Nucleic Acids Research, 2003, Vol. 31, No. 4 ***e17***
DOI: 10.1093/nar/gng017

Detection and analysis of spliced chimeric mRNAs in sequence databanks

Antonello Romani, Emanuela Guerra[1], Marco Trerotola[1] and Saverio Alberti[1,*]

Department of Experimental Medicine, Section of Molecular Pathology and Immunology, University of Parma, Italy and [1]Laboratory of Experimental Oncology, Department of Cell Biology and Oncology, Institute Mario Negri - Consorzio Mario Negri Sud, 66030 Santa Maria Imbaro (Chieti), Italy

Received September 1, 2002; Revised November 13, 2002; Accepted December 17, 2002

ABSTRACT

We have developed a databank screening procedure, the In Silico Trans-splicing Retrieval System (ISTReS), to identify heterologous, spliced mRNAs with potential origin from chromosomal translocations, mRNA *trans*-splicing and multi-locus transcription. A parsing algorithm to screen cDNA versus genome Blast outputs was implemented. Key filtering criteria were Blast scores of ⩾300, match lengths of ⩾95% of the query sequences, junction of the two partners at exon–exon borders and concordant 'sense/sense' reading orientation. ISTReS was validated by the successful identification of bona fide chromosomal translocation-derived fusion transcripts in the HGI and RefSeq databanks. The performance of ISTReS was verified against recently identified chimeric antisense transcripts, where it revealed essentially no independent proof of antisense transcription and absence of exon–exon borders at the chimeric join, consistent with an artefactual origin. Analysis of the UNIGENE database revealed 21 742 chimeric sequences overall that correspond to ~1% of the database transcripts. Novel FOP-Rho GAP and methionyl tRNA synthetase-advillin chimeric mRNAs with the canonical features of heterologous-genes spliced-transcripts were identified among 246 chimeras from the RefSeq databank. This suggests a frequency of canonically-spliced chimeras of ~1% of all the hybrid sequences in current databanks. These findings demonstrate the efficiency of ISTReS and the overall feasibility of sequence/structure-based strategies to search for chimeric mRNAs candidate to derive from the splicing of heterologous transcripts.

INTRODUCTION

Chromosomal translocations are frequently detected in hematologic and solid malignancies, where they can play a causative role or induce a cell growth selective advantage (1). At the molecular level, they act through deregulation of gene expression or through the generation of fusion oncogenes. Examples of the former are translocations where a gene lands near enhancer elements, e.g. within immunoglobulin or T-cell receptor genes, or that disrupt the promoter region of c-*myc* (2). However, chromosomal translocations in tumour cells much more frequently generate hybrid, oncogenic coding sequences (1). Often, the corresponding hybrid proteins are signaling molecules and/or transcription factors that are deranged from their normal regulatory circuits or acquire novel functional properties. Disruption of regulatory pathways appears, thus, as a major and widespread consequence of the generation of chimeric mRNAs encoding hybrid oncogenic proteins.

Hybrids between heterologous mRNAs are also generated by mRNA *trans*-splicing. The latter was first detected *in vitro* (3,4), but was subsequently shown to occur *in vivo* in several lower and higher eukaryotes (5), including mammals (6–10). Major biological functions of the *trans*-splicing of a common spliced leader in trypanosomatids and nematodes are the processing of polycistronic transcription units and the regulation of the translation efficiency and stability of the resulting mRNAs (11). On the other hand, the *trans*-splicing between heterologous transcripts in mammalian cells increases protein diversity through the joining of segments/domains originating from different genes (5). Recent findings indicate that each of the two mRNA moieties also carries specific regulatory signals that dictate the physical location of the mRNA and regulate its stability (12).

Long transcription across neighbouring genes that normally act as independent transcription units, has been demonstrated in several cases (13–16). Similarly to the cases above, this results in hybrid, multi-locus transcripts, which often increase the diversity of the exon complement of the participating genes (13–16).

Chimeric transcripts in 'antisense' orientation have also been recently identified in the RefSeq and EMBL databanks (17). The origin and role of these transcripts are not known (17). However, natural, single-gene antisense mRNAs (18) are of frequent occurrence in mammals, including man (17,19), and may play a novel role in the regulation of gene expression. Thus, a similar regulatory role was suggested for the hybrid antisense mRNA.

Hence, diverse classes of hybrid mRNAs appear to play important functional roles in normal and transformed cells (1). A whole-genome exploration for hybrid sequences is, thus, of

*To whom correspondence should be addressed. Tel: +39 0872 570 293; Fax: +39 0872 570 412; Email: alberti@negrisud.it

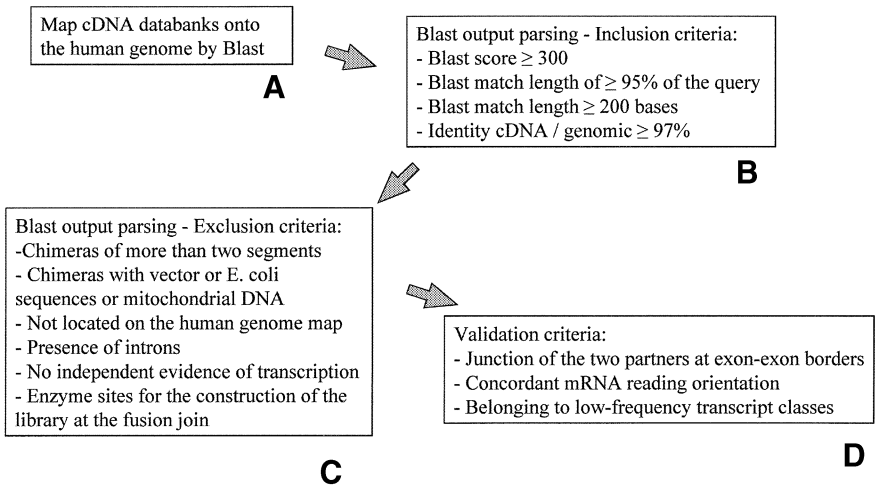

Figure 1. Schematic representation of the ISTReS analysis and retrieval strategy.

interest, as it may identify novel members of these hybrid mRNA classes, and reveal common structure/sequence characteristics. However, the construction of cDNA libraries frequently presents with cDNA fusion artifacts, linked to incorrect ligation or abnormal reverse-transcription (20,21). Thus, rigorous retrieval and analysis strategies are required to distinguish between mRNA chimeras of potential physiological origin and artifacts. In this work, we have developed the In Silico Trans-splicing Retrieval System (ISTReS) to extract 'spliced' chimeric mRNA from sequence databanks. Our findings demonstrate the overall feasibility of this sequence/structure-based strategy and the efficiency of the ISTReS procedure.

