Identifying mRNA Editing Deaminase Targets by RNA-Seq

Brad R. Rosenberg, Scott Dewell, and F. Nina Papavasiliou

Abstract

RNA editing deaminases act on a variety of targets in different organisms. A number of such enzymes have been shown to act on mRNA, with the resultant nucleotide changes modifying a transcript’s information content. Though the deaminase activity of mRNA editing enzymes is readily demonstrated \textit{in vitro}, identifying their physiological targets has proved challenging. Recent advances in ultra high-throughput sequencing technologies have allowed for whole transcriptome sequencing and expression profiling (RNA-Seq). We have developed a system to identify novel mRNA editing deamination targets based on comparative analysis of RNA-Seq data. The efficacy and utility of this approach is demonstrated for APOBEC1, a cytidine deaminase with a known and well-characterized mRNA editing target in the mammalian small intestine.

Key words: RNA editing, Cytidine deaminase, Adenosine deaminase, APOBEC1, RNA-Seq

1. Introduction

RNA editing deaminases act on diverse targets in many organisms, ranging from bacteria to mammals. Editing enzymes that act on mRNA posttranscriptionally introduce considerable diversity to the transcriptome. Single nucleotide alterations can be sufficient to alter protein products (1) or modulate gene expression (2). Though the RNA editing activity of many deaminases is readily demonstrated \textit{in vitro}, identifying physiological targets \textit{de novo} has proved a challenge. As mRNA editing often modifies an individual base, comprehensively screening for new targets effectively means detecting single nucleotide changes across the entire transcriptome.

Recent advances in DNA sequencing technology provide new platforms for studying mRNA editing. Ultra high-throughput short read sequencing now allows for whole eukaryotic transcriptome sequencing and expression profiling (RNA-Seq) (3, 4). Using a similar approach, we have developed a method to identify
novel mRNA editing targets based on the comparative analysis of whole transcriptome sequencing data.

Following generation of a transcriptome-wide short read data set, the analysis workflow is geared specifically to identify mRNA editing events. As a comparative analysis technique, this approach requires mRNA isolated from cells/tissues/animals deficient in the particular deaminase under investigation, as well as from an editing-competent control. In the examples provided here, whole intestine tissue from Apobec1−/− mice and corresponding C57/Bl6 control animals are used to demonstrate proof-of-concept. APOBEC1 deaminates a specific cytidine to uridine in its well-characterized target transcript, apoB; editing is completely ablated in the knockout animal (5). If a genetic deaminase-null model system is impractical or unavailable, mRNA derived from other comparative systems (e.g., RNAi knockdown of a deaminase, overexpression in cell line systems, etc.) is also acceptable, provided the genetic backgrounds are suitably controlled. If an editing-deficient experimental system is not available, an alternative method for detecting mRNA editing has been described (6), though it does not provide a comprehensive whole transcriptome profile and cannot establish enzyme specificity.

The experimental strategy and analysis scheme are outlined in Fig. 1. Sample preparation and ultra high-throughput sequencing are similar to RNA-Seq procedures described elsewhere (3). Beginning with total RNA from control and deaminase-deficient cells, polyA+ mRNA is isolated and used to prepare RNA-Seq libraries. These are subjected to ultra high-throughput sequencing at high coverage. Sequencing reads from both samples are mapped independently to a reference genome, thereby generating separate “consensus” alignments. For each consensus, single nucleotide mismatches to the reference genome are identified using a “quality conscious” algorithm. These mismatch datasets are then filtered against several criteria, maintaining only those mismatches representative of potential deaminase-specific mRNA editing events. The two datasets are then intersected, with read:reference mismatches present in control mRNA sequences but absent from deaminase-deficient mRNA sequences output as potential editing hits. Finally, with genome and transcript coordinates, standard Sanger sequencing may be used to validate the novel mRNA editing targets.

