Structure and functions of the rodent identifier retroposon located in the c-Ha-ras oncogene far upstream regulatory region

Asit Kumar Chakraborty

Post Graduate Department of Biotechnology and Biochemistry, Oriental Institute of Science and Technology, Vidyasagar University, Midnapore 721 102, India and Kalisita Biotech Research Pvt Ltd, Kolkata 700 075, India

Retrotransposition is developmentally programmed and such events regulate host genes creating a new trait. A ~120 bp identifier (ID) retroposon insertion was detected in the 161 bp rat c-Ha-ras oncogene far upstream regulatory element. The designed oligonucleotides was probed using gel electrophoresis mobility shift assay to assess transcription factors. The ID retroposon derived Avr1 oligonucleotide binds a nuclear protein complex present in all carcinoma cell lines tested, but not in nuclear extract from normal liver cells. However, Avr2 and Avr3 oligonucleotides (conserved Ha-ras repressor motifs) bind to protein factors found in both normal and carcinoma cells. Also, the Avr3 oligonucleotide binds a small protein in normal cells verses a large complex in carcinoma cells. Each oligonucelotide–protein binding is strong and specific. BLAST sequence analysis has demonstrated that the transposed ID sequence is conserved in too many genes. Avr1 sequence is also a part of the highly expressed HaSV and VL30-retroelements, suggesting that Avr1 factor acts as an activator of transcription. The differential expression of Avr1 trans-activator factor and ubiquitous nature of Avr1 sequence could be used as a marker in cancer diagnostics. This study supports the multiple functions of the smallest ID retroposon shaping the genomic evolution and could be studied further to understand the molecular mechanisms of transposition in cancer, stress and drug resistance.

Keywords: Cancer diagnosis, chimera genes, identifier retroposon, repressor sequence, trans-activator.

RAS superfamily of proteins are localized in the cell membrane and play a key role in the signalling network controlling the balance of proliferation and differentiation of all eukaryotic cells[1–3]. Therefore, the expression of *ras* gene must be tightly regulated to avoid upsetting this balance, which would lead to unregulated growth and neoplasia as seen by transfection of cells with acutely transforming retroviruses like Harvey Sarcoma Virus (HaSV) and also in many cancers[4,5]. Mutated *ras* genes have been proven to be associated with a variety of cancers[6], but such genes under the control of their own GC-rich TATA-less strong promoter failed to transform immortalized primary cells[7]. Recent studies on transgenic animals harbouring an inducible c-Ha-ras transgene have demonstrated that recombination to remove the upstream regulatory elements and continued overexpression of the oncogene are necessary for the genesis and maintenance of solid tumours *in vivo*[8,9]. Therefore, understanding the mechanisms underlying *ras* gene deregulation has implications for diagnosis and therapeutic intervention for cancer.

Rat Ha-*ras* gene contains four coding exons and two non-coding exons[4]. Major research interest on transcriptional regulation of *ras* genes has been focused on the GC-rich TATA-less promoter at the upstream of exon minus 1 and its downstream intron to locate the regulatory sequences (ref. 10 and references therein). Identification of untranslated exon minus 2 located 1.7 kb upstream of the strong TATA-less GC-rich promoter suggests a crucial role of far upstream sequences in the control of rat Ha-*ras* gene transcription[4]. Surprisingly, 0.6 kb BglII fragment with such promoter elements does not support transcription in CAT assay and led to the discovery of 161 bp AvrII–BglII strong repressor sequence[10]. In this article, we show that exon-2 is an inserted mobile identifier non-LTR retroelement. Identifier (ID) retroposon is also transduced into acutely transforming HaSV and MoLV-derived VL30 retroelements, abundantly expressed in rodent cells, tissues and cell lines without apparent harm[11,12]. ID retroposons are non-long terminal repeat (LTR) category of short interspersed nucleotide elements (SINE), abundantly expressed by RNA polymerase III as small polyA$^+$ RNAs (~ 70–120 bases only) in testis and brain of rodents[13]. Retrotransposition is developmentally programmed and by changing the DNA sequence simply by insertion followed by deletion, mutation or recombination, these sequences regulate many host genes creating a new trait.

Although we have characterized the potent Ha-ras repressor sequence in 161 bp far upstream sequence, the 5′ Avr1 sequence and adjoining polyA could not be a repressor sequence. It is ID retroposon origin, a part of Ha-ras

e-mail: chakraakc@gmail.com

Figure 1. Far upstream rat Ha-*ras* gene promoter region with identifier (ID) retroposon sequences and repressor motifs. The ATG start codon of ras was taken as the +1 position[4]. The TATA box was located downstream of the CAAT box but upstream of the Oct1 box (ATGCAAA) as observed in many eukaryotic promoters. Myc (CANNTG), NF1 and Ets1 (CTTGG) factor binding sites are also indicated. Three ID sequences (middle one is a derivative) are underlined. Strong repressor motifs, poly CA_{24} and 83 bp AvrII–AvrIII sequence[21] are shown in bold and located at the 5′ and 3′ ends of the sequence respectively (GenBank accession number: M61016). The forward oligonucleotides AvrF1, AvrF2 and AvrF3 are also indicated.

exon-2 and HaSV as well as VL30 retroelements which are all highly expressed in mammalian cells. This led us to analyse the functionally important transposition of ID retroelements in the DNA and protein databases. We refined the 161 bp repressor sequence[10] into 83 bp Avr2–Avr3 strong repressor sequences (Figure 1) pinpointing Avr1 sequence as transactivator which was inserted at the rat c-Ha-ras far upstream regulatory locus during evolution. We have identified a DNA–protein complex from cancer cells that binds ID retroposon-derived Avr1 oligonucleotide, suggesting a useful marker for cancer diagnosis[14–16].

Materials and methods

Plasmids

The 3.8 kb *Hin*dIII rat c-Ha-*ras* gene region (GenBank accession no. M61016) has been described earlier[4]. Plasmids, pRUF6 containing 1.9 kb rat c-Ha-ras upstream in pUMScat vector and PBBgl + F2 containing 0.6 kb c-Ha-ras repressor sequence in pBLCAT2 vector have been described previously[10].

Cell lines

HepG2 and Huh-7 human hepatocarcinoma cells were grown in MEM medium with non-essential amino acids, Earle's salts and 10% FBS, at 37°C and 5% CO_2 for three days. HT1080 and HeLa human cervical cancer cells were grown in Dulbecco's minimal Eagle medium (DMEM) containing 10% foetal bovine serum, at 37°C and 5% CO_2 for two days. SW620 and SW403 colon carcinoma cells were grown in L-15 medium supplemented with L-glutamine, sodium pyruvate, 10% FBS and 1× antibiotics for 7 days at 37°C and 5% CO_2. The confluent cells were washed in 1× PBS, treated with 0.25% Trypsin-EDTA for 2–5 min at 37°C and washed with fresh medium at 3000 rpm, re-suspended in fresh complete medium and divided in the ratio 1 : 8 for HepG2, HeLa or HT1080 cells and 1 : 4 for SW620 and SW403 cells (≤ 3 ml medium for T-25 flask, ≤ 12 ml for T-75 flask) and incubated the cells at 37°C in a humidified CO_2 incubator for few days up to 70–80% confluent.