MATERIALS AND METHODS

ISTReS strategy

Chimeric sequences from non-contiguous loci are generated through at least three different molecular mechanisms, i.e. chromosomal translocations (22), mRNA *trans*-splicing (5) and transcription of long mRNA across neighbouring loci (13–16). In these cases the two heterologous mRNAs are typically joined together by splicing at conventional donor-acceptor exon-bordering sites (see below). RNA–RNA recombination has also been demonstrated, and was shown not to follow the rules of conventional splicing (23–25). However, the *in vivo* frequency of this phenomenon is extremely low (24) and is largely limited to viral RNA sequences (25). As a consequence, a general strategy was devised to identify hybrid mRNAs that are joined together following the sequence-structure rules of canonical mRNA splicing (26,27) (Fig. 1).

As shown in Figure 1A, whole cDNA sequence databanks (RefSeq, HGI, 31. jul. 2001 release, and subsets of theirs) were mapped onto the human genome by Blast analysis. Sequence databanks and analysis programs were downloaded to and utilised in two Digital Personal Workstations with Alpha 500 Mhz processors (780 and 255 Mb RAM, respectively) and a Digital AlphaStation 255 (128 Mb RAM).

Blast outputs were parsed for sequences that mapped to different loci (Fig. 1B). A parsing algorithm written in Perl was implemented to this purpose. Filtering criteria validated during the screening were: a Blast score $\geqslant$300, a Blast match length of $\geqslant$95% of the query sequence, a minimal Blast match length of 200 nt [lower lengths were permitted in the analysis of the chimeric antisense sequence datasets (17)], identity between the cDNA and genomic sequences of $\geqslant$97%. A gap of $\leqslant$10 nt between the two matches of the chimeric sequences was permitted. These criteria were devised to search with high-sensitivity and an acceptable stringency. For example, reasonable numbers of sequencing errors were tolerated, while good matches of short length were still highlighted.

As shown in Figure 1C, the identified chimeras were excluded from further analysis if they: (i) comprised more than two segments, as they were likely to derive from the random co-ligation of unrelated cDNA fragments; (ii) could not be located on the human genome map, as this did not permit verification of their origin from distinct loci; (iii) contained introns, as these were likely to derive from genomic DNA or, less frequently, from nuclear mRNA; (iv) had no independent evidence of 'sense' transcription, i.e. no other independent mRNA or EST sequences corresponding to either of the fusion partners could be identified; these sequences were likely to correspond to intergenic/non-transcribed regions of genomic DNA (see below for the antisense chimeras); (v) the fusion join corresponded to poly-linker sequences or enzyme sites utilized in the construction of the library; the chimeras contained vector (vi), mitochondrial (vii) or *Escherichia coli* (viii) sequences.

Figure 1D shows that the selected mRNA chimeras were further analysed for junction of the two partners at exon–exon borders and for concordant 'sense/sense' (or 'antisense/antisense') reading orientation. The reading orientation of the mRNA partners of the chimeric sequences was determined by comparison with independent transcripts in the non-redundant GenBank/EMBL sequence collection. EST sequences were not utilised for this purpose, because of their unreliable orientation, due to automated sequencing and non-curated deposit in databanks.

Table 1. Fusion oncogenes retrieved from ISTReS searches

Accession number	Fusion oncogene[a]
THC480561	MLL-hCDCrel
THC482203	ETV6-NTRK3
THC558963	AML1-EVI1
THC558964	AML1-MTG8
THC519492	Rho GEF-PKA AP (LBC)
L21756, S76343	AML1-EAP
X56348	SURF-RET
X92120	EWS-CHOP
M73779, M82827	PML-RARα
Y15911, Y15912, Y16345, Y16344, Y16343, Y16342, Y16341	COL1A1-PDGFB

[a]Acronyms of the 5′ and 3′ partners of the fusion oncogenes identified in the RefSeq and HGI databanks. Search parameters of the Blast search were: Blast score ⩾300; identity between the cDNA and genomic sequence ⩾97%; match length ⩾200 bp and ⩾95% of the query sequence; gap between the two sequences ⩽10 bases.

In the case of non-characterised transcripts additional evidence was contributed by a structural analysis of the corresponding transcription units at the genomic level [promoter sequences (28), transcription start sites (29), mRNA cleavage/poly-A addition signals (30) and untranslated regions (31)].

Online sequence analysis sites

Sequences were matched against the non-redundant and human_EST datasets using Blastn (http://www.ncbi.nlm.nih.gov/BLAST/) (32). Query sequences were mapped onto the human genome using Megablast (http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html) (0.01 level expectation and default filtration). Sim4 (http://pbil.univ-lyon1.fr/sim4.html) (33) was used to define the exon–intron borders of the chimera partners.

RESULTS AND DISCUSSION

mRNA chimeras from two different genes most commonly arise from chromosomal translocations (22), mRNA *trans*-splicing (5) or transcription across neighbouring loci (13–16). As hybrid mRNA sequences play important roles in cell transformation or in regulatory pathways in normal cells (1), a whole-genome exploration, through the analysis of sequence databanks, may offer novel means to reveal other members of these classes and shared structure/sequence characteristics. Most chromosomal translocations identified in cancer cells occur in intronic regions (22,34), likely because of the longer overall length of introns and of the selective pressure for a functional protein, advantageous for cancer cell growth. The two partners in the chimera (as processed from nuclear mRNA precursors) are subsequently joined at exon–exon borders (22). *Trans*-spliced transcripts originate from the joining of independently transcribed mRNAs (5). This post-translational processing follows the rules of, and is performed by, the canonical *cis*-splicing apparatus, resulting in the joining of heterologous mRNAs at canonical exon–exon borders (3–5). Long, multi-locus transcription has been observed in several instances (13–16). The processing of the resulting long, immature mRNAs results in the joining of coding regions at exon–exon sites (13–16). Thus, joining at exon-bordering sites is a common feature of all of the above classes of heterologous mRNA chimeras.