2. Materials

2.1. RNA-Seq Library Preparation

1. MicroPoly(A)Purist Kit (Ambion, Austin, TX)
2. RNA fragmentation buffer (5×): 200 mM Tris–acetate, pH 8.2, 500 mM potassium acetate, 150 mM magnesium acetate
3. SuperScript III Reverse Transcriptase (Invitrogen, Carlsbad, CA)
Identifying mRNA Editing Deaminase Targets by RNA-Seq

4. Random Primers (Invitrogen, Carlsbad, CA)

5. SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen, Carlsbad, CA)

6. QIAquick PCR Purification Kit (Qiagen, Valencia, CA)
7. Klenow Fragment (3’ → 5’ exo-) (New England Biolabs, Ipswich, MA)
8. Quick Ligation Kit (New England Biolabs, Ipswich, MA)
9. PE Adapter Oligo Mix (Illumina, San Diego, CA)
10. HyLadder 1 kb (Denville Scientific, Metuchen, NJ)
11. QIAquick Gel Extraction Kit (Qiagen, Valencia, CA)
12. Phusion High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA)
13. PCR Primers PE 1.0 and PE 2.0 (Illumina, San Diego, CA)
14. Nanodrop 1000 spectrophotometer (Thermo Scientific, Wilmington, DE)
15. Optional Agilent 2100 Bioanalyzer (plus RNA 6000 Pico kit) (Agilent Technologies, Santa Clara, CA)

2.2. Ultra High-Throughput Sequencing

1. Single-Read Cluster Generation Kit v2 (Illumina, San Diego, CA)
2. 36-Cycle Illumina Sequencing Kit v2 (Illumina, San Diego, CA)
3. Illumina Genome Analyzer II (Illumina, San Diego, CA)
4. GA Pipeline 1.4.0 software (Illumina, San Diego, CA)

2.3. Software and Computer Requirements

2.3.3. Recommended Configuration for Mapping and Analysis Computer

3. Recommended configuration for mapping and analysis computer: 2.8 GHz Intel Core 2 Duo or equivalent (additional cores will improve mapping and alignment performance); 8 GB RAM; Linux OS (any release)

3. Methods

The methods described in Subheadings 3.1–3.5 are performed in parallel for both control and deaminase-deficient samples.

3.1. RNA-Seq Library Preparation

The following RNA-Seq library preparation protocol was adapted from Ref. (3) and Illumina product literature.

1. Total RNA can be isolated from cells/tissue by standard methods of choice (see Note 1). The starting total RNA should be of high quality as determined by gel electrophoresis...
or Agilent Bioanalyzer 2100 analysis (Fig. 2). For a typical preparation, starting with more than 25 µg of total RNA is recommended, though we have had success with less than 10 µg total RNA (see Note 2).

Fig. 2. Size distribution of RNA preparations. RNA samples used in RNA-Seq library construction were run on the Agilent 2100 Bioanalyzer. (a) Total RNA. Two sharp peaks corresponding to 18s and 28s rRNA are clearly visible. (b) PolyA⁺ mRNA. The broad range of transcript lengths is apparent and no residual small RNA or rRNA peaks are visible. (c) Fragmented polyA⁺ mRNA. Though somewhat variable in size, fragmented RNA is distributed about a peak length of approximately 200 nt, appropriate for RNA-Seq.
2. PolyA+ mRNA is isolated from total RNA using the MicroPoly(A)Purist Kit. RNA is incubated on the polyT resin at room temperature for 1 h to ensure maximum binding and recovery of polyA+ mRNA.

3. Following initial polyA+ mRNA enrichment, step 2 is repeated with fresh polyT resin to maximize depletion of non-mRNA species (see Note 3). Purity and size distribution of enriched polyA+ mRNA can be monitored by Agilent Bioanalyzer 2100 (Fig. 2).

4. The polyA+ mRNA is concentrated by standard ethanol precipitation (plus glycogen) with a 70% ethanol wash. The pellet is resuspended in an appropriate volume of RNAse-free water to 100 ng/μl by OD260.

5. 2 μl RNA fragmentation buffer (5×) is added to 8 μl polyA+ mRNA on ice. The mRNA is fragmented by incubation at 94°C for exactly 4 min 30 s (see Note 4). The tubes should be transferred immediately to ice.

6. Fragmented mRNA is concentrated by standard ethanol precipitation (plus glycogen) with a 70% ethanol wash. The RNA pellet is resuspended in 14 μl water.

7. The distribution of mRNA fragment sizes can be evaluated by the Agilent Bioanalyzer 2100 (see Fig. 2).

8. First-strand cDNA is prepared using Superscript III Reverse Transcriptase (see Note 5) with random primers (150 ng/μl in 20 μl reaction volume). The reverse transcription reaction is incubated at 51°C for 45 min prior to enzyme inactivation at 70°C.

9. Second-strand synthesis is performed using the SuperScript Double-Stranded cDNA Synthesis Kit. Following the 2 h synthesis reaction, T4 DNA polymerase is added to fill-in ends.

