ISSN 0006-2979, Biochemistry (Moscow), 2011, Vol. 76, No. 8, pp. 915-923. © Pleiades Publishing, Ltd., 2011.
Published in Russian in Biokhimiya, 2011, Vol. 76, No. 8, pp. 1120-1131.

REVIEW

Identification of A-to-I RNA Editing: Dotting the i's in the Human Transcriptome

A. Kiran[1], G. Loughran[1,2], J. J. O'Mahony[1], and P. V. Baranov[1]*

[1]Biochemistry Department and [2]BioSciences Institute, University College Cork,
Cork, Ireland; fax: +353 (21) 490-4259; E-mail: p.baranov@ucc.ie

Received February 11, 2011
Revision received March 29, 2011

Abstract—The phenomenon of adenosine-to-inosine (A-to-I) RNA editing has attracted considerable attention from the scientific community due to its potential relationship to the evolution of cognition in animals. While A-to-I editing exists in all organisms with neurons, including those with primitive neuronal systems (hydra and nematodes), it is particularly frequent in organisms with a highly developed central nervous system (primates, especially humans). Diversification of RNA transcript sequences via A-to-I editing serves a number of different functional roles, such as altering the genome-templated identity of particular amino acids in proteins or altering splice site junctions and modulating regulation of alternatively spliced mRNA variants. Here we provide an overview of current computational and experimental methods for the high-throughput discovery of edited RNA nucleotides in the human transcriptome, as well as a survey of the existing RNA editing bioinformatics resources and an outlook of future perspectives.

DOI: 10.1134/S0006297911080074

Key words: RNA editing, A-to-I, inosine, RNA processing, transcriptome diversity, RNA sequencing, human genome, editome, inosinome

DISTRIBUTION OF RNA EDITING IN THE HUMAN TRANSCRIPTOME AND VERSATILITY OF INOSINE FUNCTIONS

Broadly defined as a co- or post-transcriptional alteration of a templated sequence of RNA by a single or a few nucleotides, RNA editing is common to virtually all living organisms including viruses. Although, somewhat underappreciated, the RNA editing phenomenon has been studied for about three decades now. The observation that insertion of non-templated RNA nucleotides restores a frameshift in the cytochrome oxidase 2 (CoxII) gene from trypanosome mitochondria dates back to the 1980s [1]. The first instance of RNA editing in a human gene, a C-to-U deamination that changes a sense codon to a stop codon in the coding region of apolipoprotein B mRNA, was also discovered about thirty years ago [2]. However, interest in the phenomenon started to grow rapidly when it became apparent that a particular type of RNA editing, hydrolytic deamination of adenosines, commonly known as A-to-I editing, is widespread in the human transcriptome. Intriguingly, the levels of RNA editing seem to correlate with the complexities of animal neural systems and are highest in humans [3]. This lead to numerous speculations suggesting that RNA editing is a molecular phenomenon that is at least partially responsible for the evolution of human intelligence. The discovery of A-to-I editing abundance in the human transcriptome has become possible due to the development of high-throughput bioinformatics, sequencing, and biochemical techniques for identifying inosines in RNA. However, before giving a detailed overview of these methods, we would like to briefly outline the functional versatility of RNA editing in the human transcriptome.

A-to-I editing is a process of inosine formation as a result of hydrolytic deamination of adenosines by enzymes known as ADARs (adenosine deaminases that act on RNA) (see [4, 5] for reviews). ADARs bind to double-stranded regions of RNA (see Fig. 1), and due to their lack of strong sequence specificity, any RNA with a sufficiently long stem is a potential substrate of ADARs. Indeed, inosines have been identified in all types of RNA, including exonic and intronic regions of pre-mRNAs and in different types of ncRNAs. When RNA editing occurs

* To whom correspondence should be addressed.

in the protein coding region of mRNA, it can result in the substitution of a normal codon with a codon containing an inosine. It is believed that inosines are recognized as guanosines by the translation machinery due to the propensity of inosines to form strong base pairs with cytidines. Therefore, RNA editing in a protein coding region may lead to synonymous and non-synonymous codon substitutions, which in the latter case results in recoding of the encoded protein sequence. Such RNA editing-mediated recoding has been documented for a number of animal genes encoding proteins that function in neuronal synapses [6-8]. Recoding via RNA editing has been studied extensively in such important neurotransmitter receptors as the glutamate receptors (GluR) and the serotonin (5-hydroxytryptamine, 5-HT) receptors. RNA editing of GluR results in recoding of a DNA-encoded glutamate residue (CAG codon) to an edited arginine residue (CIG codon). This edited variant of GluR, which is normally a Ca^{2+} channel, is now impermeable to Ca^{2+} [9]. For the 5-HT$_{2c}$ receptor, there are five known RNA editing positions, all within the G-protein binding domain. All possible permutations of RNA editing sites could potentially generate 32 different 5-HT$_{2c}$ receptor mRNA variants encoding 24 protein isoforms. Differential regulation of RNA editing at the 32 sites may allow fine calibration of the serotonin response at neuronal synapses.

Since A-to-I editing occurs before splicing, modification of adenosine at a splice junction could result in either loss of a DNA-encoded splice junction or gain of a novel non-templated splice site that exists only at the RNA level. Several examples of such RNA editing-mediated regulation of alternative splicing are known [10-13]. Interestingly, ADAR2 uses this mechanism for controlling its own cellular activity, via a negative feedback mechanism [12] where RNA editing of tandem AA nucleosides generates an AI pair that mimics an AG dinucleoside motif that is characteristic of 3′ splice junctions. This alternative splice site variant is dependent on RNA editing activity and results in decreased expression of active ADAR2 forms.

