A bioinformatic screen for novel A–I RNA editing sites reveals recoding editing in BC10

D. R. Clutterbuck* A. Leroy, M. A. O’Connell and C. A. M. Semple

ABSTRACT
Motivation: Recent studies have demonstrated widespread adenosine–inosine RNA editing in non-coding sequence. However, the extent of editing in coding sequences has remained unknown. For many of the known sites, editing can be observed in multiple species and often occurs in well-conserved sequences. In addition, they often occur within imperfect inverted repeats and in clusters. Here we present a bioinformatic approach to identify novel sites based on these shared features. Mismatches between genomic and expressed sequences were filtered to remove the main sources of false positives, and then prioritized based on these features. This protocol is tailored to identifying specific recoding editing sites, rather than sites in non-coding repeat sequences.

Results: Our protocol is more sensitive for identifying known coding editing sites than any previously published mammalian screen. A novel multiply edited transcript, BC10, was identified and experimentally verified. BC10 is highly conserved across a range of metazoans and has been implicated in two forms of cancer.

Contact: daniel.clutterbuck@hgu.mrc.ac.uk

1 INTRODUCTION
RNA editing is the modification of RNA sequence that is distinct from other RNA processing events such as splicing, capping or 3’ end processing (Bass, 2002; Keegan et al., 2001). Editing is a widespread phenomenon that has been identified in numerous species including a range of metazoans, plants, protozoa and bacteria. In mammals adenosine–inosine (A-I) is the most prevalent type of editing. There are eight known mammalian genes whose amino acid sequences are recoded by A-I conversions (GluR-B, GluR-C, GluR-D, GluR-5, GluR-6, 5HT2cR, ADAR2, KCNA1 (Hoopengardner et al., 2003; Keegan et al., 2001)). Inosine is read by the translational machinery and reverse transcriptase as a guanosine; so A-I edits appear as A-G mismatches between genomic and expressed sequences. We have focused on recoding A-I mRNA editing sites. A recoding site is a site where editing alters the amino acid sequence. Many of the known recoding editing sites are critical to receptor function and are potentially implicated in a number of neurological disorders (Kawahara et al., 2004; Keegan et al., 2001). In addition, a mutation in ADAR1 (Adenosine Deaminase Acting on RNA) has recently been identified as the causative mutation behind a human pigmentation disorder (Miyamura et al., 2003). These observations underline the importance of identifying the remaining editing sites in mammals.

Most mammalian recoding sites have been discovered serendipitously, or have been identified through homology to previously discovered edited genes. KCNA1 contains the only experimentally confirmed mammalian site that has been identified through a screen, although indirectly (Hoopengardner et al., 2003). There is some biochemical evidence for additional A–I editing sites based on the amount of inosine in rat brain, compared with the amount expected based on the abundance of editing in the known sites (Paul and Bass, 1998). There have been several computational and experimental screens for mammalian mRNA editing that have identified novel targets, but none of them have managed to identify any of the known recoding sites (Kim et al., 2004; Levanon et al., 2004; Morse et al., 2002). This fact also supports the possibility of additional sites.