Preparation of nuclear extract(s)

Fresh liver (4 g) was minced and homogenized in 20 ml buffer A (10 mM Hepes KOH, pH 7.6, 25 mM KCl, 0.5 mM spermidine, 1 mM EDTA, 1 mM DTT, 0.02% Nonidet P-40, 0.1 mM PMSF, 10% glycerol, 0.25 M sucrose) and passed through a cheese cloth; then an equal volume of 2 M sucrose was added. The mixture was loaded onto a 2 ml cushion of 2 M sucrose and pure nuclei were precipitated at 3700 rpm at 4°C, using a Sorvall ultracentrifuge. The pelleted nuclei were lysed in 1 ml buffer B (10 mM Hepes KOH, pH 7.6, 1.4 M KCl, 10 mM $MgCl_2$, 1 mM DTT, 0.5 mM PMSF, 0.1 mM EGTA) and was microcentrifuged for 30 min at 4°C. The nuclear extract was dialysed in buffer C (20 mM Hepes KOH, pH 7.9, 75 mM NaCl, 0.1 mM EDTA, 0.5 mM DTT, 20% glycerol, 0.5 mM PMSF, 1 mM $MgCl_2$) and further clarified by micro-centrifugation for 10 min at 4°C and stored at –20°C.

Preparation of nuclear extract of HepG2 and HeLa cells

Approximately 70% of confluent cells were scraped (four T-75 flasks), washed in PBS (4000 rpm/10 min/4°C), and 10 ml of buffer A (above) was added; then the cells were homogenized as above. The nuclear pellet was suspended in 1 ml buffer B and kept over ice for 30 min. The extract was then microcentrifuged for 10 min in the cold room and 5–10 µl of the clear supernatant was used in gel electrophoresis mobility shift assays (GEMSA).

Sequences of oligonucleotides

Most of the oligodeoxynucleotides have been described previously[10].

AvrF1	5′-CCT AGG AAG CGC AAG GCC CTG GGT TCG GTC CCC-3′
AvrR1	5′-GAG CTG GGG ACC GAA CCC AGG GCC TTG CGC TTC C-3′
AvrF2	5′-GGT GCC TCT ACA CCT CTG GCA GGA AGC TCA TAT AC-3′
AvrR2	5′-AAC TGT ATA TGA GCT TCC TGC CAG AGG TGT AGA GCC-3′
AvrF3	5′-CTC CTT AAA CAT ACA GAG GTC TGT GTT TGG CCC CAG-3′
AvrR3	5′-GAT CTG GGG CCA AAC ACA GAC CTC TGT ATG TTT AAG-3′
AvrF2.2	5′-AGA TCT ACA CCT CTG GCA GG-3′
AvrF3.3M	5′-ATC CTT AAA CAT ACA TAT GTC TGT GTT TGG-3′

DNA sequence analysis

BLAST and seq 2 sequence analysis was performed using NCBI database. Coding prediction of the cDNA was done by Gene Jockey II software (Sigma, USA). Transcription factor binding sites were determined by PATCH TF software (BioSoft, GmbH, Germany) and GCG software version 9.1 (Wisconsin-Madison, USA).

Preparation of probes

Complementary oligonucloetides (~ 60 pmol each, AvrF1 versus AvrR1; AvrF2 versus AvrR2 and AvrF3 versus AvrR3, etc.) were annealed in 40 µl TE buffer at 60°C followed by slow cooling in the thermo-cooler. Next, 5′ over-hangings of the ds-oligonucleotides at both sides were labelled using Klenow polymerase (10 U) in a 25 µl reaction mixture containing 20 nM dTTP, dGTP, dCTP and 5 pmol of ^{32}P-dATP (3000 Ci/mmol) as described earlier[10]. About 0.5–1 ng of the diluted labelled ds-oligonucleotide (5×10^4–10^5 CPM) was used per assay.

End-labelling of ss-oligonucleotides

Five µl of ss-oligonucleotide and 5 µl kinase cocktail (10× kinase buffer, 12 µl; T4 polynucleotide kinase, 5 µl (50 units); 10 µl α-P^{32}-ATP (10 µCi/µl) in total volume of 120 µl with water) were incubated at 37°C for 30 min and heat inactivated at 65°C for 20 min. The end-labelled oligonucleotide was ethanol precipitated and 10^5 CPM was used in GEMSA assays.

Gel shift assay

The reaction mixture (30 µl) contained 5–10 µl of nuclear extract, 10 ng poly d(G–C) and poly d(A–U), 1 ng labelled oligonucleotide (~ 10^5 CPM) and 15 µl 2× DNA binding buffer (40 mM Tris-HCl, pH 7.9, 200 mM NaCl, 20% glycerol, 4 mM $MgCl_2$, 2 mM DTT, 2 mM EDTA). The reactions were incubated at 25°C for 15 min and were immediately loaded onto a 4% polyacrylamide gel in 1× TBE buffer. The gel was electrophoresed at 160 V in a room at 4°C until the bromophenol blue dye (loaded in a separate lane) ran 4 cm from the end (~ 2 h). The dried gel was auto-radiographed for 12–48 h.

Results

Sequence of the far upstream rat Ha-ras gene promoter and inserted ID retroposons

Previously we have proposed that rat Ha-ras far upstream promoter was located at 1.7 kb (ref. 4) upstream of GC-rich strong promoter. The rat genome NCBI sequence database (accession nos NT_043401 and NT_035113) fully supports our sequence analysis of rat Ha-ras upstream (accession no. M61016) except for some mismatch at the poly A sequence at nucleotide (nt) 573,604 versus 408 (AA dinucleotides addition), nt 573,311 versus 751 (AA dinucleotides deletion), nt 572,850 versus 1273 (AA dinucleotides addition), nt 57,826 versus 1292 $(CA)_3$ deletion (99.9% overall similarity). We had also described 161 bp AvrII–BglII repressor sequences[10]. The repressor sequence is dissected into potential three oligodeoxynucleotides – Avr1, Avr2 and Avr3, excluding poly A sequence to study the possible sequence-specific transcription factors in GEMSA assay. Because the Avr1 oligonucleotide does not retard any protein complex in normal nuclear extract but profoundly does in nuclear extract from many cancer cell lines (see below), we searched the genome database for Avr1 oligonucleotide. BLAST sequence analysis (> 1000 hits with high score) demonstrated that the Avr1 oligonucleotide belonged to the conserved ID sequence family, which is abundantly expressed in rodent neuron and testis. The rat Ha-ras region contained one ID sequence as part of our AvrII–BglII repressor sequence[10], but another ID derivative

