Supplementary Text:

Genome-wide Identification and Characterisation of Tissue-specific RNA Editing Events in *D. melanogaster* and their Potential Role in Regulating Alternative Splicing

Alborz Mazloomian and Irmtraud M. Meyer*

Details of the proposed pipeline.

In our analysis, we used dm3 reference genome (fasta file) from UCSC (http://genome.ucsc.edu). *D. melanogaster* annotation file (*BDGP5.74_ensembl.gtf*) was downloaded from the Ensembl web page (http://www.ensembl.org). In making the corresponding annotation file for the *OregonR* genome we employed MUMmer[1, 2] version "3.23" and Needleman-Wunsch[3] program version "0.3.5". To align short reads to the *OregonR* genome we executed: ``tophat2 -F 0 -i 40 -g 40 --library-type fr-secondstrand -r 200 --mate-std-dev 20 --segment-length 16 --read-mismatches 5 --read-edit-dist 5" using TOPHAT2[4] version "v2.0.10". Parameters such as library type and mate standard deviation were chosen based on the information provided on http://www.modencode.org/.

For each candidate position, we require at least 2 and 5 reads for each allele in the flexible and stringent threshold sets, accordingly. We employed SAMTOOLS[5, 6] mpileup to extract the reads covering each position. Sites that contain stars in SAMTOOLS mpileup tracks are also discarded (They present evidence for small insertions and deletions near a candidate site).

Additionally, at least one of the observed nucleotides from each variant should be from a high quality read (phred score of at least 20) and more than 5 nucleotides distant from the read ends. This filter can improve the results in two ways: first, random hexamer priming can cause errors in the 5' starting positions of reads[7]; and second, read ends at splice junctions are prone to being misaligned.[8] We also filter sites where two or more alleles are observed other than the reference allele.

To filter known variations, we use Ensembl fly variant file http://uswest.ensembl.org/info/data/ftp/index.html (Ensembl release 74). Because variations reported in the file only contains variations of chromosomes X, 2 and 3, we ignored all predictions from other regions.

We filter candidates with log likelihood score smaller than 3. Additionally, we require editing ratio to be between 0.03 and 0.97, in order to lower the chance of including homozygous sites in our predictions,[7] since sequencing and mapping errors are inevitable. These thresholds are equal in both sets of threshold values.

The thresholds for all four of the SAMTOOLS/BCFTOOLS tests are set to 0.15 in flexible thresholding and 0.02 for the stringent thresholding. Our results were generated using SAMTOOLS version "0.1.19".

We employ RNAFOLD[9] with default parameters and RNAPLFOLD[10] with ``-W 200 -L 150 -u 1" as suggested;[11] and for each site we calculate the average of pairing probabilities for a local region of length 5 (candidate position extended by two nucleotides from each side). A candidate site passes the structural filter if it is in a highly structured region (based on RNAFOLD[9] energy) or it shows evidence for being a part of a stem (based on RNAPLFOLD[10] energy). We set RNAFOLD thresholds to -10 and -50 for the flexible and stringent threshold sets and we set RNAPLFOLD thresholds to 0.2

and 0.7, accordingly. The analysis in the paper was carried out using RNAFOLD version "2.0.4" and RNAPLFOLD version "2.0.7".

For finding alternatively used exons, we applied DEXSEQ[12] version "1.8.0". In cases that there are transcripts with overlapping exons with different boundaries, DEXSEQ cuts the exons into multiple parts (see [12] for more details) and analyses their usage separately. Each of these exonic parts are considered as an exon in our analysis when we investigate the potential inter-relation between editing and splicing, however, we only report the ones that are longer than 10 nucleotides. Additionally, when we compare two tissues, we only consider genes that are predicted to have FPKM (fragments per kilobase of transcript per million fragments mapped) expression values greater than 2. Expression values were computed by employing CUFFLINKS[13] package version "2.2.1".

In our analysis, we classify exonic regions into two groups: for each gene, we put all the exons in all the transcripts together; then we find the union of these exonic regions. Next, for each region, if the region constitutes multiple exons that are not identical, we call the region an exonic region with multiple acceptor/donor sites. The other group contains all the other exonic regions.

When we searched for structural features using TRANSAT,[14] we only considered those helices that contain at least 8 base-pairs. The 15 fly species alignment was downloaded from USCS (http://genome.ucsc.edu) for regions of interest. We added OregonR genome to the alignments and realigned the 16 sequences in each region by employing MUSCLE[15] (version 3.8.31).

Micro-RNA target sites were downloaded from http://microrna.org (August 2010 release), and miRNA sites were downloaded from: http://www.mirbase.org (miRBase v19).

References.

1. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. Alignment of whole genomes. Nucleic Acids Research 1999; 27:2369-76.

2. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Research 2002; 30:2478-83.

3. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 1970; 48:443-53.

4. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 2013; 14:R36.

5. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011; 27:2987-93.

6. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics 2009; 25:2078-9.

7. Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB. Identifying RNA editing sites using RNA sequencing data alone. Nature Methods 2013; 10:128-32.

8. Rodriguez J, Menet JS, Rosbash M. Nascent-seq indicates widespread cotranscriptional RNA editing in Drosophila. Molecular Cell 2012; 47:27-37.

9. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research 1981; 9:133-48.

10. Bernhart SH, Hofacker IL, Stadler PF. Local RNA base pairing probabilities in large sequences. Bioinformatics 2006; 22:614-5.

11. Lange SJ, Maticzka D, Möhl M, Gagnon JN, Brown CM, Backofen R. Global or local? Predicting secondary structure and accessibility in mRNAs. Nucleic Acids Research 2012; 40:5215-26.

12. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Research 2012; 22:2008-17.

13. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols 2012; 7:562-78.
14. Wiebe NJ, Meyer IM. Transat—a method for detecting the conserved helices of functional RNA structures, including transient, pseudo-knotted and alternative structures. PLoS Computational Biology 2010; 6:e1000823.
15. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 2004; 32:1792-96.