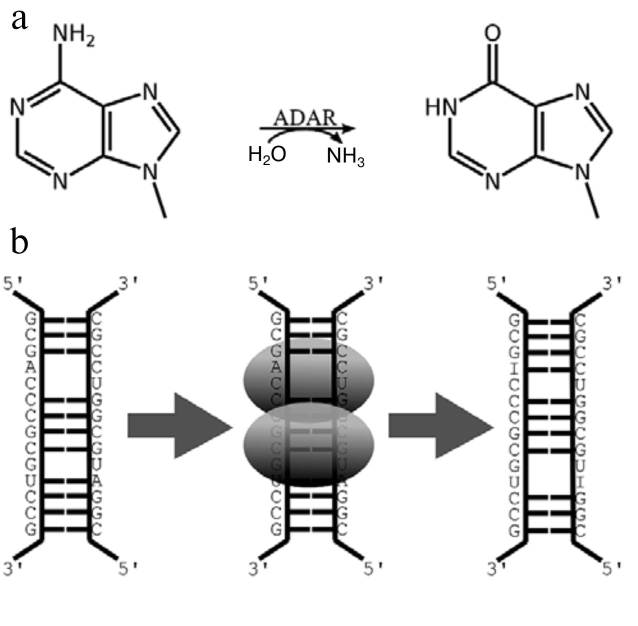


Fig. 1. A-to-I editing. a) Scheme of adenosine deamination catalyzed by ADARs. b) Substrate and the product of an ADAR catalyzed reaction. ADARs bind to double-stranded regions of RNA as homodimers and partially convert adenosines into inosines.

As both pri- and pre-miRNA are highly structured, they are also subject to RNA editing. Editing of miRNA precursors has been shown to affect their processing and thus the formation of mature miRNAs [14]. RNA editing has been reported in a number of miRNAs [15], although a recent deep sequencing study found only a few sites of inefficient A-to-I editing in miRNAs [16, 17]. Sequence alterations of mature miRNAs could result in modulation of their target repertoire thus generating an elegant mechanism for regulating the expression of a group of genes [18]. It is important to realize, that while the coding properties of inosines are similar to those of guanosines, these nucleotides have distinct biochemical properties. These distinctions allow for RNA editing functions that are specific to inosine containing RNA, such as the currently debated nuclear retention of hyper edited transcripts [19-21] and the involvement of RNA editing in the formation of heterochromatin [22]. Recent appreciation of A-to-I editing raises the anticipation that our knowledge of the functional roles of RNA editing will not only expand by attributing novel functions to this process, but also will sharpen by rejecting speculative unsupported hypothesis that emerged on the wave of the current excitement. One particularly important goal in the process of understanding the functional significance of RNA editing is the generation of an accurate and complete quantitative description of the human editome. This goal can be achieved by the systematic application of high-throughput techniques for RNA editing identification to a variety of human samples representing different physiological conditions and developmental stages. In the following subsections we review the progress on current methodologies.

HIGH-THROUGHTPUT IDENTIFICATION OF RNA EDITING SITES

Mining public sequence databases. During cDNA synthesis of inosine containing RNAs, the preferential base pairing of inosines with cytidines leads to the incorporation of cytidines at corresponding locations instead of thymidines that would be incorporated for unedited RNAs. When the sequence of a particular DNA template (such as genomic DNA) is aligned to the sequence of a cDNA produced from an edited molecule, A-to-G mismatches (A in DNA and G in RNA) can be observed within the alignments (see Fig. 2). Existence of such mismatches can be used to predict potential RNA editing sites. The problem with the blunt application of such a method for the prediction of RNA editing is that RNA editing is not the only source of mismatches between sequences of RNA and their templates. While sequence similarity between two individual human genomes is high (estimated to be about 99.9% identity), there are a substantial number of small polymorphic variants. Variations of single nucleotides between two haploid individual genomes or allelic variants within a diploid genome are termed single nucleotide polymorphisms (SNPs). The latest release of dbSNP [23] that collects information on such variants contains descriptions of about 7 million SNPs. Many of these SNPs occur in the transcribed regions of the human genome. When DNA and RNA sequences are obtained from samples of different individuals, mismatches in the alignments of such sequences can occur due to SNPs. Also, mismatches can be generated by sequencing errors that are particularly frequent during high-throughput sequencing, e.g. the level of sequencing errors in expressed sequence tags (EST) is estimated to be about 3%. Nonetheless, if A-to-I RNA editing is abundant in humans, then A-to-G mismatches should be overrepresented in DNA/RNA alignments and should occur at a higher frequency than what would be expected for the average rate of transition substitutions between the sequences of two human genomes. Also an asymmetry in the occurrence of A-to-G and G-to-A mismatches should be observable for the positive strands. Indeed, a large-scale statistical analysis of DNA and RNA discrepancies using high quality mRNA sequences derived from the NCBI Reference Sequence project (RefSeq) and the NIH Mammalian Gene Collection (MGC) confirms such expectations [24]. While such statistical analysis can be used as an argument for the abundance of RNA editing in the human transcriptome, the real issue is how to discriminate A-to-G mismatches generated by A-to-I editing from those that occur due to polymorphisms and sequencing errors. An elegant solution to this problem is to take into account a characteristic feature of ADAR substrates. ADARs bind to long double-stranded RNA and edit multiple adenosines within this structure. As a result, several inosines are generated within a relatively short region of RNA. Therefore, RNA editing-generated A-to-G mismatches between DNA and RNA should appear as clusters in corresponding alignments. Such clusters of A-to-G mismatches have been observed during the analysis of cDNA sequences from the HUGE database [25].

See Fig. 2 for a characteristic example of an A-to-G mismatch clustering due to RNA editing. In 2004 several independent groups [26-28] applied this feature as a criterion in searching for RNA–DNA discrepancies that originate due to RNA editing. While the approaches differed in detailed methodology and datasets used, the general strategy for the identification of RNA editing was similar and can be summarized as follows: RNA–DNA alignments were generated using the reference sequence of the human genome and a set of cDNA sequences. Furthermore, a number of statistical filters were applied to reduce false positives. Alignments containing multiple different types of mismatches (substitutions other than A-to-G) were tagged as sequences of poor sequencing quality. Other mismatches were filtered against known SNPs from dbSNP [23] to eliminate known human polymorphisms. Then clusters containing multiple A-to-G substitutions were selected, and the possibility of base pair formation between these regions and neighboring genomic regions was verified. Any that passed this criterion were considered as potential RNA edited candidates. To estimate the accuracy of these filters, a fraction of the predicted candidates was experimentally verified. DNA and RNA were isolated from the same tissue samples and the regions containing predicted RNA editing candidates were amplified and sequenced. *Bona fide* RNA editing candidates were confirmed only if unambiguous chromatogram peaks indicated the presence of a homozygous adenosine in the positive strand of DNA and a G or ambiguous overlapping A/G peaks in the chromatograms corresponding to RNA sequences. The experimental verification of these bioinformatics methods demonstrated a low false positive rate with a positive predictive value >90%. All three studies pointed to a very high abundance (tens of thousands) of RNA editing in transcripts containing repeated sequences, in particular Alu repeats. This observation makes sense since two inverted Alu