Table 1. Identifier retroposon insertions at the different positions of many crucial genes. BLAST analysis has suggested >500 rat gene sequences might contain inserted conserved ID retroposon sequences. Only 30 important insertions at the upstream or downstream of the coding region; or at the 3′ untranscribed region of the mRNA, or at the introns have been described. Homology was determined by Seq-2 BLAST analysis of 75 bp ID sequence (excluding Poly A) demonstrating 0–4 nucleotide point mutations in the respective gene analysed compared with conserved ID sequences of neuron (accession no. U25470) or Ha-ras (accession no. M61016). R, Complement strand inserted in the opposite orientation (reversed). All accession numbers are also not complete gene sequences (upstream with exon 1, cDNAs or part of intron–exons)

Accession no.	Gene	Length (bp)	Position	Location	Homology
M61016	Ha-ras oncoprotein	2897	1185–1283	Upstream	75/0
L18752	Glycogen phosphorylase	3184	1820–1925	Upstream	75/1
J02753	Acyl-COA oxidase	1794	390–489	Upstream	75/3
L19708	N–CH_3 aspertyl receptor	3799	417–535	Upstream	75/2
L41679	Protein kinase-elF2B	13,872	866–975	Upstream	75/2
S73569	tPA activator protein	2393	1074–1196(R)	Upstream	75/4
U30485	Asp tRNA synthetase	9330	1255–1369(R)	Upstream	75/1
X74271	Heat shock protein 70	5918	323–459	Upstream	75/1
X92751	Choline acetyl transferase	4301	1807–1915	Upstream	75/2
Z11902	Steriod 17-α-hydroxylase	3926	2019–2131	Upstream	75/1
X62889	Fatty acid synthase	23,567	1189–1250	Upstream	75/0
U03026	Proenkephalin	1359	173–264	Upstream	75/2
J05214	5′ Nuclease	3152	3065–3152	Downstream	75/4
L10073	Serotonin receptor	2585	1975–2090	Downstream	75/2
L27707	Protein kinase-elF2a	2145	2042–2145	3′-UTR	75/5
M77850	Pyruvoyl pterin synthase	1176	1022–1127	3′-UTR	75/2
X74271	Heat shock protein-70	5918	4540–4653	Downstream	75/1
X14765	GM-Sailo glycoprotein	1827	1752–1827	Downstream	75/1
U93332	Glycerol 3P dehydrogease	2636	1876–1986	Intron 2	75/1
AB003114	GATA-1 transcription factor	1425	1115–1223	Intron 1	75/3
AF040977	Muscle isoactin	4243	2462–2574	Intron 1	75/3
AJ010709	Tyrosine amino transferase	12,460	2448–2572(R)	Intron 3	75/2
K03241	Cytochrome P450d	8556	4377–4358	Intron 3	75/2
U05013	Heme oxygenase-2	14,784	4760–4880(R)	Intron 1	75/4
U30485	Asp–tRNA synthetase	9330	7741–7848(R)	Intron 1	75/3
X62889	Fatty acid synthase	23,567	18,187–18,278	Intron 32	75/0
M17091	Pyruvate kinase-R	1690	823–927(R)	Intron 4	75/2
D38556	Glutathione-S-transferase	4375	2442–2575	Intron 2	75/3
X00975	Myocin light chain-2	3361	2358–2378(R)	Intron 5	75/2
M11709	Pyruvate kinase-L	13,011	2135–2240(R)	3′-UTR	75/3

[CGG TCC CCA GCT CCG $(A)_{11}$ TAT AGC TT] flanked by CACACATT repeats, followed by a third full-length ID sequence having reverse orientation were also detected (Figure 1). There have been no reports of ID insertions at the rat Ha-*ras* gene upstream which has become a part of the ras non-coding exon minus 2 and also a part of the activated *ras* gene upstream in HaSV[4,10]. Further, ID insertion at the upstream of (i) GTP exchange protein gene – 4289 bp upstream, (ii) heat shock protein 70 – 1689 bp upstream, (iii) steroid 17 alpha-hydroxylase gene – 1256 bp upstream, (iv) aspertyl tRNA synthetase gene – 1620 bp upstream and (v) acyl COA oxidase gene – 860 bp upstream (assuming ATG codon as +1) are observed similar to Ha-*ras* gene reported in Table 1.

Gel shift assays to show differential expression of sequence-specific Avr-factors

A high salt clarified nuclear extract (0.6 mg/ml protein) of rat liver selectively retards Avr2 and Avr3, but not Avr1 (Figure 2 *a*). Avr1 oligonucleotide gives a faint non-specific smear at 25°C at low salt (lane 3) but not at high salt (lane 4). But Avr2 oligonucleotide was found to specifically bind to a higher molecular weight protein complex at both low and high salt (lanes 7 and 8), and binding was competed away by excess non-labelled Avr2-oligonucleotide (lane 9). However, Avr3 specifically binds to a lower molecular weight protein complex at 25°C and 0.4 M salt concentration (lanes 10 and 11) and binding activity is inhibited by excess non-labelled ds-oligonucleotide (lane 12).

The very specific binding of the two protein complexes by Avr2 and Avr3 oligonucleotides that are located just downstream of the exon minus 2 of rat c-Ha-ras oncogene, led us to study the status of such factors in different human cancer and primary cells. In human liver carcinoma cell line HepG2, a profound difference in the level of certain transcription factors is observed, compared to the relative absence of these factors in normal rat liver cells. For example, Avr1 oligonucleotide does not bind to any detectable proteins in rat liver nuclear extract, but