10. The double-stranded cDNA (ds-cDNA) is purified and concentrated using the QIAquick PCR Purification Kit. Product is eluted in 30 μl water.

11. Adenine overhangs are added to the ds-cDNA by Klenow Fragment (3’→5’ exo-) in the presence of 200 mM dATP. The reaction is incubated at 37°C for 30 min.

12. The ds-cDNA (with A overhangs) is purified and concentrated using the QIAquick PCR Purification Kit. The product is eluted in 30 μl water.

13. Illumina sequencing adaptors are ligated to the ds-cDNA duplexes using the Quick Ligation kit. The reaction should contain 28 μl ds-cDNA, 30 μl 2× Quick Ligase buffer, 1 μl of the Illumina PE adaptor oligo mix, and 1 μl of T4 DNA Quick Ligase. The adaptors are ligated at room temperature for 15 min.
14. The ds-cDNA (with sequencing adaptors) is concentrated using the QIAquick PCR Purification Kit. The product is eluted in 30 μl of water.

15. The adaptor-ligated ds-cDNA samples (mixed with DNA loading buffer) are run on a 1% agarose-TAE gel (see Note 6) containing ethidium bromide (or comparable nucleic acid stain). Alternating lanes should be loaded with 15 μl 1 kb HyLadder, such that each sample lane is bordered on both sides by a size marker lane. The gel is run in TAE buffer at 100 V for 1 h.

16. The DNA is visualized on a UV transilluminator. Due to the small amount of cDNA, the sample lanes are typically not visible. As such, the alternating size marker lanes serve as a guide to extract the cDNA duplex size of interest. For each sample lane, a gel slice equivalent to 200 ± 25 nt should be excised with a clean razor blade or scalpel and transferred to a clean microfuge tube.

17. The adaptor-ligated ds-cDNA is extracted and purified with the Qiagen Gel Extraction kit. Isopropanol must be added to the gel extraction buffer, as it is essential for efficient recovery of the short cDNA duplexes (~200 nt). The purified product is eluted in 30 μl water.

18. The RNA-Seq library is prepared by amplifying the adaptor-ligated ds-cDNA with Phusion high-fidelity DNA polymerase. The amplification reaction should contain 10 μl of 5× Phusion buffer, 1 μl of PCR Primer PE 1.0, 1 μl of PCR Primer PE 2.0, 1 μl of dNTPs (10 mM), 0.5 μl Phusion DNA polymerase, 6 μl water, and 30 μl of the ds-cDNA gel extraction product. Thermal cycler conditions: 98°C × 10 s, 65°C × 30 s, 72°C × 30 s for 15 cycles (see Note 2).

19. The amplified RNA-Seq library is purified using the QIAquick PCR Purification Kit. The product is eluted in 30 μl water.

20. The concentration of the RNA-Seq library is determined by Nanodrop spectrophotometer.

21. If desired, the size-based gel extraction and high-fidelity amplification steps can be checked by agarose gel electrophoresis. Approximately 5 μl of cDNA sequencing library (in DNA loading buffer) is loaded to a 1% agarose-TAE gel containing ethidium bromide (or comparable nucleic acid stain). One gel lane is reserved for DNA size markers (1 kb HyLadder). The gel is run in TAE buffer at 100 V for 1 h and the DNA is visualized by UV transillumination. A successful prep should display a broad band centered at approximately 250 bp.

3.2. Ultra High-Throughput Sequencing

Ultra high-throughput sequencing on the Illumina Genome Analyzer II (GAII) is detailed in the corresponding Illumina technical literature (7, 8). A brief summary of the technique as it applies to this protocol appears here.
1. The ds-cDNA RNA-Seq library is diluted to an appropriate concentration, typically 4–8 pM (see Note 7).

2. The RNA-Seq library is hybridized to a GAII flowcell, on which covalently linked oligomers (complimentary to the library adapter sequences) capture the ds-cDNA templates.

3. Template bridge amplification is performed at 55°C, in which hybridized templates are replicated onto the flowcell-bound oligomers.

4. Original library templates are removed by denaturing and washing.

5. Bridged templates are polymerase amplified, thereby generating flowcell-bound, sequence-matched “clusters.”

6. Clusters are “cleaned up” to standardize strand polarity, remove unbound oligomers, and block nonclustered 3’ ends (see Note 8).

