High-quality overlapping paired-end reads for the detection of A-to-I editing on small RNA

Josephine Galipon[1,2,3*], Rintaro Ishii[4*], Soh Ishiguro[1,2,3], Yutaka Suzuki[4], Shinji Kondo[5], Mariko Okada-Hatakeyama[6,7], Masaru Tomita[1,2] and Kumiko Ui-Tei[3,4]†

*equal contribution

[1] Institute for Advanced Biosciences, Keio University, 14-1 Baba-cho, Tsuruoka, Yamagata 997-0035, Japan.

[2] Systems Biology Program, Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa, Kanagawa 252-0882, Japan.

[3] Department of Biological Sciences, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

[4] Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba-ken 277-8651, Japan.

[5] Inter-university Research Institute Corporation, Research Organization of Information and Systems, w, 10-3, Midoricho, Tachikawa, Tokyo 190-8518, Japan.

[6] Laboratory for Integrated Cellular Systems, RIKEN Center for Integrative Medical Sciences (IMS), 1-7-22 Suehiro-cho Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan.

[7] Present address: Laboratory of Cell Systems, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita-shi, Osaka-fu 565-0871, Japan

†To whom correspondence should be addressed. Tel: +81 3 5841 3043; Fax: +81 3 5841 3044; Email: ktei@bs.s.u-tokyo.ac.jp

Summary

Paired-end RNA sequencing (RNA-seq) is usually applied to the quantification of long transcripts such as messenger or long non-coding RNAs, in which case overlapping pairs are discarded. In contrast, RNA-seq on short RNAs (≤ 200 nt) is typically carried out in single-end mode, as the additional cost associated with paired-end would only translate into redundant sequence information. Here, we exploit paired-end sequencing of short RNAs as a strategy to filter out sequencing errors, and apply this method to the identification of adenosine-to-inosine (A-to-I) RNA editing events on human precursor microRNA (pre-miRNA) and mature miRNA.

Combined with RNA immunoprecipitation sequencing (RIP-seq) of A-to-I RNA editing enzymes, this method takes full advantage of deep sequencing technology to identify RNA editing sites with unprecendented resolution in terms of editing efficiency.

Key words: RNA-seq, small RNA, paired-end sequencing, RNA editing, inosine, microRNA

1. Introduction

Illumina RNA sequencing (RNA-seq) is sufficient for the relative quantification of RNA, but its error rate remains problematic when the desired application is the detection of single nucleotide modifications on RNA, due to the difficulty of distinguishing between true RNA editing events and sequencing errors. Because of this, Sanger direct sequencing remains the gold standard for validating novel editing sites. In the case of adenosine deaminases acting on RNA (ADAR)-mediated adenosine-to-inosine (A-to-I) editing sites, only sites with an editing ratio above about 10% – meaning that more than 10% of transcripts are edited at that position – may be clearly distinguished from a typical Sanger sequencing background (1). For this reason, the vast majority of RNA editing studies establish an arbitrary cut-off of 5~10% for the detection of editing sites, making it extremely difficult to publish reports of lower efficiency sites. Nevertheless, editing sites with low editing ratios have the potential to be biologically important. For instance, in the case of extremely abundant transcripts such as the oncogenic miR-21, editing levels as low as 0.1~1.0% have the potential to generate a subpopulation of edited miR-21 that is abundant enough to target and downregulate a entirely novel set of genes. Indeed, editing in the miRNA seed region responsible for target recognition leads to genome-wide retargeting of the edited microRNA (miRNA) (2, 3). For example, in order to validate an editing site with a ratio of 0.1~1.0%, in theory one would need to read 2,000~20,000 individual clones by Sanger sequencing to obtain just 2 reads with an A-to-G substitution corresponding to that editing site. This is currently unrealistic in terms of time and cost, although it may be possible to overcome this by simultaneous overexpression of both the ADAR and the candidate pre-miRNA in an attempt to artificially bump up the editing ratio. In conclusion, if the accuracy problem of RNA-seq is overcome, deep sequencing is expected to reveal more of these low-efficiency editing sites with potentially important biological functions.

Here, we used paired-end sequencing on small RNAs as a trick to filter out most errors, decreasing the theoretical error frequency to about 1 in 10^6 bases. We applied this method to the detection of A-to-I editing

sites on human precursor and mature miRNA by RNA immunoprecipitation of ADAR-bound small RNA. The paired-end sequencing method reads each cDNA fragment from both the 5'- and the 3'- end, and is therefore not an essential requirement for the relative quantification of small RNAs, as the reverse and forward end of the pair would frequently overlap. However, we reasoned that since the 5'- and 3'-ends are each read by an independent sequencing reaction, paired-end sequencing may be used as an internal validation to distinguish sequencing errors from genuine RNA editing sites, by eliminating the paired-end reads for which the overlapping forward and reverse sequences do not perfectly match. Candidate editing sites are then statistically validated by the log-likelihood ratio test (LLR) using the Phred quality scores produced by both the forward and the reverse end, in an approach adapted from previous literature (4–6).

To implement this method, we aimed to identify and compare the precursor and mature miRNA editing target preferences of two constitutive nuclear isoforms ADARs: ADAR1-p110 and ADAR2. To this end, we overexpressed each isoform separately as mycGFP (mGFP)-tagged constructs as described previously (7). Two popular methods for investigating RNA-protein interactions *in vivo* are RNA immunoprecipitation (RIP) and cross-linking immunoprecipitation (CLIP) or photoactivatable-ribonucleoside-enhanced crosslinking (PAR-CLIP), a method relying on the incorporation of photoreactive ribonucleoside analogs, were successfully used to identify genome-wide miRNA-protein interactions (8, 9). However, aside from difficulties in identifying the precise location of A-to-I editing site due to crosslinking-induced mutations, CLIP-seq and its variants may introduce a bias in favor of imperfect double-stranded RNA (dsRNA) binding targets, as unpaired nucleotides are more available for UV cross-link formation due to better accessibility of the A-form helix major groove (10). Previously, Methylene blue (MB) as a photoreactive intercalating agent was successfully used to induce cross-linking of dsRNA to protein by visible light (11), but MB mediates heavy oxidative damage to purines (12). Crosslinking methods are therefore inadequate for the precise downstream identification of RNA editing sites. Out of these concerns, we opted for straightforward RNA immunoprecipitation sequencing (RIP-seq) of overexpressed mycGFP-tagged ADAR isoforms with mycGFP as a control to detect background editing by endogenous ADARs (7).

To capture both pre- and mature miRNAs, high-throughput strand-specific sequencing of small RNA (sized 15 to 110 nucleotides) was carried out in 101 nucleotide paired-end mode. The reads produced were then mapped to the human genome using a method for identifying A-to-I editing sites known as *collapse mapping* (13), which we adapted to strand-specific paired-end data. After generation of high quality super-reads from the

overlap between both sides of a pair, statistical validation by the LLR test (4–6), and discarding of sites that overlapped with annotated single nucleotide polymorphisms (SNPs) (14–18), allowed the identification of 31 unique editing sites on mature and pre-miRNA with editing ratios, 8 of which were confirmed by other studies and presented here (2, 19–25, Table 1). Among these 8 canonical sites, 4 had editing ratios below 10%, ranging from 0.21% to 5.38%, suggesting that our method successfully identifies relatively low efficiency editing sites.