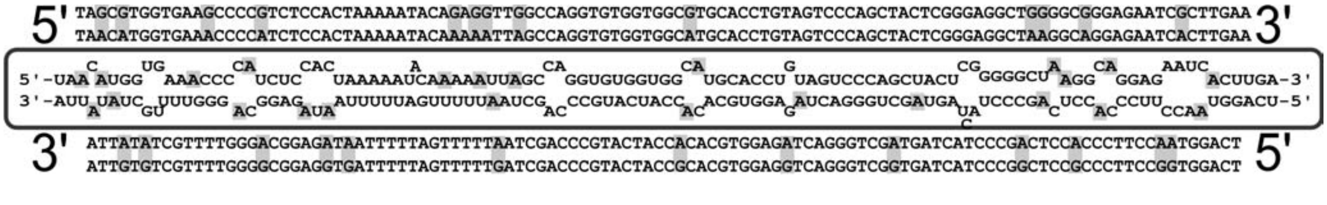

5'
TAGCGTGGTGAAGCCCCGTCTCCACTAAAAATACAGAGGTTGGCCAGGTGTGGTGGCGTGCACCTGTAGTCCCAGCTACTCGGGAGGCTGGGGCGGGAGAATCGCTTGAA
TAACATGGTGAAACCCCATCTCCACTAAAAATACAAAAATTAGCCAGGTGTGGTGGCATGCACCTGTAGTCCCAGCTACTCGGGAGGCTAAGGCAGGAGAATCACTTGAA
3'
C UG CA CAC A CA CA G CG A CA AAUC
5'-UAA AUGG AAACCC UCUC UAAAAAUCAAAAAUUAGC GGUGUGGUGG UGCACCU UAGUCCCAGCUACU GGGGCU AGG GGAG ACUUGA-3'
3'-AUU UAUC UUUGGG GGAG AUUUUUAGUUUUUAAUCG CCGUACUACC ACGUGGA AUCAGGGUCGAUGA UCCCGA UCC CCUU UGGACU-5'
A GU AC AUA AC AC G UA C AC CCAA
C
3'
ATTATATCGTTTTGGGACGGAGATAATTTTTAGTTTTTAATCGACCCGTACTACCACACGTGGAGATCAGGGTCGATGATCATCCCGACTCCACCCTTCCAATGGACT
ATTGTGTCGTTTTGGGGCGGAGGTGATTTTTAGTTTTTGATCGACCCGTACTACCGCACGTGGAGGTCAGGGTCGGTGATCATCCCGGCTCCGCCCTTCCGGTGGACT
5'

Fig. 2. A cluster of A-to-I RNA editing sites. RNA secondary structure predicted for the AluSx repeat from chromosome I along with corresponding alignments of edited and templated (unedited) RNA sequences predicted based on the sequence of the reference human genome. Sites of RNA editing inferred from A-to-G mismatches in the alignment are shaded in gray.

repeats within the same RNA are very likely to form a long double-stranded region that could be a substrate for ADARs. The analysis performed by these three research groups was a crucial landmark in the realization that RNA editing is not just a minor process that affects expression of a small proportion of human genes, but it is a much more general phenomenon that affects the expression of many human genes since Alu repeats comprise about 11% of the human genome and are usually located in or near genic regions [29]. Although these methods were very important in discovering the abundance of RNA editing in humans, they do have some serious limitations. While the positive predictive value of these methods is high and so they have high specificity, their sensitivity is probably low. Not only were there a large number of predicted candidates specific for a particular study, but most importantly, many previously known RNA editing cases were not detected. The cluster-based method is suitable only for a particular class of RNA editing, frequently termed "promiscuous RNA editing" in the literature [30]. Promiscuous RNA editing refers to those sites where the editing of a particular RNA nucleoside does not fulfill a specific significant function, and therefore such sites are under neutral evolutionary selection. A single substitution of a genomic nucleotide that either abolishes existing or creates a new RNA editing site within an Alu repeat sequence is not likely to affect the organism's fitness. In contrast, the majority of functionally significant sites, such as those mentioned in the previous section, are likely to evolve under purifying selection. RNA editing sites that are responsible for the recoding of protein coding sequences or the generation of alternative splice sites play specific functional roles. Loss of such RNA editing sites or altering editing efficiency is likely to affect organism fitness, and therefore the editing at such sites should be preserved during the course of evolution. At the same time, since such sites occur in functionally important regions, alterations in the surrounding sequence could have a negative effect on fitness, thus selecting against the generation of novel editing sites in neighboring regions. RNA editing sites where only a specific nucleoside is subject to RNA editing are often termed "selective RNA editing sites" [30]. A limited number of selective RNA editing sites have been identified using bioinformatics approaches that specifically concentrate on the identification of evolutionarily conserved sequences involved in RNA editing [31-34].

An interesting and somewhat controversial source for the discovery of RNA editing sites turned out to be the dbSNP [23]. SNPs collected in dbSNP are derived from different sources and have varying degrees of accuracy. Relatively large numbers of human genomic coordinates that are reported to have polymorphic variants in dbSNP are not supported by genomic sequences. A number of SNPs were predicted from the analysis of EST sequences; for some of the entries the information on the source of the SNP is missing. Eisenberg et al. [35] have hypothesized that a number of SNPs in dbSNP could be instances of misinterpreted RNA editing events, and they have demonstrated that this is indeed true for about a hundred SNPs. A more sophisticated procedure for mining dbSNP was developed by Gommans et al. [36] with the specific goal of identifying RNA editing sites that may result in the recoding of protein coding genes. The dbSNP was used as a starting pool of potential RNA editing candidates. The statistical filters applied by this team used only a fraction of the SNPs from dbSNP, those that were supported only by EST sequences and corresponded to A-to-G or G-to-A substitutions in the positive DNA strand. Since Gommans et al. [36] were interested only in those RNA editing instances that result in protein recoding, the candidates were further filtered leaving only those where RNA editing (if it occurs) would result in a nonsynonymous change of a codon. An additional filter was used to remove potential RNA candidates located in a particularly poor editing context (5′-adjacent G). The final filter was based on evolutionary conservation of the corresponding sequence regions as well as conservation of adenosines themselves. The remaining three hundred candidates were then scored based on several molecular features of RNA editing candidates, and a fraction of these candidates was tested experimentally by sequencing of corresponding RNA and DNA fragments. This method identified several known examples of RNA editing as well as a number of new additional RNA editing candidates.