On the other hand, cDNA construction artifacts (end-to-end joining or recombination) and the rare products of RNA–RNA recombination *in vivo* (24) are unlikely to join by chance at exact exon–exon borders. Moreover, while the *trans*-splicing of heterologous, independently transcribed mRNAs is expected to join 'sense/sense' sequences, cDNA/cDNA recombination or fusion would be expected to randomly generate 'sense/antisense' sequences in one-half of the cases. Random cDNA artifacts would also be expected to arise in proportion to the abundance of the corresponding mRNA. Hence, frequent transcripts, e.g. for ribosomal proteins, elongation factors, cytoskeletal proteins, globins (in hematopoietic cells) (30), are expected to be frequent among artefactually generated chimeric mRNA.

The concepts above were incorporated in the ISTReS screening strategy and algorithm (Fig. 1). The goal of ISTReS is to identify mRNA chimeras, whether from chromosomal translocations, mRNA *trans*-splicing or multi-locus transcription, within human transcript datasets. The criteria detailed in Material and Methods were implemented in the proof-of-principle searches presented here. Blast scores of ⩾300 were used as a cut-off in the analysis of whole sequence databanks. In combination with the sequence identity criteria this also allowed to select for good sequence matches of short length.

Table 2. Artefactual chimeric sequences identified by ISTReS in the HGI databank

AC number [a]	5′ Partner[b]	3′ Partner[b]	AC number[a]	5′ Partner[b]	3′ Partner[b]
THC480049	GBR-2-like (NM_004810)	**α-1 collagen type I** (NM_000088)	THC482361	Sepiapterin Red. (NM_003124)	FLJ22552 (AK026205)
THC480771	ADA2 (NM_001488)	Sim. BCG-CWS (BC001320)	THC481666	KIAA1844 (AB058747)	**RP 8** (BC022070)
THC482624	HP:MGC5528 (NM_024094)	Protocadherin43 (HUMPC43ABB)	THC483480	FBXL7 (NM_012304)	NRAMP2 (AB015355)
THC496279	DKFZp434M045 (HSM802409)	DKFZp451H072 (HSM803274)	THC544068	FLJ13110 (NM_022912)	SyntaxinBP (NM_003165)
THC513298	Transponase-like (AF205598)	Kinesin 2 (HUMKINESLC)	THC481841	NAPOR-2 RNA BP (AF090694)	CAGR1 (U38810)
THC482298	Sim. NPC IP (XM_053643)	Sim. BAZ2A (AK023842)	THC485014	DHPRP-2 (D78013.1)	Sim. YIL091C (NM_014388)

[a]AC: accession number of the chimeric sequences identified during the search in the HGI databank that did not pass the ISTReS inclusion/exclusion criteria (Fig. 1). The accession numbers of the partners in the chimera are in parentheses below the sequence names.
[b]BP: binding protein; IP: interacting protein; Red.: reductase; RP: ribosomal protein; Sim.: similar to. Abundant mRNA classes are in bold.

Table 3. Artefactual chimeric sequences identified by ISTReS in the RefSeq databank

AC[a]	5′ Partner[b]	3′ Partner[b]	AC[a]	5′ Partner[b]	3′ Partner[b]
BC000673	TNF-Rec. 6b (XM_056902)	**RP P0** (BC009867)	M90820	FK506 BP3 (BC020809)	KIAA0589 (AB011161)
AF132973, AF155662	**RP P0** (NM_053275)	CDA016 (AF261134)	BC003614	DAP-kinase (X76104)	**RP L30** (M94314)
X69392	**RP L26** (NM_000987)	MGC:17890 (BC015899)	AY029161	PINX1 (XM_056962)	Janus-a (AF164795)
BC015576	**E-cadherin** (Z13009)	**RP L23a** (U43701)	L10377	TIM PEAS (AB055925)	DRPLA (D38529)
X77598	Leupaxin (BC019035)	**Laminin α3A** (X85107)	BC001618	PSA (NM_058179)	SLC1A4 (XM_046668)
AK057826	Complexin 2 (NM_006650)	FLJ30540 (AK055102)	BC008038	TF ALY (AF047002)	Peroxiredoxin 3 (BC002685)
X81789	BAFF Rec. (AF373846)	SAP 617 (U08815)	BC009736	FLJ12448 (BC014661)	Scar protein (M22146)
BC000519	54 kDa protein (Y18418)	AP1G2 (NM_080545)	BC001974	KIAA0150 (D63484)	ATP BP (BC005968)
X06704	**RP L7a** (M36072)	TRK-T3 (X85960)	BC001849	Alpha NAC (X80909)	MDR / TAP (BC014081)
U60975	POZ protein (BC001269)	SORL1 (XM_006312)	U38810	NAPOR-3 (AF090693)	MAB21L1 (XM_007172)
AK027315	PPIL3 (XM_027955)	LOC122769 (XM_058657)	AF152961	ALR (AF010403)	FACTP140 (NM_007192)
AL109790	I:2960796 (BC014640)	EI: 27080 (AL109684)	BC007583	FLJ23209 (BC015692)	liver protein (L13799)
AF090896	ALP A-II (M29882)	DKFZp451J1719 (AL833081)	U02019	MGC:2158 (BC023977)	Sim. C8FW (BC019363)
BC000265	Sim. HS6-O-ST (BC001196)	DKFZp547L106 (AL512715)	U51007	I:4065996 (BC016714)	Antisecretory factor-1 (U24704)
AF118091	Sim. EF1 (BC014224)	iPP1a (AF061958)	AF135156	PCDH-gB6 (AF152522)	HSPC005 (AF070661)
L11372	PCDH-gC3 (AF152337)	FLJ25400 (AK058129)	U81554	CCPK-II (U66063)	**SRP72** (AF069765)
BC004528	I:4156703 (BC011262)	FLJ30001 (AK054563)	U97105	YIL091C (NM_014388)	DPYSL2 (NM_001386)
X73608	Ring-box 1 (BC001466)	SPOCK (NM_004598)	X56465	FLJ25091 (AK057820)	ZNF6 (NM_021998)
AK022445	**RP L7** (NM_000971)	Calponin like (BC025251)	U64876	**MHC Class II γ** (M13555)	GCNF nuclear Rec. (U80802)
BC007261	P1H12 (AF089868)	I:3344121 (BC008758)	U16258	**RP S7** (NM_001011)	NFKBIL2 (NM_013432)
U49278	UBE2V1 (NM_022442)	RNPEP (AJ242586)	L10717	KIAA1046 (AB028969)	ITK (NM_005546)
AK026712	FLJ23059 (XM_096151)	**RP S3** (BC003137)	BC007607	ATP synthase (BC019310)	**RP S3A** (NM_001006)
BC012823	FLJ22875 (AK026528)	KIAA0699 (AB014599)	BC001209	R:2810432L12 (BC006115)	**Tubulin alpha 1** (BC006379)
BC001805	**tubulin alpha 1** (BC009314)	NDRG3 (AB044943)	AK026642	**RP L35A** (NM_000996)	HSA276469 (AJ276469)

[a]AC: accession number of the chimeric sequences identified in the RefSeq databank that did not pass the ISTReS inclusion criteria (Fig. 1). The accession numbers of the partners in the chimera are in parentheses below the sequence names.
[b]BP: binding protein; EF: elongation factor; Rec.: receptor; RP: ribosomal protein; Sim.: similar to; I: IMAGE; EI: Euroimage; R: RIKEN. Abundant mRNA classes are in bold.