RNA editing has been shown to increase protein diversity by introducing altered splicing patterns, frame shifts, alternative start or stop codons, or by directly affecting the protein sequence (Schaub and Keller, 2002). However, recent evidence suggests that the large majority of human A–I editing sites are situated in introns and UTRs, which could either have regulatory functions or could be non-functional (Athanasiadis et al., 2004; Blow et al., 2004; Kikuno et al., 2002; Kim et al., 2004; Levanon et al., 2004). These analyses all used the alignment of expressed sequences to the human genome. Levanon et al. (2004) specifically looked at A-G mismatches in inverted repeats and identified more than 12,723 sites in 1,637 genes, with 95% accuracy. Similar results were obtained in other studies, although there will be significant overlap in their results (Athanasiadis et al., 2004; Kim et al., 2004). All five of the above studies showed that ~90% of the novel edits are located in Alu repeat sequences, which are unique to primates. Kim et al. (2004) also showed that the mouse had only 91 cDNAs with significant evidence of editing compared with 2,674 in human. These data suggest that this form of editing is either specific to or highly exaggerated in primates. The function of these sites remains to be determined. It is striking that none of these studies identified any of the known recoding edited sites. Indeed, the two unconfirmed recoding sites identified by Levanon are the only novel mammalian recoding sites identified by any of these screens. A more recent paper suggests that 85% of all pre-mRNAs contain edited Alu repeats and that editing of Alu repeats can result in their exonisation and incorporation into coding sequence (Athanasiadis et al., 2004). Although this recodes the predicted proteins, it is unclear whether the inclusion of Alu sequences will necessarily result in functional proteins.

*To whom correspondence should be addressed.
A screen for protein recoding sites

Using a biochemical approach, Morse et al. (2002) developed a screen to identify transcripts containing inosine residues. They identified 10 editing sites in nematode and 19 in human. All the sites are in 3' UTRs, introns or non-coding RNAs. A possible reason for this is that their protocol may have been biased towards transcripts containing multiple inosines.

A number of previous protocols have tried to identify novel editing sites. The most successful screen was based in the fruit fly. Hoopengardner et al. (2003) used comparative genomics between Drosophila melanogaster and Drosophila pseudoobscura to identify genes containing regions that were conserved almost perfectly over 50bp or more. This approach successfully identified all the known sites as well as 16 novel recoding RNA editing sites, which were all involved in rapid electrical or chemical neurotransmission (Hoopengardner et al., 2003). Looking for clusters of A–G mismatches in alignments between a Drosophila cDNA collection and the genome also identified one of these sites (Stapleton et al., 2002).

All the known recoding sites are individually functional (Keegan et al., 2001), whereas the functions of editing in non-coding repeats (Levanon et al., 2004) are proposed to be involved in more general processes such as RNA stabilisation, RNAi, protection from retrotransposition or destabilisation of viral duplexes (Kim et al., 2004; Levanon et al., 2004). We have focused our analyses on identifying only the recoding editing sites.

It has been established that many of the known recoding A–I editing sites occur in clusters, are well conserved and are found in imperfect inverted repeats termed ECSs (editing site complementary sequences) (Higuchi et al., 1993). This is illustrated in Figure 1 and Table 1. We have reviewed and analysed these features to determine their predictive power for identifying novel RNA editing sites. We have also attempted to identify other features common to these sites including similarities in local secondary structures or motifs.

We have used a combination of seven predictive features to screen a large set of expressed versus genomic sequence mismatches, including suitable filters to remove SNPs, sequencing errors and alignment errors. In contrast to previous screens this method successfully identifies most of the known A–I recoding sites as well as identifying a novel experimentally validated recoding site.

2 SYSTEMS AND METHODS

The analysis presented here is based on mouse and human, as large amounts of expressed and genomic sequence data are required. We have focused on the mouse as most of the mouse data is from a single strain, which reduces noise from single nucleotide polymorphisms. Human data are only used to confirm the putative mouse editing sites. To reduce the computational expense of this project to a manageable level, we required a collection of reference sequences that contained non-redundant sequences for the majority of the known genes for each species. For this purpose, concatenated exon sequences were obtained from Ensembl (Clamp et al., 2003) (based on mouse NCBI version 30 and human NCBI version 33). These included most of the known exonic recoding editing sites, with the exception of Glur-6 which was obtained from GenBank (Benson et al., 1999) using the accession NM_010348. Orthologous mouse–human pairs were obtained from Ensembl. Where multiple homologues were predicted, we analysed each of them for sequence conservation and putatively orthologous mismatches, and then used the best homologous gene in the remaining analyses.