It needs to be pointed out that human polymorphisms and RNA editing are not mutually exclusive phenomena. There is no fundamental reason why a polymorphic A/G site cannot also be an RNA editing site. This is likely the situation for most "promiscuous" editing sites, where a particular RNA editing site does not significantly affect an organism's fitness. This is less likely for "selective" editing sites. Therefore, we should expect that even among those SNPs validated at the genomic level, many could also be RNA editing sites. The problem with interpreting particular locations as SNPs or RNA editing is further complicated by DNA editing and somatic mutations [37].

Utilization of massively parallel sequencing for the identification and verification of RNA editing sites. Although methods based on mismatch clustering in cDNA–DNA alignments have identified tens of thousands of new RNA editing sites in the human transcriptome, the application of such approaches is unlikely to reveal the entire repertoire of RNA editing in humans. As mentioned earlier, one reason is that clusters are typical only for promiscuous RNA editing, such as seen in Alu repeats. The second reason is that the analysis of cDNA sequences may reveal only a fraction of edited nucleotides. Highly efficient RNA editing, where full deamination of a particular adenosine in the entire population of RNA produced from the same template, is rare. If the editing efficiency of a particular adenosine is low, only a small proportion of sequenced molecules will have an inosine in place of adenosine. Therefore, methods based on the analysis of mismatches between cDNA and genomic sequences are unlikely to identify every RNA editing site. Only a small number of cDNA sequences are available for any particular gene, and these sequences are not guaranteed to represent the full spectrum of edited variants. In this regard, the analysis of ESTs provides a better outlook for the distribution of edited sites at a particular locus. Yet, the higher rate of sequencing errors in ESTs requires experimental verification of RNA editing candidates predicted from such alignments.

Massively parallel sequencing provides an opportunity for verifying numerous RNA candidates at once. In 2009, Li et al. [38] developed a high-throughput experimental procedure for the verification of RNA editing candidates. The procedure is based on the use of molecular inversion probes, commonly known as padlocks [39]. Padlock probes were originally developed for rapid genotyping of genomic regions corresponding to sites of known or suspected polymorphisms. Padlocks are single-stranded molecules containing two regions of complementarity to the regions surrounding the target nucleotides (see Fig. 3). After hybridization of a probe with a target molecule, a single-stranded gap is filled with a complimentary chain by polymerase reaction, and the probe is circularized by DNA ligation. The circularized probe is then cleaved so that primers against the two regions of complementarity can be used to amplify the intervening region. Naturally, padlock probes can be used not only for the genotyping of SNPs in genomic DNA, but also for verifying the status of a nucleotide in cDNA. Li et al. [38] have used padlock probes to verify the status of ~36,000 RNA editing candidates in samples from several tissues. Candidates were generated by identifying the mismatches between genomic DNA, cDNA, and ESTs. All mismatches were further filtered through a bioinformatics pipeline for selecting the most prominent candidates. One of the filters in this pipeline was the exclusion of those regions corresponding to genomic repeats, first because there is little interest in these regions for which RNA editing has already been established. Second, molecular probes, due to their relatively small size target repetitive regions ambiguously.

One of the potential advantages of this technique is that, in principle, it could allow for the estimation of RNA editing efficiency by calculating it as a fraction of reads that support an edited base from the total number of reads aligning to its corresponding position. Li et al. [38] have reported 710 RNA editing sites that they have separated into three classes. Class III RNA editing sites are those at which RNA editing occurs with low efficiency (2.5-5%), while class II (330 candidates) and class I (239) RNA editing sites are those that occur with an efficiency ≥5%. Class I differs from Class II in that the evidence for Class I editing is found in more than one tissue sample, while Class II edits were identified from only a single tissue type. We have incorporated all of the Class I and Class II candidates into our DARNED database (described in detail in the last section of this review); however, it needs to be emphasized that these RNA editing candidates, especially those classified as Class II candidates, should be treated with caution. There are two reasons for this. For a relatively large number of candidates, even a single sequence read was considered as evidence of RNA editing; moreover, the efficiency of editing has been calculated even for those sites with only a single read, in which case it was estimated as 100%. Obviously, for statistical reasons such an estimate should be considered highly