This cut-off value was relaxed (≥80) for the analysis of the chimeric antisense transcript dataset, as this contained even shorter sequence segments (17).

We verified the capability of ISTReS to detect actual chimeric sequences in the curated databanks Human Gene Index (HGI, TIGR) (http://www.tigr.org/tdb/hgi/) and RefSeq (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html) (the latter search was kindly supported by the RefSeq curation staff). These searches were performed by Blast comparison of the HGI Tentative Human Consensus (THC) sequences and of the RefSeq candidates versus the human genome. Blast outputs were subsequently parsed for sequences that mapped to two different map locations.

Eighteen fusion sequences corresponding to ten experimentally verified translocated oncogenes, i.e. MLL-hCDCrel, ETV6-NTRK3, AML1-EVI1, AML1-MTG8, Rho GEF-PKA AP (LBC), AML1-EAP, SURF-RET, EWS-CHOP, PML-RARα, COL1A1-PDGFB, were identified (Table 1), validating our search procedure. These searches also revealed a much larger number of chimeric sequences that did not meet the structural requirements expected from 'physiological' fusion events. mRNA chimeras that failed the criteria outlined in Figure 1 are listed in Tables 2 and 3. A comprehensive analysis of the Unigene database revealed 21 742 chimeric sequences, i.e. ~1% of the total number of transcripts analysed (35). Notably, several frequent mRNAs, e.g. for ribosomal

Table 4. Chromosomal map location of the chimeric sequences identified by ISTReS in the RefSeq databank

Chimeras[a]	5′ Partner[a]	Locus[b]	3′ Partner[a]	Locus[b]
X53795	U20770	11p11.2	AF015553	7q11.23
BC007700	NM_002952	16p13.3	NM_001014	6p21
BC007583	NM_002952	16p13.3	U43701	17q11
BC007937	BC009833	19p13	AB022847	16q
BC000265	**BC001196**	**1p362**	**NM_004723**	**1q21-q22**
AF000145	BC005678	1p34-p32	AA873817	2p21
AF135156	AF152522	5q31	AF070661	11q12-q13
L11372	AF152337	5q31	AK058129	8q23
AF119855	BF929828	12q24.2	AJ007398	16p13.2
U81554	U66063	10q22	AF069765	4q11
BC005228	XM_037822	4q33	M67480	2q35-q36
BC001425	**BC007751**	**1q21.2**	**NM_032636**	**1p13**
BC007935	D83327	21q22.2	Z28407	8q24.3
U49278	NM_022442	20q13.2	AJ242586	1q32
S35959	M64241	Xq28	BC008867	3p21.3
AK026712	XM_096151	22q11.2	BC003137	11q13.3-q13.5
AK026528	BC008865	15q22	NM_000989	8q22
BC012823	AK026528	15q21	AB014599	9q22.2
BC001805	BC009314	12q12	AB044943	20q11.21-q11.23
AK026642	NM_000996	3q29-qter	AJ276469	20q131
AK026614	NM_000996	3q29-qter	BC001009	19p133
AF039747	U34846	18q11.2-q12	NM_006727	5p14-p13
BC002569	Z12962	12q13	NM_001007	Xq13
BC011860	AF156102	17q23.2	BC006483	22q13
BC002450	AB072911	19q12-13	D23660	15q22
BC007259	AK027019	1q11	BC001365	15q22
BC010079	NM_001950	16q21-q22	AB061822	3p22-p21.2
BC000673	XM_056902	20q13.3	BC009867	12q24.2
BC008791	M34539	20p13	BC021909	chr. 6
AF314817	**Y18046**	**6q27**	**AF385429**	**6q25.3**
BC003614	X76104	9q34	M94314	3q12
X69392	NM_000987	17p13	BC015899	7q35
M90309	AF332356	20p11.23	BC020809	14q21
M90820	BC020809	14q21	XM_047620	19p13.3
AY029161	XM_056962	8p23	AF164795	9q34.3
AK000572	AK026847	1p22	AJ344104	8p23
BC015576	Z13009	16q22	U43701	17q11
BC015601	BC007296	6p21	M81955	14q11.2-q12
L10377	AB055925	6p21	D38529	12p13.31
BC004134	**D84224**	**12q13.2**	**NM_006576**	**12q13.1**
BC001618	NM_058179	9q21.31	XM_046668	2p13
AK027187	XM_032319	15q13	NM_001010	9p21
BC007669	BC002959	Xp11.22	BC011615	17q25.3
AK057826	NM_006650	5q35.3	AK055102	3q22.2
AK000545	M14486	15q22	NM_079837	16q24
BC009736	BC014661	12q	M22146	Xq13
BC000519	Y18418	3q21	NM_080545	14q11.2
BC001974	D63484	8q24.3	BC005968	2q11.2
X06704	M36072	9q34	X85960	6
BC001849	X80909	12q23-q24	BC014081	6p21.3
U60975.2	BC001269	17q22	XM_006312	11q24
U38810	AF090693	10p13	XM_007172	13q12
AK027315	XM_027955	2q33	XM_058657	14q13
BC007768	NM_032891	3q29	BM045029	5q13
BC021561	AK024072	14q11	BC000502	18q21

[a]Accession number of the chimeric sequences identified from the RefSeq databank that passed the inclusion criteria depicted in Figure 1C. The accession number of the corresponding fusion partners are also indicated.
[b]Chromosomal map locations of the fusion partners of the chimeric sequences. Sequences from the same chromosome are in bold.

proteins, were rather frequent among chimeric sequences, arguing in favour of a stocastic, artefactual nature. Of interest, intra-chromosomal hybrid sequences were 7.3% of all the RefSeq chimeras (Table 4) and 6.9% of the antisense chimeras (see below). As the fraction of the genome belonging to each separate chromosome ranges from 8.5% for chromosome 1 to 1.6% for the Y chromosome (mean: 4.8%) (36), our results appear close to what would be expected on stocastic grounds only, further supporting the random nature of most of the observed events.

