

A bioinformatics approach to identify recoding events of A-to-I RNA editing

Stefan Maas¹, Daniel Lopresti², Derek Drake³, Rikhi Kaushal¹,
Stephen Hookway², Walter Scheire², Mark Strohmaier², and Christopher Wojciechowski²

¹Department of Biological Sciences and ²Department of Computer Science & Engineering, Lehigh University, Bethlehem, PA 18015

³Department of Computer Science, Purdue University, West Lafayette, IN

1 Introduction

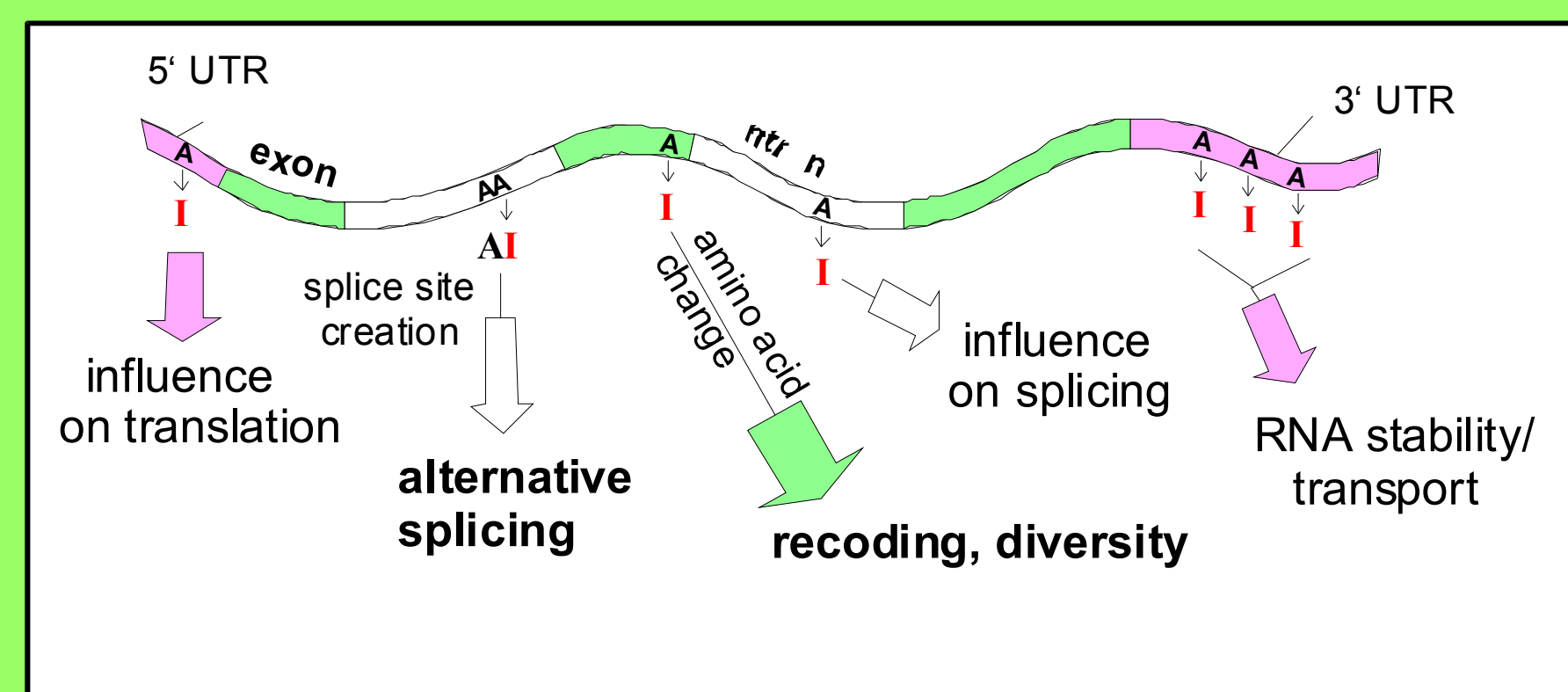
RNA editing by A-to-I modification in pre-mRNAs constitutes a major mechanism for the generation of **RNA and protein diversity** in mammals and is known to regulate important functional properties of neurotransmitter receptors in the central nervous system. A-to-I RNA editing can also **create or destroy pre-mRNA splice signals** or lead to alterations in RNA secondary structures.