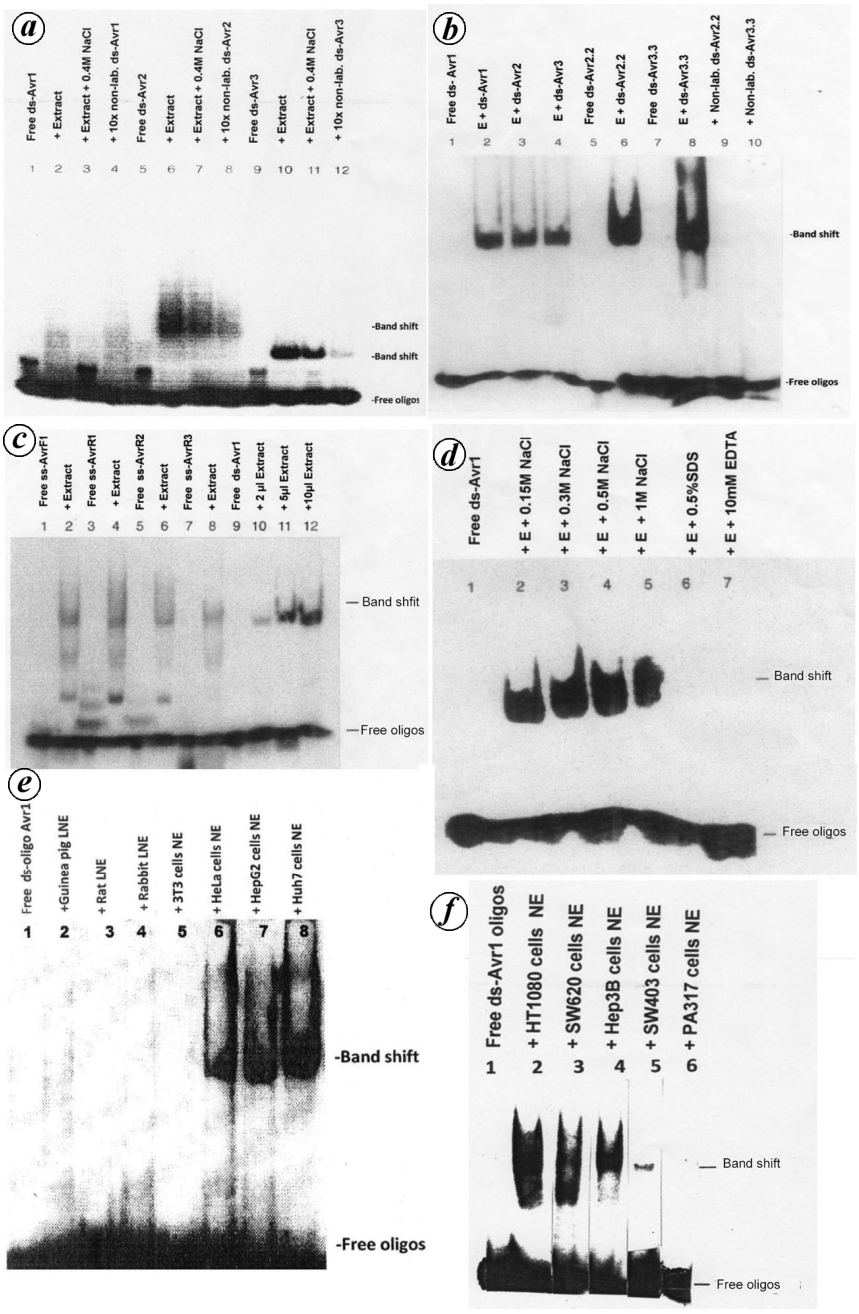

Figure 2. ***a***, Binding of rat liver nuclear factors to Avr-oligonucleotide. Lane 1, Labelled ds-Avr1 oligo; lane 2, + Extract; lane 3, + 400 mM NaCl; lane 4, + 10 times non-labelled ds-Avr1 oligonucleotide; lane 5, Labelled ds-Avr2; lane 6, + Extract; lane 7, + 400 mM NaCl; lane 8, +10 times non-labelled ds-Avr2 oligonucleotide; lane 9, Labelled Avr3 oligonucleotide; lane 10, + Extract; lane 11, + 400 mM NaCl; lane 12, + 10 times non-labelled ds-Avr3. ***b***, Binding of HepG2 nuclear factors to Avr-oligonucleotide. Lane 1, Free ds-Avr1 oligonucleotide; lane 2, + Extract; lane 3, ds-Avr2 oligonucleotide + extract; lane 4, ds-Avr3 oligonucleotide + extract; lane 5, Free ds-Avr2.2 oligonucleotide; Lane 6 + Extract; lane 7, Free ds-Avr3.3 oligonucleotide; lane 8, + Extract; lane 9, Cold 50 times ds-Avr2.2 oligos + extract; lane 10, 50 times ds-Avr3.3 oligonucleotide + extract. ***c***, Binding efficiency of ss and ds oligonucleotide to HepG2 nuclear factors. Lane 1, Free AvrF1; lane 2, + 10 µl extract; lane 3, Free AvrR1; lane 4, + 10 µl extract; lane 5, Free AvrR2 ; lane 6, +10 µl extract; lane 7, Free AvrR3; lane 8, + 10 µl extract; lane 9, Free ds-Avr1; lane 10, + 2 µl extract; lane 11, + 5 µl extract; lane 12, + 10 µl extract. ***d***, Effect of high salt on binding of HepG2 nuclear factors to ds-Avr1 oligonucleotide. Lane 1, Free ds-Avr1 oligo; lanes 2–5, + Extract with different concentrations of NaCl (150, 300, 500 and 1000 mM respectively); lane 6, Extract + 0.5% SDS; lane 7, Extract + 10 mM EDTA. ***e***, Binding efficiency of ds-Avr1 oligonucleotide to nuclear factors of normal versus cancer cells. Lane 1, Free ds-Avr1; lane 2, + Guinea pig liver nuclear extract; lane 3, + Rat liver nuclear extract; lane 4, + Rabbit liver nuclear extract; Lane 5, + 3T3 cells nuclear extract; lane 6, HeLa cells nuclear extract; lane 7, HepG2 cells nuclear extract; lane 8, Huh-7 cells nuclear extract. ***f***, Binding of ds-Avr1 oligonucleotide to nuclear factors of many human cancer cells. Lane 1, Free ds-Avr1 oligonucleotide; lane 2, + HT1080 cells nuclear extract; lane 3, + SW620 cells nuclear extract; lane 4, + Hep3B cells nuclear extract; lane 5, + SW403 cells nuclear extract; lane 6, + PA317 murine retroviral packaging cells nuclear extract.