Rather interestingly, however, ISTReS screenings also identified two novel chimeric mRNAs that demonstrated the canonical features of *trans*-spliced mRNAs (or long intergenic transcripts), FOP-Rho GAP and methionyl tRNA synthetase-advillin. As these mRNAs were selected-out from 246 RefSeq chimeras (Fig. 2), these findings suggest a 1% frequency of

canonically-spliced candidates among the hybrid sequences detected in current databanks.

Recent findings have indicated that antisense mRNAs are of frequent occurrence in human cells (17,19). Previous experimental evidence had demonstrated the existence of natural antisense mRNA in eukaryotic cells (18). However, their frequent occurrence in the human transcriptome was unexpected, raising the possibility that they may play a novel role in the regulation of gene expression (17,19). The identified antisense chimeras were analysed with the ISTReS procedure. Thirty-one 'antisense' chimeric sequences passed the screening criteria for trivial cDNA artefacts (Fig. 1A–C; Tables 5 and 6). Unexpectedly, though, independent evidence for antisense transcription proved almost nil (2 of 24 148 total hits), and in none of the latter cases were exon–exon borders detected at the chimeric join. Abundant mRNAs (for ribosomal proteins, globins, translation elongation factors, MHC invariant chain, β2-microglobulin etc.) were frequently present in the antisense chimeras, and hybrid sequences with ribosomal RNA (RNA polymerase I transcripts) (AF159295, AF095784) were also identified. Moreover, transcription initiation and cleavage/poly-adenylation sites were frequently present at the fusion joins for both 'sense' and 'antisense' mRNA segments (17) (Table 5). As 'transcription initiation' and 'cleavage/poly-adenylation' refer to the sense strand and do not have structural correlates for the 'antisense' strand, this and the analyses above strongly suggested an artifactual origin of the antisense hybrid sequences.

In summary, we have developed the ISTReS algorithm and screening procedure to identify spliced, heterologous mRNAs in sequence databanks. Our findings demonstrate the efficiency of ISTReS. They also support the overall feasibility of sequence/structure-based strategies to select for chimeric mRNAs candidate to derive from the splicing of heterologous transcripts.

ACKNOWLEDGEMENTS

We thank Bo Yuan, David Wheeler and Monica Romiti for help with sequence retrieval and databank analysis. This work

A

Dna Sequence

```
   1 cgccgaccct aagtttcggc gctcagtggt ccggcgctcc ccaaggctcg gtgtccagcg
  61 tcaaccccga ggtctctatg ccccgcctcc cgacgccagg gggcagggcc agcgcgctgc
 121 gcgtcggggc ggggctttgg ctgcgtcggc cgcgtagccc gcgcgcggag cgtaccctgc
 181 tgcggccgtt ggccgttagc gcggcttcgg cggttgtctt ggagaagcaa gatggcggcg
 241 acggcggccg cagtggtggc cgaggaggac acggagctgc gggacctgct ggtgcagacg
 301 ctggagaaca gcggggtcct gaaccgcatc aaggctgaac tccgagcagc tgtgttttta
 361 gcactagagg agcaagaaaa agtagagggtg aagacactgg tggaattcct cattgataac
 421 tgctttgaaa tatttgggga gaacattcca gtgcattcca gtatcacttc tgatgactcc
 481 ctggagcaca ctgacagttc agatgtgtcg accctgcaga atgactcagc ctacgacagc
 541 aacgaccctg atgtggaatc caacagcagc agtggcatca gctctcccag caggcagccc
 601 caggtgccca tggccacagc tgctggcttg gatagcgcgg gcccacagga tgcccgagag
 661 gtcagcccag agcccattgt gagcaccgtg gccaggctga aaagctccct cgcacagccc
 721 gataggagat actcagagcc cagcatgcca tcctcccagg agtgcctcga gagccgggtg
```

Aminoacid Sequence

```
   1    MAATAAAVVAEEDTELRDLLVQTLENSGVLNRIKAELRAAVFLALEEQEKVEVKTLVEFL
  61    IDNCFEIFGENIPVHSSITSDDSLEHTDSSDVSTLQNDSAYDSNDPDVESNSSSGISSPS
 121    RQPQVPMATAAGLDSAGPQDAREVSPEPIVSTVARLKSSLAQPDRRYSEPSMPSSQECLE
 181    SRVTNQTLTKSEGDFPVPRVGSRLESEEAEDPFPEEVFPAVQGKTKRPVDLKIKNLAPGS
 241    VLPRALVLKAFSSSSLDASSDSSPVASPSSPKRNFFSRHQSFTTKTEKGKPSREIKKHSM
 301    SFTFAPHKKVLTKNLSAGSGKSQDFTRDHVPRGVRKESQLAGRIVQENGCETHNQTARGF
 361    CLRPHALSVDDVFQGADWERPGSPPSYEEAMQGPAARLVASESQTVGSMTVGSMRARMLE
 421    AHCLLPPLPPAHHVEDSRHRGSKEPLPGHGLSPLPERWKQSRTVHASGDSLGHVSGPGRP
 481    ELLPLRTVSESVQRNKRDCLVRRCSQPVFEADQFQYAKESYI*
```

B

Dna Sequence

```
 961 gatgagtatg gtacagcaac agagaccaag gctctggagg agggactaac cccccaggag
1021 atctgcgaca agtaccacat catccatgct gacatctacc gctggtttaa catttcgttt
1081 gatatttttg gtcgcaccac cactccacag cagaccaaaa tcacccagga cattttccag
1141 cagttgctga aacgaggttt tgtgctgcaa gatactgtgg agcaactgcg atgtgagcac
1201 tgtgctcgct tcctggctga ccgcttcgtg gagggcgtgt gtcccttctg tggctatgag
1261 gaggctcggg gtgaccagtg tgacaagtgt ggcaagctca tcaatgctgt cgagcttaag
1321 aaaatggagc tggcgctggt gcctgtgagc gcccacggca acttctatga gggggactgc
1381 tacgtcatcc tctcgacccg gagagtggcc agtctcctat cccaggacat ccacttctgg
1441 atcgggaagg actcctccca ggatgagcaa agctgcgcag ccatatatac cacacagctg
1501 gacgactacc tgggaggcag ccctgtgcag caccgagagg tccagtacca tgagtcagac
1561 actttccgtg gctacttcaa gcagggcatc atctacaagc agggggggtgt cgcctctggg
1621 atgaagcacg tggagaccaa tacctacgac gtgaagcggc tgctacatgt gaaagggaaa
1681 agaaacatca gggctaccga ggtggaaatg agctgggaca gtttcaaccg aggtgatgtc
1741 ttcttgctgg accttgggaa agtcatcatc caatggaatg gcccagagag caacagtggg
```