2. Materials

2.1. Recommendations for handling RNA

The following general precautions should be followed at all times to avoid RNA degradation. If these steps are respected, it is not necessary to prepare buffers and reagents with DEPC-treated water.

1. Work on ice and use gloves; while wearing the gloves avoid contact with things that are regularly in contact with human skin (such as door knobs or garbage lids);
2. Use a set of pipettes that is specific for RNA work and have never been used for handling cell cultures; if unavailable, the use of filter pipette tips is recommended;
3. Use commercial RNase-free water for resuspension, dilution, and enzymatic reaction steps involving RNA;
4. All buffers and reagents should be 0.22 µm filter-sterilized, as autoclaving would release the RNases contained in micro-organic contaminants, an undetermined portion of which may renature after the solution cools down;
5. Samples should be harvested quickly with a method that disrupts cells and inactivates nucleases.

2.2. Materials for cell culture and transfection

1. 9-cm cell culture dishes;
2. HeLa cells or cell line of interest;
3. 37ºC CO_2 incubator;
4. Dulbelcco's modified Eagle Medium (DMEM), 4.5 g/L glucose, without antibiotics, supplemented with sodium pyruvate and 10% heat-inactivated fetal bovine serum (FBS).
5. Opti-MEM™ I Reduced Serum Medium (Thermo Fisher Scientific, Cat. # 31985070)
6. Lipofectamine™ 2000 Transfection Reagent (Thermo Fisher Scientific, Cat. # 11668019)

7. pcDNA3.1 or equivalent expression vector containing the coding region for the tagged protein of interest, in this case mycGFP-ADAR1-p150, -p110 and mycGFP-ADAR2 (not covered here). For the method to clone the gene of interest into the expression vector for mycGFP-ADAR isoform cloning, please refer to (7);
8. DAPI (4',6-diamidino-2-phenylindole) staining reagent.

2.3. Materials for RNA immunoprecipitation and deep sequencing (RIP-seq)

1. An antibody against the tag of interest, in this case rabbit polyclonal anti-GFP antibody (antigen: Full Length *Aequorea victoria* GFP);
2. Goat anti-rabbit immunoglobulin G (IgG) conjugated with alkaline phosphatase;
3. 1X PBS(-) [137 mM NaCl, 2.7 mM KCl, 10 mM Na2HPO4, 1.8 mM KH2PO4 (pH7.4 adjusted with HCl)];
4. RiboCluster Profiler™ RIP-Assay Kit (Code No. RN1001) or equivalent;
5. TURBO™ DNase (2 U/μL) (Thermo Fisher Scientific, Cat. # AM2238) or equivalent;
6. SuperSep™ Ace 5-20% gradient pre-casted polyacrylamide gel (Wako, Cat. # 197-15011) or equivalent;
7. BioTrace™ PVDF membrane (Pall Corporation) or equivalent;
8. Tris/Glycine/SDS electrophoresis buffer [0.3% (w/v) Tris base, 1.44 (w/v) glycine, 0.1% (w/v) SDS, pH8.3 (no pH adjustment required);
9. TBS-T buffer [20 mM Tris-HCl (pH 7.5), 150 mM NaCl, 0.2% Triton X-100] for washing, supplemented with 5% skim milk for blocking;
10. Alkaline phosphatase buffer [100 mM Tris-HCl (pH9.5), 100 mM NaCl, 5 mM $MgCl_2$];
11. CDP-Star® Detection Reagent (GE Healthcare Life Sciences) or equivalent;
12. RiboZero™ Magnetic Gold Kit (epicentre, Cat. # MRZE706);
13. Illumina® TruSeq Small RNA Library Prep Kit Reference Guide (Part # 15004197 Rev. C); newer versions may be applicable;

2.3. Hardware and equipment

1. Bio-Rad Trans-Blot® Turbo™ Transfer System, iBlot® Dry Blotting System or equivalent;
2. NanoDrop™ 2000 Spectrophotometer, QuBit Fluorometer or equivalent;

3. Thermocycler for polymerase chain reaction (PCR);
4. Fluorescence microscope with filters compatible with GFP and DAPI;
5. Illumina® HiSeq 2000 Sequencing System.

2.4. Software and database versions

Although the method can be implemented with later versions without any major problems, the versions that were used for this study are as follows:

1. CASAVA v1.8.2 (provided by Illumina);
2. Human genome sequence: Genome Reference Consortium GRCh37, UCSC human hg19;
3. Alignment software: bowtie 0.12.7 (26);
4. Non-coding RNA database: NONCODE v3.0 (27);
5. Repetitive sequences: RepeatMasker3.3.0 (http://www.repeatmasker.org);
6. microRNA database: miRBase v19 (28);
7. SNP databases: 1000 genomes (14), dbSNP (15), HapMap (16), miRNASNP (17), and a publication from Han *et al.* (18)

3. Methods

3.1. Cell culture and transfection

Pre-warm all cell media and 1X PBS unless otherwise specified. Note that the cell dishes for fluorescence imaging involving DAPI staining of the nucleus were separate from those for RNA immunoprecipitation, and that antibiotics were not used.

1. Plate 3 x 10^6 HeLa cells per 9-cm cell culture dish in DMEM supplemented with 10% FBS and culture overnight in a CO_2 incubator at 37ºC;
2. Rinse in pre-warmed 1X PBS and change medium to DMEM without FBS;
3. Prepare transfection mix following instructions for Lipofectamine™ 2000, but using 5.5 µg of tagged protein expression vector per 9-cm dish in Opti-MEM™ I Reduced Serum Medium;
4. Add the transfection mix to each cell culture;
5. After incubating 4 hours in a 37ºC CO_2 incubator, remove the transfection media and replace by DMEM supplemented with 10% FBS;
6. Incubate in a CO_2 incubator at 37ºC for 2 days;

7. To observe cell confluency and GFP-tagged protein expression by fluorescence microscopy, remove culture media and add 1X PBS(-);
8. Observe localization of mycGFP-tagged proteins by DAPI staining of nuclear DNA (Fig. 1A);
9. For RNA immunoprecipitation, proceed to Section 3.2.

