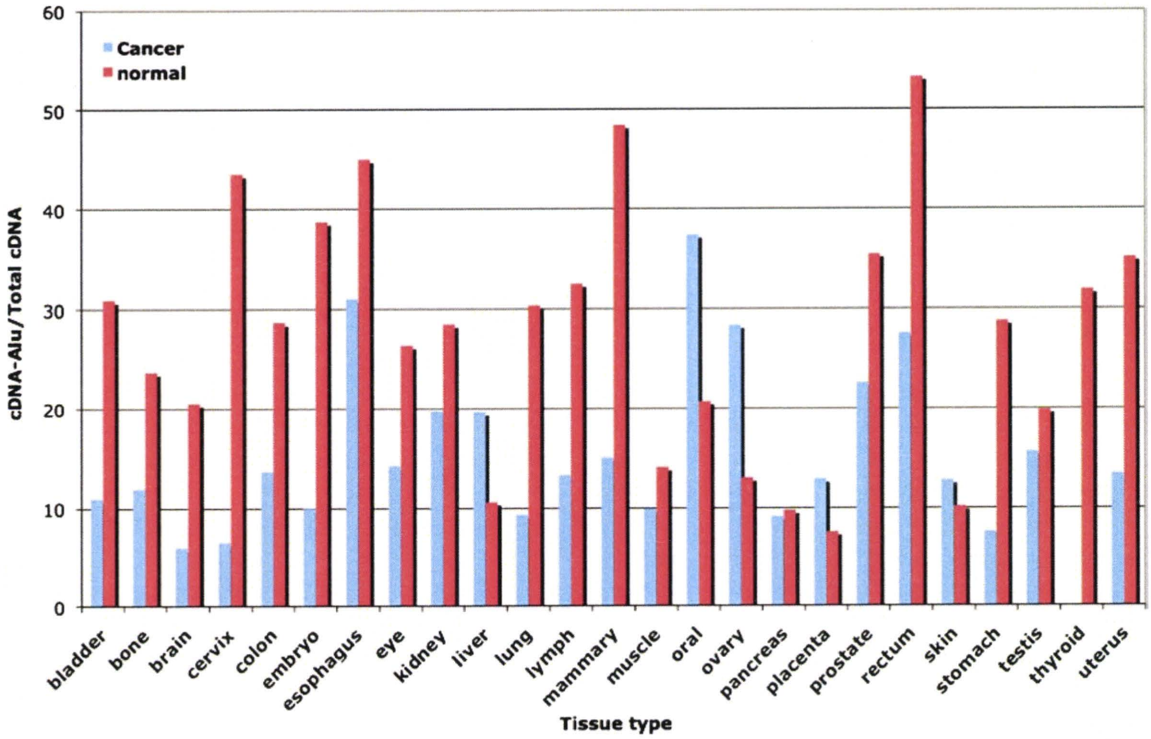


Fig. 1 The frequency of Alu-containing transcript (Alu-transcript) in human normal and cancer tissue types. The frequency of Alu-transcript per transcript sequences represented in normal (*blue bars*) or cancerous tissues (*red bars*). The statistical significance of the differences between the normal and cancerous tissues is indicated by the symbols *single asterisk*, *double asterisk* and *commercial at* next to the tissue names. Statistical significance Fisher's exact $P<0.0001$, *single asterisk* $P<0.001$, *double asterisk* $P<0.005$, *triple asterisk* $P<0.01$, *commercial at* not significant

functional categories and then assigned into 35 GO Slim categories for the Homo sapiens taxa identifier 9636 (http://www.geneontology.org/GO.downloads.shtml#ont). Over 70% of the Alu-transcript loci grouped into binding and catalytic activity molecular functions (Supplementary Figure 1), and more than 80% were involved in biological processes, such as physiological process, nucleotide metabolism, cell communication and transport. Less than 3% of the transcript and Alu-transcript loci were designated 'molecular function unknown' or 'biological process unknown'. The molecular function, biological process and cellular component of loci unique to tissue traits for the full transcript and Alu-transcript subsets showed no Alu-transcript affect on the GO slims (http://www.geneontology.org/GO_slims/) category proportions. The proportion of loci unique to tissue traits for a few molecular functions did however vary between unique normal and cancer tissues for all transcripts and Alu-transcript. The molecular functions of both transcript and Alu-transcript were associated mainly with binding (40%) and catalytic activity (20%). The number of loci products involved in the function of nucleic acid binding increased nearly twofold for the disease loci, whereas loci associations with signal transducer activity and transporter activity decreased in the cancerous trait three- and twofold, respectively. These shifts in function were significant by a Fisher's exact test at $P<0.0001$ (Supplementary Figure 1).

Alu families and subfamilies

When Alu elements were categorised into their particular families and subfamilies, the contribution of the Alu families in transcripts was similar to the proportion of Alu families within the genome: Alu-S 54%, Alu-J 26%, Alu-Y 10%, monomeric 8% and Alu 2% (Price et al. 2004). The most frequent Alu subfamily was AluSx. No family bias was detected between the cancerous and normal tissue sets.

Alu locations within the transcript

The relative position and size of Alu insertions within the transcript sequences was determined for 17,819 of 17,861 distinct transcript sequences with an annotated CDS that contained at least one Alu. The Alu positions were partitioned into three groups according to the regions in which they were found: (1) overlapping the CDS, (2) exclusively in the 5′ UTR or (3) exclusively in the 3′ UTR. Supplementary Figure 2 shows the number of fragments and distinct transcripts in normal or cancerous tissues, and the transcripts with known functions are indicated in red.

The transcript Alu content was measured based on the proportion of transcripts containing an Alu fragment represented within the CDS, 5′ UTR or 3′ UTR or represented by sequence length (base pairs) to the total sum of sequence represented in these three regions (Table 2).

Alu sequence quantification within transcripts

Greater than 28,000 Alu fragmented or full-length elements were found in the 17,000 Alu-transcript sequences surveyed, at an average of 1.6 Alu per sequence. The Alu elements ranged in size from ten nucleotides in length to full-sized. All these Alu fragments were within the RepeatMasker recommended RM score to be outside the threshold for a false positive match (http://www.repeatmasker.org/). Supplementary Figure 3a shows the number of Alu fragments plotted as a percentage of the consensus Alu sequence length for transcripts from cancerous and normal tissues, with known and unknown (hypothetical) functions. The plots show that the number of Alu-transcript in normal tissues is clearly higher than in cancerous tissues. In the Alu-transcript plot, minor peaks represent a slightly increased number of the monomeric form of the Alu structure (110–130 bp), and major peaks above 85% of the Alu consensus full-length sequence represent a greatly increased number of the dimeric form of the Alu structure (>250 bp). This trend of minor and major peaks was similar for the Alu-transcript in both the normal and diseased tissues. Supplementary Figure 3b shows the number of Alu fragments plotted as a percentage of the consensus Alu sequence length for transcripts from the different Alu families; Alu-S repeats are highly represented by lengths greater than 85%.