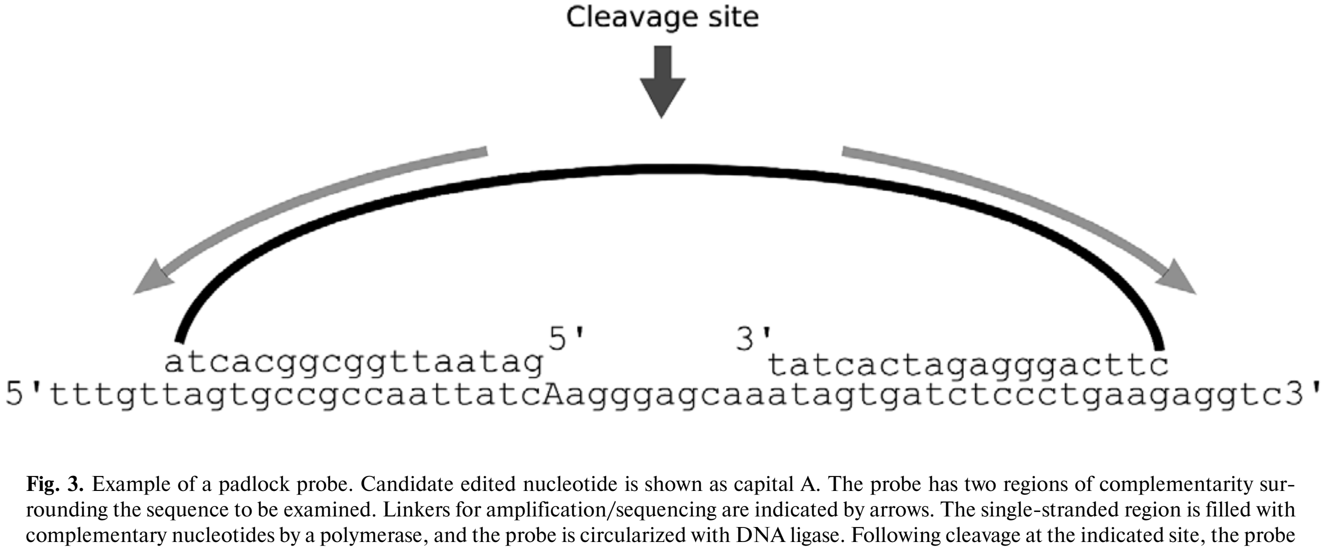


Fig. 3. Example of a padlock probe. Candidate edited nucleotide is shown as capital A. The probe has two regions of complementarity surrounding the sequence to be examined. Linkers for amplification/sequencing are indicated by arrows. The single-stranded region is filled with complementary nucleotides by a polymerase, and the probe is circularized with DNA ligase. Following cleavage at the indicated site, the probe is ready for PCR amplification and sequencing.

inaccurate. Therefore, the data reported for RNA editing candidates vary substantially in terms of statistical confidence, which was not taken into account during the classification of RNA editing sites into these three classes. Another reason to be cautious is that genomic sequences were not verified for all tissue samples. While all tissue samples were derived from the same individual (and the authors validated this by genotyping), genomic DNA was sequenced only from a single tissue sample. Thus instances of tissue-specific DNA editing or somatic mutations [37] in one of the tissues would mimic RNA editing behavior.

Despite these caveats regarding the quality of some of the candidates reported by Li et al. [38], this pioneering work is a key landmark towards developing high-throughput experimental techniques for the identification of novel RNA editing sites. With massively parallel sequencing platforms becoming widely available to the research community at affordable prices, we anticipate that in future many more studies will use similar approaches for identifying novel RNA editing sites and for quantifying RNA editing efficiency. Interestingly, as deep sequencing transcriptomics becomes increasingly more popular, useful information on RNA editing is being generated, even though identification of novel RNA editing sites is not a primary goal. Examples include the work on the characterization of mammalian microRNAs already mentioned in the first section of this review [16, 17].

Chemical methods. Although inosine has properties similar to those of guanosine with respect to its preferential base pairing with cytidines – inosine and guanosine should not be treated as the same nucleosides. Distinct chemical properties of inosine could also be used for its identification. Morse and Bass [40] have developed a technique based on the differential reaction of these two nucleosides with glyoxalin. Glyoxalin reacts with guanosine, but not with inosine. Both inosine and guanosine are recognized by RNase T1, which cleaves single-stranded regions of RNA 3′ of these nucleotides. However, glyoxalination blocks guanosine from RNase T1 recognition. Therefore, RNA treated with glyoxaline can be cleaved by RNase T1 only 3′ of inosines. In the Morse and Bass technique, RNA fragments are polyadenylated at their 3′-end after RNase T1 treatment, then converted to cDNA and amplified by PCR. Gel electrophoresis of cDNA derived from inosine-containing RNA can then be compared with cDNA derived from RNA that has not been treated with RNase T1. Subsequent sequencing of cDNA derived from treated and untreated RNA allows identification of the RNA edited locations. Morse and Bass have been able to identify several novel RNA editing sites in *C. elegans* using this technique [41, 42]. However, to our knowledge this technique has not been applied for the large scale identification of editing in mammalian samples.

Last year Sakurai et al. [43] developed a high-throughput method that they termed ICE (inosine chemical erasing) for the chemical identification of inosines. The method is based on the reaction of inosine with acrylonitrile to produce a derivative called cyanoethylate that cannot base pair with cytidine [44]. Sakurai et al. [43] reasoned that, during primer extension reactions, reverse transcriptase should arrest on RNA templates at cyanoethylated inosines. Therefore, RNA containing cyanoethylated inosines cannot be amplified efficiently by PCR. This allows discrimination of a population of RNA containing A or G at a particular position (e.g. due to DNA polymorphisms) from a population of RNA containing A or I (due to editing). In the former, the RT-PCR products of both samples (treated and untreated with acrylonitrile) will have ambiguous base calling of A/G at the examined position and have the same A/G ratio. Where there is RNA editing, untreated samples would have ambiguous A/G base calling, while treated samples would have unambiguous A base calling or a highly reduced A/G ratio, since inosine containing fragments will not be efficiently amplified. One very significant advantage of this method in comparison with the bioinformatic and high-throughput sequencing methods described in the previous sections is that it does not require prior sequencing information of the genomic templates.

Sakurai et al. [43] have used this method to verify 1277 predicted RNA editing sites. A total of 716 sites were confirmed as at least 10% efficient RNA editing sites. A total of 39 of the examined cases had the same A/G ratio for treated and untreated samples, suggesting differential allelic origin. Most interestingly, Sakurai et al. found over 2000 new sites within examined regions that were not previously predicted as RNA editing sites by bioinformatics approaches. A total of 79 of the new editing sites had previously been reported as SNPs, consistent with reports that dbSNP contains RNA editing sites misinterpreted as SNPs [35, 36].