3.2. RNA immunoprecipitation, silver staining, and Western blot

1. Wash HeLa cells 5 times in pre-warmed 1X PBS(-);
2. Cell lysis and RNA immunoprecipitation (IP) was carried out following the maker's instructions for the RiboCluster Profiler™ RIP-Assay Kit (Code No. RN1001). Cell lysates were immunoprecipitated using 15 ng of either rabbit anti-GFP antibody or control IgG;
3. 2 µL of total cell lysate and 100 µL of the final elution mixture for anti-GFP and control IgG IP samples were set aside for silver staining and Western blotting before proceeding to RNA extraction;
4. Two identical 5-20% gradient polyacrylamide gels were purchased for SDS-PAGE electrophoresis of the samples obtained in Step 3 in Tris/Glycine/SDS electrophoresis buffer at 20 mA for 80 min;
5. One gel may be used for silver staining (not described here), while the protein content of the other gel was transferred to a methanol-washed PVDF membrane by electroblotting in an iBlot® Dry Transfer System with default parameters (P0 7:00 min: 20V for 1 min 23V for 4 min 25V for the remainder);
6. Blocking of the PVDF membrane was performed in TBS-T buffer supplemented with 5% skim milk at room temperature for 1 hour;
7. The membrane was washed once for 10 min in TBS-T by gentle shaking at room temperature;
8. Primary antibody (rabbit polyclonal anti-GFP) was diluted 1/2000th in 1 mL of TBS-T, added to the membrane in a hybridization bag, and incubated overnight at 4ºC;
9. The membrane was washed 3 times 10 min in TBS-T buffer by gentle shaking at room temperature;
10. Secondary antibody (alkaline phosphatase-conjugated goat anti-rabbit IgG from CAPPEL # 59298) was pre-diluted 1/10,000th in 1 mL and added to the membranes in a new hybridization bag, and incubated at 4ºC overnight;
11. The membrane was washed 3 times 10 min in TBS-T buffer by gentle shaking at room temperature;
12. Lay the membrane onto a piece of Saran wrap, and cover it with alkaline phosphatase (AP) buffer;
13. Discard AP buffer, cover with CDP-Star Detection Reagent, and proceed to imaging (Fig. 1B);

14. The RNA extracted from input and IP samples at Step 3 was kept for small RNA sequencing in Section 3.3.

3.3. Library preparation and strand-specific paired-end deep sequencing

1. The concentration of RNA extracted in Section 3.2. for input and IP samples was measured using a NanoDrop™ 2000 spectrophotometer or Qubit fluorometer;
2. DNA was digested using 0.88U of TURBO™ DNase;
3. Ribosomal RNA was removed using the RiboZero™ Magnetic Gold Kit;
4. Strand-specific libraries for deep sequencing were prepared according to instructions in the Illumina® TruSeq Small RNA Library Prep Kit Reference Guide (Part # 15004197 Rev. C), and the amplified cDNA templates were size-selected on a polyacrylamide gel to obtain fragments that correspond to cDNAs that would be 15 to 110 nucleotides long without the adapters. This length range is expected to recover most mature and precursor miRNAs (pre-miRNAs);
5. 101 nucleotide paired-end sequencing with HiSeq 2000 yielded between 14,539,145 and 25,450,999 paired-end reads depending on the sample. Given the read length of 101 nucleotides, and the fact that cDNAs ranging 15 to 110 nucleotides were size-selected, the full-length mature miRNA and at least part of the pre-miRNA regions are expected to be read twice, once in each direction.

3.4. Data pre-processing

An overview of the bioinformatics workflow from Sections 3.4 to 3.7 is available in Figure 2.

1. `CASAVA` v1.8.2 was used to generate `FASTQ` files in Illumina 1.8+ / Phred+33 format, for which raw read quality values (QV) range between 0 and 41. These FASTQ files contain a label for each read specifying whether or not a read has "passed filtering" or not, indicated by Y or N, with Y meaning that the read failed the test. Therefore, only reads labeled with an N were kept for further analysis;
2. Illumina TruSeq® Universal Adapter (5'-adapter) and Indexed Adapter (3'-adapter) were trimmed. This may be achieved using custom software or tools that handle paired-end reads, such as `cutadapt` or `trimmomatic`. For reads originating from shorter transcripts such as mature miRNA, the sequencing reaction may bleed through into the adapter regions on both sides. The adapter sequences were as follows, according to the TruSeq Small RNA kit; please note that most tools do not automatically look for the reverse complement of the 5' Adapter, in which case the reverse

complement of RA5 should be provided as it will be detected on the reverse reads:

RNA 5' Adapter (RA5) 5'-GTT CAG AGT TCT ACA GTC CGA CGA TC-3'

RNA 3' Adapter (RA3) 3'-TGG AAT TCT CGG GTG CCA AGG-3'

3. 5'-end trimming: due to relatively poor quality, the 5'-most nucleotide was deleted by trimming the forward reads by 1 base at the 5'-end, and the reverse reads by 1 base at the 3'-end as displayed in the `FASTQ` file;
4. 3'-end trimming: miRNAs are often subject to the addition of one or several nucleotides at their 3'-end (29), which could interfere with mapping. Thus, forward reads were trimmed by 2 base at the 3'-end, and reverse reads were trimmed by 2 bases at the 5'-end as displayed in the `FASTQ` file;
5. To ensure a theoretical error of less than 1 in 1000 bases (10^{-3}) for all positions on the read, the read sequences were scanned from the 5'-end and any downstream sequences were trimmed upon encountering a base with a quality value (QV) lower than 30;
6. Reads that were less than 10 bases long following the above steps were discarded. Depending on the sample, 33 ~ 65% of reads remained.

3.5. Collapse mapping and calling of RNA editing sites

Most mapping algorithms are not optimal for detecting mismatches in long reads, due to limitations in the number of allowed mismatches. Here, we chose to adopt a method called *collapse mapping* that enables the mapping of edited reads regardless of the number or distribution of a given mismatch type within the read (13), and adapted it as below to make it compatible with strand-specific paired-end data.

1. First, an index is generated for the human genome (GRCh37, UCSC human hg19) using `bowtie-index`, which comes with `bowtie` (we used version 0.12.7 at the time, but any other version may be employed) (26);
2. ***Normal mapping***: `bowtie` was used to map reads with unique hits and no mismatches to the human genome with help from the index generated in Step 1;
3. The overlapping region was extracted from pairs for which the forward and reverse overlapped without any mismatches; here we obtained between 0.9 and 1.8 million high quality *super-reads* while also retaining strand information;
4. ***Collapse mapping****:* `bowtie` mapping was performed again but this time allowing multiple hits and no mismatches, and the unmapped reads were output to a separate file;

5. The unmapped reads from Step 3 were then collapsed to a three-base system: all A's were converted to G's, on both the forward and the reverse reads. These reads, that now contain only 3 types of bases, are referred to as *collapsed reads*;
6. Two collapsed versions of the human genome (+) strand were prepared: one that was collapsed from A to G, and the other from T to C (Fig. 3A). These are hereby referred to as *collapsed genomes*;
7. Using **`bowtie`**, the collapsed forward and reverse reads were mapped to the A-to-G and T-to-C collapsed genomes, respectively; no mismatches were allowed, and reads that mapped to more than 20 locations are discarded;
8. Reverse and forward reads of a same pair that fail to overlap on the same genomic location are discarded, and the overlapping region was extracted from the remaining pairs; in our case,
9. Paired-end reads for which the overlapping sequence do not perfectly match, meaning that it contains one or more mismatches between the forward and the reverse, are discarded, as this is most likely indicative of a sequencing error (Fig. 3B); in our case, this eliminated 0.02 ~ 0.05% of pairs;
10. The overlapping part of the paired-end read was revert to its original 4-base sequence, and compared to the sequence of the original 4-base genome to which it mapped. Among these super-reads, those containing mismatches relative to the genome other than the "A-to-G" type were discarded. Here, 12 ~ 17% of reads contained "A-to-G" type only;
11. Out of the remaining A-to-G super-reads, those that mapped to multiple positions in the genome (around 80%) were discarded. In the end, we obtained 74,113 to 290,561 collapsed super-reads which may contain genuine editing sites;
12. In this study, we focus specifically on reads that map to pre-miRNA sequences retrieved from miRBase (version v19) (27); reads were separated into categories according to length: "mature" (24 nucleotides or less), or "non-mature" (25 nucleotides or more), the latter corresponding for the most part to non-mature miRNA species (mostly pre-miRNA, and possibly some primary miRNA or pri-miRNA). The distribution of read lengths for normal and collapse reads that mapped to miRNA regions can be found in Figure 4.