While multiple fragments of Alu elements can overlap the CDS within a single transcript, the largest number of Alu contained in the CDS was four copies in one particular transcript sequence. Full-length dimeric Alu (>280 bp) fragments overlapping the CDS that represent half of the transcripts potentially have cryptic splicing sites provided by the Alu sequences.

In an examined subset, the majority of the transcript isoforms containing full-length Alu within the coding regions of the transcripts were categorised as hypothetical proteins rather than known proteins. As the function of these Alu-transcripts is not known or is labelled 'hypothetical', it is proposed that they are either untranslated or destined for rapid degradation or they serve a binding role against excessive retrotranspositional activity within the genome (Table 1 in Supplementary document 1).

Incorporation of Alu elements as part of transcribed genes

The size of the Alu within transcripts is surprisingly mostly full-length (Supplementary Figure 3) without ORF disruption. The high prevalence of expressed full-length Alu, greater than 85% of the full element, in more than half available loci, supports the view that the Alu within transcripts might have some important function rather than random insertions within transcripts with no molecular or biological function (Chu et al. 1998; Kim et al. 2007; Paz et al. 2007). One important role for the complementary Alu-transcript sequences may be to participate in RNA sense/antisense hybridisations (Hagan et al. 2003) or dilution effects to protect the genome from unbridled Alu retrotransposition events by inhibiting or limiting the expression of the active Alu master copies involved with Alu retrotransposition. Overall, Alu within transcripts contribute significantly to various characteristics and patterns in transcriptome activity.

Some examples of genes that have transcripts with Alu sequences embedded in the 3′ UTR are glucose-6-phosphatase, placental alkaline phosphatase and receptors for platelet activating factor, vitamin D, interferon-alpha,

Table 2 The proportion of Alu length and transcript number represented in available transcript in the 5′ untranslated region (UTR), coding sequence and 3′ UTR regions

	Total transcript	Normal	Cancer	Alu-transcript	%	Normal Alu-transcript	%	Cancer Alu-transcript	%
Count transcript sequences represented in each region									
5′UTR	99,671	67,868	31,803	6,585	6.61	5,619	8.28	966	3.04
CDS	106,415	72,645	33,770	3,930	3.69	3,203	4.41	727	2.15
3′UTR	101,292	69,163	32,129	14,228	14.05	11,224	16.23	3,004	9.35
Sum transcript (bp) represented in each region									
5′UTR	35,897,377	29,061,960	6,835,417	1,716,495	4.78	1,526,474	5.25	190,021	2.78
CDS	103,506,598	70,328,008	33,178,590	507,946	0.49	426,050	0.61	81,896	0.25
3′UTR	82,313,000	61,388,458	20,924,542	4,250,911	5.16	3,445,289	5.61	805,622	3.85

epidermal growth factor and various other cytokines and interleukins. Consistently a larger proportion of Alu-transcript is found in those mRNA derived from normal than cancerous tissues regardless of the Alu sequence position within the mRNA (Table 2). Some of the genes with Alu in their 5′ UTR are METTL8, LIPT1, BDKRB1, GBA, KCNH5 and SLC39A1. The orientation of Alu fragment within the transcript is predominately antisense in the 5′ UTR and CDS regions and sense in the 3′ UTR.

In an earlier study of 87 Alu-transcripts, only four Alu were found to overlap the CDS (Yulug et al. 1995). We found in our review of the H-Inv transcript database that 3,639 of the 17,861 Alu-transcript sequences contained at least one Alu fragment that overlapped with the CDS. The distribution of the Alu within the CDS was determined as any overlap in the 5′ end of the CDS or the 3′ end of the CDS, spanning the entire CDS or contained entirely within the CDS (internal). From these distributions, over 200 transcripts contained an Alu that spanned the entire CDS.

The impact of Alu elements within the transcript UTR

The relative proportion of Alu that overlapped the CDS in our study was increased approximately tenfold above the numbers previously provided (Yulug et al. 1995), and over half of the Alu sequences overlapping the CDS were greater than 85% of full-length elements. A significantly higher number of Alu sequences were present in the 5′ and 3′ UTR regions than the CDS (Table 2). Therefore, the Alu location within the transcript is strongly biassed to the 5′ and 3′ UTR end of the transcript (Table 2). This finding is in contrast to an expressed sequence tag (EST) study on cell lines, where 82% of TEs that are found within the EST derived from cancerous tissues were located in the CDS (Kim et al. 2007). The Alu sequences within the 3′ UTR may affect mRNA stability or degradation by adenosine to inosine (A-to-I) editing in Alu sequences (Levanon et al. 2004) or by contributing adenine and uracil-rich elements to the transcript (An et al. 2004). This Alu-associated RNA editing is a potential mechanism for marking non-standard transcripts for degradation rather than for translation (Kim et al. 2004). The Alu sequences within the 3′ UTR might also affect translational efficiency by providing secondary and tertiary structures to hinder translational editing or translational rates (An et al. 2004; Hagan et al. 2003). The Alu sequences within the 5′ UTR of transcripts provide cryptic promoter sites, steroid binding sites or other regulatory elements and also secondary or tertiary structures that may hinder or enhance translation of the transcript (Hagan et al. 2003). The BRCA1 Alu-rich gene, an anti-oncogene involved with a hereditary predisposition to ovarian and breast cancer, is a well-studied example of a gene that expresses variable forms of a transcript with and without an Alu insertion in its 5′ UTR sequence (Sobczak and Krzyzosiak 2002). Some of the genes with Alu in their 5′ UTR sequences, such as METTL8, LIPT1, BDKRB1, GBA, KCNH5 and SLC39A1, however, might exploit the same Alu regulatory mechanisms as BRCA1 in disease (Sobczak and Krzyzosiak 2002).

The impact of Alu elements within transcribed exons

The Alu sequences within genes can contribute to different transcriptional isoforms by providing intron–exon recognition sites, exonisation (Claverie-Martin et al. 2005; Lei et al. 2005; Lei and Vorechovsky 2005). Approximately 5% of alternately spliced internal exons in the human genome were found to originate from an Alu sequence (Sorek et al. 2002). Most Alu-derived exons are alternately spliced, and only a segment of the Alu contributes to the new ORF (Makalowski 2003). In this review, we found that 16% of the Alu-transcripts provided potential splice sites within CDS. CDS from some functioning human genes exist that are almost entirely derived from Alu elements, such as the AD7C gene that encodes a neuronal thread protein and has 99% of its transcript composed of four to five Alu fragment elements (Britten 2004; De La Monte et al. 1997), although the validity of these findings has been questioned on the basis of EST and genomic sequence analysis (Kriegs et al. 2005). It was further argued that functional proteins are unlikely to contain transposable cassettes derived from young TEs, but if so, then their role is probably limited to regulatory functions (Gotea and Makalowski 2006). Some of the genes previously found to have Alu sequences contributing to exonisation include ADARB1, DSCRB8, ITCH, CDK5RAP1 (Dagan et al. 2004), RPE2-1, C-rel-2, MTO1-3 and PKP2b-4 (Krull et al. 2005). Potentially, numerous new cryptic splice sites exist in the human transcriptome (Yura et al. 2006), and many of them await full structural and functional characterization.

