S and Liang J. (2003). *J. Mol. Biol.*, **327,** 1021–1030.
Strausberg RL, Simpson AJ and Wooster R. (2003). *Nat. Rev. Genet.*, **4,** 409–418.
Sunyaev S, Lathe III W and Bork P. (2001a). *Curr. Opin. Struct. Biol.*, **11,** 125–130.
Sunyaev S, Ramensky V and Bork P. (2000). *Trends Genet.*, **16,** 198–200.
Sunyaev S, Ramensky V, Koch I, Lathe III W, Kondrashov AS and Bork P. (2001b). *Hum. Mol. Genet.*, **10,** 591–597.
Syvanen AC, Landegren U, Isaksson A, Gyllensten U and Brookes A. (1999). *Eur. J. Hum. Genet.*, **7,** 98–101.
Turk V, Kos J and Turk B. (2004). *Cancer Cell*, **5,** 409–410.
Vadlamudi RK, Bagheri-Yarmand R, Yang Z, Balasenthil S, Nguyen D, Sahin AA, den Hollander P and Kumar R. (2004). *Cancer Cell*, **5,** 575–585.
Wall SJ, Jiang Y, Muschel RJ and DeClerck YA. (2003). *Cancer Res.*, **63,** 4750–4755.
Wang W, Lo KW, Kan HM, Fan JS and Zhang M. (2003). *J. Biol. Chem.*, **278,** 41491–41499.
Wang Z and Moult J. (2001). *Hum. Mutat.*, **17,** 263–270.
Xi T, Jones IM and Mohrenweiser HW. (2004). *Genomics*, **83,** 970–979.

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