3.6. Probability of false positives, and editing ratio threshold

In the previous section, the overlapping sequence of each overlapping pair was retrieved and considered as one single "high-confidence" super-read. Because forward and the reverse of a pair are generated by two

independent sequencing reactions, any mismatch between them should be indicative of an error. Assuming that each error event is independent from one another, the probability of obtaining an error at the same position i on both the forward and the reverse is the product of their probability of error, defined by the QV of the base encountered at position i. In the present scenario, only bases with a QV of 30 or above are retained, so that the probability of double error occurrence should never go above 10^{-6} (1 in 1,000,000 bases). Furthermore, the probability of the same error type occurring on both strands at the same position should be even lower than this. It ensues that the probability of a false positive is negligeable. For instance, in order for such a false positive to count as a candidate RNA editing site, an A-to-G error on the forward end should correspond to the complementary T-to-C error at the exact same genomic position on the overlapping reverse end of the paired-end read.

Given this information, the editing ratio (ER) of a candidate editing site, defined as the percentage of reads at a harboring an A-to-G type mismatch at a given genomic position, is given by the following formula which uses the coverage of each nucleotide type at that particular position:

$$ER = ([G]_{collapse}+[G]_{normal}) / ([A,G]_{collapse}+[A,G]_{normal}) \times 100 \quad (1)$$

Sites with an ER ≥ 0.1 % and a total coverage of 10 or more (5 × 2 ends of a paired-end read) were kept as candidate A-to-I editing sites for statistical validation in Section 3.7. It should be noted that the threshold defined as ER ≥ 0.1 % (1 in 1,000 transcripts edited at a particular position) ensures that this A-to-G type mismatch is at least 1,000 times more frequent than the theoretical per-read error occurrence (1 in 1,000,000 reads). A method to statistically validate this is presented in the Section 3.7.

3.7. Statistical validation of editing sites

Over the past decade, there has been much debate about the validity of deep sequencing-based identification of RNA editing sites. False positives may arise both from sequencing errors, from the assignment of reads to the wrong genomic location during mapping, and from overlapping with single nucleotide polymorphisms (SNPs). Since the quality value (QV) of a base provides an indication of the expected error rate for that base, using the QV of that base on all reads that overlap with that genomic position, it is possible to calculate the average local theoretical error rate, and therefore the probability that a given mismatch is an error rate, versus the probability that it is a true editing site. The logarithm of the ratio of these two probabilities is known as the log-likelihood ratio (LLR). This method adapted from (4–6) to accommodate the fact that each position in our super-read is read twice independently:

1. Given n reads mapping to a given position i, if x of reads contain a mismatch at position i, we assume that x follows the binomial distribution:

$$f(p\,;n,x,i) = {}_nC_x \times p^x \times (1-p)^{(n-x)} \quad (2)$$

2. The maximum likelihood p of the binomial distribution in (1) is then:

$$p = x / n \quad (3)$$

3. To account for sequencing errors in overlapping paired-end reads, the observed error rate ε is defined by the sum Φ of the two quality values at position i originating from both ends of the paired-end read, such as $\varepsilon = 10^{-\Phi}$. From there, the maximum likelihood estimator r taking ε into account can be defined as:

$$r = \{\Sigma[i \in G](1-\varepsilon_i) + \Sigma[i \in A](\varepsilon_i)\} / n \quad (4)$$

4. Then, the maximum likelihood $\max_r$ defined in (3) is compared to the probability $P(n|r{=}0, \varepsilon, i)$ of all reads containing a sequencing error at position i (in which case, $r = 0$) using the following log likelihood ratio (*LLR*):

$$LLR = \log_{10} \{\max_r [P(n|r,\varepsilon,i)] / P(n|r{=}0,\varepsilon,i)\} \quad (5)$$

5. Finally, we selected the candidate editing sites for which LLR ≥ 2, meaning that the mismatch at position i is 100 times more likely to be a true editing site than a sequencing error, given two quality values per position i from the paired-end sequencing.

4. Notes

1. Epicentre's RiboZero™ Gold Machetic Kit contains enough reagents for 6 reactions, but we were successful in using the reagents to a dilution of 6/7 in order to accommodate 7 samples.
2. Bowtie 1 (bowtie) and Bowtie 2 (bowtie2) are two very different implementations, and their indexes are not interchangeable. While bowtie2 is generally the better choice when aligning reads that are longer than 50 bp, bowtie is sometimes faster and/or more sensitive when aligning short reads, but ignores unpaired alignments. Even when 101 nt in paired-end mode, reads originating from mature miRNA reads are very short (18 to 24 bases). Furthermore, trimming steps and merging of paired-end overlaps generates reads for pre-miRNA that are shorter than the initial read length generated by the sequencer (peak between 40 and 50 bases). For these reasons, we recommend Bowtie 1, but the choice is up to the reader.

3. The possibility remains that true editing sites overlap with known SNPs. Similarly, the opposite case where novel stable SNPs, present only in this cell line and absent from public databases, were falsely interpreted as editing sites. However, it is reasonable to expect that in this case, the editing ratio would be either very close to 50% in the case of a heterozygote allele, or very close to 100% in case of a homozygote. Here, one editing site on pre-let-7g had an editing ratio of 100%. Nonetheless, this site was confirmed by several other independent studies (2, 22, 24) and is conserved in mouse (22), suggesting that this is a genuine editing site. Combining the paired-end sequencing trick presented here with DNA-seq to compare the genome and transcriptome of the cell line of interest will be ideal for the future experimental setup of RNA editing studies.
4. Sequence quality filtering results in a trade-off between quality and coverage. In this method, we prioritized quality over coverage, but the thresholds suggested here may be tuned to optimize coverage. Indeed, if the general quality of the sequencing data is lower, a full-length QV $\geq$ 30 cut-off may filter out most of the reads. For instance, one may decide on a QV cut-off of 25 (1 error every $10^{2.5}$ or 316 bases), resulting in a maximum error on the paired-end overlapping regions of 10^{-5} (1 error every 100,000 bases). In this case, one may consider editing sites with an ER $\geq$ 1% as valid candidates, or keep it at ER $\geq$ 0.1% and decide on more stringent coverage and / or LLR test thresholds.
5. The editing ratio is strongly affected by the relative level of expression of ADAR protein and their target miRNAs, respectively. Therefore, transfection efficiency of the expression construct is expected to influence the editing ratio. This paired-end sequencing-based strategy is useful for the identification of editing sites with a wide range of editing ratios, including low efficiency sites.

5. Acknowledgements

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan to KU-T. Keio University Institute for Advanced Bioscience affiliates were supported by research funds from the Yamagata prefectural government and the City of Tsuruoka.