We have recently identified widespread editing in 5'- and 3'-untranslated mRNAs involving Alu repeat elements in the human transcriptome (Athanasiadis et al., *PLoS Biology* 2004) using a combined bioinformatics and experimental screening and validation strategy.

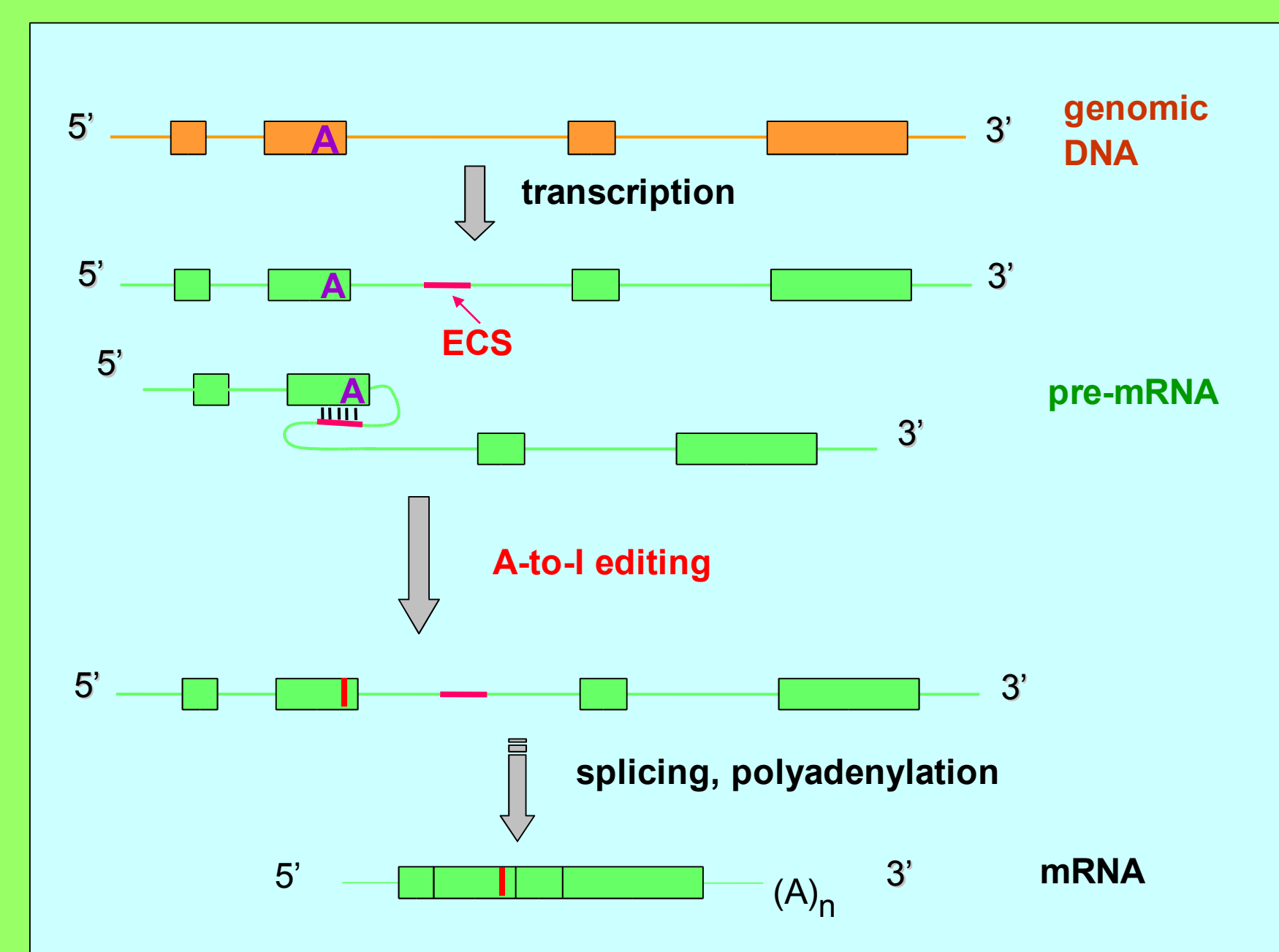
In all known and characterized cases of **recoding by A-to-I editing** the ensuing amino acid substitutions have been linked to alterations in protein function. A few additional cases of recoding due to RNA editing were recently identified through bioinformatics approaches but the total number of targets is still low despite evidence that many more should exist.