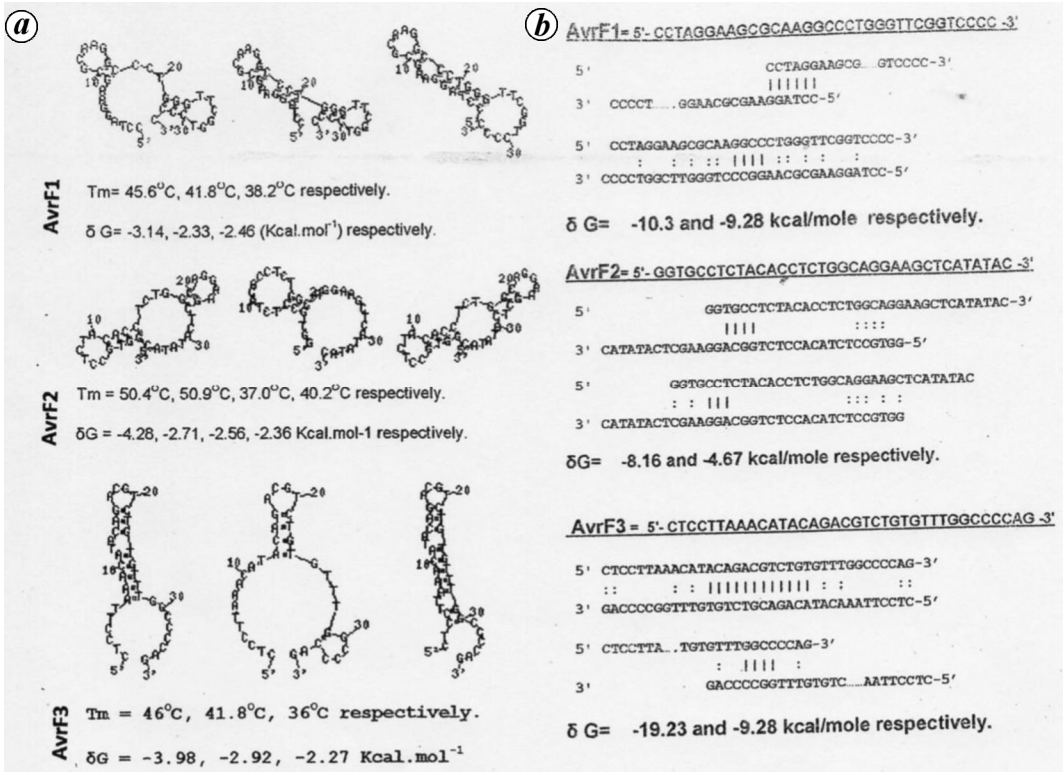

Figure 3. Analysis of Avr-oligonucleotides for dimer and hairpin structures using OligoAnalyzer 3.1 software. The oligonucleotides (AvrF1, AvrF2 and AvrF3) for GEMSA are forcefully made without software analysis from 161 bp Ha-ras oncogene far upstream repressor elements, excluding poly A sequence (see Figure 1). Usually ds-DNA oligonucleotides are used for GEMSA assay for transcription factors, but many ss-DNA binding factors have a pivotal role in transcription and mRNA stability or transport. ***a***, Hairpin structures formed by AvrF1, AvrF2 and AvrF3 oligonucleotides at 50 mM salt at standard temperature and in the presence of 10 mM $MgCl_2$. ***b***, Potential dimer structures could be formed during GEMSA assay using ss oligonucleotides, particularly as for AvrF3 ($\delta G = -19.23$ kcal/mol). This information is pivotal for kinasing of ss-oligos or to make ds-oligos and its labelling with DNA polymerase as well as to pinpoint minimal sequence required for DNA–protein binding to discover new transcription factors.

efficiently and specifically binds to a protein complex present in HepG2 cells nuclear extract (Figure 2 *b*, lane 2 versus Figure 2 *a*, lane 2). Also, the Avr3 oligonucleotide binds a high molecular weight complex (Figure 2 *b*, lane 4) compared to a smaller one using normal rat liver extract (Figure 2 *a*, lane 10). Binding of transcription factor(s) is specific, and is competed away by 50 times non-labelled oligonucleotide (Figure 2 *b*, lanes 9 and 10).

We have further studied whether Avr1-binding factor(s) binds both ss and ds oligonucleotides. We observed (Figure 2 *c*) that both forward and reverse ss-Avr1 oligonucleotide (lane 2 forward) and (lane 4 reverse) have gel retarded using HepG2 nuclear extract with more or less the same specificity and comparable to the ds-Avr1 oligo (lane 12). Also, there is gradual increase of binding on increasing nuclear factor concentrations, 2 µl (lane 10), 5 µl (lane 11) and 10 µl (lane 12). The ss-oligonucleotides, however, have retarded some non-specific bands (Figure 2 *c*, lanes 2 and 4) than ds-oligonucleotide (lane 12) due to lariat or dimer formation in the assay condition. Analysis of hairpin or dimer formation is presented in Figure 3. AvrF1, AvrF2 and AvrF3 oligonucleotides (33–35 nt) all form hairpin structures with melting temperatures 40–51°C and δG at 2–4 kcal/mol. Also, AvrF1 and AvrF2 form stable dimer with δG minus 8–10 kcal/mol, but a more stable dimer is formed by ss-AvrF3 ($\delta G = -19.23$). Thus analysis for dimer or hairpin structures is absolutely necessary for forced oligonucleotides used in GEMSA techniques.

The DNA–protein complex also appears very stable at high salt (Figure 2 *d*, lane 4) and is only partially inhibited at 1 M NaCl (lane 5). The type of clear gel retardation at higher salt is a high quality result in GEMSA techniques. Also, 0.5% of SDS or 10 mM EDTA completely inhibits the reaction in the presence of 150 mM NaCl. This interaction suggests a specific and strong DNA–protein binding useful for further characterization of the DNA-binding protein.

Also, the abundant presence of Avr1 factors in HepG2 cells prompted us to study whether such activity is present in other carcinoma cells compared to normal liver cells. Our results have indicated (Figure 2 *e*) that liver nuclear extract from guinea pig (lane 2), rat (lane 3) and rabbit (lane 4) has rarely recognized Avr1 oligonucleotide

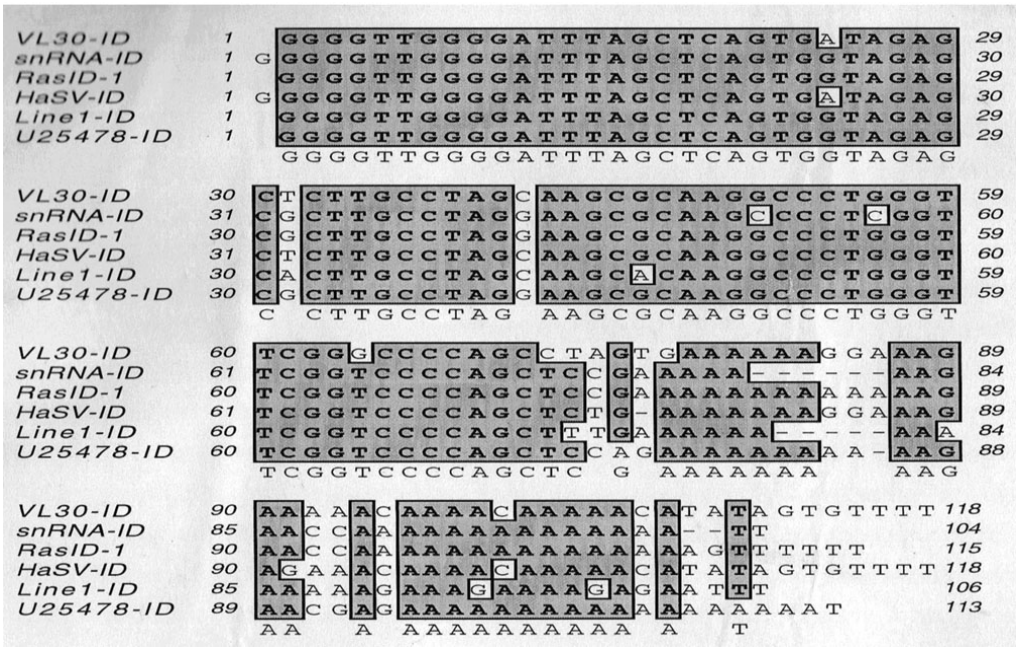

Figure 4. Sequence similarity among the different ID retroposons. VL30-ID, snRNA-ID, Ras-ID, HaSV-ID, LINE-ID and Brain-ID sequences have been derived from the rat VL30 transcript cDNA (accession no. M91235, nucleotides 561–678), small nuclear RNA gene (accession no. K02430, nucleotides 1030–1133, reverse oriented), Ha-*ras* gene upstream (accession no. M61016, –1730 to –1837 assuming 'ATG' initiation codon as +1), HaSV transcript cDNA (accession no. X00740, nucleotides 66–183), LINE1 element gene (accession no. M60824, nucleotides 462–567, reverse oriented) and brain ID transcript cDNA (accession no. U25478, nucleotides 1–113) respectively. The possible splice acceptor/donor (SA/SD) sites are located at nucleotide positions 21 (AGTGT) and 111 (AGTGT) of the Ha-ras-ID. Few human cDNA containing conserved ID sequences (accession nos BQ188034, BQ186207, BM932101, AK098165) are identified but were not compared here.