References

1. Ihle MA, Fassunke J, König K, Grünewald I, Schlaak M, Kreuzberg N, Tietze L, Schildhaus HU, Büttner R, Merkelbach-Bruse S (2014) Comparison of high resolution melting analysis,

pyrosequencing, next generation sequencing and immunohistochemistry to conventional Sanger sequencing for the detection of p.V600E and non-p.V600E BRAF mutations. BMC Cancer 14:13

2. Kawahara Y, Zinshteyn B, Sethupathy P, Iizasa H, Hatzigeorgiou AG, Nishikura K (2007) Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. Science 315(5815):1137-1140
3. Kume H, Hino K, Galipon J, Ui-Tei K (2014) A-to-I editing in the miRNA seed region regulates target mRNA selection and silencing efficiency. Nucleic Acids Res 42:10050-10060
4. Li JB*, Levanon EY*, Yoon JK, Aach J, Xie B, LeProust E, Zhang K, Gao Y, Church GM (2009) Genome-wide identification of human RNA editing sites by massively parallel DNA capturing and sequencing. Science 324:1210-1213 *equal contribution
5. Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X (2012) Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res 22:142-150
6. Chepelev I (2012) Detection of RNA editing events in human cells using high-throughput sequencing. Methods Mol Biol 815:91-102
7. Galipon J, Ishii R, Suzuki Y, Tomita M, Ui-Tei K (2017) Differential Binding of Three Major Human ADAR Isoforms to Coding and Long Non-Coding Transcripts. Genes (Basel) 8(2):68
8. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Scott Dewell S, Zavolan Z, Tuschl T (2010) J Vis Exp 41:2034-2039
9. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature 460:479-486
10. Weeks KM, Crothers DM (1993) Major groove accessibility of RNA. Science 261:1574-1577
11. Liu ZR, Sargueil B, Smith CW (2000) Methylene blue-mediated cross-linking of proteins to double-stranded RNA. Methods Enzymol 318:22-33
12. Tuite EM, Kelly JM (1993) Photochemical interactions of methylene blue and analogues with DNA and other biological substrates. J Photochem Photobiol B 21:103-124
13. Wu D, Lamm AT, Fire AZ (2011) Competition between ADAR and RNAi pathways for an extensive class of RNA targets. Nat Struct Mol Biol 18:1094-1101
14. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56-65

15. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311
16. The International HapMap Consortium (2003) The International HapMap Project. Nature 426:789-796
17. Gong J, Liu C, Liu W, Wu Y, Ma Z, Chen H, Guo AY (2015) An update of miRNASNP database for better SNP selection by GWAS data, miRNA expression and online tools. Database (Oxford) 2015:bav029
18. Han M, Zheng Y (2013) Comprehensive analysis of single nucleotide polymorphisms in human microRNAs. et al. PLoS One 8:e78028
19. Kawahara Y, Megraw M, Kreider E, Iizasa H, Valente L, Hatzigeorgiou AG, Nishikura K (2008) Frequency and fate of microRNA editing in human brain. Nucleic Acids Res 36:5270-5280
20. Luciano DJ, Mirsky H, Vendetti NJ, Maas S (2004) RNA editing of a miRNA precursor. RNA 10:1174-1177
21. García-López J, Hourcade Jde D, Del Mazo J (2013) Reprogramming of microRNAs by adenosine-to-inosine editing and the selective elimination of edited microRNA precursors in mouse oocytes and preimplantation embryos. Nucleic Acids Res 41:5483-5493
22. Ekdahl Y, Farahani HS, Behm M, Lagergren J, Öhman M (2012) A-to-I editing of microRNAs in the mammalian brain increases during development. Genome Res 22:1477-1487
23. Alon S, Mor E, Vigneault F, Church GM, Locatelli F, Galeano F, Gallo A, Shomron N, Eisenberg E (2012) Systematic identification of edited microRNAs in the human brain. Genome Res 22:1533-1540
24. Ramaswami G, Li JB (2014) RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res 42(Database issue):D109-113
25. Warnefors M, Liechti A, Halbert J, Valloton D, Kaessmann H (2014) Conserved microRNA editing in mammalian evolution, development and disease. Genome Biol 15:R83
26. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
27. Bu D, Yu K, Sun S, Xie C, Skogerbø G, Miao R, Xiao H, Liao Q, Luo H, Zhao G, Zhao H, Liu Z, Liu C, Chen R, Zhao Y (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic Acids Res 40:D210-215
28. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39:D152-157

29. Neilson JR, Sandberg R (2010) Heterogeneity in mammalian RNA 3' end formation. Exp Cell Res. 316:1357-1364

Figure & Table Legends

Figure 1. Subcellular localization of full-length mycGFP-ADAR proteins.
[A] Representative images of fluorescence microscopy of HeLa cells 48 hours after transfection. GFP signals were merged with DAPI staining of nuclear DNA. This data shows that mycGFP-ADAR-p150 is mostly in the cytoplasm, whereas mycGFP-ADAR1-p110 and mycGFP-ADAR2 localize in the nucleus, concentrated in nucleolar foci. Control mycGFP expression is ubiquitous. [B] Western blot using anti-GFP antibody confirms expression and successful immunoprecipitation of full-length mycGFP-ADAR proteins. From left to right, GFP: immunoprecipitation using an anti-GFP antibody, IgG: control immunoprecipitation using goat anti-rabbit immunoglobulin G (IgG), input: total cell lysate before immunoprecipitation. The expected lengths for mycGFP-ADAR1-p150, mycGFP-ADAR1-p110, mycGFP-ADAR2 and mycGFP are 165 kDa, 133 kDa, 110 kDA, and 29 kDA, respectively. A background signal observed at around 50 kDa is consistent with the size of the antibody heavy chain and is not present in the total cell lysate. Immunoprecipitation resulted in a unique band at the expected size for all four isoforms, which also corresponded exactly to the most intense band in the respective total cell lysate. This means that overexpression and immunoprecipitation of ADAR isoforms were successful.

Figure 2. Bioinformatics pipeline for paired-end sequencing-based identification of A-to-I editing sites on small RNA.
This figure summarizes the steps described in Sections 3.4 to 3.7.

Figure 3. Collapse mapping of paired-end reads and generation of high-quality *super-reads* from the overlapping sequences of paired-ends.
[A] Upper half: A-to-G collapsed forward reads are mapped to a three-base A-to-G collapsed (+) strand genome. Lower half: A-to-G collapsed reverse reads are mapped to a T-to-C collapsed (+) strand genome. [B] Left: the overlap between two reads of a pair is retrieved. Middle: overlaps that contain mismatches are discarded to filter out sequencing errors. Right: the overlaps' forward and reverse strand are merged into one high-quality sequence while conserving strand-specific information. These images were modified from the *Handbook on RNA-Seq Experiments*, Chapter 2, Part 3 (Galipon et al. 2016) with permission from Yodosha, Co., Ltd.