Here we present our ongoing work to develop an algorithm that comprehensively identifies site-selective A-to-I editing events in human mRNAs. This approach will be optimized to identify genes that harbor single or few editing events, such as the well characterized codon changes in glutamate and serotonin receptor transcripts. High scoring candidates will then be experimentally validated.

2 Known and proposed functions of A-to-I editing



3 RNA fold back structures are essential for A-to-I editing



4 Bioinformatics approach for identification of recoding events due to RNA editing

The flow chart summarizes our initial transcriptome-wide search strategy for A-to-I edited genes integrating the following parameters:

1) The exonic strand of the predicted fold-back structure must exhibit at least one A/G-discrepancy between genomic and cDNA sequence that

2) has not been annotated as a SNP [according to the dbSNP database; excluding SNPs that lack genomic confirmation since SNPs based solely on expressed data might represent RNA editing events instead of polymorphisms (49)].

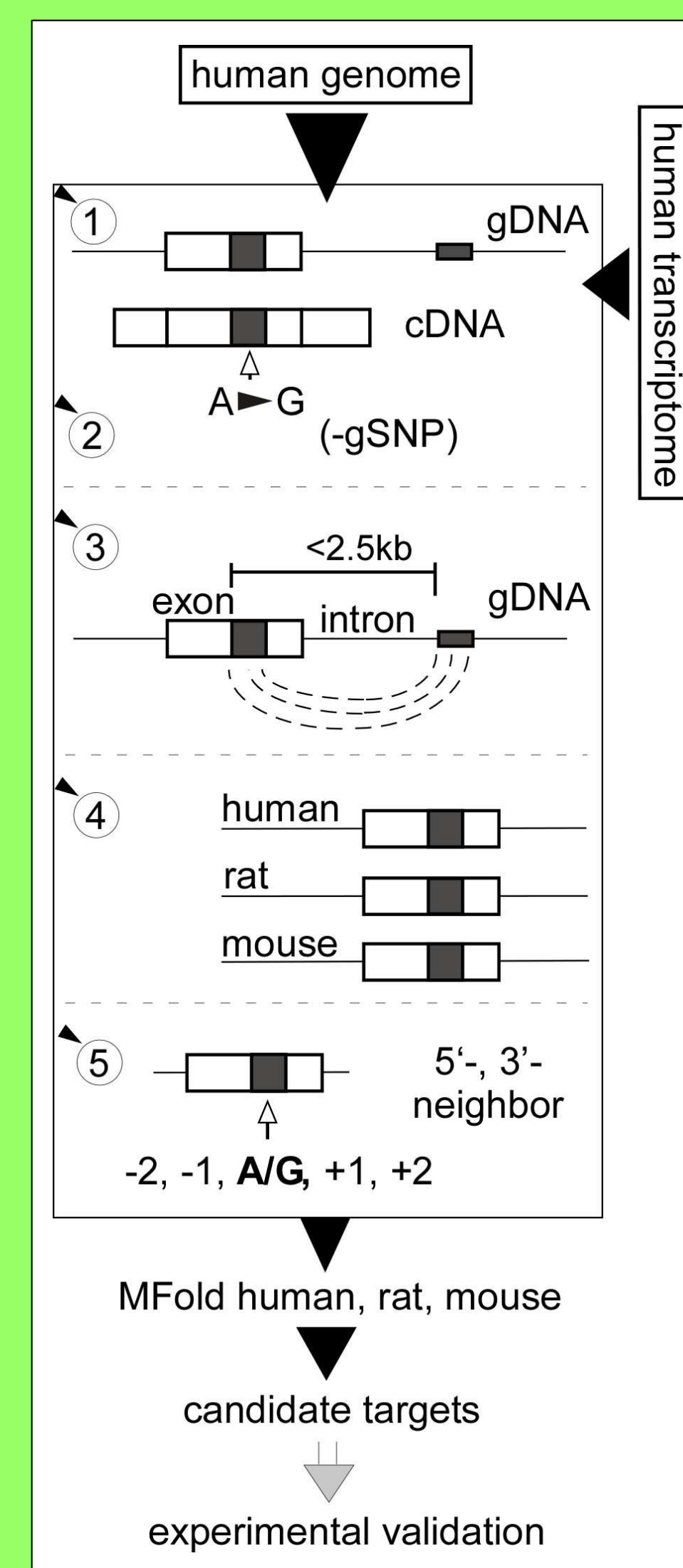
3) Our preliminary results outlined in section A indicate that the distance between oppositely oriented repeat sequences is critical for editing and we found that significant editing levels are only observed for distances below ca. 2.5kb (38). One of the pairing sequences in each case must be contained within an exon, whereas the other can be in any region of the primary transcript, which statistically will be most likely an intron. As for our preliminary studies, we will use the UCSC annotation database for this experiment.