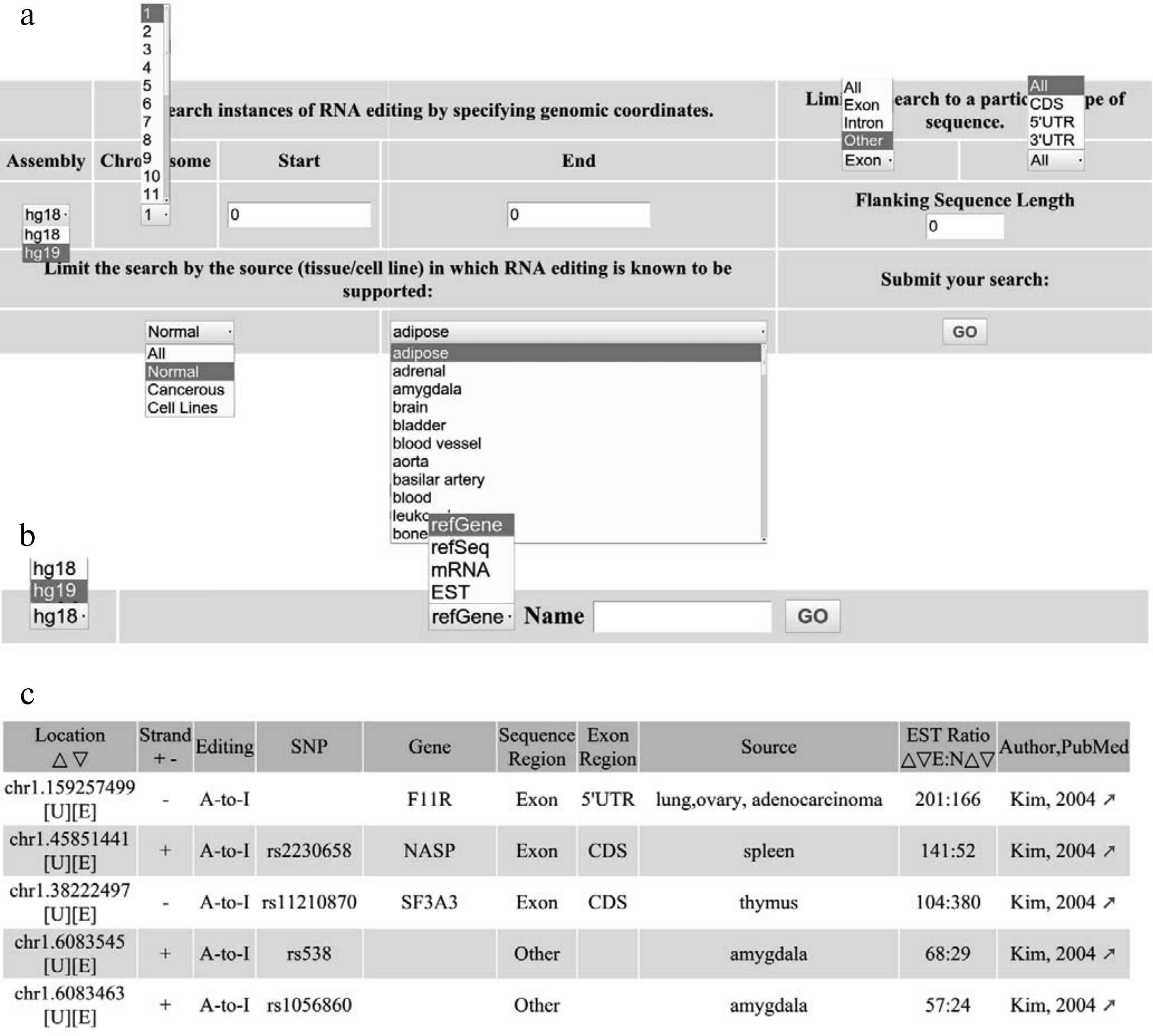
In summary, there is growing interest in developing methods for high-throughput detection of RNA editing. Current methodologies are broadly in agreement that A-to-I editing is abundant in humans. However, the datasets of predicted RNA editing sites produced by each individual study do not entirely overlap. This is not necessarily due to limitations of the methodologies and in principle could be explained by sample differences. It is probable and even likely that RNA editing efficiency at promiscuous sites (where editing at a particular site does not effect organism fitness efficiently) differs among individuals. The issue of polymorphic RNA editing should be addressed by future studies. Another explanation of inconsistency between the precise RNA editing locations derived from different studies is the likely stochastic nature of RNA editing at promiscuous sites. If, among a large number of sites each is stochastically edited with a limited efficiency, the diversity of resultant differentially edited RNA will be enormous [45]. It is unlikely that any of the methods described in this review would be capable of capturing such a diversity of differentially edited RNA.

RNA EDITING COMPUTATIONAL RESOURCES

The RNA editing phenomenon has been known for decades, and a number of databases have been designed in an attempt to provide information about RNA editing on the World Wide Web. Some of these databases are no longer supported [46] or updated [47]. Among those that have survived the test of time are dbRES [48] and REDIdb [49, 50]. The recently developed RESOPS [51] database is another notable tool dedicated to RNA editing.

However, despite the existence of several RNA editing depositories on the web, none of them are useful as a source of information on RNA editing in humans. REDIdb and RESOPS focus on RNA editing in organelles. While dbRES has a wider scope than REDIdb and RESOP, its internal structure and dependency on manual annotation make it unsuitable for accommodating large influxes of high-throughput data.

The rapid growth of information on RNA editing in the human transcriptome and the lack of adequate com-



earch instances of RNA editing by specifying genomic coordinates.

Limi earch to a partic pe of sequence.

Assembly	Chro some	Start	End		
hg18	1	0	0	Exon	All

Flanking Sequence Length: 0

imit the search by the source (tissue/cell line) in which RNA editing is known to be supported: Normal adipose

Submit your search: GO

hg18 refGene Name GO

Location △▽	Strand + -	Editing	SNP	Gene	Sequence Region	Exon Region	Source	EST Ratio △▽E:N△▽	Author,PubMed
chr1.159257499 [U][E]	-	A-to-I		F11R	Exon	5'UTR	lung,ovary, adenocarcinoma	201:166	Kim, 2004 ↗
chr1.45851441 [U][E]	+	A-to-I	rs2230658	NASP	Exon	CDS	spleen	141:52	Kim, 2004 ↗
chr1.38222497 [U][E]	-	A-to-I	rs11210870	SF3A3	Exon	CDS	thymus	104:380	Kim, 2004 ↗
chr1.6083545 [U][E]	+	A-to-I	rs538		Other		amygdala	68:29	Kim, 2004 ↗
chr1.6083463 [U][E]	+	A-to-I	rs1056860		Other		amygdala	57:24	Kim, 2004 ↗

Fig. 4. I/O interface of DARNED. a) Coordinate based query form. b) Gene based query form. c) Example of output.

putational resources where this information is organized prompted us to design yet another RNA editing database which we termed DARNED [52] (database of RNA editing, http://darned.ucc.ie). We have taken advantage of the fact that the human genome reference sequence is almost complete [53], and we used the reference sequence as a backbone for the organization of RNA editing data. We reasoned that, although RNA editing takes place at the RNA level, to our knowledge, all RNA synthesis in humans is DNA-template based (apart from humans infected with RNA viruses), and therefore all of the RNA in human cells can be linked to its template in the genome. Thus, every RNA editing case reported in the literature could be unambiguously assigned to a particular coordinate of the human genome. One drawback to such classification is a scenario in which two different types of RNA are produced from the same template, but only one of these is edited. As we are currently unaware of such examples, mapping RNA editing sites to the human genome was considered as the most reasonable approach for organizing RNA editing information in DARNED.