Figure 4. Distribution of read lengths for reads that mapped to annotated miRNA regions.
[A–E] Read length distribution of reads mapped to miRNA regions by collapse (red) and normal (dark gray) mappings of input (A, C, E) and IP (B, D) samples from the cells expressing mycGFP-ADAR1-p110 (ADAR1-p110; A, B), mycGFP-ADAR2 (ADAR2; C, D), and mycGFP control (GFP; E), respectively; X-axis: read length in nt, Y-axis: percentage of read number shown in the $\log_{10}$ scale. Pie charts summarize the respective percentages of longer reads assumed to originate from miRNA precursors ($\geq$ 25 nt) and mature miRNA ($\leq$ 24 nt) in the collapse and normal mappings of input and IP samples. [F] Percentage of collapse reads in longer ($\geq$ 25 nt) and shorter reads ($\leq$ 24 nt) for each sample, showing a clear enrichment in precursor miRNA-like collapse reads in the IP fractions of mycGFP-ADAR1-p110 and mycGFP-ADAR2. This may be indicative of binding and editing by ADARs on the double-stranded regions of pre-miRNA stem loops.

Table 1. Statistically validated candidate A-to-I editing sites that overlapped with previous studies.
Name: name of the miRNA precursor or mature miRNA on which the candidate editing site was identified. For precursors, only reads longer than 25 nt were taken into account; reads that were 24 nt or shorter were considered as coming from mature miRNAs, and the position of the editing site relative to the mature sequences was indicated as (5p) or (3p) in the case of miRNA precursors; Chr: chromosome; Str: strand; hg19: position on the corresponding chromosome of the human genome UCSC hg19 release. Pos: position relative to the +1 nucleotide of the corresponding mature miRNA. A+G: total coverage of A's and G's at the editing site. G: coverage of G's at the editing site. ER (%): editing ratio. LLR: log-likelihood ratio calculated as described in Section 3.7. Sample: sample in which the editing site was detected.

Table 1

Editing sites confirmed by other studies

Name	Chr	Str	GRCh37/hg19	Pos	A+G	G	ER (%)	LLR	Sample
pre-let-7g (5p)	3	–	52302364	10	6	6	100.000	6.6	ADAR1 IP
pre-mir-21 (3p)	17	+	57918679	8	674	2	0.297	2.7	ADAR1 input
pre-mir-140 (5p)	16	+	69967021	16	14	2	14.286	5.0	ADAR2 IP
pre-mir-455 (5p)	9	+	116971745	17	10	4	40.000	13.0	ADAR1 input
pre-mir-1304 (3p)	11	–	93466874	5	18	8	44.444	7.1	ADAR1 IP
let-7d-3p	9	+	96941181	5	186	10	5.376	25.5	ADAR2 input
miR-22-3p	**17**	**–**	**1617215**	**15**	**23978**	**62**	**0.259**	**59.1**	**ADAR1 IP**
miR-22-3p	**17**	**–**	**1617215**	**15**	**13992**	**30**	**0.214**	**11.6**	**ADAR2 IP**
miR-22-3p	**17**	**–**	**1617215**	**15**	**105860**	**548**	**0.518**	**35.2**	**ADAR2 input**
miR-24-2-5p	19	–	13947156	6	84	4	4.762	4.4	ADAR1 IP
miR-151a-3p	8	–	141742704	3	290	2	0.690	10.4	ADAR1 input
miR-210-5p	11	–	568122	12	486	2	0.412	2.3	ADAR2 input
miR-340-3p	5	–	179442328	13	34	2	5.882	4.4	ADAR2 input
miR-1301-3p	2	–	25551539	5	24	2	8.333	5.1	ADAR2 input

Figure 1

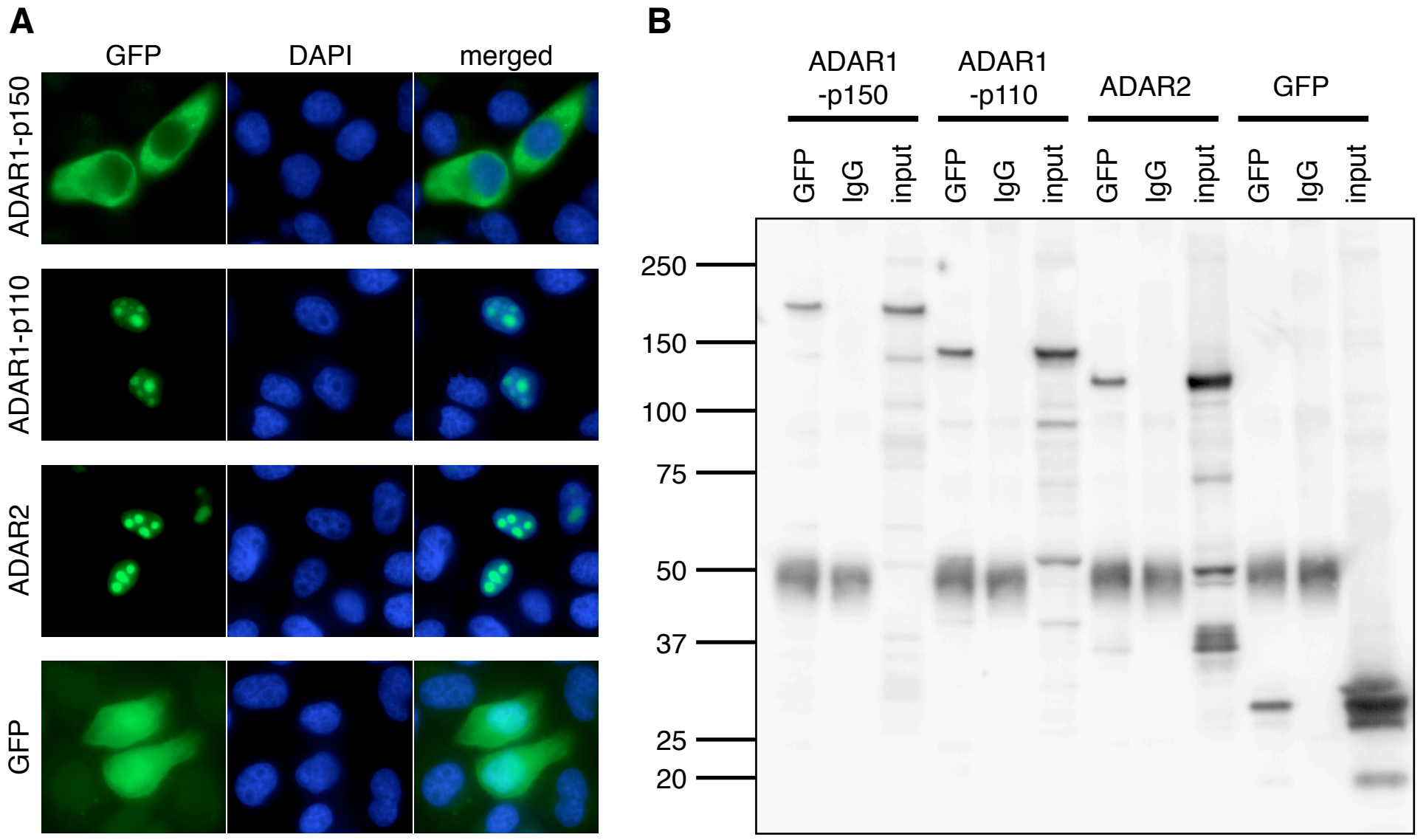

A
GFP
DAPI
merged
ADAR1-p150
ADAR1-p110
ADAR2
GFP
B
ADAR1 -p150
ADAR1 -p110
ADAR2
GFP
GFP
IgG
input
250
150
100
75
50
37
25
20