7. Ultra high-throughput sequencing data is acquired by standard sequencing-by-synthesis reaction (see Note 9). This is typically performed in parallel for eight flowcell lanes. In the present example, acquisition is cycled for 36 nt reads (see Note 10). For RNA-Seq, one lane should be reserved for the Illumina-provided phiX174 library control.

8. Raw image data are processed using the standard Illumina software pipeline (SCS2.4). Real-time analysis and base calling generates files containing data on each sequencing read (*qseq.txt), intensities (*.cif) and noise profiles (*.cnf).

9. Spectral crosstalk and phasing (see Note 11) should be corrected using the Bustard (GA Pipeline 1.4.0) program and the phiX174 control lane as follows:

   ```bash
   $ bustard.py --CIF --matrix=auto4 --phasing=auto4 Data/Intensities --make
   $ nohup make -j N &
   where N is the number of processors for analysis.
   ```

10. The resulting *qseq.txt files contain read IDs, sequence, and quality scores for each flowcell “tile.” Standard FASTQ files are generated using the Gerald (GA Pipeline 1.4.0) program.

### 3.3. Mapping RNA-Seq Reads to Reference Genome

1. Trim 5’ bases of RNA-Seq reads

   Empirical evidence suggests that priming reverse transcription with random primers leads to an overrepresentation of G and C residues in the initial sequencing cycles. To eliminate this potential source of error, prior to mapping, each short read is “trimmed” – the 5’ first two bases (and associated quality information) are digitally removed from the FASTQ data files. Such trimming can be accomplished through a variety
Identifying mRNA Editing Deaminase Targets by RNA-Seq

of text edit script commands. We have had success with a simple and publicly available solution: the FASTQ/A Trimmer tool, part of the FASTX Toolkit software package.

General syntax:

\$ fastx_trimmer [-f First base to keep] [-i INFILE] [-o OUTFILE]

Example (trim 5' first 2 bases):

\$ fastx_trimmer -f 3 -i C57Bl6_Int.fq –o C57Bl6_Int.trim2.fq

2. Map RNA-Seq reads to reference genome

Mapping RNA-Seq reads to a reference genome presents a number of computational challenges. Sequencing libraries derived from randomly fragmented mature mRNAs will generate a significant proportion of reads spanning exon–exon junctions. At the time of writing, there are several academic and commercial software packages available for aligning RNA-Seq short read data. Many are compatible with this workflow, though alignment parameters should be adjusted to reflect the demands of detecting nucleotide variations in mRNA libraries:

(a) Alignments should allow for mismatches in reads relative to the reference genome

(b) Alignments should be “quality conscious.” Due to the relatively high error rate of ultra high-throughput sequencing and relatively low probability of an mRNA editing event, base quality scores should be taken into account for mapping and mismatch calling algorithms.

(c) Alignments should be unique. As an mRNA editing event is detected as a read mismatch to reference, it is imperative that mismatches occur only at “real” editing sites and not as a result of sequencing errors. Therefore, the mapping algorithm should suppress all reads for which an alignment is not “unique,” i.e., it can be satisfactorily mapped to more than one location in the genome while still satisfying mismatch and quality limits. Though such an approach will dramatically reduce the number of potentially “good” alignments, such stringency is recommended to ensure high confidence in mismatch hits.

(d) Reads spanning exon–exon junctions should be mapped accordingly. Different alignment algorithms approach the problem of mapping reads derived from mature, spliced mRNAs to genomic reference with various strategies; these include alignment to an artificial “splice-ome” reference sequence of all predicted exon–exon junctions, and \textit{ab initio} mapping of reads to junctions predicted by read
distribution and pileup. Most strategies are appropriate for mRNA editing studies. Most importantly, mapping results should be written to a single comprehensive output consisting of both exonic and splice junction alignments; separate outputs for these alignments can lead to difficulties in mismatch analysis.

Taking the above criteria into account, we use the “Tuxedo Tools,” TopHat, (9) and Bowtie (10) for mapping RNA-Seq reads. A command line example with appropriate parameters set appears below.