Aminoacid Sequence

```
   1    MRLFVSDGVPGCLPVLAAAGRARGRAEVLISTVGPEDCVVPFLTRPKVPVLQLDSGNYLF
  61    STSAICRYFFLLSGWEQDDLTNQWLEWEATELQPALSAALYYLVVQGKKGEDVLGSVRRA
 121    LTHIDHSLSRQNCPFLAGETESLADIVLWGALYPLLQDPAYLPEELSALHSWFQTLSTQE
 181    PCQRAAETVLKQQGVLALRPYLQKQPQPSPAEGRAVTNEPEEEELATLSEEEIAMAVTAW
 241    EKGLESLPPLRPQQNPVLPVAGERNVLITSALPYVNNVPHLGNIIGCVLSADVFARYSRL
 301    RQWNTLYLCGTDEYGTATETKALEEGLTPQEICDKYHIIHADIYRWFNISFDIFGRTTTP
 361    QQTKITQDIFQQLLKRGFVLQDTVEQLRCEHCARFLADRFVEGVCPFCGYEEARGDQCDK
 421    CGKLINAVELKKMELALVPVSAHGNFYEGDCYVILSTRRVASLLSQDIHFWIGKDSSQDE
 481    QSCAAIYTTQLDDYLGGSPVQHREVQYHESDTFRGYFKQGIIYKQGGVASGMKHVETNTY
 541    DVKRLLHVKGKRNIRATEVEMSWDSFNRGDVFLLDLGKVIIQWNGPESNSGERLKAMLLA
 601    KDIRDRERGGRAEIGVIEGDKEAASPELMKVLQDTLGRRSIIKPTVPDEIIDQKQKSTIM
 661    LYHISDSAGQLAVTEVATRPLVQDLLNHDDCYILDQSGTKIYVWKGKGATKAEKQAAMSK
 721    ALGFIKMKSYPSSTNVETVNDGAESAMFKQLFQKWSVKDQTMGLGKTFSIGKIAKVFQDK
 781    FDVTLLHTKPEVAAQERMVDDGNGKVEVWRIENLELVPVEYQWYGFFYGGDCYLVLYTYE
 841    VNGKPHHILYIWQGRHASQDELAASAYQAVEVDRQFDGAAVQVRVRMGTEPRHFMAIFKG
 901    KLVIFEGGTSRKGNAEPDPPVRLFQIHGNDKSNTKAVEVPAFASSLNSNDVFLLRTQAEH
 961    YLWYGKVGWLGPGSDQPLGAQTCTPLLLSARSKDLE*
```

Figure 2. (**A**) Structure, sequence and translation product of the FOP-Rho GAP chimeric mRNA (Accession number: AF314817) (Table 4). The 5′ partner is FOP (translocated to the FGFR1 oncogene partner), whereas the 3′ partner is T-cell activation Rho GTPase activating protein (GAP). The junction between the two partners is at exon–exon borders (exon II for FOP and exon IX for Rho GAP). The large distance between the two loci (6q27 versus 6q25) suggests a *trans*-splicing origin. However, as transcription is directed toward the centromere in both transcription units, this is formally consistent also with long, intergenic transcription. (Top) DNA sequence of the first 780 bases of the chimeric mRNA; (bottom) sequence of the encoded chimeric protein. DNA bases and aminoacids surrounding the splice site are boxed. (**B**) Structure, sequence and translation product of the methionyl tRNA synthetase-advillin chimeric mRNA (Accession number: BC004134) (Table 4). The 5′ partner is the methionyl tRNA synthetase, whereas the 3′ partner is advillin. The junction between the two partners is at exon–exon borders, (exon X for methionyl tRNA synthetase and exon II for advillin). The chromosomal position of methionyl tRNA synthetase is at 12q13.2, whereas that of advillin is at 12q13.1, and might be compatible with intergenic transcription. (Top) DNA sequence flanking the junction of the chimeric mRNA; (bottom) sequence of the encoded chimeric protein. DNA bases and aminoacids surrounding the splice site are boxed.

Table 5. Structural features of the antisense chimeric sequences in the RefSeq databank