The CDS coverage by Alu is nearly 40% in over a half of the Alu-transcript with an Alu size greater than 280 bp. Of the 3,639 transcripts, 2,473 were annotated as hypothetical proteins. To determine if the Alu content changed the location or structure of the ‘normal’ CDS, a subset of transcript with full-length Alu sequences internal to the CDS were further investigated. In this subset, 85% of the Alu-transcript was annotated as translating hypothetical products with no known function. The remainder was either identical to the known genes glucose-inhibited division protein A family protein and penicillin-binding protein, dimerisation domain-containing protein or similar to the genes N4BP2, Alpha-COP, A1BG, dNT-1, GPI transamidase component PIG-U, RPIP8, PAOX, programmed cell death 6, enkurin, CRL2, G protein-coupled receptor 43, selenoprotein N precursor, CTAGE family member 5 isoform and a solute carrier family 24 (sodium/potassium/calcium exchanger), member 5 (Table 1 in Supplementary document 1).

The variation in the A1BG gene transcripts provided an example of how the Alu content of the transcripts might change their structure, function and tissue specificity of expression. The protein encoded by the eight exons of the A1BG gene is a plasma alpha-1 glycoprotein with sequence similarity to the variable regions of some immunoglobulin supergene family member proteins. A1BG interacts in the plasma with the cysteine-rich secretory protein 3 that is secreted by neutrophilic granulocytes, and it is believed to play a role in innate immunity. Some variants of the A1BG transcripts were transformed by the Alu insertions from full-length coding forms (1,645–1,810 bp) composed of eight exons into longer transcripts (1,951–3,466 bp), usually with shorter ORF of one to three exons in the coding region. The Alu-containing A1BG transcripts were found in the amygdala, cerebellum and tetracarcinoma, while the Alu-free A1BG transcripts were expressed by foetal and adult livers, primary hepatoblastoma and ovary.

Alu-siRNA-mediated feedback model in cancer

The transcriptome-wide survey of human transcript, carried out in this review, represented nearly 30,000 gene loci, which is close to the full complement of known gene loci of the human genome. Overall, 13,240 loci (44%) expressed a transcript that contains a partial or full-length Alu sequence. Alu fragments are far more abundant in the transcriptome than previously reported. No discernable bias for Alu families was identified between the normal and cancerous tissues, which confirmed that the Alu families in the transcriptome reflect the proportion of Alu families in the genome. For all the total transcript analysed, 17% contain an Alu fragment which is four times greater than previous estimates of Alu-transcript content made from smaller data sets and based on different search methods (Yulug et al. 1995).

Loci-based transcriptome investigation of Alu sequence has highlighted that Alu fragments are proportionally more abundant in normal tissue transcript than the corresponding cancerous tissues, many with hypothetical functions. In our feedback model (Fig. 2), we propose that a high proportion of the non-functional (hypothetical) Alu-transcript expressed in normal tissues or during cellular stress and infection does not undergo degradation, whereas the Alu-transcripts are degraded in cancerous tissues due to increased Alu small RNA activity.

Loss of genome-wide methylation is a common feature of cancer. It has been hypothesised that DNA methylation initially evolved as a defence mechanism against viral and other DNA pathogens as a way to silence foreign DNA sequences (Bird 1993; Liu et al. 1994; Yoder et al. 1997) This is consistent with the observation that LINE and SINE (Alu) TEs are heavily methylated in normal cells (Richards et al. 2009). Bollati et al. (2009) found that global hypomethylation of L1, Alu and SAT-alpha is significantly associated with tumour progression in *Mus musculus* and may contribute toward a more extensive stratification of the disease (Bollati et al. 2009). Our model proposes that in cancer tissues, the hypomethylation of intergenic Alu-loci may result in transcription of these elements to generate Alu-dsRNA molecules that are further cleaved into siRNAs. The generated Alu-derived siRNA (Alu-siRNAs) guide the RNA silencing complex to Alu-containing mRNAs and mediate their post-transcriptional degradation (Fig. 2).

Methylation and chromatin structure together play a role between retroelements and their host. Hypomethylation and expression in developing germ cells opens a 'window of opportunity' for retrotransposition and recombination that contribute to human evolution, but also inherited diseases. In somatic cells, the presence of retroelements may be exploited to organise the genome into active and inactive regions, to separate domains and functional regions within one chromatin domain, to suppress transcription noise and to regulate transcript stability. Retroelements, particularly Alu, may fulfil physiological and protective roles during responses to stress and infections (Schulz et al. 2006).

In support of our model, it has been previously reported that reduction of the methylation index of L1 and Alu following treatment of three lung cell lines with 5-aza-2′-deoxycitidine consistently resulted in increased expression of both elements. This study demonstrated the strong link between hypomethylation of TEs with genomic instability in non-small cell lung cancer and provides early evidence for a potential active role of these elements in lung neoplasia. As demethylating agents are now entering lung cancer trials, it was viewed imperative to gain a greater insight into the potential reactivation of silent retrotransposons in order to advance the clinical utilisation of epigenetics in cancer therapy (Daskalos et al. 2009).

To investigate the role of miRNAs on epigenetic therapy of gastric cancer, the miRNA expression profile was analysed in human gastric cancer cells treated with 5-aza-20-deoxycytidine (5-Aza-CdR) and 4-phenylbutyric acid (PBA). Microarray miRNA analysis shows that most of miRNAs activated by 5-Aza-CdR and PBA in gastric cancer cells are located at Alu repeats on chromosome 19 (Saito et al. 2009). Analyses of chromatin modification showed that DNA demethylation and HDAC inhibition at Alu repeats activates silenced miR-512-5p by RNA polymerase II. These results suggest that chromatin remodelling at Alu repeats