Example (mapping 36 nt raw reads trimmed to 34 nt):

```
$tophat -n 2 -g 1 -a 12 -m 1 --segment-length 34 mm9_genome C57Bl6_Intestine.trim2.fq
```

- n 2 allow for up to 2 mismatches to reference in seed region (first 28 nt); quality conscious
- g 1 suppress all alignments for reads that map to >1 location in reference
- a 12 for exon–exon junction reads, require at least 12 bases (“anchor”) on either side of the junction
- m 1 for exon–exon junction reads, allow for up to one mismatch in anchor segment
- segment-length 34 read length is 34 nt (do not split reads)

TopHat alignments are output to a single comprehensive SAM file, which can be used in downstream analyses.

### 3.4. Call Read: Reference Mismatches

Once suitable RNA-Seq alignments are generated, single nucleotide mismatches to the reference sequence are identified. There are few (if any) software options specifically designed to call mismatches generated by RNA editing. However, several analysis packages incorporate SNP calling algorithms, which can often be implemented for RNA mismatch analysis. Most importantly, as for the alignment step, mismatch calling should be “quality conscious” to ensure high confidence in read:reference discrepancies.

Beginning with a TopHat generated SAM file, we use the publicly available, open source SAMTools software package (11) from the Sanger Institute for mismatch calling. The complete workflow can be found at [http://samtools.sourceforge.net/samtools.shtml](http://samtools.sourceforge.net/samtools.shtml); an abridged example appears below.

Example of SAMTools workflow

1. TopHat output SAM file to SAMTools Binary (BAM) conversion

```
$.samtools view -S -b -t mm9_genome.fa.fai C57Bl6_Intestine.accepted_hits.sam -o C57Bl6_Intestine.bam
```
Identifying mRNA Editing Deaminase Targets by RNA-Seq

2. BAM file sort

$./samtools sort C57Bl6_Intestine.bam C57Bl6_Intestine.sorted

3. BAM file indexing for rapid lookup functions

$./samtools index C57Bl6_Intestine.sorted.bam

4. Pileup conversion and consensus calling

$./samtools pileup -f mm9_genome.fa -c C57Bl6_Intestine.sorted.bam > C57Bl6_Intestine.CNS

Generates a “pileup” file relative to the reference genome. This file contains reads and their qualities on a reference base-by-base scale and a corresponding consensus base call at each position.

5. Mismatch calling and preliminary filter

$./samtools.pl varFilter -d 3 -D 1000 -Q 25 C57Bl6_Intestine.CNS > C57Bl6_Intestine.filtered.MMtoRef

Filters the pileup consensus file for variations in reads relative to reference. This dataset will serve as the starting point for identifying those mismatches that resulted from mRNA editing. In this example, a relatively stringent filter is applied requiring a minimum of three mismatch-containing reads (-d 3) to call a variant. The maximum parameter (-D 1000) should be adjusted based on the read depth achieved in a given experiment. Finally, the root mean squared mapping quality must exceed 25 (-Q 25) to call a mismatch. Additional parameters can be adjusted as needed.

3.5. Read:Reference Mismatch Filters

The list of variations called from consensus will often contain large numbers of read:reference mismatches unrelated to mRNA editing. These mismatches may be a result of genomic SNPs, sequencing errors, misaligned reads, reverse transcription/amplification errors, and unrelated mRNA modification processes. As such, the initial variation list is filtered on several criteria appropriate to the editing enzyme under investigation, including error probability, sequence type (exons of known or predicted genes only), read:reference mismatch base calls, known SNPs, and repetitive elements. Filters can be applied by many different bioinformatics packages and/or standard Linux shell commands. We use Galaxy, a web-based genomics suite (http://galaxy.psu.edu/), largely because of its user-friendly interface and powerful genomic interval operations. A particular series of filters will be unique to each mRNA editing enzyme studied. A sample filter workflow for an mRNA-editing cytidine deaminase appears below. This example uses a particularly stringent set of filters to minimize false-positive hits; parameters should be adjusted to suit a given experiment.
1. Error probability filter
   Keep only those mismatches with a PHRED-scaled probability score $\geq 40$ (as calculated by SAMtools pileup function).

2. Sequence type filter
   Keep only those mismatches within mRNA exons (from RefSeq database, http://genome.ucsc.edu/cgi-bin/hgTables).

3. Base call filter
   Keep only C:T (reference:read) mismatches for + strand transcripts (G:A mismatches for – strand transcripts) (see Note 12).

4. Known SNP filter
   Keep only those mismatches absent from SNP databases (see Note 13).

5. Repetitive element filter
   Keep only those mismatches outside of repetitive regions (see Note 14).

Subsequent analysis steps are performed on a single, merged dataset derived from the above control and deaminase-deficient mismatch results.