We have processed the published data on RNA editing as described in Kiran and Baranov [52] to identify genomic coordinates for each reported RNA editing site. At present, RNA editing data are available for two human genome assemblies, hg18 and hg19. Figure 4 shows the input menu and an example of a DARNED output. Currently DARNED provides two types of search methods, coordinate based and gene based. In coordinate based mode (Fig. 4a) users need to specify a coordinate range in the human genome, which they can also limit by inputting information on the functional location of the RNA editing site (e.g. exon/intron, CDS/UTR). In the gene mode (Fig. 4b), a search can be performed for a human genome locus corresponding to a particular sequence entity, such as annotation, e.g. refGene, refSeq, mRNA, or EST. The output (Fig. 4c) contains a list of RNA editing sites along with additional information on the literature source from which the information was obtained, number of ESTs supporting an edited or non-edited base, functional location, status in the dbSNP, etc. Nonetheless, the additional information provided by DARNED is rather limited. Therefore, in order to allow users more flexibility we have designed DARNED custom tracks that links users to remote genome browsers, either UCSC [54] or Ensembl [55], and provides RNA editing information in the context of these genome browsers. In fact, the latest update of the UCSC genome browser [54] provides DARNED tracks within a set of their standard tracks for the latest two genome assemblies. Therefore, users do not even need to visit the DARNED database in order to gain access to the data. However, it should be noted that editing sites collected in DARNED are continuously updated, while UCSC tracks are updated less frequently. Consequently, when users are interested in the latest and most complete information on the human editome, it is recommended to use DARNED database instead of UCSC local tracks.

Combining the output from DARNED with the UCSC genome browser data and interface provides a variety of ways in which RNA editing data can be explored. Figure 5 (see color insert) illustrates a few examples: comparison of RNA editing with SNPs (Fig. 5a), dense RNA editing regions in Alu repeats (Fig. 5b); presence of evolutionary RNA secondary structures predicted by EvoFold [56] for selective RNA editing sites in exons (Fig. 5c) and non-coding RNAs (Fig. 5d).

In the future, we plan to continue adding and revising RNA editing sites in the DARNED database and to improve its interface. We hope that the database will become a central computational resource for the quantitative description of the human editome.

The authors are grateful to the members of RNA editing community for their welcoming reception of the DARNED database and to the UCSC Genome Browser technical team for their help in designing the RNA editing tracks.

The authors would like to express their gratitude to Science Foundation Ireland for its support of the DARNED database development, by principal investigator award to P. V. B. (grant number 06/IN.1/B81).

REFERENCES

1. Benne, R., van den Burg, J., Brakenhoff, J. P., Sloof, P., van Boom, J. H., and Tromp, M. C. (1986) *Cell*, **46**, 819-826.
2. Driscoll, D. M., Wynne, J. K., Wallis, S. C., and Scott, J. (1989) *Cell*, **58**, 519-525.
3. Paz-Yaacov, N., Levanon, E. Y., Nevo, E., Kinar, Y., Harmelin, A., Jacob-Hirsch, J., Amariglio, N., Eisenberg, E., and Rechavi, G. (2010) *Proc. Natl. Acad. Sci. USA*, **107**, 12174-12179.
4. Nishikura, K. (2010) *Annu. Rev. Biochem.*, **79**, 321-349.
5. Bass, B. L. (2002) *Annu. Rev. Biochem.*, **71**, 817-846.
6. Reenan, R. A. (2005) *Nature*, **434**, 409-413.
7. Kawahara, Y., Ito, K., Sun, H., Aizawa, H., Kanazawa, I., and Kwak, S. (2004) *Nature*, **427**, 801.
8. Brusa, R., Zimmermann, F., Koh, D. S., Feldmeyer, D., Gass, P., Seeburg, P. H., and Sprengel, R. (1995) *Science*, **270**, 1677-1680.
9. Higuchi, M., Single, F. N., Kohler, M., Sommer, B., Sprengel, R., and Seeburg, P. H. (1993) *Cell*, **75**, 1361-1370.
10. Wang, Q., O'Brien, P. J., Chen, C. X., Cho, D. S., Murray, J. M., and Nishikura, K. (2000) *J. Neurochem.*, **74**, 1290-1300.
11. Grohmann, M., Hammer, P., Walther, M., Paulmann, N., Buttner, A., Eisenmenger, W., Baghai, T. C., Schule, C., Rupprecht, R., Bader, M., Bondy, B., Zill, P., Priller, J., and Walther, D. J. (2010) *PLoS One*, **5**, e8956.
12. Rueter, S. M., Dawson, T. R., and Emeson, R. B. (1999) *Nature*, **399**, 75-80.
13. Laurencikiene, J., Kallman, A. M., Fong, N., Bentley, D. L., and Ohman, M. (2006) *EMBO Rep.*, **7**, 303-307.