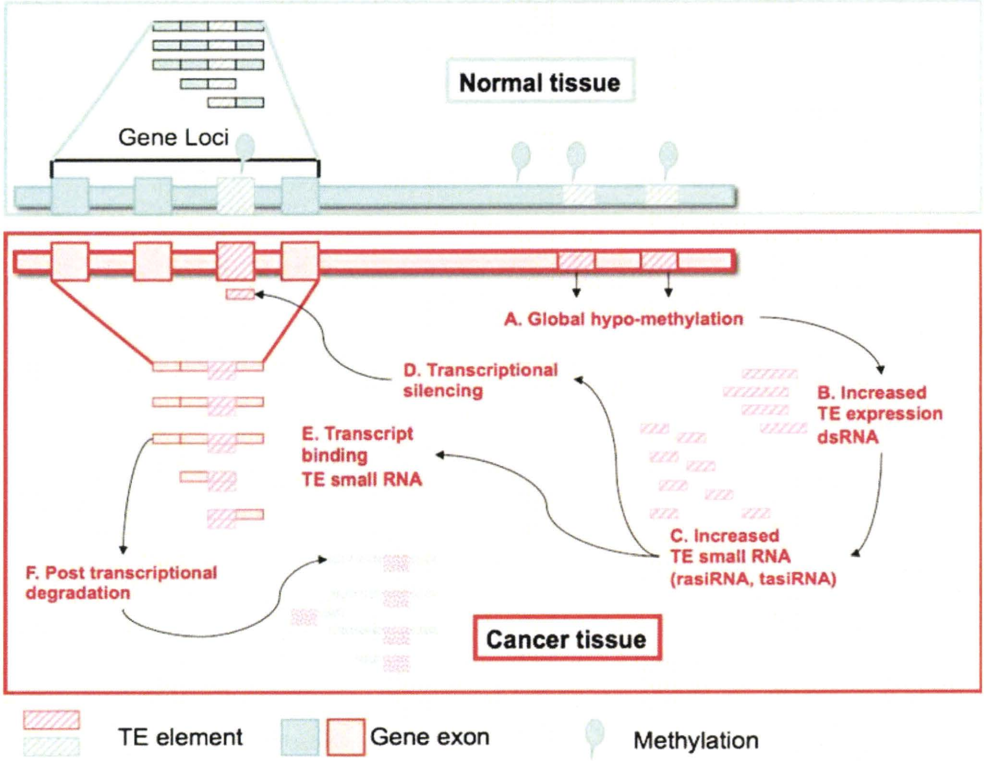


Fig. 2 Feedback model for Alu-mediated siRNA interference of Alu-containing transcripts in cancerous tissues. In the feedback model, cancer transposable element (TE) transcripts are suppressed by a feedback mechanism due to increased TE expression from genomic hypomethylation. The genome (*solid*), TE (*diagonal shaded*) and loci show the TE-containing transcripts sourced from normal tissues (*boxed section in blue above the genomic regions*) and loci transcripts sourced from cancerous tissue (*boxed sectioned in red below the genome*). In the cancer feedback model, the genomic hypomethylation (**a**) and the increase in TE sequence (**b**) may result in the formation of Alu-dsRNAs, which in turn are processed further into siRNAs (Alu-siRNAs; **c**). Alu-siRNAs guide the RNA silencing complex to Alu-containing transcripts to mediate their post-transcriptional down-regulation. The small RNA complementation to loci-containing TE may then block transcription and the production of mRNA (**d**) and/or post-transcriptional-binding TE-containing transcripts (**e**) leading to their degradation (**f**)

plays critical roles in the regulation of miRNA expression and that epigenetic activation of silenced Alu-associated miRNAs could be a novel therapeutic approach for gastric cancer (Saito et al. 2009).

Hypomethylation of the genome largely affects the intergenic and intronic regions of the DNA, particularly repeat sequences such as TEs, and believed to result in chromosomal instability and increased mutation events. It has been considered that global demethylation of repeat sequences including TEs and the site-specific hypomethylation of certain genes might contribute to the deleterious effects that ultimately result in the initiation and progression of cancer and other diseases (Wilson et al. 2007). In our model, the increased hypomethylation of TEs is shown in the lower section of Fig. 2 highlighted in red to represent the cancerous tissue state, while normal state is represented in the upper half of the diagram in blue.

In the feedback model, genomic hypomethylation (Fig. 2) and the increase in TE sequences are shown correlated with the increased production of small RNA. The small RNA complementation to loci-containing TE may then silence transcription and the production of mRNA (rasiRNA) and/or post-transcriptional-binding TE-containing transcripts (tasiRNA) leading to their degradation, mRNA cleavage.

Mature miRNA sequences of approximately 50 additional human miRNAs have been shown to lie within Alu and other known repetitive elements, extending the current view of miRNA origins and the transcriptional machinery driving their expression (Borchert et al. 2006). Further, base-pair complementation can be demonstrated between the seed sequence of a subset of human miRNAs and Alu repeats that are integrated parallel (sense) in mRNAs (Lehnert et al. 2009).

The proportion of Alu-transcript was significantly higher in 18 of the 22 normal tissue types than in the corresponding cancerous tissues (Fig. 1). In general, these tissue results are consistent with previous studies that reported that the RNA embedded with Alu showed variable patterns between tissue types, but that there was an overall twofold increase of edited Alu-transcript in normal cells than malignant cells (Maas et al. 2001). Paz et al. (2007) identified significant hypoediting of the Alu elements within the transcripts of tumours from the brain, prostate, lung, kidney and the testis and suggested that A-to-I RNA editing was an epigenetic mechanism relevant to cancer development and progression (Paz et al. 2007). In their tissue analysis, they found that the placental tissue was an exception in that there was more editing of Alu within the transcripts of cancerous than normal tissue. Recently based on the results obtained for nine different editing sites, it was determined that RNA editing is an epigenetic mechanism that does not participate in the evolution of urinary bladder cancer (Zilberman et al. 2009). A significantly higher proportion of Alu-transcript to transcript in normal tissues than in the corresponding cancerous tissues of the bladder may suggest another mechanism for bladder cancers. Further, there is a significantly higher proportion of Alu-transcript to transcript in cancerous tissues than in the corresponding normal tissues of the liver, oral cavity, ovary and placenta. This may suggest that the Alu-transcript oncogenic affect may occur in most, but not all, tissues and as reported may not be completely reliant on A–I editing in certain cases (Supplementary document 1).

The hybridised complementation between the small RNA and the repeat containing transcripts in hypomethylated diseased tissue may lead to reduced RNA Alu-transcript levels due to their degradation and/or interference (Fig. 2).

In this review, it is proposed that TEs influence gene regulation/expression on both the transcriptional and post-transcriptional level. TEs are further found to be an abundant source for small RNA and hence potential gene interference activity. In reviewing the available human transcript data in H-Inv, the Alu element was found with a much higher abundance in transcript than previously reported, serving as possible gene targets for repeat derived small RNA (Alu-siRNAs). Due to the curated nature of H-Inv, transcripts could be assigned to their tissue source, disease condition and function. Transcripts from cancerous tissues revealed an under-representation of transcript containing Alu elements, and the majority of full-length Alu-transcripts derived from normal tissues were highly represented by hypothetical proteins. It is proposed that Alu-derived small RNAi activity increases in certain cancerous tissue with increased genome hypomethylation, thus suppressing the abundance of Alu-derived transcripts.

Acknowledgments We wish to acknowledge Professor Rudi Appels for his guidance in compiling this review. The H-Inv human transcript data was analysed mostly by Paula Moolhuijzen as part of her studies for a Ph.D. The Australian Research Council (ARC) grant number DP0559547 provided financial support. The authors are grateful to Mr. Adam Hunter (CCG) and Chisato Yamasaki (Integrated Database Group, Japan Biological Information Research Centre, JBIRC) for facilitating access to a broad range of databases.