3.6. Identifying RNA Editing Targets

1. Mismatches specific to deaminase activity are identified by intersecting the control and deaminase-deficient analyses. Read:reference mismatches present in the control dataset and absent from the deaminase-deficient dataset are flagged as potential mRNA editing hits.

2. Potential editing events are further refined by comparing read consensus between control and deaminase-deficient datasets at each mismatch hit site. Consensus pileup data (Subheading 3.4, step 4) from each library are juxtaposed at each mismatch hit site. Potential hits in the control dataset should have a multiple read, reference-matching consensus with low error probability at the corresponding site in the deaminase-deficient dataset. Using SAMTools, we define this as a minimum read-depth $\geq 3$ with a PHRED-scaled probability score $\geq 30$. Mismatches that do not satisfy these criteria are discarded.

3. Potential editing hits are tabulated and organized with associated consensus data and error probabilities for both datasets. As hits are reported with coordinates on the genomic reference, the results can be additionally annotated with information from a wide variety of genome resource databases. Pileup data can also be used to calculate “editing frequency,” as determined by the percentage of reads representing the edited base in the control dataset. Results for APOBEC1 in mouse small intestine appear in Table 1.
### Table 1
**APOBEC1 mRNA editing targets**

<table>
<thead>
<tr>
<th>Site annotation</th>
<th>C57/Bl6</th>
<th>APOBEC1&lt;sup&gt;−/−&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gene name</strong></td>
<td><strong>Strand</strong></td>
<td><strong>Consensus base</strong></td>
</tr>
<tr>
<td><strong>Reference base</strong></td>
<td><strong>Read consensus</strong></td>
<td></td>
</tr>
<tr>
<td>ApoB</td>
<td>+</td>
<td>C</td>
</tr>
<tr>
<td>2010106E10Rik</td>
<td>+</td>
<td>C</td>
</tr>
<tr>
<td>Cyp4v3</td>
<td>–</td>
<td>G</td>
</tr>
<tr>
<td>B2m</td>
<td>+</td>
<td>C</td>
</tr>
<tr>
<td>App</td>
<td>–</td>
<td>G</td>
</tr>
<tr>
<td>Rnf128</td>
<td>+</td>
<td>C</td>
</tr>
</tbody>
</table>

RNA-Seq libraries from the small intestine of control C57/Bl6 and *Apobec1<sup>−/−</sup>* mice were prepared and screened for mRNA editing events. Sorted by PHRED-scaled mismatch probability score, the “top hit” is a C–T change at position 6666 of the apoB transcript, the well-characterized physiological substrate of APOBEC1. Additional, previously unknown APOBEC1 editing targets were also identified; a partial list of these hits is provided here.
3.7. Validating Potential RNA Editing Targets

Newly identified mRNA editing targets should be validated by standard Sanger sequencing.

1. Regions of interest (containing editing hits) are amplified by high-fidelity PCR. Identical regions are amplified from control genomic DNA, control cDNA, deaminase-deficient genomic DNA, and deaminase-deficient cDNA.

2. PCR products are sequenced by standard Sanger methods.

3. mRNA editing hits are validated by comparing the four sets of Sanger sequencing results. A *bona fide* mRNA editing hit is confirmed by reference-matching sequence in control genomic DNA, deaminase-deficient genomic DNA, and deaminase-deficient cDNA, but reference-mismatch sequence in control cDNA (see Note 15). An example is shown in Fig. 3.

Fig. 3. mRNA editing target validation by Sanger sequencing. Regions spanning a potential editing site (chr8:46391931, NM_133969, 32.4% editing frequency by RNA-Seq) were PCR amplified from genomic DNA and cDNA isolated from control C57/Bi6 and *Apobec1<sup>−/−</sup>* mouse small intestine. PCR products were sequenced by standard Sanger methods. The C:T modification characteristic of APOBEC1 editing is present only in the control C57/Bi6 cDNA.
4. Notes

1. Total RNA isolated by most standard molecular biology techniques is appropriate for prepping RNA-Seq libraries. However, certain spin column-based total RNA kits may limit the recovery of very long and/or very short transcripts. Total RNA isolated with such kits remains acceptable for library prep, but may bias against certain transcript types.