14. Yang, W., Chendrimada, T. P., Wang, Q., Higuchi, M., Seeburg, P. H., Shiekhattar, R., and Nishikura, K. (2006) *Nat. Struct. Mol. Biol.*, **13**, 13-21.
15. Blow, M. J., Grocock, R. J., van Dongen, S., Enright, A. J., Dicks, E., Futreal, P. A., Wooster, R., and Stratton, M. R. (2006) *Genome Biol.*, **7**, R27.
16. Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A. L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., Eaves, C. J., and Marra, M. A. (2008) *Genome Res.*, **18**, 610-621.
17. De Hoon, M. J., Taft, R. J., Hashimoto, T., Kanamori-Katayama, M., Kawaji, H., Kawano, M., Kishima, M., Lassmann, T., Faulkner, G. J., Mattick, J. S., Daub, C. O., Carninci, P., Kawai, J., Suzuki, H., and Hayashizaki, Y. (2010) *Genome Res.*, **20**, 257-264.
18. Kawahara, Y., Zinshteyn, B., Sethupathy, P., Iizasa, H., Hatzigeorgiou, A. G., and Nishikura, K. (2007) *Science*, **315**, 1137-1140.
19. Zhang, Z., and Carmichael, G. G. (2001) *Cell*, **106**, 465-475.
20. Prasanth, K. V., Prasanth, S. G., Xuan, Z., Hearn, S., Freier, S. M., Bennett, C. F., Zhang, M. Q., and Spector, D. L. (2005) *Cell*, **123**, 249-263.
21. Hundley, H. A., Krauchuk, A. A., and Bass, B. L. (2008) *RNA*, **14**, 2050-2060.
22. Wang, Q., Zhang, Z., Blackwell, K., and Carmichael, G. G. (2005) *Curr. Biol.*, **15**, 384-391.
23. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001) *Nucleic Acids Res.*, **29**, 308-311.
24. Furey, T. S., Diekhans, M., Lu, Y., Graves, T. A., Oddy, L., Randall-Maher, J., Hillier, L. W., Wilson, R. K., and Haussler, D. (2004) *Genome Res.*, **14**, 2034-2040.
25. Kikuno, R., Nagase, T., Waki, M., and Ohara, O. (2002) *Nucleic Acids Res.*, **30**, 166-168.
26. Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z. Y., Shoshan, A., Pollock, S. R., Sztybel, D., Olshansky, M., Rechavi, G., and Jantsch, M. F. (2004) *Nat. Biotechnol.*, **22**, 1001-1005.
27. Athanasiadis, A., Rich, A., and Maas, S. (2004) *PLoS Biol.*, **2**, e391.
28. Kim, D. D., Kim, T. T., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S., and Gabriel, A. (2004) *Genome Res.*, **14**, 1719-1725.
29. Grover, D., Majumder, P. P., Rao, C-B., Brahmachari, S. K., and Mukerji, M. (2003) *Mol. Biol. Evol.*, **20**, 1420-1424.
30. Wulff, B. E., Sakurai, M., and Nishikura, K. (2011) *Nat. Rev. Genet.*, **12**, 81-85.
31. Levanon, E. Y., Hallegger, M., Kinar, Y., Shemesh, R., Djinovic-Carugo, K., Rechavi, G., Jantsch, M. F., and Eisenberg, E. (2005) *Nucleic Acids Res.*, **33**, 1162-1168.
32. Hoopengardner, B., Bhalla, T., Staber, C., and Reenan, R. (2003) *Science*, **301**, 832-836.
33. Enstero, M., Akerborg, O., Lundin, D., Wang, B., Furey, T. S., Ohman, M., and Lagergren, J. (2010) *BMC Bioinformatics*, **11**, 6.
34. Clutterbuck, D. R., Leroy, A., O'Connell, M. A., and Semple, C. A. (2005) *Bioinformatics*, **21**, 2590-2595.
35. Eisenberg, E., Adamsky, K., Cohen, L., Amariglio, N., Hirshberg, A., Rechavi, G., and Levanon, E. Y. (2005) *Nucleic Acids Res.*, **33**, 4612-4617.
36. Gommans, W. M., Tatalias, N. E., Sie, C. P., Dupuis, D., Vendetti, N., Smith, L., Kaushal, R., and Maas, S. (2008) *RNA*, **14**, 2074-2085.
37. Zaranek, A. W., Levanon, E. Y., Zecharia, T., Clegg, T., and Church, G. M. (2010) *PLoS Genet.*, **6**, e1000954.
38. Li, J. B., Levanon, E. Y., Yoon, J. K., Aach, J., Xie, B., Leproust, E., Zhang, K., Gao, Y., and Church, G. M. (2009) *Science*, **324**, 1210-1213.
39. Nilsson, M., Malmgren, H., Samiotaki, M., Kwiatkowski, M., Chowdhary, B. P., and Landegren, U. (1994) *Science*, **265**, 2085-2088.
40. Morse, D. P., and Bass, B. L. (1997) *Biochemistry*, **36**, 8429-8434.
41. Morse, D. P., and Bass, B. L. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 6048-6053.
42. Morse, D. P., Aruscavage, P. J., and Bass, B. L. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 7906-7911.
43. Sakurai, M., Yano, T., Kawabata, H., Ueda, H., and Suzuki, T. (2010) *Nat. Chem. Biol.*, **6**, 733-740.
44. Yoshida, M., and Ukita, T. (1968) *Biochim. Biophys. Acta*, **157**, 455-465.
45. Barak, M., Levanon, E. Y., Eisenberg, E., Paz, N., Rechavi, G., Church, G. M., and Mehr, R. (2009) *Nucleic Acids Res.*, **37**, 6905-6915.
46. Hinz, S., and Goringer, H. U. (1999) *Nucleic Acids Res.*, **27**, 168.
47. Simpson, L., Wang, S. H., Thiemann, O. H., Alfonzo, J. D., Maslov, D. A., and Avila, H. A. (1998) *Nucleic Acids Res.*, **26**, 170-176.
48. He, T., Du, P., and Li, Y. (2007) *Nucleic Acids Res.*, **35**, D141-144.
49. Picardi, E., Regina, T. M., Brennicke, A., and Quagliariello, C. (2007) *Nucleic Acids Res.*, **35**, D173-177.
50. Picardi, E., Regina, T. M., Verbitskiy, D., Brennicke, A., and Quagliariello, C. (2010) *Mitochondrion*, **11**, 360-365.
51. Yura, K., Sulaiman, S., Hatta, Y., Shionyu, M., and Go, M. (2009) *Plant Cell Physiol.*, **50**, 1865-1873.
52. Kiran, A., and Baranov, P. V. (2010) *Bioinformatics*, **26**, 1772-1776.
53. Consortium, I. H. G. S. (2004) *Nature*, **431**, 931-945.
54. Fujita, P. A., et al. (2011) *Nucleic Acids Res.*, **39**, D876-882.
55. Flicek, P., et al. (2011) *Nucleic Acids Res.*, **39**, D800-806.
56. Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. (2006) *PLoS Comput. Biol.*, **2**, e33.