Mismatches were identified using MegaBlast (Altschul et al., 1997) (version 2.2.6) searches of mouse and human Ensembl genes against all publicly available expressed and genomic sequences for their respective species. BLAST matches that were <100 bp long, or <98% nucleotide identity, were discarded. This threshold removed low quality sequences and matches from homologous genes, but it also removed one edited EST from a known editing site. Unknown edited transcripts may also have been removed. Edited sequences were obtained from dbEST (Boguski et al., 1993) and GenBank (Benson et al., 1999). Genomic sequences were obtained from the EMBL hgt repository for both mouse and human. Mouse shotgun trace repository sequences were obtained from Ensembl (Clamp et al., 2003). All sequences were up to date as of September 2003. Clone and strain data were obtained from GenBank (Benson et al., 1999). Clone identifiers were used to remove redundancy from the set of expressed sequences. An editing region is defined here as an exon with one or more editing site. A necessary limitation of this protocol is that editing sites would not be observed if they are intronic or do not occur within an Ensembl transcript. This was the case for three of the known RNA editing regions (ADAR2 and Glur-6 QR and I/V sites). The remaining known editing sites constitute the positive control set for this protocol.

An initial set of 28,992 A–G mismatches found between the genomic and expressed sequences, but not between genomic sequences was constrained. Every mismatch was then analysed for seven features; (1) Number of putatively edited mouse cDNAs or ESTs with the same mismatch as the same position (Allowed values: 1, 2, >2); (2) Number of non-edited mouse cDNAs or ESTs combined with the number of publicly available genomic sequences for each given mismatch (Allowed values: 1, 2, >2); (3) Where possible the human homologues were aligned using Lagan (Brudno et al., 2003). We considered putative mouse sites to be conserved in human if they were also observed as a putative editing site at the equivalent location in human expressed sequences (Allowed values: Y;N); (4) We calculated the effect of the edit on the amino acid sequence by BLAST (Altschul et al., 1997) searching the Ensembl nucleotide sequence against the equivalent protein sequence, then mapping the putative editing site onto the alignment. This allowed us to distinguish between edits that alter the amino acid sequence and those that do not (Allowed values: Synonymous, Non-synonymous); (5) Sequence conservation was analysed using the same Lagan mouse/human alignments, from which the best conserved 120 bp window overlapping each putative editing site was selected (Continuous variable); (6) Putative mouse ECSs were identified by scanning for inverted repeats using a Smith–Waterman alignment algorithm from EMBOSS (Rice et al., 2000), Water, based on a scoring matrix modified for RNA base pairing specificities (available on
request). The alignment was generated between a 70 bp region flanking the editing site and a reversed flanking 4kb region. This test is unavoidably biased against edits that occur towards the end of inverted repeats (Continuous variable); (7) Clusters of sites were defined by the observation of more than one putative editing site within an exon (Continuous variable).

The results of these analyses were combined using a relative entropy approach (Lim et al., 2003). For a given feature i with a value \( x_i \), we assign a log-odds (LOD) score:

\[
s_j(x_i) = \log 2 \left( \frac{f_i(x_i)}{g_i(x_i)} \right), \quad S = \sum_{i=1}^{7} s_j(x_i),
\]

where \( f_i(x_i) \) is the proportion of all the positive controls in an interval containing feature value \( x_i \) and \( g_i(x_i) \) is the proportion of the remaining ~29,000 A–G mismatches in the same interval. The proportions used are smoothed to avoid over-fitting to the limited number of positive controls (see Supplementary methods for details of this and the calculation of intervals). The overall score assigned to a putative editing site is the sum of the LOD scores for the seven features. The mismatches were then ranked by their total scores. Methods to reduce sources of false positives, avoid over-fitting to the positive controls and the experimental validation techniques are described in the Supplementary methods.