Figure 2

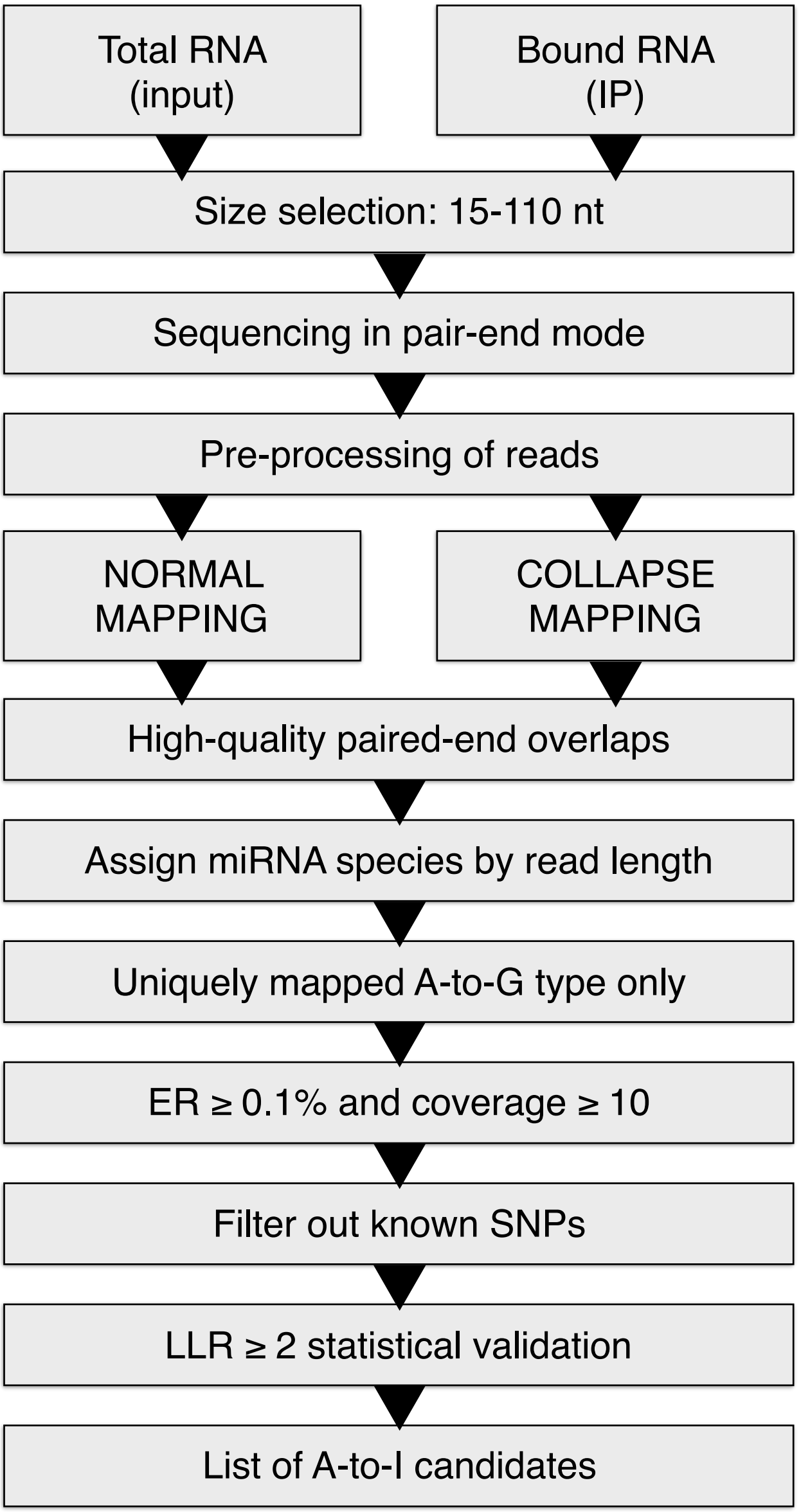
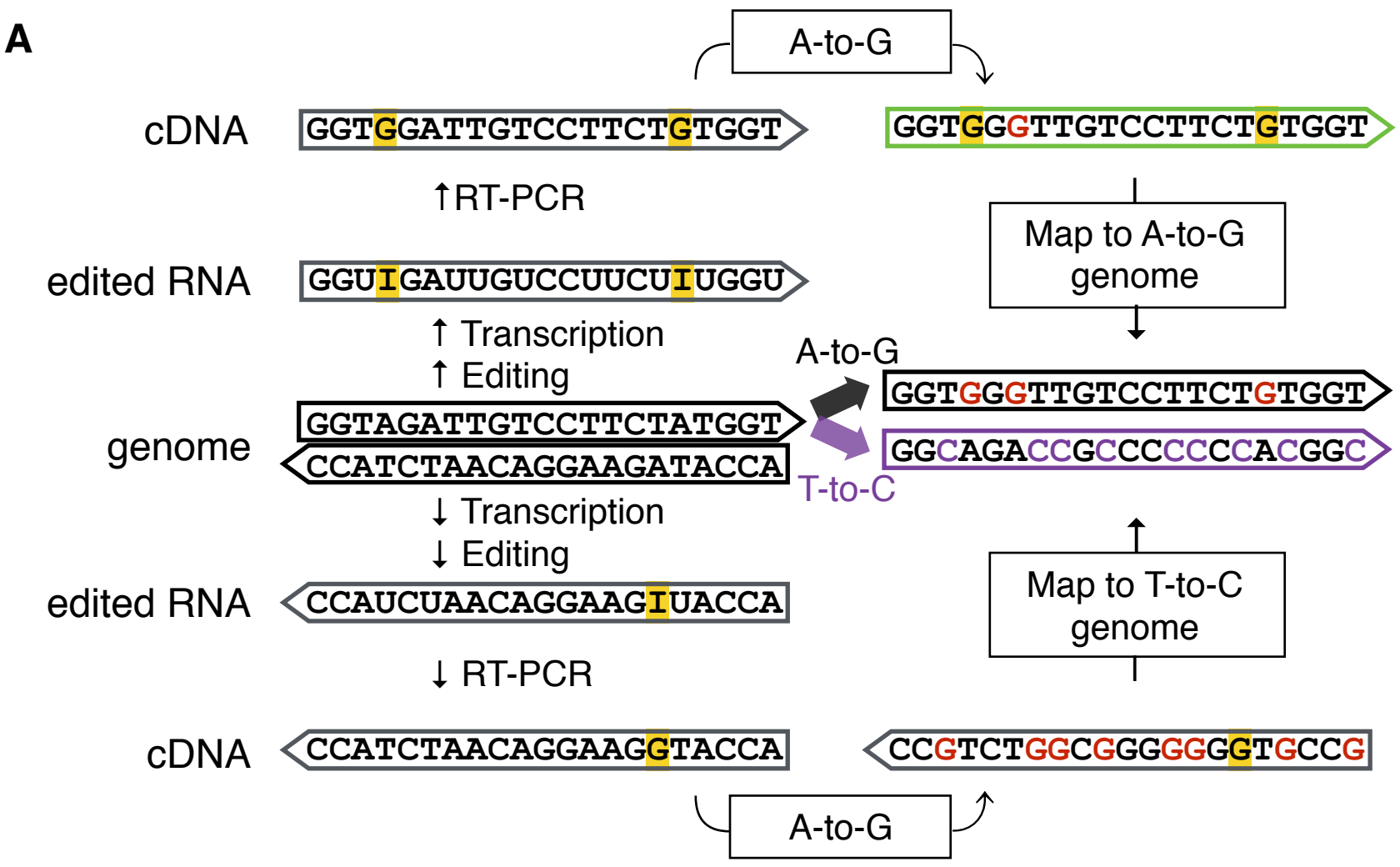