2. If less than 10 μg total RNA is used, additional PCR amplification cycles may be added at step 18 (Subheading 3.1). We have had success amplifying 15 cycles for >10 μg starting total RNA, 16 cycles for 5–10 μg starting total RNA and 17 cycles for 4 μg starting total RNA. Though additional amplification cycles can provide sufficient library concentration from limited starting material, it also increases polymerase errors and may bias the RNA-Seq dataset.

3. Though the final yield of polyA+ mRNA is somewhat reduced, a second round of polyA+ enrichment is necessary to completely remove rRNA, tRNA, and other species that interfere with RNA-Seq.

4. The ideal mRNA fragment size for a standard RNA-Seq run on the Illumina GA II is 200 nt. If required for alternative applications, the fragment size can be adjusted by altering the duration of the reaction. Longer fragments can be achieved by reducing incubation time at 94°C and vice versa.

5. Though the SuperScript Double-Stranded cDNA Synthesis Kit includes a reverse transcriptase (Superscript II), this protocol substitutes a different enzyme (Superscript III). Superscript III can be used with the same buffers described in the ds-cDNA synthesis kit but functions at a higher temperature (51°C), thereby minimizing potential secondary structure in RNA fragments.

6. RNA-Seq ultra high-throughput sequencing is most effective with cDNA fragments of relatively short (~200 nt) and defined size. As fragmentation can be variable and somewhat heterogeneous, ds-cDNA libraries are separated by agarose gel electrophoresis and extracted based on DNA size standards.

7. An appropriate concentration range allows each ds-cDNA template to be spatially separated on the flowcell at a distance sufficient for cluster amplification and resolution while maximizing total template number.

8. The clean up steps remove all extended nucleotides from one of the flowcell-bound oligomers, leaving only the sequence extended from the other oligomer. This guarantees that only one of the strands will be sequenced during acquisition.
9. The sequencing reaction consists of stepwise cycles that proceed iteratively; total cycle number dictates read length. Polymerase extends the 3’ end of primer (annealed in the final amplification step) through the incorporation of reversibly terminating nucleotides (fluorescently labeled by base). After the incorporation of one chain-terminating nucleotide, the entire flowcell is imaged in small “tiles,” 120 tiles per lane. Four images are captured per tile, one for each nucleotide laser/filter combination. A cleavage step removes the fluorophore and reverses chain termination. After a wash step, the cycle is repeated until the desired read length has been reached.

10. RNA-Seq read depth required for mismatch/editing analysis will vary based on the size of the genome/transcriptome under investigation and the read length. For mammalian transcriptomes (mouse, human), at 36 nt read lengths, approximately 50–60 million raw unmapped reads should provide comprehensive transcriptome coverage (>95% of all bases in expressed transcripts).

11. Due to the nature of the sequencing methodology, certain calculations and corrective measures are implemented to account for spectral crosstalk from the fluorophores and the fact that not all strands in each cluster are extended in perfect synchrony; some will not be extended and others will be extended by two bases at certain cycles, leading to what is termed “phasing.” For RNA-Seq it is important to include a sequencing control sample in one lane of the flowcell; the phiX174 library provided is appropriate for this purpose.

12. As RNA-Seq reads are mapped to a reference genome, all mismatch bases and coordinates are provided in the context of the genomic forward strand. Therefore, for an analysis of mRNA editing, transcript “strandedness” must be taken into account. When filtering for cytidine deaminase editing, both C–T mismatches for + strand transcripts AND G–A mismatches for – strand transcripts pass. Similarly, for adenosine deaminases, A–G mismatches for + strand transcripts AND T–C mismatches for – strand transcripts pass.

13. Genomic heterozygosity and SNPs are detected as read:reference mismatches and can lead to false positive mRNA editing hits. To minimize this error source, known SNPs are removed from the dataset. Known SNP databases for a variety of genomes are available from UCSC: http://genome.ucsc.edu/cgi-bin/hgTables.

14. Repetitive element databases (such as RepeatMasker) for a variety of genomes are available from UCSC: http://genome.ucsc.edu/cgi-bin/hgTables.
15. Hit validation requires standard Sanger sequencing of all four samples to rule out several potential non-mRNA editing sources of read:reference mismatch. First, genomic DNA from each library is sequenced to rule out heterozygosity. Next, genomic DNA sequences from control and deaminase-deficient samples are compared to rule out a SNP. Control cDNA sequence is compared to control genomic DNA sequence to confirm mRNA editing. Finally, deaminase-deficient cDNA sequence is compared to control cDNA sequence to validate specific deaminase activity.

References