4) Cross-species conservation. This property is based on the observation that the sequences surrounding the to be edited adenosine are often more strongly conserved between mammalian species than exon sequences that are not involved in a functionally relevant secondary structure (80).

5) 5'-, 3'-neighbor preferences of the editing machinery (obtained from our preliminary studies).

6) The strongest hits are folded using the Mfold algorithm (81).

The algorithm will be trained and refined using 11 well characterized editing sites within the pre-mRNA sequences of glutamate receptor subunits GluR-B, -C, -D, -5, -6, serotonin receptor subunit 5-HT2C as well as the potassium channel gene KCNA1 (12, 80). Four recently identified recoding targets (49, 50), BC10, IGFBP7, FLNA, and CYFIP2, and newly validated recoding targets from our preliminary work will also be included resulting in a **total number of 17 control sequences** to date. Since currently there is no edited KCNA1 cDNA sequence in the GenBank database, we will introduce one edited sequence into our source database. Additional targets will be included in the training set as they become identified and validated.



5 Computational screen for potential RNA editing sites

Our computational screen for potential RNA editing sites is implemented as a pipeline involving of a number of processing stages. The current code consists of a set of programs written in the C language running on a Unix-compatible operating system, and makes use of a number of standard built-in utilities and open source software, including the MySQL database server. All benchmark timings given below are for runs performed on a Sun Microsystems Sun-Blade-150 workstation with an UltraSPARC-IIe 650MHz CPU, 512 megabytes of main memory, and a 512 kbyte cache.

At the first stage, the locations of all discrepancies of a specified type between a genomic DNA sequence and its corresponding mRNA's is output. For example, the sites of all A G mismatches that occur in regions marked as exons might be recorded. Database tables from the UCSC Genome Browser (<http://genome.ucsc.edu/>) are downloaded locally and queried using the MySQL server to produce all mRNA's associated with a given chromosome. This requires approximate 0.2 seconds of CPU time and yields 6,901 mRNA's in the case of human Chromosome 4, for example. This list is then filtered, leaving only the sites for which the discrepancy between the genomic and mRNA sequence reflects an A→G substitution (including proper handling of indeterminate nucleotides) and for which the A is located in a marked exon. For human Chromosome 4, this yields 2,832 potential editing sites and requires approximately 10 seconds of CPU time. These differences are stored in a file for further processing.

This list is then processed by the second stage of the computational pipeline, which is designed to correlate the same RNA editing event appearing in multiple, overlapping mRNA sequences. The output from this stage enumerates all discrepancies corresponding to a given position in the genomic sequence, what the discrepancy was, the percentage of mRNA's having that discrepancy, and a list of the associated accession numbers. This stage requires 0.15 seconds to run. All of the resulting data is stored in a file for future use – there are 2,102 distinct sites for the Chromosome 4 example we have been using to illustrate this discussion – but this data is not currently employed by the later stages in our pipeline (the use of this information is a subject for future research).