Total RNA (input)
Bound RNA (IP)
Size selection: 15-110 nt
Sequencing in pair-end mode
Pre-processing of reads
NORMAL MAPPING
COLLAPSE MAPPING
High-quality paired-end overlaps
Assign miRNA species by read length
Uniquely mapped A-to-G type only
ER ≥ 0.1% and coverage ≥ 10
Filter out known SNPs
LLR ≥ 2 statistical validation
List of A-to-I candidates

Figure 3

A


A-to-G
cDNA
GGTGGATTGTCCTTCTGTGGT
GGTGGGTTGTCCTTCTGTGGT
↑RT-PCR
Map to A-to-G genome
edited RNA
GGUIGAUUGUCCUUCUIUGGU
↑ Transcription
↑ Editing
A-to-G
genome
GGTAGATTGTCCTTCTATGGT
GGTGGGTTGTCCTTCTGTGGT
CCATCTAACAGGAAGATACCA
GGCAGACCGCCCCCCCACGGC
T-to-C
↓ Transcription
↓ Editing
Map to T-to-C genome
edited RNA
CCAUCUAACAGGAAGIUACCA
↓ RT-PCR
cDNA
CCATCTAACAGGAAGGTACCA
CCGTCTGGCGGGGGGGTGCCG
A-to-G

B

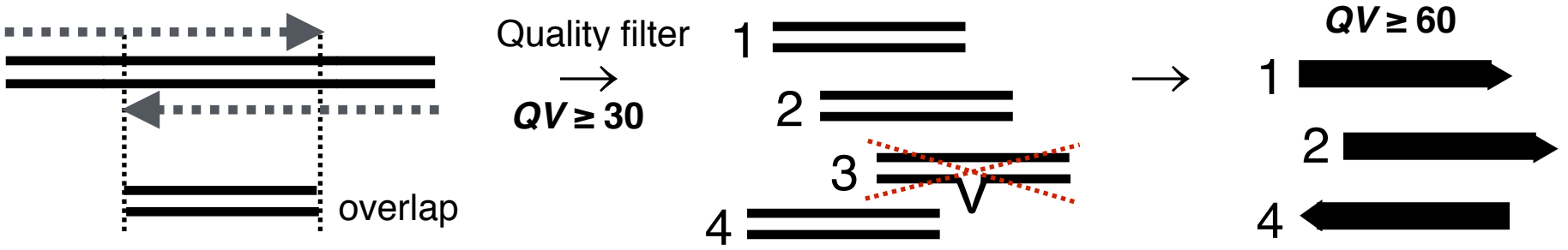
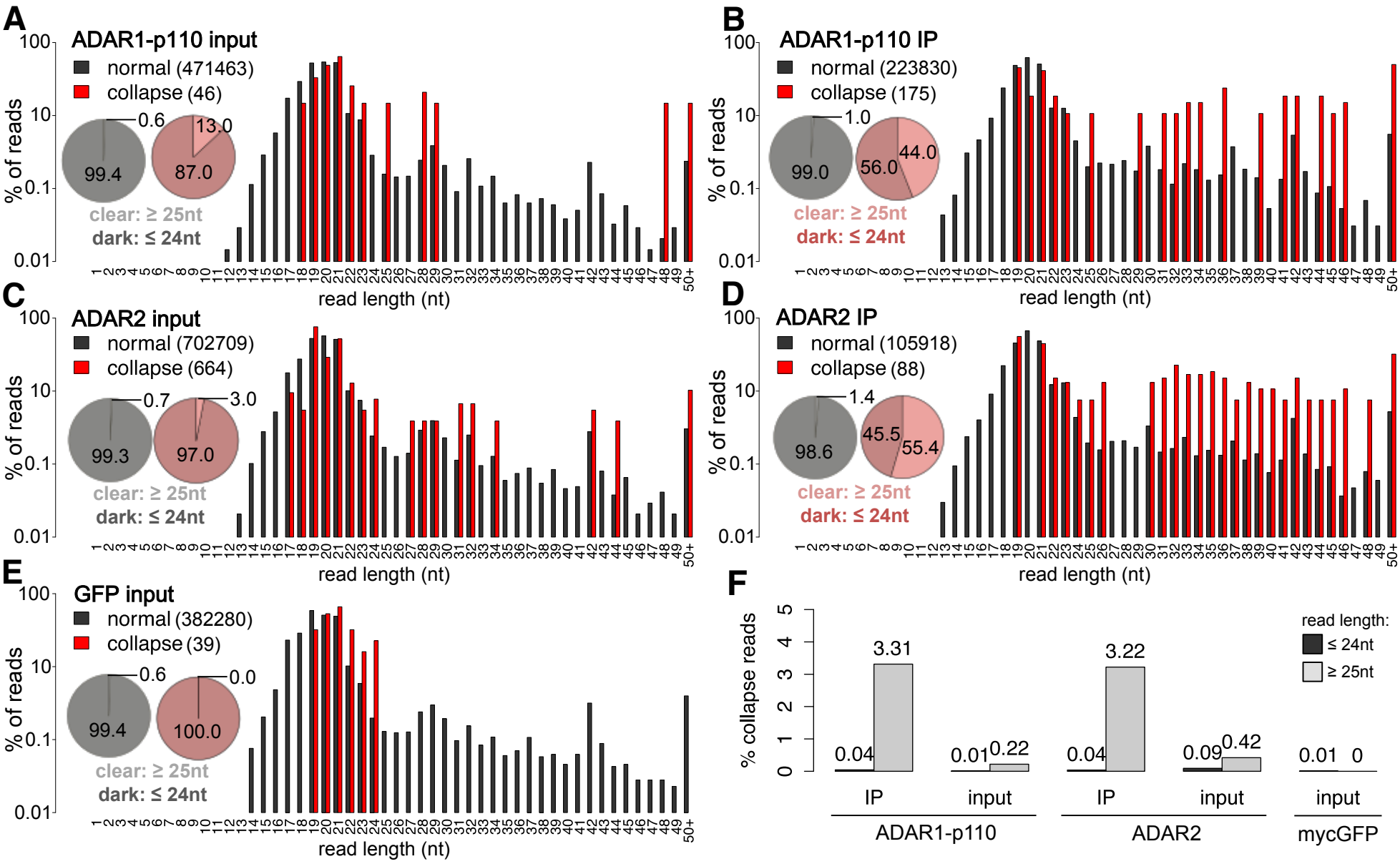

Paired-end small RNA-seq
overlap
Discard mismatched pairs
Quality filter
QV ≥ 30
1
2
3
4
Merge overlap
QV ≥ 60
1
2
4

Figure 4



View publication stats