Accession numbers[a]	5′ Partner[b]	3′ Partner[b]	Exon borders[c]	Antisense transcription[d]
X82540 (NM_005538/XM_087061)	Inhibin beta C	Sim. hnRNP	CP/anti	None of 423 hits
U64876 (M13555/U80802)	**MHC Class II γ**	hGCNF	Anti/break	None of 173 hits
X13227 (J03910/NM_001917)	Metallothionein1G	D-aminoacid oxidase	Anti/EB	None of 532 hits
X77777 (XM_113730/X75299)	hCSDA	VIP receptor	Anti/break	None of 406 hits
D49372 (BC032589/NM_002986)	**β2 microglobulin**	Eotaxin precursor	Anti/TI	None of 180 hits
L10717 (AB028969/NM_005546)	HP KIAA1046	ITK	Anti/TI	None of 141 hits
AB045369 (Y15059/NM_007232)	Neurogranin	Histamine Rec. H3	Anti/break	None of 97 hits
M60725 (AF036892/M60724)	NCOA3	P70 S6Ka1	Anti/break	None of 231 hits
AF003522 (L10335/XM_035684)	Reticulon 1	DLL1	Anti/break	None of 213 hits
M31520 (U57847/NM_001026)	**RP S27**	**RP S24**	Anti/TI	None of 183 hits
AB017111 (NM_003933/XM_046457)	BAIAP3	Sim. subtilisin	CP/anti	None of 179 hits
AF116719 (BC010054/BC010913) **U38979**	**RP SA**	**Globin γ2**	Anti/TI Break	**1 of 808 hits**[e]
X73608 (BC001466/NM_004598)	RBX1	Testican	(TI)/TI	None of 76 hits
AF031379 (NM_006053/XM_007328)	TCIRG1	CNIL	Anti/break	None of 362 hits
X56465 (NM_006601/NM_021998)	Inactive progesteron Rec.	ZNF6	Anti/EB	None of 381 hits
L29126 (NM_002055/NM_004067)	**Glial fibrillary acidic protein**	β2 chimaerin	(TI)/TI	None of 479 hits
U50079 (BC015405/NM_004964)	Sim. **RP S5**	HDAC1	(TI)/break	None of 193 hits
L35253 (AF112299/AF286697)	MAN1	CS BP	Anti/TI	None of 184 hits
U90236 (AB029290/NM_004999)	Actin BP 620	Myosin VI	Anti/break	None of 572 hits
AF152961 (AF010403/NM_007192)	MLL2	**EF FACT P140**	Anti/break	None of 282 hits
U90176 (AF160973/NM_004730)	Sim. p53 inducible protein	**eRF1**	Anti/break	None of 210 hits
AB032254 (XM_085473/XM_048948)	LOC146452	BAZ2A	Anti/break	None of 1444 hits
AF118065 (AF057352/S95936)	IGF-II mRNA BP	Transferrin	Anti/break	None of 257 hits
AF159295 (X03205/XM_040900)	**18S rRNA**	Sim. CTAK 75a	Anti/TI	None of 416 hits
L40904 (BC008572/NM_005037)	**Globin α2**	PPARγ	(TI)/TI	None of 356 hits
J03909 (AK055875/BC031020)	Sim. NADH-ubiquinone Red.	IFI30	Anti/TI	None of 14 322 hits
U17899 (AF054185/U53454)	Sim. **proteasome α7**	CLCI	Anti/break	None of 140 hits
M34182 (XM_030914/AJ001597)	E1B-AP5	PKA catalytic γ	Anti/break	None of 250 hits
X52195 (NM_017657/NM_001629) **AK074373**	FLJ20080	5-Lipoxygenase AP	Anti/break Break	**1 of 60 hits**
AF116625 (AF031548/ XM_056260)	**Erythrocyte membrane GP**	Rb BP6	Anti/break	None of 182 hits
AF095784 (X03205/ AF056085)	**18S rRNA**	GPCR51	Anti/break	None of 416 hits

[a]Accession numbers of the antisense chimeric sequences selected as indicated in the text. Accession numbers are indicated for the chimeras and the 5′/3′ partners. In the two cases where independent antisense transcription was detected, the corresponding accession number is indicated below in bold.
[b]AP: activating protein; BP: binding protein; EF: elongation factor; GP: glycoprotein; HP: hypothetical protein; IP: interacting protein; Rec.: receptor; Red.: reductase; RF: release factor; RP: ribosomal protein; RNP: ribonucleoprotein. Sim.: similar to; Abundant mRNA classes are in bold.
[c]Occurrence of the mRNA/mRNA junction at an exon–exon border; anti: antisense orientation; break: junction within an exon; EB: exon border; TI: transcription initiation; CP: cleavage and poly-adenylation sites. As these terms refer to the 'sense' strand they are in parentheses for the 'antisense' strand.
[d]Independent evidence for antisense transcription in the non-redundant GenEMBL databank.
[e]Twelve additional hits were detected (H19 mRNA, BC006831), but these did not overlap with the joining region of the chimera.

was performed with the support of the Italian Association for Cancer Research (AIRC, Italy). The financial support of Telethon - Italy (Grant no. GGP02353) is also gratefully acknowledged.

REFERENCES

1. Mitelman,F. (2000) Recurrent chromosome aberrations in cancer. *Mutat. Res.*, **462**, 247–253.
2. Croce,C.M., Erikson,J., ar-Rushdi,A., Aden,D. and Nishikura,K. (1984) Translocated c-*myc* oncogene of Burkitt lymphoma is transcribed in plasma cells and repressed in lymphoblastoid cells. *Proc. Natl Acad. Sci. USA*, **81**, 3170–3174.
3. Solnick,D. (1985) Trans splicing of mRNA precursors. *Cell*, **42**, 157–164.
4. Konarska,M.M., Padgett,R.A. and Sharp,P.A. (1985) Trans splicing of mRNA precursors *in vitro*. *Cell*, **42**, 165–171.
5. Bonen,L. (1993) Trans-splicing of pre-mRNA in plants, animals and protists. *FASEB J.*, **7**, 40–46.
6. Bruzik,J.P. and Maniatis,T. (1992) Spliced leader RNAs from lower eukaryotes are trans-spliced in mammalian cells. *Nature*, **360**, 692–695.
7. Bruzik,J.P. and Maniatis,T. (1995) Enhancer-dependent interaction between 5′ and 3′ splice sites in trans. *Proc. Natl Acad. Sci. USA*, **92**, 7056–7059.
8. Puttaraju,M., Jamison,S.F., Mansfield,S.G., Garcia-Blanco,M.A. and Mitchell,L.G. (1999) Spliceosome-mediated RNA trans-splicing as a tool for gene therapy. *Nat. Biotechnol.*, **17**, 246–252.
9. Li,B.L., Li,X.L., Duan,Z.J., Lee,O., Lin,S., Ma,Z.M., Chang,C.C., Yang,X.Y., Park,J.P., Mohandas,T.K. *et al.* (1999) Human acyl-CoA:cholesterol acyltransferase-1 (ACAT-1) gene organization and evidence that the 4.3-kilobase ACAT-1 mRNA is produced from two different chromosomes. *J. Biol. Chem.*, **274**, 11060–11071.
10. Caudevilla,C., Serra,D., Miliar,A., Codony,C., Asins,G., Bach,M. and Hegardt,F.G. (1998) Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. *Proc. Natl Acad. Sci. USA*, **95**, 12185–12190.
11. Nilsen,T.W. (1993) Trans-splicing of nematode premessenger RNA. *Annu. Rev. Microbiol.*, **47**, 413–440.
12. Hyde,M., Block-Alper,L., Felix,J., Webster,P. and Meyer,D.I. (2002) Induction of secretory pathway components in yeast is associated with increased stability of their mRNA. *J. Cell Biol.*, **156**, 993–1001.
13. Magrangeas,F., Pitiot,G., Dubois,S., Bragado-Nilsson,E., Cherel,M., Jobert,S., Lebeau,B., Boisteau,O., Lethe,B., Mallet,J. *et al.* (1998) Cotranscription and intergenic splicing of human galactose-1-phosphate uridylyltransferase and interleukin-11 receptor alpha-chain genes generate a fusion mRNA in normal cells. Implication for the production

Table 6. Chromosomal map location of the antisense chimeric sequences in the RefSeq databank