References

An HJ, Lee D, Lee KH, Bhak J (2004) The association of Alu repeats with the generation of potential AU-rich elements (ARE) at 3′ untranslated regions. BMC Genomics 5:97–102

Apweiler R, Attwood TK, Bairoch A (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucl Acids Res 29:37–40

Aravin AA, Hannon GJ, Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318:761–764

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29

Bird AP (1993) Functions for DNA methylation in vertebrates. Cold Spring Harb Symp Quant Biol 58:281–285

Bollati V, Fabris S, Pegoraro V, Ronchetti D, Mosca L, Deliliers GL, Motta V, Bertazzi PA, Baccarelli A, Neri A (2009) Differential repetitive DNA methylation in multiple myeloma molecular subgroups. Carcinogenesis 30:1330–1335

Borchert GM, Lanier W, Davidson B (2006) RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol 13:1097–1101

Britten RJ (1996) Cases of ancient mobile element DNA insertions that now affect gene regulation. Mol Phylogenet Evol 5:13–17

Britten RJ (2004) Coding sequences of functioning human genes derived entirely from mobile element sequences. Proc Natl Acad Sci 101:16825–16830

Britten RJ, Davidson EH (1969) Gene regulation for higher cells: a theory. Science 165:349–357

Britten RJ, Davidson EH (1971) Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q Rev Biol 46:111–138

Chandler V, Alleman M (2008) Paramutation: epigenetic instructions passed across generations. Genetics 178:1839–1844

Chandler VL, Stam M (2004) Chromatin conversations: mechanisms and implications of paramutation. Nat Rev Genet 5:532–544

Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet 8:93–103

Chu WM, Ballard R, Carpick BW, Williams BRG, Schmid CW (1998) Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR. Mol Cell Biol 18:58–68

Claverie-Martin F, Flores C, Anton-Gamero M, Gonzalez-Acosta H, Garcıa-Nieto V (2005) The Alu insertion in the CLCN5 gene of a patient with Dent's disease leads to exon 11 skipping. J Hum Genet 50:370–374

Dagan T, Sorek R, Sharon E, Ast G, Graur D (2004) AluGene: a database of Alu elements incorporated within protein-coding genes. Nucleic Acids Res 32:D489–D492

Daskalos A, Nikolaidis G, Xinarianos G, Savvari P, Cassidy A, Zakopoulou R, Kotsinas A, Gorgoulis V, Field JK, Liloglou T (2009) Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. Int J Cancer 124:81–87

De La Monte SM, Ghanbari K, Frey WH, Beheshti I, Averback P, Hauser SL, Ghanbari HA, Wands JR (1997) Characterization of the AD7c-NTP cDNA expression in Alzheimer' disease and measurement of a 41-kD protein in cerebrospinal fluid. J Clin Invest 100:3093–3104

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397–405

Gotea V, Makalowski W (2006) Do transposable elements really contribute to protease? Trends Genet 22:260–267

Grewal SI, Jia S (2007) Heterochromatin revisited. Nat Rev Genet 8:35–46

Hagan CR, Sheffield RF, Rudin CM (2003) Human Alu element retrotransposition induced by genotoxic stress. Nat Genet 35:219–220

Hambor JE, Mennone J, Coon ME, Hanke JH, Kavathas P (1993) Identification and characterization of an Alu-containing, T-cell-specific enhancer located in the last intron of the human CD8 alpha gene. Mol Cell Biol 13:7056–7070

Hasler J, Strub K (2006) Alu elements as regulators of gene expression. Nucleic Acids Res 34:5491–5497

Houck CM, Rinehart FP, Schmid CW (1979) A ubiquitous family of repeated DNA sequences in the human genome. J Mol Biol 132:289–306

Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, Yura K (2004) Integrative annotation of 21, 037 human genes validated by full-length cDNA clones. PLoS Biol 2:856–875

Jamalkandi SA, Masoudi-Nejad A (2009) Reconstruction of *Arabidopsis thaliana* fully integrated small RNA pathway. Funct Integr Genomics 9:419–432

Johnson R, Gamblin RJ, Ooi L, Bruce AW, Donaldson IJ, Westhead DR, Wood IC, Jackson RM, Buckley NJ (2006) Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. Nucleic Acids Res 34:3862–3877

Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20:119–121

Kapitonov V, Jurka J (1996) The age of Alu subfamilies. J Mol Evol 42:59–65

Kim VN (2005) Small RNAs: classification, biogenesis, and function. Mol Cells 19:1–15

Kim DDY, Kim TTY, Walsh T, Kobayashi Y, Matise TC, Buyske S, Gabriel A (2004) Widespread RNA editing of embedded Alu elements in the human transcriptome. Genome Res 14:1719–1725

Kim DS, Huh JW, Kim HS (2007) Transposable elements in human cancers by genome-wide EST alignment. Genes Genet Syst 82:145–156

Kondo Y, Issa JP (2003) Enrichment for histone H3 lysine 9 methylation at Alu repeats in human cells. J Biol Chem 278 (30):27658–27662

Kriegs JA, Schmitz J, Makalowski W, Brosius J (2005) Does the AD7c-NTP locus encode a protein? BBA Gene Struct Expr 1727:1–4

Krull M, Brosius J, Schmitz J (2005) Alu-SINE exonization: en route to protein-coding function. Mol Biol Evol 22:1702–1711

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lehnert S, Van Loo P, Thilakarathne PJ, Marynen P, Verbeke G, Schuit FC (2009) Evidence for co-evolution between human microRNAs and Alu-repeats. PLoS ONE 4:e4456

Lei H, Vorechovsky I (2005) Identification of splicing silencers and enhancers in sense Alus: a role for pseudoacceptors in splice site repression. Mol Cell Biol 24:6912–6920

Lei H, Day INM, Vorechovsky I (2005) Exonization of AluYa5 in the human ACE gene requires mutations in both 3′ and 5′ splice sites and is facilitated by a conserved splicing enhancer. Nucleic Acids Res 33:3897–3906

Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallengger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, Olshansky M, Rechavi G, Jantsch MF (2004) Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol 22:1001–1005

Lev-Maor G, Ram O, Kim E, Sela N, Goren A, Levanon EY, Ast G (2008) Intronic Alus influence alternative splicing. PLoS Genet 4:e1000204

Liu WM, Maraia RJ, Rubin CM, Schmid CW (1994) Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. Nucleic Acids Res 22:1087–1095

Maas S, Patt S, Schrey M, Rich A (2001) Underediting of glutamine receptor GluR-B mRNA in malignant gliomas. Proc Natl Acad Sci 98:14687–14692

Maas S, Rich A, Nishikura K (2003) A to I RNA editing: recent news and residual mysteries. J Biol Chem 278:1391–1394

Makalowski W (2000) Genomic scrap yard: how genomes utilize all that junk. Gene 259:61–67

Makalowski W (2003) Not junk after all. Science 300:1246–1247

Mattick JS (2007) A new paradigm for developmental biology. J Exp Biol 210:1526–1547

Medstrand P, van de Lagemaat LN, Dunn CA, Landry JR, Svenback D, Mager DL (2005) Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenet Genome Res 110:342–352

Morris KV, Chan SW, Jacobsen SE, Looney DJ (2004) Small interfering RNA-induced transcriptional gene silencing in human cells. Science 305:1289–1292

Novikova O (2009) Chromodomains and LTR retrotransposons in plants. Commun Integr Biol 2(2):158–162

Paz N, Levanon EY, Amariglio N, Heimberger AB, Ram Z, Constantini S, Barbash ZS, Adamsky K, Safran M, Hirschberg A, Krupsky M, Ben-Dov I, Cazacu S, Mikkelsen T, Brodie C, Eisenberg E, Rechavi G (2007) Altered adenosine-to-inosine RNA editing in human cancer. Genome Res 17:1586–1595

Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell 7:597–606

Piriyapongsa J, Marino-Ramirez L, Jordan IK (2007) Origin and evolution of human microRNAs from transposable elements. Genetics 176:1323–1337

Price AL, Eskin E, Pevzner PA (2004) Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Res 14:2245–2252

Quentin Y (1992a) Fusion of a free left Alu monomer and a free right Alu monomer at the origin of the Alu family in the primate genomes. Nucleic Acids Res 20:487–493

Quentin Y (1992b) Origin of the Alu family: a family of Alu-like monomers gave birth to the left and the right arms of the Alu elements. Nucleic Acids Res 20:3397–3401

Richards K, Zhang B, Baggerly K, Colella S, Lang J, Schuller D, Krahe R (2009) Genome-wide hypomethylation in head and neck cancer is more pronounced in HPV-negative tumors and is associated with genomic instability. PLoS ONE 4:e4941

Saito Y, Suzuki H, Tsugawa H, Nakagawa I, Matsuzaki J, Kanai Y, Hibi T (2009) Chromatin remodeling at Alu repeats by epigenetic treatment activates silenced microRNA-512-5p with downregulation of Mcl-1 in human gastric cancer cells. Oncogene 28:2738–2744

Schulz WA, Steinhoff C, Florl AR (2006) Methylation of endogenous human retroelements in health and disease. Curr Top Microbiol Immunol 3:211–250

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8:272–285

Smalheiser NR, Torvik VI (2005) Mammalian microRNAs derived from genomic repeats. Trends Genet 21:322–326

Smalheiser NR, Torvik VI (2006) Alu elements within human mRNAs are probable microRNA targets. Trends Genet 22:532–536

Smit A, Hubley R, Green P (1996–2004) RepeatMasker Open-3.0. http://www.repeatmasker.org

Sobczak K, Krzyzosiak WJ (2002) Structural determinants of BRCA1 translational regulation. J Biol Chem 277:17349–17358

Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. Genome Res 12:1060–1067

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka ED, Wilkinson M, Birney E (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12:1611–1618

Stam M, Belele C, Dorweiler JE, Chandler VL (2002) Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation. Genes Dev 16:1906–1918

Ullu E, Tschudi C (1984) Alu sequences are processed 7SL RNA genes. Nature 312:171–172

Vila MR, Gelpi C, Nicolas A, Morote J, Schwartz SJ, Schwartz S, Meseguer A (2003) Higher processing rates of Alu-containing sequences in kidney tumours and cell lines with overexpressed Alu-mRNAs. Oncol Rep 10:1903–1909

Wilson AS, Power BE, Molloy P (2007) DNA hypomethylation and human diseases. Biochim Biophys Acta 1775:138–162

Yamasaki C, Koyanagi KO, Fujii Y, Itoh T, Barrero R, Tamura T, Yamaguchi-Kabata Y, Tanino M, Takeda J, Fukuchi S, Miyazaki S, Nomura N, Sugano S, Imanishi T, Gojobori T (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). Gene 364:99–107

Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13:335–340

Yulug IG, Yulug A, Fisher EMC (1995) The frequency and position of Alu repeats in cDNAs, as determined by database searching. Genomics 27:544–548

Yura K, Shionyu M, Hagino K, Hijikata A, Hirashima Y, Nakahara T, Eguchi T, Shinoda K, Yamaguchi A, Takahashi K, Itoh T, Imanishi T, Gojobori T, Go M (2006) Alternative splicing in human transcriptome: functional and structural influence on proteins. Gene 380:63–71

Zhou YH, Zheng JB, Gu X, Saunders GF, Yung WK (2002) Novel PAX6 binding sites in the human genome and the role of repetitive elements in the evolution of gene regulation. Genome Res 12:1716–1722

Zilberman DE, Safran M, Paz N, Amariglio N, Simon A, Fridman E, Kleinmann N, Ramon J, Rechavi G (2009) Does RNA editing play a role in the development of urinary bladder cancer? Urol Oncol (in press)

doi:10.1093/nar/gkq1173

Towards BioDBcore: a community-defined information specification for biological databases

Pascale Gaudet[1,2,*], Amos Bairoch[1], Dawn Field[3], Susanna-Assunta Sansone[4], Chris Taylor[5], Teresa K. Attwood[6,7], Alex Bateman[8], Judith A. Blake[9], Carol J. Bult[9], J. Michael Cherry[10], Rex L. Chisholm[2], Guy Cochrane[5], Charles E. Cook[4], Janan T. Eppig[9], Michael Y. Galperin[11], Robert Gentleman[12,13], Carole A. Goble[7], Takashi Gojobori[14,15], John M. Hancock[16], Douglas G. Howe[17], Tadashi Imanishi[14], Janet Kelso[13,18], David Landsman[13], Suzanna E. Lewis[19], Ilene Karsch-Mizrachi[11], Sandra Orchard[5], B. F. Francis Ouellette[13,20], Shoba Ranganathan[21,22], Lorna Richardson[23], Philippe Rocca-Serra[4], Paul N. Schofield[24], Damian Smedley[5], Christopher Southan[25], Tin Wee Tan[22], Tatiana Tatusova[11], Patricia L. Whetzel[26], Owen White[27] and Chisato Yamasaki[14] on behalf of the BioDBCore working group

[1]Swiss Institute of Bioinformatics, CMU, 1 Rue Michel Servet, 1211 Geneva 4, Switzerland, [2]Feinberg School of Medicine, Northwestern University, Chicago, IL, 60611, USA, [3]NERC Center for Ecology and Hydrology, Oxford, OX1 3SR, [4]Oxford e-Research Centre, University of Oxford, Oxford, OX1 3QG, [5]European Molecular Biology Laboratory (EMBL) Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, [6]Faculty of Life Sciences, The University of Manchester, Manchester M13 9PT, [7]School of Computer Science, The University of Manchester, Manchester M13 9PT, [8]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK, [9]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, [10]Department of Genetics, Stanford University, Stanford, California 94305-5120, [11]NCBI, NLM, National Institutes of Health, Bethesda, MD 20894, [12]Genentech, 1 DNA Way, South San Francisco, CA 94080, USA, [13]DATABASE, The Journal of Biological Databases and Curation, Oxford University Press, Oxford OX2 6DP, UK, [14]Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi Koto-ku, Tokyo 135-0064, [15]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan, [16]MRC Harwell, Mammalian Genetics Unit, Harwell Science and Innovation Campus, Oxfordshire, OX11 0RD, UK, [17]The Zebrafish Model Organism Database, 5291 University of Oregon, Eugene, Oregon 97401-5291, USA, [18]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [19]Genomics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road 64R0121 Berkeley, California 94720, USA, [20]Ontario Institute for Cancer Research, Suite 800, 101 College Street, Toronto, Ontario, M5G 0A3, Canada, [21]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney NSW 2109, Australia, [22]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, [23]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh, EH4 2XU, [24]Department of Physiology, Development and Neuroscience, University of Cambridge, Doning Street, Cambridge CB2 3EG, UK, [25]ChrisDS Consulting, Göteborg, Sweden, [26]Stanford Center for Biomedical Informatics Research, National Center for Biomedical Ontology, Stanford University, Stanford, CA, 94305 and [27]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

Received and Accepted November 1, 2010

*To whom correspondence should be addressed. Tel: +41 22 379 5050; Fax: +41 22 379 5858; Email: pascale.gaudet@isb-sib.ch

This paper is also being published in the DATABASE journal, http://www.database.oxfordjournals.org/cgi/doi/10.1093/database/baq027

ABSTRACT

The present article proposes the adoption of a community-defined, uniform, generic description of the core attributes of biological databases, BioDBCore. The goals of these attributes are to provide a general overview of the database landscape, to encourage consistency and interoperability between resources and to promote the use of semantic and syntactic standards. BioDBCore will make it easier for users to evaluate the scope and relevance of available resources. This new resource will increase the collective impact of the information present in biological databases.