3 RESULTS

3.1 Analysis of known editing sites

Recoding A–I edits tend to be conserved across related species (Emeson and Singh, 2000). Twelve of the thirteen A–I recoding mouse edits shown in Table 1 were supported by A–G mismatches in the public databases, and eight of these were also observed in human expressed sequences (although four of these are from the same cluster in the 5HT2R gene). The levels of sequence conservation observed between the A–I mouse and human editing sites varied between 99 and 82% identity over 120 bp around the editing site (Table 1). This high level of conservation agrees with previous observations (Hoopengardner et al., 2003) and suggests that sequence conservation is a useful predictor of recoding editing sites.

Several of the known recoding sites have published ECSs in nearby introns (Burns et al., 1997; Dawson et al., 2004; Herb et al., 1996; Higuchi et al., 1993; Lomeli et al., 1994; Sommer et al., 1991). Our novel method for finding mouse ECSs was able to correctly identify four out of seven of these ECSs. Putatively, orthologous ECSs could also be observed in human for these four ECSs (data not shown). The remaining ECSs, overlapping the GluR-B, GluR-5 and GluR-6 Q/R sites, were not identified owing to the identification of higher scoring putative ECSs nearby. We also identified putative ECSs for several of the other positive controls, although they were generally weaker and did not tend to occur in the 3′ introns as with most previously characterized ECSs (Table 1). We analysed the regions incorporated between the editing sites and their putative ECSs, seeking any common features in the secondary structure or common motifs. Neither analysis identified any convincing results (data not shown).

One feature of many recoding sites is that in addition to one highly edited adenosine, other nearby adenosines are also edited, although to a lesser extent. We define an editing region as an exon that contains one or more known editing sites. Table 1 demonstrates the usefulness of a cluster analysis as 5 of the 11 A–I regions contain a cluster. Finally, all the A–I editing regions contain at least one recoding edit, many of which have been shown to be functional and some have been implicated in disease.

Through simple BLAST searches of the publicly available cDNA and EST databases, we were able to identify edited sequences expressed in the nervous system for all the mammalian A–I recoding edits, except for the KCNA1 site, the ADAR2 site and the GluR-D N/S site. This observation agrees with previous reports on mammalian (Emeson and Singh, 2000) and Drosophila recoding editing sites (Hoopengardner et al., 2003) which suggest that most A–I recoding edits are specific to the brain and associated tissues. Although this could be a powerful predictor of novel editing sites, it would introduce bias against editing in other tissues and therefore, we have not included it in this analysis.