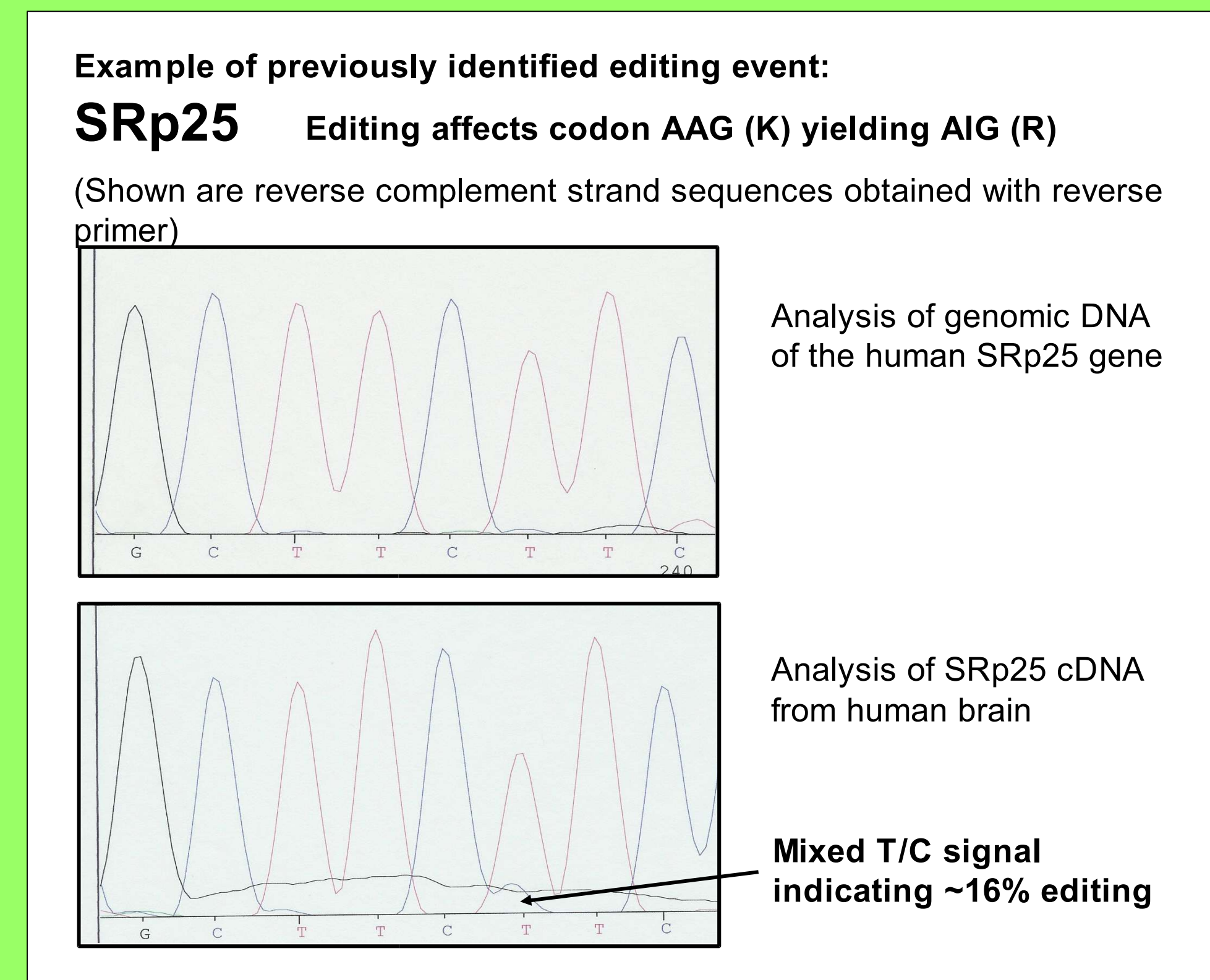
In the next stage, we filter the list of differences recorded so far to remove known SNP's, again using database tables downloaded from the UCSC Genome Browser website. To be conservative about not excluding potential RNA editing sites that have been mistakenly annotated as SNP's, we can limit the filter to exclude, for example, only SNP's that are labeled in the database as having been derived through genomic-level analysis. Rather than discard the filtered output, we separate the discrepancies into two files: those that pass this stage of the screen (i.e., the A→G differences that are not explained as arising from genomic SNP's), and those that do not. In the case of our experiment using human Chromosome 4, 1,719 of the purported editing sites pass this stage of the screen, while 1,113 fail and are excluded. This portion of the computational pipeline requires approximately 10.5 seconds to complete.