INTRODUCTION

The world of public biological databases is constantly evolving, as attested by the ever-growing size of the 'Nucleic Acids Research' (NAR) annual database issue and online Molecular Biology Database Collection, as well as by the creation of a new journal dedicated to databases and biocuration, 'DATABASE' (1,2). A wealth of new technologies is responsible for the exponential increase in the quantity, complexity and diversity of data generated in the life sciences. The need to store and share this data helps explain the explosion in the number and variety of resources that cater to the needs of biological research. Many researchers have commented that this increased volume of data has not yet yielded proportional improvements in biological knowledge (3–5). To a great extent this is owing to the widespread and unconnected distribution of data through databases scattered around the world. Clearly, adherence to open standards, as well as powerful and reliable tools, have become a necessity to support data sharing, integration and analysis (6). The available databases can be broadly placed into three categories: (i) archival repositories, (ii) curated resources, hence the rise of biocuration described in (7), and (iii) data integration warehouses. All three offer a range of querying and mining tools to explore the data and enable knowledge discovery. In addition, databases range from well-established repositories to burgeoning, innovative resources that cover emerging scientific areas or use novel technologies. While some databases are intended as long-term, consistently maintained community resources, others are intentionally temporary in nature, their existence being limited to the lifetime of the underlying grant or research project.

As in any emerging field, standardization across the biological databases is still inadequate at many levels. Consequently, there is still unnecessary and costly duplication of efforts, poor interoperability between resources and loss of valuable data and annotations when a resource is no longer supported. Most critically, the large number and variety of resources available are major hurdles for users, who are often unable to locate the resource(s) that best fits their specific needs. Even when appropriate resources are located, combining data from different resources can be a very difficult task. Having a uniform system for describing biological databases available in a single, centralized location would benefit both users and database providers: it would be much easier for users to find appropriate resources, while publicizing specialized resources and lesser known functionality of established databases more widely.

To address some of these issues we propose the adoption of a community-defined, uniform, generic description of the 'core attributes of biological databases', which we will name BioDBCore. Such minimum information checklists are now being developed for a wide range of data types. For example, the MIBBI (Minimum Information for Biological and Biomedical Investigations) portal [http://mibbi.org; (8)] contains over 30 MI checklists. BioDBCore will contain essential descriptors common to all databases.

GOALS OF THE BioDBCore ATTRIBUTES

The goals of the proposed BioDBCore checklist are to:

(i) Gather the necessary information to provide a general overview of the database landscape, and compare and contrast the various resources.
(ii) Encourage consistency and interoperability between resources.
(iii) Promote the uptake and use of semantic and syntactic standards.
(iv) Provide guidance for users when evaluating the scope and relevance of a resource, as well as details of the data access methods supported.
(v) Ensure that the collective impact of these resources is maximized.

This working group is open to all interested parties, and has started to collect a list of attributes of the BioDBCore checklist. Proposed core attributes are presented in Table 1. BioDBCore is registered with MIBBI, the umbrella organization that works to promote minimal information reporting in biomedical and biological research (8).

THE BioDBCore WORKING GROUP

To achieve widespread uptake and adoption of the BioBDCore guidelines, these recommendations must be developed as a community effort. To get the initiative started, we have formed a working group encompassing representatives from a wide range of existing life sciences resources. This includes representatives from MIBBI, editors from key journals publishing database descriptions, staff from model organism, sequences and protein databases, members of the Asia-Pacific Bioinformatics network (APBioNet, http://www.apbionet.org/), the Bioinformatics Links Directory (http://www.bioinformatics.ca/links_directory/) (9), developers from the ELIXIR survey of European databases and leaders of the Database Description Framework (DDF) from the CASIMIR project (10). One of the working group participants, APBioNet, has developed a framework for Minimum Information about a Bioinformatics Investigation (MIABi) (11) that aims to cover all aspects of bioinformatics studies. We plan to coalesce the BioDBCore with the relevant aspects of MIABi. This is an important opportunity to build a combined framework for advancing bioinformatics standards in a coordinated manner.

The BioDBCore checklist is overseen by the International Society for Biocuration (ISB) (http://biocurator.org/), in collaboration with the BioSharing forum [http://www.biosharing.org/, (12)]. The ISB was created in 2009 to promote and support the work of biocurators and bio-programmers. One of its goals is to foster interactions between these professionals to maximize the usefulness of all resources by encouraging the interoperability of databases and supporting data sharing. The BioSharing forum works at the global level to build stable linkages between funders, implementing data-sharing policies, and well-constituted standardization efforts in the biosciences domain to expedite communication and achieve harmonization and mutual support. A 'one-stop shop' portal is under development for those seeking data sharing policy documents and information about the standards (checklists, ontologies and file-formats), linking to exiting resources, such as MIBBI.

Table 1. Proposed core descriptors for inclusion in the BioDBCore specification

Proposed core descriptors for a biological database	
(1)	Database name
(2)	Main resource URL
(3)	Contact information (e-mail; postal mail)
(4)	Date resource established (year)
(5)	Conditions of use (free, or type of license)
(6)	Scope: data types captured, curation policy, standards used
(7)	Standards: MIs, Data formats, Terminologies
(8)	Taxonomic coverage
(9)	Data accessibility/output options
(10)	Data release frequency
(11)	Versioning policy and access to historical files
(12)	Documentation available
(13)	User support options
(14)	Data submission policy
(15)	Relevant publications
(16)	Resource's Wikipedia URL
(17)	Tools available