Chimeras[a]	5′ Partner[a]	Locus[b]	3′ Partner[a]	Locus[b]
X82540	NM_005538	17q13.1	XM_087061	1p32
U64876	M13555	5q33	U80802	9q33-q34.1
X13227	J03910	16q13	NM_001917	12q24
X77777	XM_113730	12p13	X75299	3p22
D49372	BC032589	15q13	NM_002986	17q21.1-q21.2
L10717	AB028969	4q31.1	NM_005546	5q31-32
AB045369	Y15059	11q24	NM_007232	20q13.33
M60725	AF036892	20q12	M60724	17q23.1
AF003522	L10335	14q21-q22	XM_035684	6q27-qter
M31520	U57847	1q21	NM_001026	10q22-q23
AB017111	NM_003933	16p13.3	XM_046457	20p11.2-p12
AF116719	BC010054	16q22.1	BC010913	11p15-pter
X73608	BC001466	22q13.2	NM_004598	5q31
AF031379	NM_006053	11q13.4-q13.5	XM_007328	14q21
X56465	NM_006601	12	NM_021998	Xq13-q21.1
L29126	NM_002055	17q21	NM_004067	7p15.3
U50079	BC015405	19q13.4	NM_004964	1p34
L35253	AF112299	12q14.3	AF286697	19p13.2
U90236	AB029290	1p34.3	NM_004999	6q13
AF152961	AF010403	12q12-q14	NM_007192	14q11.1
U90176	**AF160973**	**5q34**	**NM_004730**	**5q31.1**
AB032254	XM_085473	16q22.3	XM_048948	12q13
AF118065	**AF057352**	**3q28**	**S95936**	**3q21**
L40904	BC008572	16p13.3	NM_005037	3p25
J03909	AK055875	11q13	BC031020	19p13.1
U17899	AF054185	20q13.3-qter	U53454	11q13.5-q14
M34182	XM_030914	19q13.1	AJ001597	9q13
X52195	NM_017657	2p16.7	NM_001629	13q12
AF116625	AF031548	6p12.3	XM_056260	16p12-p11.2

[a]Accession number of the antisense chimeric sequences listed in Table 5 and of the corresponding fusion partners.
[b]Chromosomal map location of the fusion partners of the chimeric sequences. Sequences from the same chromosome are in bold.

of multidomain proteins during evolution. *J. Biol. Chem.*, **273**, 16005–16010.
14. Communi,D., Suarez-Huerta,N., Dussossoy,D., Savi,P. and Boeynaems,J.-M. (2001) Cotranscription and intergenic splicing of human P2Y11 and SSF1 genes. *J. Biol. Chem.*, **276**, 16561–16566.
15. Moore,R.C., Lee,I.Y., Silverman,G.L., Harrison,P.M., Strome,R., Heinrich,C., Karunaratne,A., Pasternak,S.H., Chishti,M.A., Liang,Y. *et al.* (1999) Ataxia in prion protein (PrP)-deficient mice is associated with upregulation of the novel PrP-like protein doppel. *J. Mol. Biol.*, **292**, 797–817.
16. Finta,C. and Zaphiropoulos,P.G. (2000) The human cytochrome P450 3A locus. Gene evolution by capture of downstream exons. *Gene*, **260**, 13–23.
17. Lehner,B., Williams,G., Campbell,R.D. and Sanderson,C.M. (2002) Antisense transcripts in the human genome. *Trends Genet.*, **18**, 63–65.
18. Vanhee-Brossollet,C. and Vaquero,C. (1998) Do natural antisense transcripts make sense in eukaryotes? *Gene*, **211**, 1–9.
19. Fahey,M.E., Moore,T.F. and Higgins,D.G. (2002) Overlapping antisense transcription in the human genome. *Comp. Funct. Genom.*, **3**, 244–253.
20. Brakenhoff,R.H., Schoenmakers,J.G. and Lubsen,N.H. (1991) Chimeric cDNA clones: a novel PCR artifact. *Nucleic Acids Res.*, **19**, 1949.
21. Chang,J. and Taylor,J. (2002) *In vivo* RNA-directed transcription, with template switching, by a mammalian RNA polymerase. *EMBO J.*, **21**, 157–164.
22. Rabbitts,T.H. (1994) Chromosomal translocations in human cancer. *Nature*, **372**, 143–149.
23. Chetverin,A.B., Chetverina,H.V., Demidenko,A.A. and Ugarov,V.I. (1997) Nonhomologous RNA recombination in a cell-free system: evidence for a transesterification mechanism guided by secondary structure. *Cell*, **88**, 503–513.
24. Chetverina,H.V., Demidenko,A.A., Ugarov,V.I. and Chetverin,A.B. (1999) Spontaneous rearrangements in RNA sequences. *FEBS Lett.*, **450**, 89–94.
25. Gmyl,A.P., Belousov,E.V., Maslova,S.V., Khitrina,E.V., Chetverin,A.B. and Agol,V.I. (1999) Nonreplicative RNA recombination in poliovirus. *J. Virol.*, **73**, 8958–8965.
26. Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.
27. Aebi,M., Hornig,H., Padgett,R.A., Reiser,J. and Weissmann,C. (1986) Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. *Cell*, **47**, 555–565.
28. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.
29. Kristiansen,T.Z. and Pandey,A. (2002) A database of transcriptional start sites. *Trends Biochem. Sci.*, **27**, 174.
30. Lewin,B. (2000) *Genes VII*. 8th Edn. Wiley, New York.
31. Pesole,G., Liuni,S., Grillo,G., Licciulli,F., Mignone,F., Gissi,C. and Saccone,C. (2002) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res.*, **30**, 335–340.
32. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
33. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
34. Zhang,Y., Strissel,P., Strick,R., Chen,J., Nucifora,G., Le Beau,M.M., Larson,R.A. and Rowley,J.D. (2002) Genomic DNA breakpoints in AML1/RUNX1 and ETO cluster with topoisomerase II DNA cleavage and DNase I hypersensitive sites in t(8;21) leukemia. *Proc. Natl Acad. Sci. USA*, **99**, 3070–3075.
35. Zhuo,D., Zhao,W.D., Wright,F.A., Yang,H.Y., Wang,J.P., Sears,R., Baer,T., Kwon,D.H., Gordon,D., Gibbs,S. *et al.* (2001) Assembly, annotation and integration of UNIGENE clusters into the human genome draft. *Genome Res.*, **11**, 904–918.
36. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.