Finally, each potential editing site that remains is individually scored by determining a 50-nucleotide window around it and another 50-nucleotide window within 2,500 bases upstream or downstream for which a foldback score is computed. This score is determined by weighing the matches between different base pairings as follows: C→G = 3, A→U = 2, and G→U = 1. Only the optimal score for all possible window "slides" is saved. This is by far the most computationally intensive step in our current pipeline and requires over 5 hours to complete, yielding a ranked list of 1,719 potential RNA editing sites.

Additional parameterized stages for scoring candidate sequences are being added and genes that receive the highest overall scores will then be validated in the laboratory for evidence of RNA editing in vivo and potential downstream effects on the function of the gene products.

6 Experimental validation

Top-scoring candidate genes are currently validated by designing PCR oligonucleotide primers for analysis by RT-PCR and direct sequencing from human brain total RNA (Maas et al., *PNAS* 2001).



Total RNA from human brain tissue reverse transcribed with random primers and paired genomic DNA from the same tissue specimen were used for gene-specific PCR. The amplicons are gel-purified and directly subjected to cycle-sequencing.

The approximate editing extents are determined by comparing the peak heights of mixed signals. This method has been shown to be reliable and reproducible for editing extents >5-10% and <95%. Identified editing events are then further analyzed in more detail by subcloning and multiple clone-sequencing of PCR amplicons. Experimentally validated genes will then be characterized functionally for consequences of editing.

7 Conclusions

- Our computational screen is able to identify known editing positions as high-scoring candidates.
- The screen is unbiased and will allow the delineation of site-selective, recoding editing events.
- The flexible design of this approach permits the user to set feature parameters interactively to identify diverse sets of editing targets.

References

- Athanasiadis, A., Rich, A. and Maas, S. (2004). Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biology*, 2 (12) e391, 1-15.
- Bass, B. L. (2002). RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem* 71, 817-846.
- Batzer MA., and Deininger, PL., 2002: Alu repeats and human genomic diversity. *Nature Rev. Genetics* 3, 370-380
- Higuchi, M. et al., 1993: RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. *Cell* 75, 1361-1370
- Hoopengardner, B., Bhalla, T., Staber, C. and Reenan, R. (2003). Nervous system targets of RNA editing identified by comparative genomics. *Science* 301, 832-836
- Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z. Y., Shoshan, A., Pollock, S. R., Szybel, D., Olshansky, M., Rechavi, G. and Jantsch, M. F. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol* 22, 1001-1005.
- Levanon, E. Y., Hallegger, M., Kinar, Y., Shemesh, R., Djinic-Carugo, K., Rechavi, G., Jantsch, M. F. and Eisenberg, E. (2005). Evolutionarily conserved human targets of adenosine to inosine RNA editing. *Nucleic Acids Research* 33, 1162-1168.
- Maas, S., Rich, A. and Nishikura, K. (2003). A-to-I RNA editing: recent news and residual mysteries. *J Biol Chem* 278, 1391-1394.
- Maas, S., Patt, S., Schrey, M. and Rich, A. (2001). Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. *Proc Natl Acad Sci U S A* 98, 14687-14692
- Bass, BL, 2002: RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* 71, 817-846
- Morin, S., F. Charron, L. Robitaille, and M. Nemer, (2000) GATA-dependent recruitment of MEF2 proteins to target promoters. *Embo J*, 19(9): 2046-55.