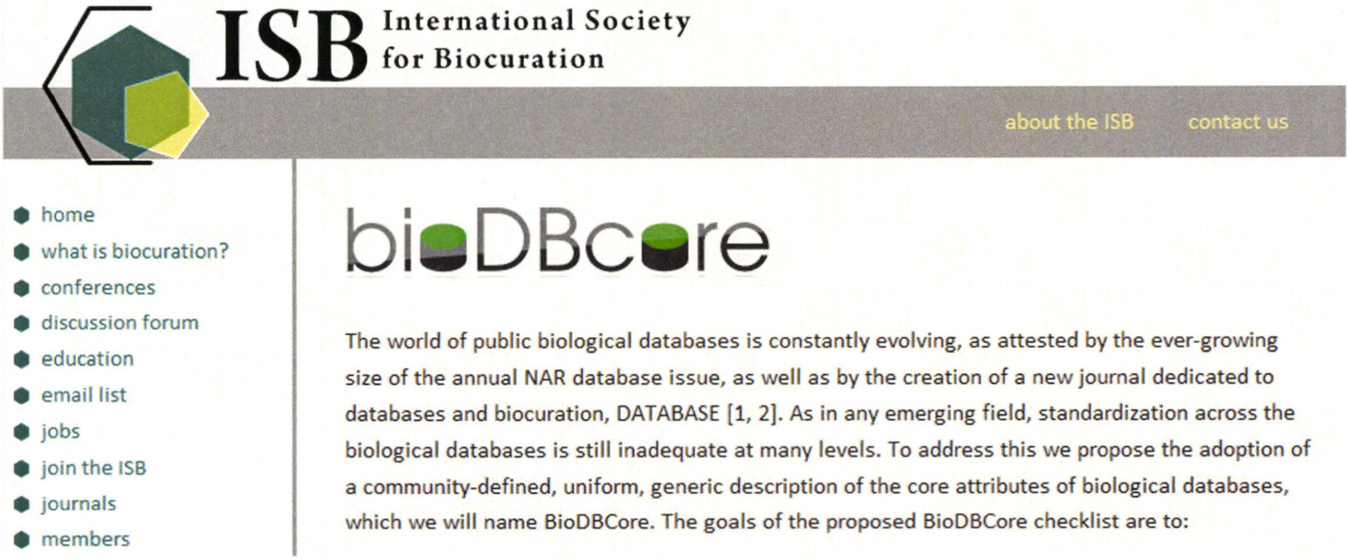
The BioDBCore will be used to collect information about databases for use in online browsing, searching and classification. The current specification can be found as an online survey and users are encouraged to join the project and leave feedback (http://biocurator.org/biodbcore.shtml; Figure 1). Examples can be found in the Supplementary Data and at the BioDBCore web site.

PARTICIPATION OF THE BIOCURATION COMMUNITY IN THE BioDBCore INITIATIVE

With this editorial, we announce the launch of this initiative and present for discussion an initial draft version of the specification of information to be captured. We welcome and encourage representatives of resources, included those listed in this NAR database issue, NAR Molecular Biology Database Collection (1) and the DATABASE journal to actively participate in the development of BioDBCore.

LONG TERM VISION AND POTENTIAL IMPACT

The BioDBCore implementation will take place in three phases: (i) consultation with interested parties; (ii) collaborative development of the minimal information list. To help establish requirements, some examples can be found on the BioDBCore page of the ISB, and moreover the APBioNet's BioDB100 initiative will be used to develop further working examples (11) and (iii) in the longer term, completion of stable guidelines and their implementation as a public submission website that will allow data entry and easy update by database providers, in collaboration



The world of public biological databases is constantly evolving, as attested by the ever-growing size of the annual NAR database issue, as well as by the creation of a new journal dedicated to databases and biocuration, DATABASE [1, 2]. As in any emerging field, standardization across the biological databases is still inadequate at many levels. To address this we propose the adoption of a community-defined, uniform, generic description of the core attributes of biological databases, which we will name BioDBCore. The goals of the proposed BioDBCore checklist are to:

Figure 1. A screenshot of the BioDBCore discussion page on the ISB web site (http://biocurator.org/biodbcore.shtml).

with the existing database collections and the BioSharing standards portal to reduce duplication of effort. Many of the members of the BioDBCore working group have experience and expertise in establishing such services.

We are aware that the adoption of this specification requires significant effort from all participating groups. However, the long-term benefits, both for the specific adopters and for the community as a whole, provides considerable compensation for this effort. The complete, uniform and centralized descriptions of databases should benefit both users and data providers by providing easy access to the scope of each resource. This will be particularly valuable for specialized resources that are only used within with a restricted research community. We envisage that having such rich information readily available may facilitate collaboration between resources currently outside each other's immediate networks. We expect the BioDBCore guideline to be useful not only to users of life sciences resources, but also to drive the evolution of databases themselves. For example, the initial version of BioDBCore includes a field to describe data-submission policies. Currently, many databases do not provide such documents. We hope that by including such a field in BioDBCore, they will be encouraged to develop them. A longer term application of the information captured by BioDBCore is to allow bird's eye views of the database world to emerge by drawing connections between them into a resource network, showing the flow of data between different sites and how each complements the other.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: Invited paper.

Conflict of interest statement. None declared.

REFERENCES

1. Cochrane,G.R. and Galperin,M.Y. (2010) The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acids Res.*, **38**, D1–D4.
2. Landsman,D., Gentleman,R., Kelso,J. and Ouellette,B.F.F. (2009) DATABASE: a new forum for biological databases and curation. *DATABASE*, doi:10.1093/bap002 (Advance acess published online, 26 March 2009).
3. Attwood,T.K., Kell,D.B., McDermott,P., Marsh,J., Pettifer,S.R. and Thorne,D. (2009) Calling International Rescue: knowledge lost in literature and data landslide! *Biochem. J.*, **424**, 317–333.
4. Seringhaus,M.R. and Gerstein,M.B. (2007) Publishing perishing? Towards tomorrow's information architecture. *BMC Bioinform.*, **8**, 17.
5. Philippi,S. and Kohler,J. (2006) Addressing the problems with life-science databases for traditional uses and systems biology. *Nat. Rev. Genet.*, **7**, 482–488.
6. Goble,C. and Stevens,R. (2008) State of the nation in data integration for bioinformatics. *J. Biomed. Inform.*, **41**, 687–693.
7. Howe,D., Costanzo,M., Fey,P., Gojobori,T., Hannick,L., Hide,W., Hill,D.P., Kania,R., Schaeffer,M., St Pierre,S. *et al.* (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.
8. Taylor,C.F., Field,D., Sansone,S.A., Aerts,J., Apweiler,R., Ashburner,M., Ball,C.A., Binz,P.A., Bogue,M., Booth,T. *et al.* (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.*, **26**, 889–896.
9. Brazas,M.D., Yamada,J.T. and Ouellette,B.F.F. (2009) Evolution in bioinformatic resources: 2009 update on the Bioinformatics Links Directory. *Nucleic Acids Res.*, **37**, W3–W5.
10. Smedley,D., Schofield,P., Chen,C.K., Aidinis,V., Ainali,C., Bard,J., Balling,R., Birney,E., Blake,A., Bongcam-Rudloff,E. *et al.* (2010) Finding and sharing: new approaches to registries of databases and services for the biomedical sciences. *DATABASE*, doi:10.1093/baq014 (Advance acess published online, 2 July 2010).
11. Tan,T.W., Tong,J.C., De Silva,M., Lim,K.S. and Ranganathan,S. (2010) Advancing standards for bioinformatics activities: persistence, reproducibility, disambiguation and Minimum Information about a Bioinformatics Investigation (MIABi). *BMC Genomics*, **11(Suppl. 4)**, S27.
12. Field,D., Sansone,S.A., Collis,A., Booth,T., Dukes,P., Gregurick,S.K., Kennedy,K., Kolar,P., Kolker,E., Maxon,M. *et al.* (2009) 'Omics Data Sharing. *Science*, **326**, 234–236.