### Table 1. Sequence conservation and ECS predictions for the known recoding edited regions

<table>
<thead>
<tr>
<th>Edit type</th>
<th>Edited gene</th>
<th>Protein modification</th>
<th>Edits in cluster (Recoding5)</th>
<th>Seq. cons.(^a) over 120 bp (%)</th>
<th>ECS score(^b)</th>
<th>Relative ECS position(^c)</th>
<th>Agrees with known ECS location(^d)</th>
<th>Identified by our protocol(^e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A–G GluR-B</td>
<td>Q/R</td>
<td>4 (I)</td>
<td>99.2</td>
<td>112</td>
<td>+324</td>
<td>N (Higuchi et al., 1993)</td>
<td>Yes—Ranked 6th</td>
<td></td>
</tr>
<tr>
<td>A–G GluR-B</td>
<td>R/G</td>
<td>1 (I)</td>
<td>92.5</td>
<td>106</td>
<td>+68</td>
<td>Y (Lomeli et al., 1994)</td>
<td>Yes, but ranked very poorly.</td>
<td></td>
</tr>
<tr>
<td>A–G GluR-5</td>
<td>Q/R</td>
<td>1 (I)</td>
<td>95.8</td>
<td>69</td>
<td>−300</td>
<td>N (Herd et al., 1996)</td>
<td>Yes, but ranked very poorly.</td>
<td></td>
</tr>
<tr>
<td>A–G 5HT2R</td>
<td>Various</td>
<td>4 (I)</td>
<td>95.8</td>
<td>121</td>
<td>+143</td>
<td>Y (Burns et al., 1997)</td>
<td>Yes—Ranked 1st, 2nd, 4th and 5th</td>
<td></td>
</tr>
<tr>
<td>A–G GluR-C</td>
<td>R/G</td>
<td>1 (I)</td>
<td>98.3</td>
<td>140</td>
<td>+58</td>
<td>Y (Lomeli et al., 1994)</td>
<td>Yes—Rank 8th</td>
<td></td>
</tr>
<tr>
<td>A–G GluR-D</td>
<td>R/G</td>
<td>1 (I)</td>
<td>95.0</td>
<td>98</td>
<td>+61</td>
<td>Y (Lomeli et al., 1994)</td>
<td>Yes—Ranked 9th</td>
<td></td>
</tr>
<tr>
<td>A–G GluR-D</td>
<td>Various</td>
<td>4 (I)</td>
<td>88.3</td>
<td>74</td>
<td>+6</td>
<td>—</td>
<td>Yes, but ranked poorly.</td>
<td></td>
</tr>
<tr>
<td>A–G KCNA1</td>
<td>I/V</td>
<td>1 (I)</td>
<td>99.2</td>
<td>76</td>
<td>−1537</td>
<td>—</td>
<td>No—mismatch not observed</td>
<td></td>
</tr>
<tr>
<td>A–G ADAR2</td>
<td>Alt. Splice</td>
<td>8 (—)</td>
<td>—</td>
<td>NA</td>
<td>NA</td>
<td>—</td>
<td>No—it is intronic</td>
<td></td>
</tr>
<tr>
<td>A–G GluR-6</td>
<td>I/V</td>
<td>2 (1)</td>
<td>83.3</td>
<td>72</td>
<td>−1490</td>
<td>—</td>
<td>No—it was not in mouse Ensembl 30</td>
<td></td>
</tr>
<tr>
<td>A–G GluR-6</td>
<td>Q/R</td>
<td>1 (I)</td>
<td>81.7</td>
<td>69</td>
<td>+1717</td>
<td>N (Herd et al., 1996)</td>
<td>No—it was not in mouse Ensembl 30</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Sequence conservation is generated between 120 bp surrounding the mouse site and the orthologous human sequence. \(^b\)ECS score is generated by an alignment algorithm using RNA base-pairing specificities (see systems and Methods Section). \(^c\)Location of the 5′-end of the ECS, relative to the editing site (+ indicates downstream). \(^d\)If the predicted ECS agrees with the published ECS this is recorded with a Y. References are provided for the initial publications identifying the known ECSs. \(^e\)Brief comments on how the sites ranked or why they were not found. \(^f\)Editing of ADAR2 results in alternative splicing by creating an alternative 3′ splice site. \(^g\)Highlighted recoding sites are those used as a positive control set (see System and methods section).
The sum of all seven LOD scores for each mismatch. The distribution of the positive controls skews heavily towards the right. The cluster size is the number of A–G mismatches in the exon containing the given mismatch. Sequence conservation over 120 bp overlapping the mismatch. The cluster size demonstrates that the three variable features and the total LOD score combined are capable of predicting novel A–G mismatches. This scoring system identified 7 of 11 positive control edits in the 10 top ranked mismatches (including the GluR-B Q/R, GluR-C R/G, GluR-D R/G, and all 4 5HT2C-R edits). Of the four remaining positive controls, KCNA1 was missed as we did not find any matching edited sequences, while the rest were ranked badly owing to poor sequence conservation, poor ECS predictions, the lack of clusters or the lack of orthologous mismatches in human. These results are also relatively robust to modifications in the smoothing and binning of the score distributions (data not shown). Complete results for the 20 top ranked edits are given in the Supplementary data (Table S1).