in GEMSA, but similar extracts from cancer cell lines, HeLa (lane 6), HepG2 (lane 7) and Huh7 (lane 8) have significantly retarded it. Interestingly, mouse 3T3 cells nuclear extract also does not recognize the Avr1 oligonucleotide, suggesting that 3T3 cells are an immortalized cell line (it does not follow the contact inhibition like primary cells) and not cancer cells. This prompted us to test few more cancer cell lines available in our laboratory (because we are unable to get sufficient tumour tissue from patients for this study). The result is presented in Figure 2*f*. It can be seen that ds-Avr1 retards the protein complex using nuclear extract of HT1080 (lane 2), SW620 (lane 3), Hep3B (lane 4) and SW403 (lane 5) cells respectively. SW403 colon carcinoma cells grow slowly and give a poor but sharp band. Interestingly, like 3T3 murine cells (Figure 2*e*, lane 5), PA317 retroviral packaging cells nuclear extract does not produce a good complex in our standard GEMSA assay (Figure 2*f*, lane 6). However, after long exposure of the film (> 3 days), a faint DNA–protein complex was detected in PA317 cells as well as to a lower extent in 3T3 cells. These observations are important while considering a relation of cell transformation verses abundance of Avr1 protein factor in cells. These results suggest that further characterization of Avr1-factors may be important to identify new diagnostic methods for cancer progression in humans.

Comparison of different ID retroposon sequences

Abundant expression of Avr1-binding transcription factors in cancer cells prompted us to study the Avr1 sequence in the genome database. BLAST sequence analysis has demonstrated that highly conserved Avr1 sequence belongs to the conserved ID family of retroposon with poly A tail, not only present in neuron and testis (100% homology), but also in many crucial cellular genes. Multiple alignments of different ID retroelements (Figure 4) have suggested that such small mobile DNAs are highly conserved in nature. In fact, the Avr1 sequence in most sequences has only one nucleotide mismatch with the VL30 ID or HaSV ID, or a two-nucleotide mismatch with LINE1-ID and 100% similarity with brain ID (Figure 4). Guinea pig and mouse genomes have numerous ID sequences with 90–95% sequence similarity (data not shown). The ID sequences are highly conserved among the 500 genes analysed, with 0–4 nucleotides substituted within 75 bases ID sequences, excluding the poly A tail (score 347–163, $E = 1.3^{e-57}$–0.00087) (Table 1).

Deregulation of DNA-binding factors in human cancer cells has prompted us to study the possible ID elements or Avr1 sequences in the human genome and protein database. BLAST analysis has resulted in identification of ID mobile DNA in a few human mRNAs having both known

and unknown function(s). For example, bmp-1 proto-oncogene (AAA19873) exon 1 contains a small part of the ID sequence (accession no. L13689). Many contig of human chromosome 6 hit with a higher score than any other chromosome. Among these, one unknown cDNA (accession no. AK098165), isolated from human trachea has a full-length ID sequence with poly A tail located at the 3′-untranslated region[16]. A few differentially displayed unknown cDNAs expressed in human eyes[17] also have ID sequence with substitutions in the poly A sequence (accession nos. BQ188034 and BM932101). The BLAST2 sequence analysis has shown that sequence similarity of ras-ID to human eye-ID or trachea-ID is about 80% due to variation in the poly A sequence, but with respect to Avr1 sequence the similarities are 95–100% (data not shown). The low abundance of ID retroposons in humans has supported the Greally's hypothesis that SINE retroelements are evolutionarily lacking in the imprinted human genome[18].

ID retroposons may encode protein chimeras

The unique nature of sequence conservancy and widespread localization and expression of ID sequences have led us to consider their other possible biological functions. Surprisingly, we found that inserted ID sequence can encode peptides from zero to all six reading frames and might form protein chimeras of novel function due to the unexpected presence of 5′ and 3′ splice acceptor/donor sites (e.g. AGTGGT at position nt 20, Figure 4). Indeed, a search in the short peptide and protein database has identified such chimera proteins (Figure 5), but of unknown function (protein IDs: AAP85371, AAH83623 and XP_346693). Most importantly, 24 amino acids of an unknown liver regeneration protein (LAR, 246aa) are found to be encoded by reverse-oriented ID sequence. We also hypothesize a mechanism of P^{29} v-ras oncogenic formation *in vivo* by HaSV due to ID retroposition whose molecular mechanism has not been resolved during the past 30 years of research on ras oncogene. We have proposed a hypothetical model of P^{29} formation assuming that in the HaSV genome one nucleotide at position 779 bp is absent (sequence error) creating an ATG codon and 185 bp GC-rich loop structure (exon-1) is removed (909–1068 bp, AN: X00740) by alternating splicing; this will result in a hypothetical P^{29} protein keeping the P21 coding sequence in place. The suggested N-terminal 43 AA peptide sequence (originating from ID sequence) in P^{29} ras will be then as follows: H_2N MCF GGW GFS SVI ELL PSK RKA LGS VPS SEK KER ETK QKH IVF Y-CO_2H. In fact, many other retroviral oncogenes like v-myb, v-abl, v-mos, v-sis, v-ets are chimera oncogenes. Overall, a role for Avr1 ID sequence and associated factor in transcriptional and oncogenic activation has been suggested, but it needs further experimental proof.

Discussion

This study shows that an inserted ID retroposon at the rat Ha-*ras* gene far upstream promoter is a potential target for protein factor(s) that are highly expressed in carcinoma cell lines but not in normal liver tissues. Indeed the ID sequence is highly expressed in rodent tissues and cell lines[15] as well as in HaSV[4] and VL30 retroelements[14]. However, their binding factors have not been reported so far. Thus expression of Avr1 transcription factor only in cancer cells is surprising, because *ras* gene expression from the HaSV promoter is 100 times more abundant than that is found in normal cells[19]. We have suggested Avr1 factor act as a transactivator protein and Avr2/Avr3 factors as repressors. Also, we have postulated here that ID sequence insertions may have occurred during evolution between strong Avr2–Avr3 and poly CA repressor sequences possibly to allow normal transcription of the Ha-*ras* gene from the far upstream TATA promoter elements[4]. However, such an ID sequence is absent at the immediate vicinity of the human or mouse Ha-*ras* genes (see accession nos NW_043401 and NT_035113). However, the repressor motifs (Avr2–Avr3), including poly CA or poly GT sequences are much conserved between rat and mouse c-Ha-*ras* genes (data not shown).

Cellular Ha-*ras* gene overexpression has been found in human carcinomas in the absence of genetic reorganization or amplification and point mutations, suggesting that increased expression may be due to point mutations in cellular genes which regulate *ras* gene expression, or in regulatory elements close to this locus[20]. We have shown that proto-Ha-*ras* genes and their associated regulatory proteins are influenced by specific, negative-acting, far upstream regulatory elements[10]. Other recent studies suggest that retrotransposition is developmentally programmed and retrotransposons appear as the parasite of host regulatory information. By changing the DNA sequence simply by insertion followed by deletion, mutation or recombination, these sequences regulate many host genes creating a new trait[21,22].

Avr1 retroposon-derived oligonucleotides bind factors present in human carcinoma cells (Figure 2 *b*, *e* and *f*). We have also identified that ID retroposon insertion can form novel chimera proteins (Figure 5) apart from transcriptional activation as in HaSV or promoting alternate splicing as in pyruvate kinase gene (Table 1). Also, high affinity for ss-oligonucleotides (Figure 2 *c*) warns a role for such proteins in mRNA transport or stability[23]. Indeed hairpin and self-dimer formation may greatly affect the protein binding patterns (Figure 3). Recent observations from other studies also suggest an epi-genetic role of HERVs, L1 and Alu retroelements in shaping new genetic traits[24–27]. Abundant expression of Avr1-nuclear factors in human carcinoma cells is indeed important, but needs further purification and characterization of the factors for *in vitro* assay.

Figure 5. ***a***, Generation of protein chimeras by inserted ID retroposon. BLAST searching (www.ncbi.nlm.nih.gov/blast) of the short peptide sequence and protein database was performed using amino acid sequence coded by ID retroposons. A rat liver regeneration cDNA (accession no. CD670559) encodes an unknown protein of 246 amino acids (AAs) having similarity with a 24 AAs stretch at position 111 (in reverse orientation) and many unknown cDNAs, XM_346693 (88 AAs), BC08323 (356 AAs), XM_346711 (825 AAs) have coded unknown protein chimeras (protein IDs are AAP85371, AAQ96231, XP346693 and AAH83623). Part of the amino acid sequences with ID chimera are shown (*N*-terminal AA numbers are in parentheses). ***b***, Derived peptide sequences of the c-Ha-ras ID sequence (sense and anti-sense), HaSV ID and Testis ID are shown (for sequence, see Figure 4). The hypothetical chimera proteins are coded (***a***) from the ID anti-sense strand (underlined).

Sequence analysis suggests that Avr1 sequence could be a potential binding site of many known transcription factors like Ets1/Pea3 (AGG AAG A), v-Myb/SF-1 (GCA AGG CCC T) and Zic1/GAL4 (GGG TTC GGT C). However, localization of a new factor binding site cannot be ruled out. Our future work will involve site-directed mutagenesis as well as to make a cascade of smaller oligonucleotides to address the issue. Interestingly, evidences suggest that the Avr1 protein factor might act as a trans-activator of transcription. This hypothesis is supported by the following observations: (i) the ID sequence is present in the upstream region of the *ras* gene in HaSV genome (accession nos X00740 and M24154) which highly expresses ras protein in cells and is a single-hit carcinogen; (ii) sequences are present in the rodent VL30 retroelements (accession no. M91235) that are highly expressed *in vivo*[13] and as VL30 vectors (accession no. AY260553) in human primary and transformed cell lines[12]; (iii) preliminary work by Ariazi *et al.*[28] has indicated that the rat Ha-ras upstream (–2876 to –2110) region was as activator of transcription in primary mammary epithelial cells; (iv) a preliminary report by Kim *et al.*[22] has suggested that ID sequence upstream of the SV40 promoter increased CAT gene activity, and (v) our preliminary result using a concatamerized Avr1 ds-oligonucleotide upstream of the thymidine kinase promoter has increased the luciferase gene activity (data not shown).

The ID sequences are highly conserved among the 500 genes analysed (Table 1) with 0–4 nucleotides substituted within 75 bases ID sequences, excluding the poly A tail (score 347–163, $E = 1.3^{e-57} - 0.00087$). Short nucleotide and STS database searches have identified such sequences in rat and human chromosomes (Figure 5 *a*) and Avr1-binding factor actively expressed in human cancer cells (Figure 2 *e* and *f*). That Avr1 factor does not express in normal rodent liver cells and mouse 3T3 or PA317 immortalized cells is an important observation. Similarly, expression of Avr1-factor in HeLa, HepG2, SW620, HT1080, SW403, Hepatoma3B and Huh-7 diverse human cancer cell lines has suggested a role for such factors in cell transformation[29–31]. Therefore, Avr1-binding factors are potential targets for therapeutic intervention in cancer.

Specific inhibitor of DNA–protein interaction and RNAi technology could be developed disrupting ras or Avr (unknown) gene transcription. For example, R115777, an inhibitor of Cys186 isoprenylation of ras protein, is used in patients with advanced breast cancer by disrupting the estrogen and EGF receptors-mediated growth signals through ras[32]. Recently, signal transducing membrane-bound oncogenes such as Ha-ras and Src have been implicated in hepatitis B virus and hepatitis C virus-mediated hepatocellular carcinoma[33]. We also suggest that Avr2 repressor expression in cancers may repress the activated oncogene(s) and suppress tumour growth. The bigger complex formation by Avr3 oligonucleotide in cancer cells than normal cells may also be used as diagnostic tool for cancer. Also, evidences suggest that micro-RNAs (~ 20–30 bp only) originating from the SINE or ALU retroelements greatly influence the retro-transposition events and genomic stability[34–36]. Taken together, the evidence suggests that Avr-factors could be used either as markers of cancer, particularly hepatocellular carcinoma, or as therapeutic targets, or both. Minimal binding site for Avr-factors and purification of the same are underway. McClintock's mobile DNA in eukaryotic cells is emerging out rapidly pinpointing no junk DNA in cells[37].