The top highest scoring novel candidate recoding edits were experimentally tested for evidence of editing. We tested mouse brain and heart RT–PCR products for evidence of editing at the predicted sites. The brain was chosen as all the known A–I recoding sites are edited in this tissue, and the heart was chosen as a control. Athanasidis et al. (2004) tested brain and lung for the same reasons. For 9 of the top 10 novel candidates, there was no evidence of editing. It is possible that these genes are edited in other tissues. However, exhaustive testing of these sites in every tissue and developmental stage was beyond the scope of this work. The top ranking novel edit, BC10, contains a novel editing region, which we have experimentally verified (Fig. 3).

We tested 17 other novel candidate editing sites, in addition to the 10 highest scoring novel candidates. These candidates were randomly selected and vary widely in their ranks. None of these candidates showed experimental evidence of editing, which suggests that recoding editing is only common at the top end of the distribution, and confirms the reliability of the scoring system.

### 3.3 Confirmation of BC10—a novel A–I editing region

The top novel candidate, BC10, shows all the features of the positive controls. It has a very high scoring putative ECS 480 bp 5′ of the most frequently edited site; its edited region is 99% identical between mouse and human; it shows orthologous A–G mismatches in human; it is highly edited in brain tissue, affects the amino acid sequence and the editing sites occur in a tight cluster. The three recoding sites are supported by multiple ESTs/cDNAs in both species (Fig. 3). The putative ECS is the second highest scoring ECS from the ∼30,000 A–G mismatches. The region containing these sites has been tested and editing has been verified in the lab. Interestingly, most of the editing was observed in the intron across the length of the predicted ECS and was specific to brain. The low number of edited RT–PCR products from the exonic region was partially due to the frequency distributions before generating LOD scores. In addition, we used a conservative jack-knife approach which ensured that each positive control was scored using only the non-related positive controls (see Systems and methods section).

This is the first mammalian screen to identify any of the known recoding editing sites. Figure 2 shows the distributions for the three continuous variables (sequence conservation, clustering and ECS score) and the total LOD scores of all A–G mismatches. This figure demonstrates that the three variable features and the total LOD score are useful and efficient for distinguishing the positive controls from the remaining mismatches.

This scoring system identified 7 of 11 positive control edits in the 10 top ranked mismatches (including the GluR-B Q/R, GluR-C R/G, GluR-D R/G, and all 4 5HT2C-R edits). Of the four remaining positive controls, BC10 was identified as the top ranking novel candidate recoding edit. BC10 is a novel gene that is specifically expressed in the brain and is predicted to encode a small globular protein containing two transmembrane helices. All the editing sites are found in either the 5′-UTR or the N-terminal section of the protein, which is predicted to be outside the membrane. The three coding edits are all non-synonymous and predicted to encode exposed residues. A species alignment shows that this region of BC10 is exceptionally well conserved from human down to Caenorhabditis elegans, suggesting that it is a fundamentally important gene (data not shown). Notably, the three recoding edited adenosines are conserved in all the species as well as most of the adjacent bases. This suggests that editing of this region may be conserved across all of these species.
The underlined nucleotides show the start of the coding region. The intervening 448 bp are shown by the arrow on the left hand side. The supporting evidence and the coding modification of the putative edits are given above/below the sites. There are four sources of edited sequences.

The number of intronic brain RT–PCR products edited at this site (out of 50).

The number of exonic brain RT–PCR products edited at this site (out of 23).

The number of publicly available mouse brain expressed sequences edited at this site (out of 28).

The number of publicly available human expressed sequences edited at the orthologous sites (out of 85).

4 DISCUSSION

Our protocol is sensitive, given that it identified 7 of the 11 positive control edits. In contrast, no other screen for mammalian editing sites has identified any of the known recoding editing sites. Our protocol is also specific as five of the top eight genes contain known or experimentally verified sites. One of these is a confirmed novel editing region, BC10, which is an extremely well conserved gene implicated in renal and bladder cancer (Gromova et al., 2002; Rae et al., 2000). These results demonstrate that this is a useful and efficient method for identifying both known and novel A–I editing sites.

Using other computational protocols two groups have shown that there are over 1500 genes edited in introns or non-coding regions (Kim et al., 2004; Levanon et al., 2004). This shows a clear difference in magnitude between the number of coding and non-coding editing sites. The results of Levanon et al. (2004) suggest that the occurrence of strong a ECS is a very good predictor of editing sites in non-coding sequences. Here, we find that our ECS predictions are only moderate predictors for recoding sites and that a combination of features must be used to identify these sites. A better ECS prediction method could improve this situation. In contrast, we found that sequence conservation and the observation of orthologous mismatches in human are strong indicators of recoding sites. Notably, these features cannot be applied to editing sites in Alu repeats, as they are not conserved in the mouse.

Despite the success of this protocol for identifying many of the known recoding sites, it is possible that there are many more sites to identify in addition to BC10 and the previously known sites. For example, it is possible that some of the genes we experimentally tested may be edited, but that the degree of editing was too low or restricted to particular tissues or developmental stages other than adult brain and heart. The frequency of editing is 100% for only the GluR-B Q/R site, whereas the frequency of editing for other sites is often much lower (Keegan et al., 2001). It can also be developmentally regulated as with the AMPA receptor R/G sites (GluR-B,C and D), which are poorly edited early in mouse development (Lomeli et al., 1994). Although all the known mammalian recoding sites are edited in adult brain (Keegan et al., 2001), there may be some ascertainment bias towards brain editing. A large proportion of the publicly available expressed sequence data is from brain libraries, adding to this bias. Both the editing enzymes and edited Alu elements appear in a range of tissues (Keegan et al., 2001; Kim et al., 2004; Levanon et al., 2004), supporting the possibility of recoding editing in these tissues. Indeed, the disease phenotype of the ADAR1 null allele in humans is a skin condition, rather than neurological (Miyamura et al., 2003). It is also possible that there is an additional class of recoding sites that do not conform to the features used in this analysis; however, there is no evidence for this.

Environmental factors, such as the disease state of the tissue, may also be important. One form of ADAR1 is known to be interferon inducible (Patterson and Samuel, 1995), suggesting the existence of sites that are edited only during inflammation (Yang et al., 2003). Aberrant A–I editing of the endothelin receptor has been implicated in Hirschsprung disease (Tanoue et al., 2002), while aberrant C–U editing has been shown to induce liver dysplasia and carcinomas (Yamanaka et al., 1995, 1997). It is possible that there are many more sites that could be aberrantly edited, some of which could contribute to disease. When looking for ECSs we identified >1500 mismatches with inverted repeats that scored better than half of the known recoding sites. This suggests that there are a lot of potential hairpins formed between exons and their introns, which the editing enzymes could potentially bind. It is interesting to ask whether aberrant editing of these potential hairpins could be involved in disease.

Our results demonstrate that sequence conservation between mouse and human and the observation of identical orthologous mismatches are powerful predictors of recoding editing sites. As a result, this scoring system will be less useful for identifying putative species-specific sites. However, most of the positive controls have been shown to be widely conserved throughout mammals (Keegan et al., 2001).

In addition to recoding sites, there are many non-coding sites remaining to be discovered or characterized (Kim et al., 2004; Levanon et al., 2004). Understanding the functions of these non-coding sites is vital for a more complete understanding of RNA editing. This work has identified many of the known mammalian recoding editing sites and one novel edited region in BC10 and it is clear that there may be further sites to be identified. However, the present data suggest that recoding editing is a rare phenomenon, both as a proportion of total editing activity and as a proportion of affected exons in the mammalian genome.

ACKNOWLEDGEMENT

This work was funded by the Medical Research Council, UK.
REFERENCES


Paul,M.S. and Bass,B.L. (1998) Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. EMBO J., 17, 1120–1127.