1. Barbacid, M., Ras genes. *Annu. Rev. Biochem*., 1997, **56**, 779–827.
2. Weiss, R., Teich, N., Varmus, H. and Coffin, J., *Molecular Biology of RNA Rumor Viruses*, Cold Spring Harbor Lab., Cold Spring Harbor, 1995.
3. Solanas, M. and Escrich, E., Ha-ras in normal and tumoral tissues: structure, function and regulation. *Rev. Esp. Fisiol*., 1996, **52**, 173–192.
4. Chakraborty, A. K., Cichutek, K. and Duesberg, P. H., Transforming function of proto-ras genes depends on heterologous promoters and is enhanced by specific point mutations. *Proc. Natl. Acad. Sci. USA*, 1991, **88**, 2217–2221.

5. Nomura, K., Kanemura, H., Satoh, T. and Kataoka, T., Identification of a novel domain of Ras and Rap1 that directs their differential subcellular localizations. *J. Biol. Chem*., 2004, **279**, 22664–22673.
6. Harada, N., Oshima, H., Katoh, M., Tamai, Y., Oshima, M. and Taketo, M. M., Hepatocarcinogenesis in mice with beta-catenin and Ha-ras gene mutations. *Cancer Res*., 2004, **64**, 48–54.
7. Newbold, R. E. and Overell, R. W., Fibroblast immortality is a prerequisite for transformation by EJ-c-Ha-ras oncogene. *Nature*, 1983, **304**, 648–651.
8. Finney, R. E. and Bishop, J. M., Predisposition to neoplastic transformation caused by gene replacement of H-ras1. *Science*, 1991, **260**, 1524–1527.
9. Makris, A., Patriotis, C., Bear, S. E. and Tsichlis, P. N., Structure of a Moloney murine leukemia virus-like 30 recombinant: implications for transduction of the c-Ha-ras proto-oncogene. *J. Virol*., 1993, **67**, 1286–1291.
10. Chakraborty, A. K. and Hodgson, C. P., Role of far upstream repressor elements controlling proto-Ha-ras gene transcription. *Biochem. Biophys. Res. Commun*., 1998, **252**, 716–722.
11. Kim, J., Kass, D. H. and Deininger, P. L., Transcription and processing of the rodent ID repeat family in germline and somatic cells. *Nucleic Acids Res.*, 1995, **23**, 2245–2251.
12. Chakraborty, A. K., Zink, M. A. and Hodgson, C. P., Expression of VL30 vectors in human cells that are targets for gene therapy. *Biochem. Biophys. Res. Commun*., 1995, **209**, 677–683.
13. Kim, J., Martignetti, J. A., Shen, M. R. and Brosius, J., Transcription and processing of the rodent ID repeat family in germline and somatic cells. *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 3607–3611.
14. Shen, M. R., Brosius, J. and Deininger, P. L., BC1 RNA, the transcript from a master gene for ID element amplification, is able to prime its own reverse transcription. *Nucleic Acids Res.*, 1997, **25**, 1641–1648.
15. Pascale, E., Liu, C., Valle, E., Usdin, K. and Furano, A. V., The evolution of long interspersed repeated DNA (L1, LINE 1) as revealed by the analysis of an ancient rodent L1 DNA family. *J. Mol. Evol*., 1993, **36**, 9–20.
16. Strausberg, R. L. *et al.*, Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 16899–16903.
17. Bonaldo, M. F., Lennon, G. and Soares, M. B., Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res*., 1996, **6**, 791–806.
18. Greally, J. M., Short interspersed transposable elements (SINES) are excluded from imprinted regions in the human genome. *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 327–332.
19. Hua, V. Y., Wang, W. K. and Duesberg, P. H., Dominant transformation by mutated human ras genes *in vitro* requires more than 100 times higher expression than is observed in cancers. *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 9614–9619.
20. Theillet, C. *et al*., Loss of a c-H-ras-1 allele and aggressive human primary breast carcinomas. *Cancer Res.*, 1986, **46**, 4776–4781.
21. Whitelaw, E. and Martin, D. I., Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nature Genet.*, 2001, **27**, 361–365.
22. Kim, D. D., Kim, T. T., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S. and Gabriel, A., Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Res.*, 2004, **14**, 1719–1725.
23. Kazazian Jr, H. H., Mobile elements: drivers of genome evolution. *Science*, 2004, **303**, 1626–1632.
24. Buzdin, A. A., Retroelements and formation of chimeric retrogenes. *Cell Mol. Life Sci.*, 2004, **61**, 2046–2059.
25. Cordaux, R. and Batzer, M. A., The impact of retrotransposons on human genome evolution. *Nature Rev. Genet.*, 2009, **10**, 691–703.
26. Babatz, T. D. and Burns, K. H., Functional impact of the human mobilome. *Curr. Opin. Genet. Dev*., 2013, **23**, 264–270.
27. Piriyapongsa, J., Marino-Ramirez, L. and Jordan, I. K., Origin and evolution of human microRNAs from transposable elements. *Genetics*, 2007, **176**, 1323–1337.
28. Ariazi, E. A., Thompson, T. A., Burkholder, J. K., Yang, N. S. and Gould, M. N., Transcriptional regulatory and response mapping of the rat Ha-ras upstream sequence using primary mammary epithelial cells. *Cancinogenesis*, 1995, **32**, 965–968.
29. Gunguly, T., Dunber, P., Chen, L. and Godmilow, T., Exon skipping caused by an intronic insertion of a young Alu Yb9 element leads to severe hemophilia A. *Hum. Genet*., 2003, **113**, 348–352.
30. Cruickshanks, H. A. and Tufarelli, C., Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter. *Genomics*, 2009, **94**, 397–406.
31. Polak, P. and Domany, E., Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics*, 2007, **7**, 133.
32. O'Regan, R. M. and Khuri, F. R., Farnesyl transferase inhibitors: the next targeted therapies for breast cancer? *Endocr. Relat. Cancer*, 2004, **11**, 191–205.
33. Been, J. and Schneider, R., Hepatitis B virus HBx protein activates Ras–GTP complex formation and establishes a Ras, Raf, MAP kinase signaling cascade. *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 10350–10354.
34. Spengler, R. M., Oakley, C. K. and Davidson, B. L., Functional microRNAs and target sites are created by lineage-specific transposition. *Hum. Mol. Genet.*, 2014, **23**, 1783–1793.
35. Brennecke, J. *et al.*, An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science*, 2008, **322**, 1387–1392.
36. Atanabe, T. W. *et al.*, Identification and characterization of two novel classes of small RNAs in the germline in testes. *Genes Dev.*, 2006, **20**, 1732–1743.
37. McClintock, B., The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. USA*, 1950, **36**, 344–355.

ACKNOWLEDGEMENTS. I thank Prof. Peter Duesberg (UC Berkeley) and Dr Clague Hodgson (Nature Technology Corporation, USA) for assistance. I also thank Dr Takis Papas (Hollings Cancer Center, USA), Dr Hemanta Majumder (IICB, Kolkata) and Dr J. B. Medda (OAER, Burdwan) for providing laboratory facilities and support. Part of the work was presented at the International Symposium on Translational Research in Cancer held at Bhubaneswar on 19 January 2007.

Received 10 February 2014; revised accepted 11 September 2014